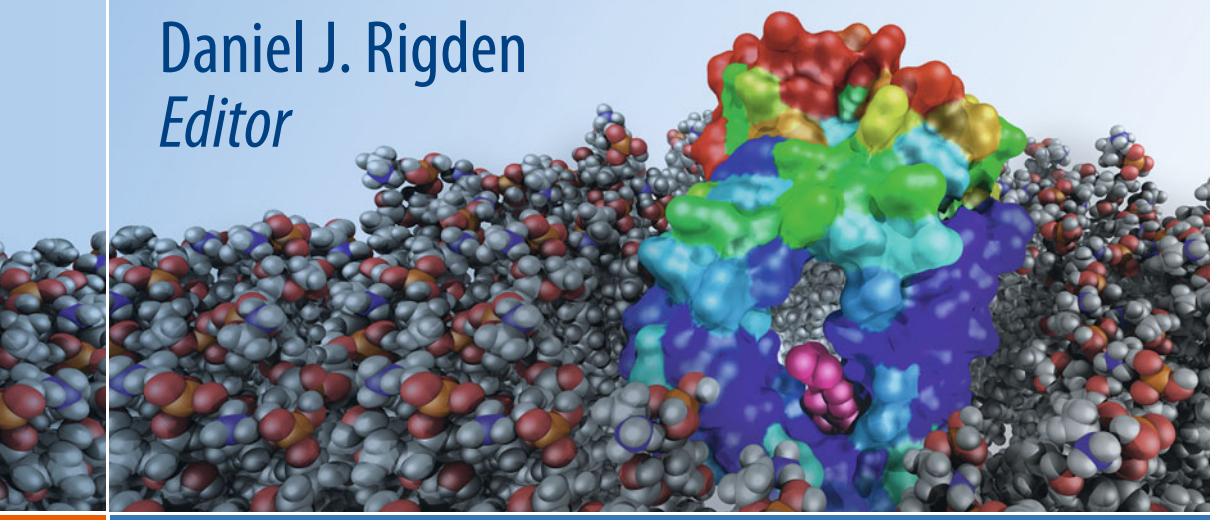
Daniel J. Rigden
*Editor*



# From Protein Structure to Function with Bioinformatics

*Second Edition*

Springer

# From Protein Structure to Function with Bioinformatics

Daniel J. Rigden
Editor

# From Protein Structure to Function with Bioinformatics

Second Edition

Springer

*Editor*
Daniel J. Rigden
Institute of Integrative Biology
University of Liverpool
Liverpool
UK

ISBN 978-94-024-1067-9 ISBN 978-94-024-1069-3 (eBook)
DOI 10.1007/978-94-024-1069-3

Library of Congress Control Number: 2017932416



Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer Science+Business Media B.V.
The registered company address is: Van Godewijckstraat 30, 3311 GX Dordrecht, The Netherlands

# Preface to the Second Edition

Welcome to the second edition! Since the publication of the first edition, the research area of protein structural informatics has continued to grow in volume and significance. A search of PubMed for 'protein structural bioinformatics' shows around 1000 papers in 2009 when the first edition was published, doubling to over 2000 in 2015. In the same period, the Protein Data Bank has similarly almost doubled, breaching 100,000 entries in 2014. Nevertheless, the gap between the protein sequences and structures continues to grow, as new technologies allow cheap and facile sequencing of previously intractable organisms and even of entire environments. Protein structural bioinformatics offers a computational route to bridge this gap by predicting structures for uncharacterised families. Those structures can then be analysed by a wide variety of further bioinformatics algorithms to shed light on their function. These two interlinking research areas are the topic of this book.

This second edition contains three chapters addressing areas not covered in the first edition. Each is contributed by world-leading experts in the field. The remaining chapters are all revised, many dramatically, to reflect seven years of fast-moving bioinformatics research with one chapter being entirely replaced. As previously, there are two sections covering first methods to generate or infer structure and secondly structure-based function annotation. Naturally, such a division is never clear-cut as prediction of a structure may simultaneously inform about its likely functions. For example, annotation of an intrinsically disordered region would immediately suggest, in eukaryotes at least, a role in protein-protein interaction since such stretches frequently harbour linear motifs bound by recognition modules on partner proteins.

The first new chapter, Chap. 2, covers arguably the most exciting development in protein bioinformatics of recent years, namely the new-found ability to accurately predict contacting residue pairs through covariance analysis of large multiple sequence alignments. These contact predictions have a wide and still expanding range of applications. Most obviously, the data allow for protein structure prediction in conjunction either with protein distance geometry methods or, more effectively, by synergistic incorporation into fragment assembly ab initio modelling

methods. The contact predictions also inform on the likely harmfulness of single amino acid polymorphisms (SAPs) and allow for better prediction of protein-protein interactions. Prediction of protein-protein complex structures, both between globular domains and between a domain and a short linear motif, is the subject of the new Chap. 8. A full accounting of protein-protein interactions in cells is crucial for the future prospects of integrative systems-level methods, while structural knowledge of interfaces again contributes to prediction of the consequences of SAPs. The third new arrival, Chap. 7, covers predictions of amyloid structure in proteins. Such structure is of huge biomedical interest, underlying diseases such as Parkinson's and Alzheimer's, but is equally intriguing for the normal physiological roles of 'functional amyloids'. Finally, the new Chap. 10 text covers the fascinating variety of means by which structural bioinformatics can mark up a structure, experimental or modelled, for likely functional pockets and patches on the protein surface.

The methods covered in this book comprise a comprehensive toolkit to address future challenges in protein structure, function and evolution. Recent papers open up new viewpoints on protein evolution (Alva et al. 2015; Edwards and Deane 2015) and on the amenability of different folds to functional innovation (Toth-Petroczy and Tawfik 2014), treat the biophysical consequences of protein ageing (de Graff et al. 2016) and even reveal oversights in our accounting of molecular interactions (Newberry and Raines 2016). Clearly, exciting times lie ahead for protein bioinformaticians!

Liverpool, UK Daniel J. Rigden

## References

Alva V, Soding J, Lupas AN (2015) A vocabulary of ancient peptides at the origin of folded proteins. Elife 4:e09410

de Graff AM, Hazoglou MJ, Dill KA (2016) Highly charged proteins: the achilles' heel of aging proteomes. Structure 24(2):329–336

Edwards H, Deane CM (2015) Structural bridges through fold space. PLoS Comput Biol 11(9): e1004466

Newberry RW, Raines RT (2016) A prevalent intraresidue hydrogen bond stabilizes proteins. Nat Chem Biol 12(12):1084–1088

Toth-Petroczy A, Tawfik DS (2014) The robustness and innovability of protein folds. Curr Opin Struct Biol 26:131–138

# Contents

**Part II From Structures to Functions**

# Part I
# Generating and Inferring Structures

# Chapter 1
# Ab Initio Protein Structure Prediction

**Jooyoung Lee, Peter L. Freddolino and Yang Zhang**

**Abstract** Predicting a protein's structure from its amino acid sequence remains an unsolved problem after several decades of efforts. If the query protein has a homolog of known structure, the task is relatively easy and high-resolution models can often be built by copying and refining the framework of the solved structure. However, a template-based modeling procedure does not help answer the questions of how and why a protein adopts its specific structure. In particular, if structural homologs do not exist, or exist but cannot be identified, models have to be constructed from scratch. This procedure, called ab initio modeling, is essential for a complete solution to the protein structure prediction problem; it can also help us understand the physicochemical principle of how proteins fold in nature. Currently, the accuracy of ab initio modeling is low and the success is generally limited to small proteins (<120 residues). With the help of co-evolution based contact map predictions, success in folding larger-size proteins was recently witnessed in blind testing experiments. In this chapter, we give a review on the field of ab initio structure modeling. Our focus will be on three key components of the modeling algorithms: energy function design, conformational search, and model selection. Progress and advances of several representative algorithms will be discussed.

**Keyword** Protein structure prediction · Ab initio folding · Contact prediction · Force field

J. Lee
School of Computational Sciences,
Korea Institute for Advanced Study, Seoul 130-722, Korea

P.L. Freddolino · Y. Zhang
Department of Biological Chemistry, University of Michigan,
Ann Arbor, MI 48109, USA

P.L. Freddolino · Y. Zhang (✉)
Department of Computational Medicine and Bioinformatics,
University of Michigan, Ann Arbor, MI 48109, USA
e-mail: zhng@umich.edu



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_1

## 1.1 Introduction

With the success of an expanding array of genome sequencing projects, the number of known protein sequences has been increasing exponentially. However, the sequences on their own cannot tell what each protein does in cell. Although protein structure information is essential for understanding the function, the speed of protein structure determination lags far behind the increase of sequences, due to the technical difficulties and laborious nature of structural biology experiments. By the end of 2015, about 90 million protein sequences were deposited in the UniProtKB database (Bairoch et al. 2005) (http://www.uniprot.org/). However, the corresponding number of protein structures in the Protein Data Bank (PDB) (Berman et al. 2000) (http://www.rcsb.org) is only about 100,000. The gap is rapidly widening as indicated in Fig. 1.1, where the ratio of sequences over structure increased from less than 1 magnitude to around 3 magnitudes in the last two decades. Thus, developing efficient computer-based algorithms that can generate high-resolution 3D structure predictions becomes probably the only avenue to fill up the gap.

Depending on whether similar proteins have been experimentally solved, protein structure prediction methods can be grouped into two categories. First, if proteins of a similar structure are identified from the PDB library, the target model can be constructed by copying and refining framework of the solved proteins (templates). The procedure is called "template-based modeling (TBM)" (Sali and Blundell 1993; Karplus et al. 1998; Jones 1999; Skolnick et al. 2004; Soding 2005; Wu and Zhang 2008a; b; Yang et al. 2011), and will be discussed in the subsequent chapters. Although high-resolution models can often be generated by TBM, the procedure cannot help us understand the physicochemical principle of protein folding.

If protein templates are not available, we have to build the 3D models from scratch. This procedure has been given different names, e.g. ab initio modeling (Klepeis et al. 2005; Liwo et al. 2005; Wu et al. 2007; Taylor et al. 2008; Xu and

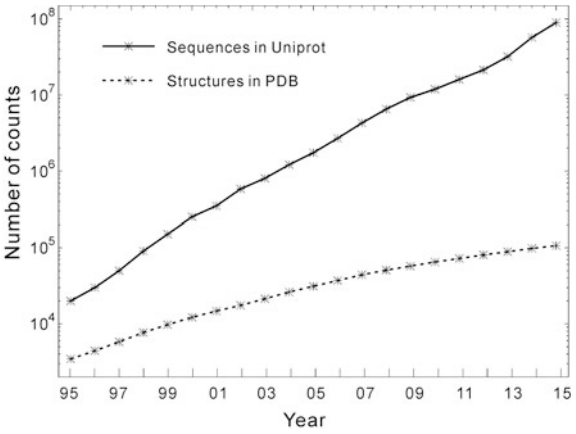



**Fig. 1.1** The numbers of available protein sequences and solved protein structures are shown for the last 20 years. The ratio of sequences over structures increases from less than 10 in 1995 to three orders of magnitude in 2015. Data are taken from UniProtKB (Bairoch et al. 2005) and PDB (Berman et al. 2000) databases

Zhang 2012); de novo modeling (Bradley et al. 2005a, b), physics-based modeling (Oldziej et al. 2005), or free modeling (Jauch et al. 2007; Kinch et al. 2015). In this chapter, the term ab initio modeling is uniformly used to avoid confusion. Unlike the template-based modeling, a successful ab initio modeling procedure could help address the basic questions on how and why a protein adopts the specific structure out of many possibilities.

Typically, ab initio modeling conducts a conformational search under the guidance of a designed energy function. This procedure usually generates a number of possible conformations (also called structure decoys), and final models are selected from them. Therefore, a successful ab initio modeling depends on three factors: (1) an accurate energy function with which the native structure of a protein corresponds to the most thermodynamically stable state, compared to all possible decoy structures; (2) an efficient search method which can quickly identify the low-energy states through conformational search; (3) a strategy that can select near-native models from a pool of decoy structures.

This chapter gives a review on the most recent progress in ab initio protein structure prediction. This review is neither sufficiently complete to include all available ab initio methods nor sufficiently in depth to provide all backgrounds/motivations behind them. For a quantitative comparison of the state-of-the-art ab initio modeling methods, readers are suggested to read the assessment articles on template-free modeling in the recent CASP experiments (Kinch et al. 2011; Tai et al. 2014; Kinch et al. 2015). The rest of the chapter is organized as follows. First, the three major issues of ab initio modeling, i.e. energy function design, conformational search engine and model selection scheme, will be described in detail. New and promising ideas to improve the efficiency and effectiveness of the prediction are then discussed. Finally, current progress and challenges of ab initio modeling are summarized.

## 1.2 Energy Functions

In this section, we discuss energy functions used for ab initio modeling. It should be noted that in many cases energy functions and the search procedures are intricately coupled to each other, and as soon as they are decoupled, the modeling procedure often loses its power and/or validity. We classify the energy functions into two groups: (a) physics-based energy functions and (b) knowledge-based energy functions, depending on whether they make use of statistics from the existing protein 3D structures in the PDB. A few promising methods from each group are selected to discuss according to their uniqueness and modeling accuracy. A list of ab initio modeling methods is provided in Table 1.1 along with their properties about energy functions, conformational search algorithms, model selection methods and typical running times.

**Table 1.1** A list of ab initio modeling algorithms reviewed in this chapter is shown along with their energy functions, conformational search methods, model selection schemes and typical CPU time per target

| Algorithm and server address | Force-field type | Search method | Model selection | Time cost per CPU |
|---|---|---|---|---|
| AMBER/CHARMM/OPLS (Brooks et al. 1983; Weiner et al. 1984; Jorgensen and Tirado-Rives 1988; Duan and Kollman 1998; Zagrovic et al. 2002) | Physics-based | Molecular dynamics (MD) | Lowest energy | Years |
| UNRES (Liwo et al. 1999; Liwo et al. 2005, Oldziej et al. 2005) | Physics-based | Conformational space annealing (CSA) | Clustering/free-energy | Hours |
| ASTRO-FOLD (Klepeis et al. Klepeis and Floudas 2003; Klepeis et al. 2005) | Physics-based | αBB/CSA/MD | Lowest energy | Months |
| ROSETTA (Simons et al. 1997, Das et al. 2007) http://www.robetta.org | Physics- and knowledge-based | Monte Carlo (MC) | Clustering/free-energy | Days |
| TASSER/Chunk-TASSER (Zhang et al. 2004, Zhou and Skolnick 2007) http://cssb.biology.gatech.edu/skolnick/webservice/MetaTASSER | Knowledge-based | MC | Clustering/free-energy | Hours |
| I-TASSER (Roy et al. 2010; Yang et al. 2015a, b) http://zhanglab.ccmb.med.umich.edu/I-TASSER | Knowledge-based | MC | Clustering/free-energy | Hours |
| QUARK (Xu and Zhang 2012) http://zhanglab.ccmb.med.umich.edu/QUARK | Physics- and knowledge-based | MC | Clustering/free-energy | Hours |

### 1.2.1 Physics-Based Energy Functions

In a strictly-defined physics-based ab initio method, interactions between atoms should be based on quantum mechanics and the Coulomb potential with only a few fundamental parameters such as the electron charge and the Planck constant; all atoms should be described by their atom types where only the number of electrons is relevant (Hagler et al. 1974; Weiner et al. 1984). However, there have not been serious attempts to start from quantum mechanics to predict structures of (even small) proteins, simply because the computational resources required for such calculations are far beyond what is available now. Without quantum mechanical treatments, a practical starting point for ab initio protein modeling is to use a force field treating atoms as point particles interacting through a defined potential form, with the parameters governing inter-atomic interactions obtained through the comparisons of the force field with a combination of experimental and quantum mechanical data (Hagler et al. 1974; Weiner et al. 1984). Well-known examples of such all-atom physics-based force fields include AMBER (Weiner et al. 1984; Cornell et al. 1995; Duan and Kollman 1998), CHARMM (Brooks et al. 1983; Neria et al. 1996; MacKerell et al. 1998), OPLS (Jorgensen and Tirado-Rives 1988; Jorgensen et al. 1996), and GROMOS96 (van Gunsteren et al. 1996). These potentials contain terms associated with bond lengths, angles, torsion angles, van der Waals, and electrostatics interactions. The major difference between them lies in the selection of atom types and the interaction parameters.

**Coupling Physics-Based Potentials With Molecular Dynamics Simulations** For the study of protein folding, these classical force fields were often coupled with molecular dynamics (MD) simulations. The obvious appeal of such an approach is that the prediction of protein folding via MD simulations provides not only information on the folded structure, but also the folding process itself, which must be fully simulated *en route*. However, the results, from the viewpoint of protein structure prediction, have until quite recently been disappointing. (See Chap. 12 for the use of MD in elucidation of protein function from known structures).

The first milestone in MD-based ab initio protein folding was probably the 1997 work of Duan and Kollman, who simulated the villin headpiece subdomain (a 36 amino acid protein) in explicit solvent for 6 months on parallel supercomputers. Although the authors did not fold the protein with high resolution, the best of their final models was within 4.5 Å RMS deviation of the native state (Duan and Kollman 1998). With Folding@Home, a worldwide-distributed computer system, this small protein was later folded by Pande and coworkers (Zagrovic et al. 2002) to 1.7 Å with a total simulation time of 300 μs or approximately 1000 CPU years. The years since then have seen an increasing number of successful ab initio folding simulations using molecular dynamics (Chowdhury et al. 2003; Ensign et al. 2007; Lei et al. 2007; Freddolino and Schulten 2009), although all have required heroic amounts of computing time either through supercomputing centers or distributed community projects. During the same period, ab initio folding simulations also revealed secondary structure biases in several physics-based force fields that

hampered their general applicability to different folds (Best et al. 2008; Freddolino et al. 2008, Best and Hummer 2009; Freddolino and Schulten 2009; Lindorff-Larsen et al. 2012).

A flurry of force field development efforts spurred by these shortcomings have resulted in a new generation of parameter sets that are able to reliably fold a wide variety of protein structures (Lindorff-Larsen et al. 2010; Mittal and Best 2010; Piana et al. 2011; Lindorff-Larsen et al. 2012), leaving simulation timescales as the main barrier for MD ab initio folding simulations. Even this barrier has begun to crumble in the face of recent advances in computing hardware. The special purpose Anton machine, designed by Shaw and co-workers specifically for extreme-performance molecular dynamics simulations, has allowed complete, reversible folding simulations of proteins up to $\sim$100 residues long in explicit solvent (Lindorff-Larsen et al. 2011; Piana et al. 2012, Piana et al. 2013a, b; Piana et al. 2014). Following a separate path, the use of GPU acceleration in most major molecular dynamics packages has enabled ab initio folding simulations on commodity hardware to reach performances of 1 microsecond per GPU-day for small proteins with implicit solvent (Nguyen et al. 2014), and allowed successful folding of 16 out of 17 test proteins (10–100 residues). Despite these remarkable efforts, the all-atom physics-based MD simulation is far from being routinely used for structure prediction of typical-size proteins ($\sim$100–300 residues), and it is instead primarily used to provide additional information on folding pathways or equilibriums.

**Application to Atomic-Level Structure Refinement** Another protein structure niche where physics-based MD simulation can contribute is structure refinement. Starting from low-resolution protein models, the goal is to draw the structure closer to the native by refining the local side-chain and peptide-backbone packing. When the starting models are not very far away from the native, the intended conformational change is relatively small and the simulation time would be much shorter than that required in ab initio folding. One of the early MD-based protein structure refinements was for the GCN4 leucine zipper (33-residue dimer) (Nilges and Brunger 1991; Vieth et al. 1994), where a low-resolution coiled-coil dimer structure (2–3 Å RMS deviation from native) was first assembled by Monte Carlo (MC) simulation before the subsequent MD refinement. With the help of helical dihedral-angle restraints, Skolnick and coworkers (Vieth et al. 1994) were able to generate a refined structure of GCN4 with below 1 Å backbone RMSD using CHARMM (Brooks et al. 1983) with the TIP3P water model (Jorgensen et al. 1983).

Later, using AMBER 5.0 (Case et al. 1997) and the TIP3P water model (Jorgensen et al. 1983; Lee et al. 2001) attempted to refine 360 low-resolution models generated by ROSETTA (Simons et al. 1997) for 12 small proteins (<75 residues); but they concluded that no systematic structure improvement was achieved (Lee et al. 2001). Fan and Mark (Fan and Mark 2004) tried to refine 60 ROSETTA models for 11 small proteins (<85 residues) using GROMACS 3.0 (Lindahl et al. 2001) with explicit water (Berendsen et al. 1981) and they reported that 11/60 models were improved by 10% in RMSD, but 18/60 got worse in RMSD

after refinement. Similarly, Chen and Brooks (Chen and Brooks 2007) used CHARMM22 (MacKerell et al. 1998) to refine five CASP6 CM targets (70–144 residues). In four cases, refinements with up to 1 Å RMSD reduction were achieved. In this work, an implicit solvent model based on the generalized Born (GB) approximation (Im et al. 2003) was used, which significantly speeded up the computation. In addition, the spatial restraints extracted from the initial models were used to guide the refinement procedure (Chen and Brooks 2007).

More recently, Zhang et al. (2011) proposed to use analogous fragments from known structures to bias the physics-based force field and improve structure refinement. In this work, the initial structure model was split into segments of 2–4 secondary structure elements, which are structurally matched through the PDB library by TM-align (Zhang and Skolnick 2005a, b) to identify analogous fragments. The distance map from the analogous fragments is then used as restraints to reshape the MD energy funnel. The protocol was tested on 181 benchmarking and 26 CASP targets. It was found that structure models of correct folds with TM-score >0.5 can be often pulled closer to native with higher GDT-HA score, but improvement for the models of incorrect folds (TM-score <0.5) were much less pronounced. The previous experiments have shown that the physics-based force field can often recognize the native but lacks middle-range correlation to the RMSD in the high RMSD region (Bradley et al. 2005a, b; Jagielska et al. 2008), which leads to a golfcourse like energy landscapes with a deep basin around the native that cannot help for refining low-resolution models. The data by Zhang et al. seemed to indicate that template-based fragmental distance maps reshaped the MD energy landscape from golfcourse-like to funnel-like in the successfully refined targets with an approximate radius of TM-score $\sim$0.5. Similarly, Feig and coworkers used the C$\alpha$ maps collected from initial structure models to guide the MD based structure refinement simulations (Mirjalili and Feig 2013). In the recent CASP experiment (Feig and Mirjalili 2015), the approach showed a small but consistent improvement on the structural models, with average RMSD improvement by 0.13 Å for the first submitted models and 0.52 Å for the best in top five models.

**Molecular Mechanics Approaches** A noteworthy observation was made by Summa and Levitt (2007) who exploited various molecular mechanics (MM) potentials (AMBER99 (Wang et al. 2000; Sorin and Pande 2005), OPLS-AA (Kaminski et al. 2001), GROMOS96 (van Gunsteren et al. 1996), and ENCAD (Levitt et al. 1995)) to refine 75 proteins by *in vacuo* energy minimization. They found that a knowledge-based atomic contact potential outperformed the MM potentials by moving almost all test proteins closer to their native states, while the MM potentials, except for AMBER99, essentially drove decoys further away from their native structures. The vacuum simulation without solvation may be partly the reason for the failure of the MM potentials. This observation demonstrates the possibility of combining knowledge-based potentials with physics-based force fields for more successful protein structure refinement.

While the physics-based potential driven by MD simulations was not particularly successful in structure prediction due to the immense computational cost of MD

simulations on the timescales of folding processes, fast search methods (such as Monte Carlo simulations and genetic algorithms) combined with similar physics-based potentials have been shown to be promising in both structure prediction and structure refinement. One example is the effort by Scheraga and coworkers (Liwo et al. 1999; Liwo et al. 2005; Oldziej et al. 2005) who have been developing a physics-based protein structure prediction method solely based on the thermodynamic hypothesis. The method combines the coarse grained potential UNRES with a global optimization algorithm called conformational space annealing (Oldziej et al. 2005). In UNRES, each residue is described by two interacting off-lattice united atoms, $C_\alpha$ and the side-chain center. This effectively reduces the number of atoms by 10, enabling one to handle polypeptide chains of larger than 100 residues. The resulting prediction time for small proteins can be then reduced to 2–10 h. The UNRES energy function (Liwo et al. 1993) consists of pair-wise interactions between all interacting parties and additional terms such as local energy and correlation energy. The low energy UNRES models are then converted into all-atom representations based on ECEPP/3 (Nemethy et al. 1992). Although many of the parameters of the energy function are calculated by quantum-mechanical methods, some of them are derived from the distributions and correlation functions calculated from the PDB library. For this reason, one might question classifying it as a truly physics-based approach. Nevertheless, this method is one of the most faithful ab initio methods available (in terms of the application of a thorough global optimization to a physics-based energy function) and has been systematically applied to many CASP targets since 1998. The most notable prediction success by this approach was for T061 from CASP3, for which a model of 4.2 Å RMSD for a 95-residue α-helical protein was generated with an accuracy gap between it and the models of others. It was shown in a clear-cut fashion that the ab initio method can sometime provide better models for the targets where the template-based methods fail. In CASP6, a structure genomics target of TM0487 (T0230, 102 residues) was folded to 7.3 Å by this approach. However, it seems that the scarcity and the best-but-still-low accuracy of such models by a pure ab initio modeling failed to draw much attention from the protein science community, where accurate protein models are in great demand.

Another example of the physics-based modeling approaches is the multi-stage hierarchical algorithm ASTRO-FOLD, proposed by Floudas and coworkers (Klepeis and Floudas 2003; Klepeis et al. 2005). First, secondary structure elements (α-helices and β-strands) are predicted by calculating a free energy function of overlapping oligopeptides (typically pentapeptides) and all possible contacts between 2 hydrophobic residues. The free energy terms used include entropic, cavity formation, polarization, and ionization contributions for each oligopeptide. After transforming the calculated secondary structure propensity into the upper and lower bounds of backbone dihedral angles and the distant restraints between Cα atoms, the final tertiary structure of the full length protein is modeled by globally minimizing the energy using the ECEPP/3 all-atom force field. This approach was successfully applied to an α-helical protein of 102 residues in a double-blind fashion (but not in an open community-wide way for relative performance

comparison to other methods). The RMSD of the predicted model was 4.94 Å away from the experimental structure. The global optimization method used in this approach is a combination of α branch and bound (αBB), conformational space annealing, and MD simulations (Klepeis and Floudas 2003; Klepeis et al. 2005). The relative performance of this method on larger number of proteins is yet to be examined.

Taylor and coworkers (Taylor et al. 2008) proposed a novel approach which constructs protein structural models by enumerating possible topologies in a coarse-grained form, given the secondary structure assignments and the physical connection constraints of the secondary structure elements. The top scoring conformations, based on the structural compactness and element exposure, are then selected for further refinement (Jonassen et al. 2006). The authors successfully folded a set of five αβ sandwich proteins with length up to 150 residues with the first model having 4–6 Å RMS deviation from the known experimental structure. Again, although appealing in methodology, the performance of the approach in open blind experiments and on proteins of various fold-types is yet to be seen.

### *1.2.2 Knowledge-Based Energy Function Combined with Fragments*

The term knowledge-based potential refers to a set of empirical energy terms derived from the statistics and regularities of the solved structures in deposited PDB. Such potentials can be divided into two types as described by Skolnick (Skolnick 2006). The first covers generic and sequence-independent terms such as the hydrogen bonding and the local backbone stiffness of a polypeptide chain (Zhang et al. 2003). The second contains amino-acid or protein-sequence dependent terms, e.g. pair-wise residue contact potential (Skolnick et al. 1997), distance dependent atomic contact potential (Samudrala and Moult 1998; Lu and Skolnick 2001; Zhou and Zhou 2002; Shen and Sali2006; Zhang and Zhang 2010), and secondary structure propensities (Zhang et al. 2003, Zhang and Skolnick 2005a, b; Zhang et al. 2006).

Although most knowledge-based force fields contain secondary structure propensities, it may be that local protein structures are rather difficult to reproduce in the reduced modeling. That is, in nature a variety of protein sequences prefer either helical or extended structures depending on the subtle differences in their local and global sequence environment, yet we have not yet developed force fields that can reproduce this subtlety properly. One way to circumvent this problem is to use secondary structure fragments, obtained from sequence or profile alignments, directly into 3D model assembly. One additional advantage of the fragment-based approach is that the use of excised secondary structure fragment can significantly reduce the entropy of the conformational search.

Here, we introduce several representative methods utilizing knowledge-based energy functions, which have proven to be the most successful in ab initio protein structure prediction methods in recent community competitions (Simons et al. 1997; Zhang and Skolnick 2004a, b; Xu and Zhang 2012).

**ROSETTA** One of the best-known ideas for *ab initio*, pioneered by Bowie and Eisenberg, involves generating protein models by assembling small fragments (mainly 9-mers) taken from the PDB library (Bowie and Eisenberg 1994). Based on a similar idea, Baker and coworkers developed ROSETTA (Simons et al. 1997), which has been very successful for the free modeling (FM) targets in the CASP experiments, and which has greatly boosted the popularity of the fragment assembly approach in the field. In recent versions of ROSETTA (Bradley et al. 2005a, b; Das et al. 2007; Ovchinnikov et al. 2015), the authors first generated models in a reduced form with conformations specified with heavy backbone and Cβ atoms. In the second phase, a set of selected low-resolution models were subject to all-atom refinement procedure using an all-atom physics-based energy function, which includes van der Waals interactions, pair-wise solvation free energy, and an orientation-dependent hydrogen-bonding potential. The flowchart of the two-phase modeling is shown in Fig. 1.2 and details on the energy functions can be found in references (Bradley et al. 2005a, b; Das et al. 2007). For the conformational search, multiple rounds of Monte Carlo minimization (Li and Scheraga 1987) are carried out. One of the notable examples for this two-step protocol is the blind prediction of a FM target (T0281 from CASP6, 70 residues), whose Cα RMSD from its crystal structure is 1.6 Å (Bradley et al. 2005a, b), where a very extensive sampling was carried out using the distributed computing network of Rosetta@home allowing about 500,000 CPU hours for each target domain. Despite the significant success, the computational cost of the procedure is rather expensive for routine use.

Partially because of the notable success of the ROSETTA algorithm, as well as the limited availability of its energy functions to others, several groups initiated developments of their own energy functions following the idea of ROSETTA. Derivatives of ROSETTA include Simfold (Fujitsuka et al. 2006) and Profesy (Lee et al. 2004); their energy terms include van der Waals interactions, backbone dihedral angle potentials, hydrophobic interactions, backbone hydrogen-bonding potential, rotamer potential, pair-wise contact energies, beta-strand pairing, and a term controlling the protein radius of gyration. However, their predictions seems to be only partially successful in comparison to ROSETTA (Lee et al. 2004; Fujitsuka et al. 2006).

**TASSER/I-TASSER** Another successful free modeling approach, TASSER by Zhang and Skolnick (Zhang and Skolnick 2004a, b), constructs 3D models based on a purely knowledge-based approach. The target sequence is first threaded through a set of representative protein structures to search for possible folds. Contiguous fragments (>5 residues) are then excised from the threaded aligned regions and used to reassemble full-length models, while unaligned regions are built by a lattice-based ab initio modeling (Zhang et al. 2003). The protein conformation in TASSER is represented by a trace of Cα atoms and side-chain centers of mass,

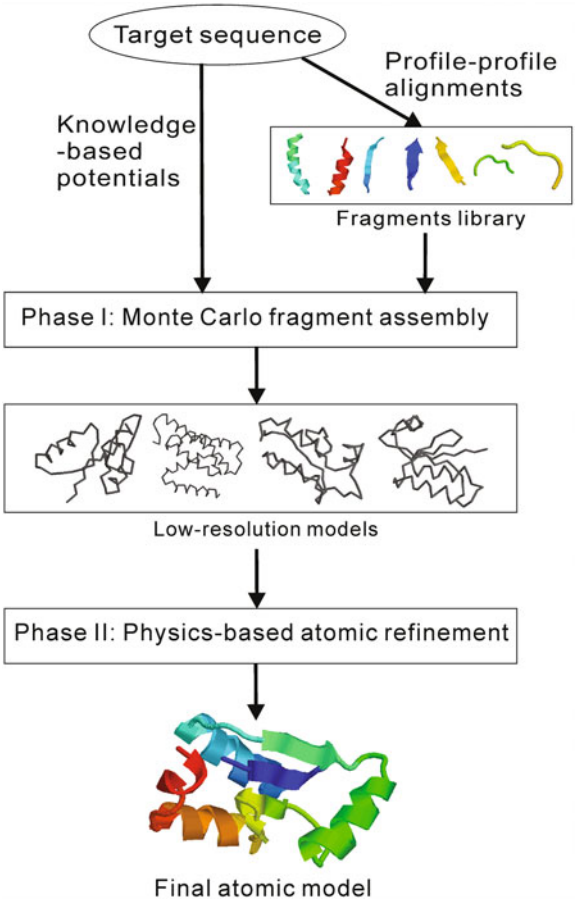



**Fig. 1.2** Flowchart of the ROSETTA protocol (Simons et al. 1997). Fragments are first created from unrelated protein structures in the PDB, which are used to assemble full-length models by simulated annealing simulations guided by a knowledge-based force field. In the second phase, selected models are refined at atomic level using a physics-based potential

and the reassembly process is conducted by parallel-hyperbolic Monte Carlo simulations (Zhang et al. 2002). The energy terms of TASSER include information about predicted secondary structure propensities, backbone hydrogen bonds, a variety of short- and long-range correlations and hydrophobic energy based on the structural statistics from the PDB library. Weights of knowledge-based energy terms are optimized using a large-scale structure decoy set (Zhang et al. 2003) which coordinates the complicated correlations between various interaction terms.

Several derivatives of the TASSER approach have also found independent success. One is Chunk-TASSER (Zhou and Skolnick 2007), which first splits the target sequences into subunits (or "chunks"), each containing 3 consecutive regular secondary structure elements (helix and strand). These chunks are then folded separately. Finally, the spatial restraints are extracted from the chunk models and used for the subsequent TASSER simulations.

Another notable development is I-TASSER by Zhang and coworkers (Wu et al. 2007; Roy et al. 2010, Yang et al. 2015a, b), which refines TASSER cluster centroids by iterative fragment assembly simulations. The spatial restraints are extracted from the first round TASSER models and the template structures searched by TM-align (Zhang and Skolnick 2005a, b) from the PDB library, which are exploited in the second round simulations (Zhang and Skolnick 2013). The purpose is to remove the steric clashes from the first round models and refine the topology. Although the procedure uses structural fragments and spatial restraints from threading templates, it often constructs models of correct topology even when topologies of constituting templates are incorrect. From CASP7 to the latest CASP11 experiments, I-TASSER was consecutively ranked as one of the best methods for automated protein structure prediction (Battey et al. 2007; Cozzetto et al. 2009; Mariani et al. 2011; Montelione 2012; Kinch et al. 2015). As an independent test, Helles carried out a comparative study on 18 ab initio prediction algorithms and concluded that I-TASSER is about the best method in terms of the modeling accuracy and CPU cost per target (Helles 2008). Figure 1.3a shows an example of successful ab initio structure modeling by I-TASSER that constructed a correct model for the FM target T0604, which has a TM-score = 0.701 and RMSD = 2.66 Å from the X-ray structure.

Recently, many efforts have been made to improve the I-TASSER force field by the integration of sequence-based contact prediction (Wu and Zhang 2008a, b), short- and medium-range contact maps derived from segmental threading (Wu and Zhang 2010) and structure alignments (Zhang et al. 2011); these components have been proven particularly important for modeling distant-homology proteins in the CASP experiments (Zhang 2009; Xu et al. 2011; Zhang 2014; Zhang et al. 2015). The flowchart of current I-TASSER protocol is depicted in Fig. 1.4.

**QUARK** QUARK is a recently developed ab initio structural prediction method built on continuous fragment assembly using both knowledge and physics based energy terms (Xu and Zhang 2012). The flowchart of QUARK is shown in Fig. 1.5, which starts from position-specific fragment structure generation. At each residue position, 4000 (=200 × 20) structural fragments are generated, with lengths ranging from 1 to 20 residues, based on gapless threading of the fragment sequence through a non-redundant set of 6023 high-resolution PDB structures. The scoring function of the gapless threading consists of profile-profile, secondary structure, torsion angle and solvent accessibility matches (Wu et al. 2008a, b). Two types of information are derived from the fragments to assist next step of structure folding simulations. First, a torsion angle ($\varphi$, $\Psi$) distribution is collected from the 10-mer fragments at each residue position. Second, a residue-residue contact map is derived from the distance profiles between fragments. Here, a distance ($d_{ij}$) is recorded for each pair of fragments at two positions ($i$ and $j$) if these two fragments come from the same PDB structure. A histogram is then generated for $d_{ij}$ counting distances for all such fragment pairs. If the histogram of $d_{ij}$ has a non-trivial peak below 9 Å, a contact between residue $i$ and $j$ will be predicted (Xu and Zhang 2013).

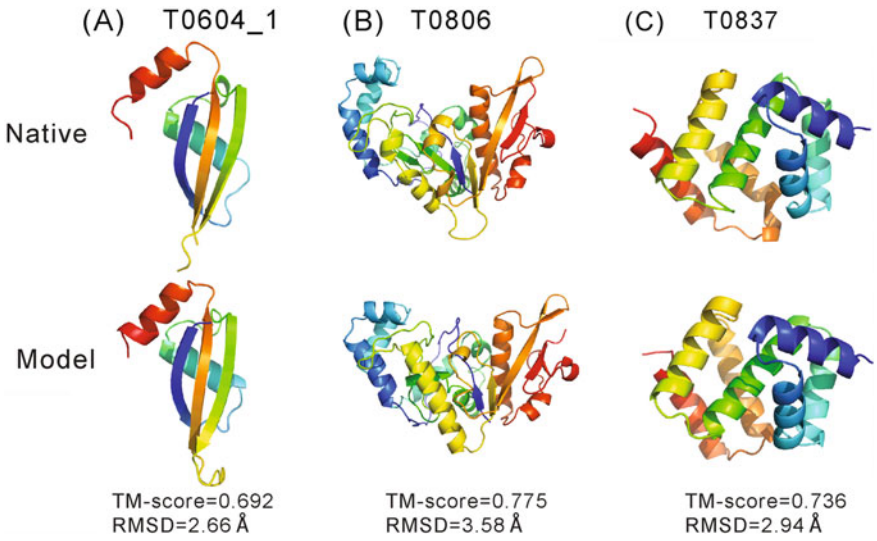



**Fig. 1.3** Three examples of successful free modeling (FM) in recent CASP experiments. **a** T0604_1 is the first domain of the VP0956 protein from *Vibrio parahaemolyticus* in CASP9 that has 79 residues. The first model by the I-TASSER server has a TM-score = 0.692 and Cα-RMSD = 2.66 Å to the native. The success of this target was partially due to the sequence-based contact map prediction (Xu et al. 2011). **b** T0806 is the YaaA protein from *E. coli* K-12 in CASP11 that has 258 residues. The Rosetta human group (Ovchinnikov et al. 2015) constructed a correct model, using a co-evolution based contact prediction derived from >1100 homologous sequences, which has a TM-score = 0.775 and Cα-RMSD = 3.58 Å to the experimental structure. **c** T0837 is a hypothetical protein (YPO2654) from *Yersinia pestis* CO92 with 128 residues. The QUARK server generated a correct model with a TM-score = 0.736 and Cα-RMSD = 2.94 Å to the native, the success of which was attributed to the distance-profile based contact map prediction (Zhang et al. 2015). According to the assessors (Kinch et al. 2011; Kinch et al. 2015), there were no proteins in the PDB with a similar fold to any of these three targets at the time the predictions were made

In the next step, replica-exchange Monte Carlo (REMC) simulations are performed to assemble the fragments into full-length models under a composite physics- and knowledge-based potential, containing hydrogen bonding, van der Waals, solvation, Coulomb, backbone-torsion, bond-length and bond-angle, atomic distance, and strand pairing. The conformational changes are driven by 11 local and global movements shown in the top-right panel of Fig. 1.5. While the first feature, the torsion-angle distribution as collected from the fragments, is used to constrain local torsion movement selection, the second feature, the contact map derived from the fragment distance profiles, is used as a restraint to guide the simulations. The final models are selected by SPICKER (Zhang and Skolnick 2004a, b), which clusters all the decoys generated in the REMC simulations and ranks models by the size of the clusters.

Since its development, QUARK has been consistently ranked as one of the best methods in CASP for ab initio structure prediction (Kinch et al. 2011; Tai et al. 2014;

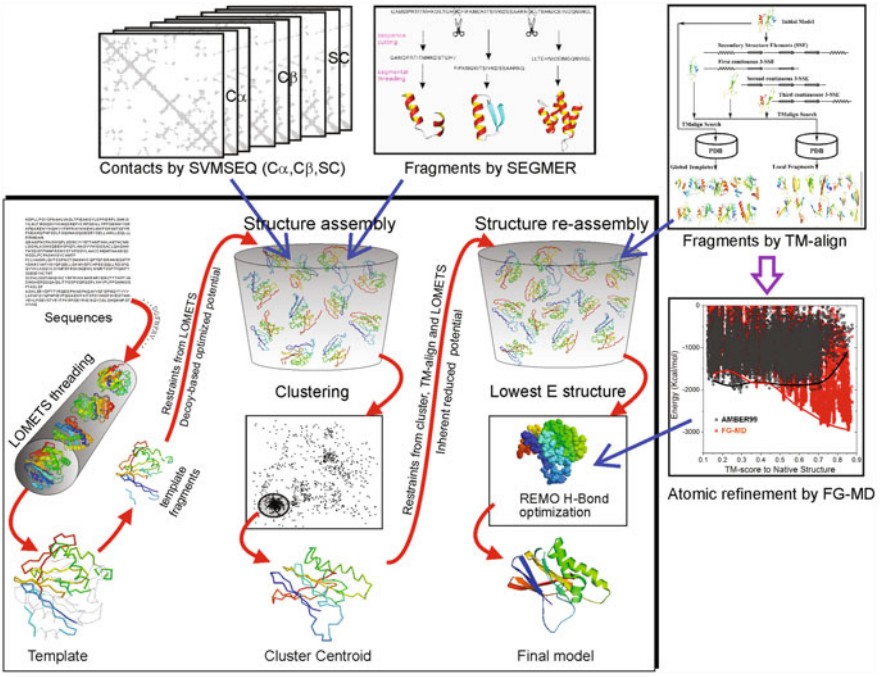



**Fig. 1.4** Flowchart of I-TASSER protein structure modeling (Yang et al. 2015a, b). Multiple threading programs are used to identify templates and super-secondary structure fragments. Segments excised from the continuously aligned regions are used to reassemble the full-length models with the threading-unaligned regions built by lattice-based ab initio simulations. In the next step, templates structurally similar to the first-round models are identified from the PDB by structure alignments, with spatial restraints extracted from the templates to assist the second-round refinement simulations. In recent developments, sequence-based contact predictions and segmental threading were developed for improving results for distant homology modeling

Kinch et al. 2015). Figure 1.3c shows an example of the QUARK server modeling on T0837 in CASP11, where the distance profiles provided correct contacts for some of the critical medium-range contacts, which resulted in the first predicted models with a TM-score = 0.736 and RMSD = 2.94 Å to the experimental X-ray structure.

**Coupling of Contact Prediction And Ab Initio Structure Prediction** Sequence-based contact predictions have recently been investigated for improving ab initio modeling (Wu and Zhang 2008a, b; Wu et al. 2011; Marks et al. 2012; Kosciolek and Jones 2014). Unlike template-based protein structure prediction where high accuracy contacts can be derived from homologous structural templates, the CASP experiments for hard free-modeling (FM) protein targets show that purely sequence-based contact predictions can be more helpful than those collected from the best template-based models because the latter often have low quality for FM (Ezkurdia et al. 2009).

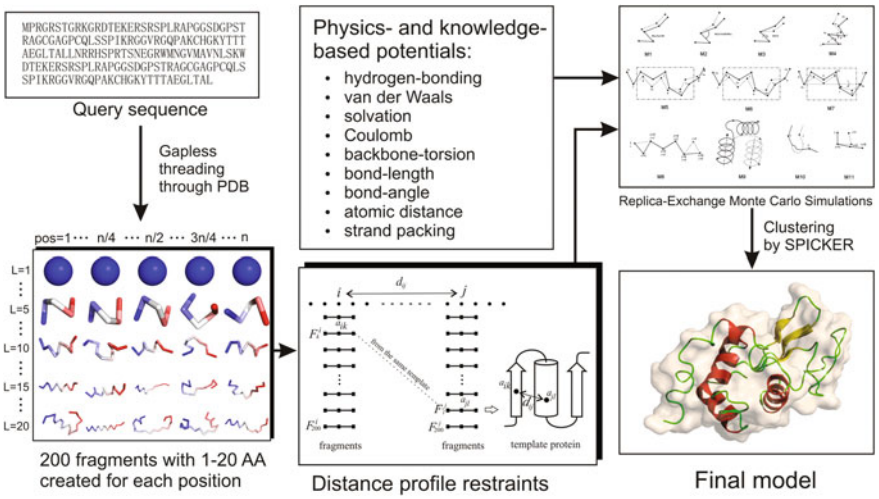



**Fig. 1.5** Flowchart of QUARK protein structure modeling (Xu and Zhang 2012). Multiple fragments with continuously distributed lengths are identified at each position from unrelated protein structures. Contact maps are then collected from distance profiles of the structural fragments, which are used to assist the fragment assembly simulations. Decoys are generated by replica-exchange Monte Carlo simulations under the guide of a composite physics and knowledge-based force field, with the final model selected by structure clustering

Some improvement of final models, with an average TM-score increase by 4.6%, was previously observed by Wu et al. after integrating nine SVM-based contact predictors (3 distance cutoffs multiplying 3 different contact atoms) into the I-TASSER force field (Wu et al. 2011). A handful of targets were converted from "nonfoldable" to "foldable" by several critical contacts when incorporated with the state-of-the-art structure assembly simulation methods. Similarly, Marks et al. (2011) showed that by integrating co-evolution based contact predictions with distance geometry programs, correct folds with RMSD values of 2.7–4.8 Å were generated for 15 test proteins with lengths between 50 and 260 residues. Later, Jones and coworker combined PSICOV (Jones et al. 2012), a co-evolution based contact predictor, with the fragment assembly program (Fragfold) and demonstrated the ability to fold 80% of cases with a TM-score above 0.5, when tested on a set of 150 proteins up to 266 amino acids in length (Kosciolek and Jones 2014).

One of the issues in applying co-evolution based contact predictions to ab initio structure prediction is that the accuracy of contact predictions depends on the number of homologous sequences that can be retrieved from the sequence databases, whereas hard FM targets often have few closely homologous sequences. Most recently, Baker and coworkers (Ovchinnikov et al. 2015) demonstrated an exciting achievement in the blind CASP11 experiment, where 4.6 $L$ homologous sequences (with $L$ being the protein length) were detected for a 256-residue FM target. The combination of the contact map with Rosetta simulations resulted in a

first predicted model with the correct fold, with a TM-score = 0.775 and RMSD = 3.58 Å to the experimental structure (Fig. 1.3b). This probably represents the largest target that has been successfully folded in the CASP experiments, demonstrating the power of coupling contact map prediction and knowledge-based structure modeling.

## 1.3 Conformational Search Methods

Successful ab initio modeling of protein structures depends on the availability of a powerful conformation search method which can efficiently find the global minimum energy structure for a given energy function with a complicated energy landscape. Historically, Monte Carlo and molecular dynamics are two popular simulation methods to explore the conformational space of macromolecules such as proteins. For complicated systems like proteins, canonical MD/MC methods usually require a huge amount of computational resources for a complete exploration of the conformational space. The record for direct application of MD to obtain the protein native structure is not so impressive. One explanation for the failure could be that the simulation time required to fold a small protein takes as long as milliseconds, $10^{12}$ times longer than the usual incremental time step of femtoseconds ($10^{-15}$ s). The technical difficulty of MC simulations mainly comes from that the energy landscape of protein conformational space is typically quite rugged containing many energy barriers, which may easily trap the Metropolis-based MC simulation procedures (Metropolis et al. 1953).

In this section we discuss recent development in conformational search methods to overcome these problems. We intend to illustrate the key ideas of conformational search methods used in various ab initio and related protein-modeling procedures. Unlike various energy functions used in ab initio modeling, the search methods should be, in principle, transferable between protein modeling methods, as well as other problems in science and technology. Currently, there exists no single omni-powerful search method that outperforms the others for all cases, and the investigation and systematic benchmarking on the performance of various search methods has yet to be carried out.

### *1.3.1 Monte Carlo Simulations*

Simulated annealing (SA) (Kirkpatrick et al. 1983) is probably the most popular conformational search method. SA is general in that it is easy and straightforward to apply to any kind of optimization problem. In SA, one typically applies the Metropolis MC algorithm to generate a series of conformational states following the canonical Boltzmann energy distribution for a given temperature. SA initially executes high temperature MC simulation, followed by a series of simulations

subject to a temperature-lowering schedule, hence the name simulated annealing. As much as SA is simple, its conformational search efficiency is not so impressive compared to other more sophisticated methods discussed below.

When the energy landscape of the system under investigation is rugged (due to numerous energy barriers), MC simulations are prone to get stuck in meta-stable states that will distort the distribution of sampled states by breaking the ergodicity of sampling. To avoid this malfunction, many simulation techniques have been developed, and one of the successful approaches is based on the generalized ensemble approach in contrast to the usual canonical ensemble. This kind of method was initially called by different names including multi-canonical ensemble (Berg and Neuhaus 1992) and entropic ensemble (Lee 1993). The underlying idea is to expedite the transition between states separated by energy barriers by modifying the transition probability so that the final energy distribution of sampling becomes more or less flat rather than bell-shaped. A popular method similar in this spirit is the replica exchange MC method (REM) (Swendsen and Wang 1986) where a set of many canonical MC simulations with temperatures distributed in a selected range are simultaneously carried out. From time to time one attempts to exchange structures (or equivalently temperatures) from neighboring simulations to sample states in a wide range of energy spectrum as the means to overcome energy barriers. Parallel hyperbolic sampling (PHS) (Zhang et al. 2002) further extends the REM by dynamically deforming energy using an inverse hyperbolic sine function to lower the energy barrier.

Monte Carlo with minimization (MCM), proposed by Li and Scheraga (1987), was successfully applied for the conformational search by several structure prediction programs (Simons et al. 1997). In MCM, one performs MC moves between local energy minima after local energy minimization of each perturbed protein structure. For a given local energy minimum structure A, a trial structure B is generated by random perturbation of A and is subsequently subject to local energy minimization. The usual Metropolis algorithm is used to determine the acceptance of B over A by calculating the energy difference between the two.

### *1.3.2 Molecular Dynamics*

MD simulation (discussed in detail in Chap. 12) propagates physically realistic trajectories by applying Newton's equations of motion iteratively to allow atom movement, and is thus the most faithful method depicting atomistically what is occurring in proteins. The method is therefore often used for the study of protein folding pathways (Duan and Kollman 1998; Freddolino et al. 2010). The massive computational cost of long simulations is a major challenge with this method, since the incremental time scale is usually in the order of femtoseconds ($10^{-15}$ s) while the fastest folding time of small proteins are on timescales of several microseconds (for folding model systems) or in the millisecond range (more typically). From the standpoint of search efficiency, molecular dynamics simulations are guaranteed to

propagate some motion after each energy/force evaluation, but the steps that are taken are very small; in contrast, as described in the preceding section, Monte Carlo simulations may make larger steps, but not all steps will be accepted after energy evaluation. The relative sampling efficiency of the methods is thus dependent on the acceptance rate of Monte Carlo moves; with modern move sets (see, e.g., Fig. 1.5) Monte Carlo sampling of protein conformational space tends to be much more efficient. Thus, the application of molecular dynamics simulations using atomistic models is reserved for cases where the topic of interest is the folding process, rather than the folded structure per se. One unusual strength of MD sampling compared with Monte Carlo is that MD can accommodate the presence of explicit water much more readily, which might prove useful in the rare cases where implicit solvent models are directly responsible for failed structure predictions (Zhou 2003).

In addition, molecular dynamics simulations have been successfully applied in protein structure prediction using a variety of coarse-grained models, in which the computational complexity is substantially reduced and the folding accelerated due to the simulation of a smaller system with a less rugged energetic landscape, but of course with reduced resolution (Tozzini 2005; Hills and Brooks 2009). In addition, when a low-resolution model is available, MD simulations are often carried out for structure refinement since the conformational changes are assumed to be small (Zhang et al. 2011; Mirjalili and Feig 2013). Sampling in molecular dynamics simulations of protein folding may be enhanced using similar methods to those in Monte Carlo simulations, *e.g.* through the use of replica exchange simulations (Sugita and Okamoto 1999), but at the price of complicating the interpretation of folding kinetics and pathways. One particularly promising enhanced sampling method for future protein folding simulations and structure prediction is accelerated molecular dynamics (aMD) (Hamelberg et al. 2004), which applies a bias to lower the relative height of barriers on the potential energy surface. In a recent application, aMD allowed the prediction of the folded structures and folding free energy landscapes of a set of four commonly used model proteins with 10–100 fold less computational effort than unbiased simulations (Miao et al. 2015), providing promise for future applications to study folding pathways and equilibriums.

### *1.3.3 Genetic Algorithm*

A genetic algorithm (GA) is a heuristic approach to the optimization problems based on a natural selection process mimicking the biological evolution. GA is designed to repeatedly modify a population of individual solutions. At each step, the algorithm randomly selects individuals from the current population, which are used as parents to produce the children for the next generation. Over successive generations, the population "evolves" toward the optimal solutions (Mitchell 1996).

Conformational space annealing (CSA) (Lee et al. 1998) is one of the most successful genetic algorithms developed for protein conformational search. By utilizing a local energy minimizer as in MCM and the concept of annealing in

conformational space, it searches the whole conformational space of local minima in its early stages and then narrows the search to smaller regions with low energy as the distance cutoff is reduced. Here the distance cutoff is defined as the similarity between two conformations, and it controls the diversity of the conformational population. The distance cutoff plays the role of temperature in the usual SA, and initially its value is set to a large number in order to force conformational diversity. The value is gradually reduced as the search progresses. CSA has been successfully applied to various global optimization problems including protein structure prediction separately combined with ab initio modeling in UNRES (Oldziej et al. 2005) and ASTRO-FOLD (Klepeis and Floudas 2003; Klepeis et al. 2005), and with fragment assembly in Profesy (Lee et al. 2004).

### *1.3.4 Mathematical Optimization*

The conformational searching approach by Floudas and coworkers, α branch and bound (αBB) (Klepeis and Floudas 2003; Klepeis et al. 2005), is unique in the sense that the method is mathematically rigorous, while all the others discussed here are stochastic and heuristic methods. The search space is successively cut into two halves while the lower and upper bounds of the global minimum (LB and UB) for each branched phase space are estimated. The estimate for the UB is simply the best currently obtained local minimum energy, and the estimate for the LB comes from the modified energy function augmented by a quadratic term of the dissecting variables with the coefficient α (hence the name αBB). With a sufficiently large value of α, the modified energy contains only one energy minimum, whose value serves as the lower bound. While performing successive dissection of the phase space accompanied by estimates of LB and UB for each dissected phase space, phase spaces with LB higher than the global UB can be eliminated from the search. The procedure continues until one identifies the global minimum by locating a dissected phase space where LB becomes identical to the global UB. Once the solution is found, the result is mathematically rigorous, but large proteins with many degrees of freedom are yet to be addressed by this method.

## 1.4 Model Selection

Ab initio modeling methods typically generate many non-native structure conformations (also called decoys) during the simulation. How to select appropriate models structurally close to the native state is an important issue. The development of algorithms for selection of protein models has been emerged as a new field called Model Quality Assessment Programs (MQAP) (Fischer 2006). In general, modeling selection approaches can be classified into two types, the energy based and the free energy based. In the energy-based methods, one designs a variety of specific

potentials and identifies the lowest-energy state as the final prediction. In the free-energy based approaches, the free energy of a given conformation $R$ can be written as

$$F(R) = -k_B T \ \ln Z(R) = -k_B T \ \ln \int_{\Omega \in R} \mathrm{e}^{\frac{-E(R)}{k_B T}} \mathrm{d}\Omega \tag{1}$$

where $Z(R)$ is the restricted partition function which is proportional to the number of occurrences of the structures in the neighborhood of $R$ during the simulation. This can be estimated by a clustering procedure at a given RMSD cutoff (Zhang and Skolnick 2004a, b).

For the energy-based model selection methods, we will discuss three energy/scoring functions: (1) physics-based energy function; (2) knowledge-based energy function; (3) scoring function describing the compatibility between the target sequence and model structures. In MQAP, there is another popular method which takes the consensus conformation from the predictions generated by different algorithms (Wallner and Elofsson 2007), also known as the meta-server approach (Ginalski et al. 2003; Wu et al. 2007). The essence of this method is similar to the clustering approach since both assume the most frequently occurring states to be the near-native structures. This approach has been mainly used for selecting models generated by threading-servers (Ginalski et al. 2003; Wu et al. 2007); but it has recently become popular for full-length model selection in the CASP experiments (Larsson et al. 2009; Kryshtafovych et al. 2015).

### *1.4.1 Physics-Based Energy Function*

For the development of all-atom physics-based energy functions, Lazaridis and Karplus (1999a, b) exploited CHARMM19 (Neria et al. 1996) and EEF1 (Lazaridis and Karplus (1999a, b)) solvation potential to discriminate the native structure from decoys that are generated by threading on other protein structures. They found the energy of the native state is lower than those of decoys in most cases. Later, Petrey and Honig (Petrey and Honig 2000) used CHARMM and a continuum treatment of the solvent, Brooks and coworkers (Dominy and Brooks 2002; Feig and Brooks 2002) used CHARMM plus GB solvation, Felts et al. (2002) used OPLS plus GB, Lee and Duan (Lee et al. 2004) used AMBER plus GB, and Hsieh and Luo (2004) used AMBER plus Poisson-Boltzmann solvation potential on a number of structure decoy sets (including the Park-Levitt decoy set (Park and Levitt 1996), Baker decoy set (Tsai et al. 2003), Skolnick decoy set (Kihara et al. 2001; Skolnick et al. 2003), I-TASSER decoy set (Wu et al. 2007; Zhang and Zhang 2010), and CASP decoys set (Moult et al. 2001)). All these authors obtained similar results: the native structures have lower energy than decoys in their potentials. The claimed success of model discrimination of the physics-based potentials seems contradicted by other

less successful physics-based structure prediction results. Wroblewska and Skolnick (Wroblewska and Skolnick 2007) showed that the AMBER plus GB potential can only discriminate the native structure from roughly minimized TASSER decoys (Zhang and Skolnick 2004a, b). After a 2-ns MD simulation on the decoys, none of the native structures were lower in energy than the lowest energy decoy, and the energy-RMSD correlation was close to zero. This result partially explains the discrepancy between the widely reported decoy discrimination ability of physics-based potentials and the less successful folding/refinement results.

Another related issue is that many of the decoy selection approaches are focused on the discrimination of the native structures from the decoy pools. However, such ability is of no practical usefulness in real cases of structure prediction because no structure prediction simulation could generate decoys exactly matching the native structure. Furthermore, the native structure has usually a nearly perfect local secondary structure packing, in addition to the fitness of global topology arrangement, whereas the computer generated decoys often have various flaws in the local structure packing and steric clashes. This makes it much more challenging to recognize the near-native structure decoys that are structurally closest to the native, compared to the task of discriminating the native structure from a set of computer-generated, flawed structure decoys (Deng et al. 2016).

### *1.4.2 Knowledge-Based Energy Function*

Sippl proposed a pair-wise residue-distance based potential (Sippl 1990) using the statistics of known PDB structures in 1990 (its newest version is PROSA II (Sippl 1993; Wiederstein and Sippl 2007)). Since then, a variety of knowledge-based potentials have been developed, which include atomic interaction potential, solvation potential, hydrogen bond potential, torsion angle potential, etc. In the coarse-grained potentials, each residue is represented either by a single atom or by a few atoms, e.g., Cα-based potentials (Melo et al. 2002), Cβ-based potentials (Hendlich et al. 1990), side-chain-center-based potentials (Bryant and Lawrence 1993; Kocher et al. 1994; Thomas and Dill 1996; Skolnick et al. 1997; Zhang and Kim 2000; Zhang and Skolnick 2004a, b), side-chain and Cα-based potentials (Berrera et al. 2003).

One of the most widely-used knowledge-based potentials is a residue-specific, all-atom, distance-dependent potential, which was first formulated by Samudrala and Moult (RAPDF) (Samudrala and Moult 1998); it counts the distances between 167 amino acid specific pseudo-atoms. Following this, several atomic potentials with various reference states have been proposed, including those by Lu and Skolnick (KBP) (Lu and Skolnick 2001), Zhou and Zhou (DFIRE) (Zhou et al. 2002), Wang et al. (self-RAPDF) (Wang et al. 2004), Tostto (victor/FRST) (Tosatto 2005), Shen and Sali (DOPE) (Shen and Sali 2006), Zhang and Zhang (RW) (Zhang and Zhang 2010), and Zhou and Skolinck (GOAP) (Zhou and Skolnick 2011). All these potentials claimed that native structures could be

distinguished from decoy structures in their tests. Deng et al. (2012) recently conducted a comparative investigation on all these potentials. To eliminate biases from the datasets and computing environments, they re-derived the potentials from a unified PDB structure dataset but based on the same original reference states. It was found that the performance varies with the tested decoy datasets and no potential could clearly outperform the others for all decoy sets.

The task of selecting the near-native models out of many decoys remains a challenge for these potentials (Skolnick 2006). Based on the CAFASP4-MQAP experiment in 2004 (Fischer 2006), the best-performing energy functions were Victor/FRST (Tosatto 2005) which incorporates an all-atom pair-wise interaction potential, solvation potential and hydrogen bond potential, and MODCHECK (Pettitt et al. 2005) which includes Cβ atom interaction potential and solvation potential. From CASP7-MQAP in 2006, the consensus-based method, Pcons developed by Elofsson group, showed the best performance (Wallner and Elofsson 2007). In the most recent CASP experiments, the consensus-based model selection scheme has kept ranking higher than any of the physics or knowledge-based scoring functions (Kryshtafovych et al. 2011; Kryshtafovych et al. 2014; Kryshtafovych et al. 2015). Several of the advanced structure modeling approaches in the CASP experiment have exploited a combined consensus and statistics scoring system to select models in the recent CASP (Cao et al. 2015; Yang et al. Yang et al. 2015a, b; Zhang et al. 2015).

### *1.4.3 Sequence-Structure Compatibility Function*

In the third type of MQAPs, selection of the best models is not purely based on energy functions. Instead, they are selected based on the compatibility of target sequences to model structures. The earliest and still successful example is that by Luthy et al. (1992), who used threading scores to evaluate structures. Colovos and Yeates (1993) later used a quadratic error function to describe the non-covalently bonded interactions among atom pairs CC, CN, CO, NN, NO and OO, showing that near-native structures have fewer errors than other decoys. Verify3D (Eisenberg et al. 1997) improves the method of Luthy et al. (Luthy et al. 1992) by considering local threading scores in a 21-residue window. Jones developed GenThreader (Jones 1999) and used neural networks to classify native and non-native structures. The inputs of GenThreader include pairwise contact energy, solvation energy, alignment score, alignment length, and sequence and structure lengths. Similarly, based on neural networks, Wallner and Ellofsson built ProQ (Wallner and Elofsson 2003) for quality prediction of decoy structures. The inputs of ProQ include contacts, solvent accessible area, protein shape, secondary structure, structural alignment score between decoys and templates, and the fraction of protein regions to be modeled from templates. Later, McGuffin developed a consensus MQAP

(McGuffin 2007) called ModFold that includes ProQ (Wallner and Elofsson 2003), MODCHECK (Pettitt et al. 2005) and ModSSEA. The author showed that ModFold outperforms its component MQAP programs.

### *1.4.4 Clustering of Decoy Structures*

For the purpose of identifying the lowest free-energy state, structure clustering techniques were adopted by many ab initio modeling approaches. In the work by Shortle et al. (1998), for all 12 cases tested, the cluster-center conformation of the largest cluster was closer to native structures than the majority of decoys. Cluster-center structures were ranked as the top 1–5% closest to their native structures.

Zhang and Skolnick developed an iterative structure clustering method, called SPICKER (Zhang and Skolnick 2004a, b). Based on 1489 representative benchmark proteins each with up to 280,000 structure decoys, the best of the top 5 models was ranked in the top 1.4% of all decoys. For 78% of the 1489 proteins, the RMSD difference between the best of the top 5 models and the most native-like decoy structure was less than 1 Å.

In ROSETTA ab initio modeling (Bradley et al. 2005a, b), structure decoys are clustered to select low-resolution models and these models are further refined by all-atom simulations to obtain final models. In the case of TASSER/I-TASSER (Zhang and Skolnick 2004a, b; Yang et al. 2015a, b) and QUARK (Xu and Zhang 2012), thousands of decoy models from MC simulations are clustered by SPICKER (Zhang and Skolnick 2004a, b) to generate cluster centroids as final models. In the approach by Scheraga and coworkers (Oldziej et al. 2005), decoys are clustered and the lowest-energy structures among the clustered structures are selected.

## 1.5 Remarks and Discussions

Successful ab initio modeling from amino acid sequence alone is considered the "Holy Grail" of protein structure prediction (Zhang 2008), since this will mark an eventual and complete solution to the problem. In addition to the generation of 3D structures, ab initio modeling can also help us understand the underlying principles of how proteins fold in nature; this could not be done by the template-based modeling approaches which build 3D models by copying and refining the framework of other solved structures.

An ideal approach to ab initio modeling would be to treat atoms in a protein as interacting particles according to an accurate physics-based potential, and fold the protein by solving Newton's equations of motion in each step of movements. A number of molecular dynamics simulations were carried out along this line of approach by using the classic CHARMM and AMBER force fields. Although the

MD based simulation is very important for the study of protein folding, the success in the viewpoint of structure prediction is quite limited. One reason is the prohibitive computing demand for a normal size protein. On the other hand, knowledge-based (or hybrid knowledge- and physics-based) approaches making use of Monte Carlo sampling schemes appear to be progressing rapidly, producing many examples of successful low-to-medium accuracy models often with correct topology for small and medium size proteins. Although very rare, successful higher resolution models (<2–3 Å in Cα-RMSD) have been witnessed in blind experiments (Bradley et al. 2005a, b; Xu et al. 2011; Zhang et al. 2015).

The current state-of-the-art ab initio protein structure prediction methods often utilize as much information as possible from known structures, in several different ways. First, the use of local structure fragments directly excised from the PDB structures helps reduce the degrees of freedom and the entropy of the conformational search and yet keep the fidelity of the native protein structures. Second, the knowledge-based potential derived from the statistics of a large number of solved structures can appropriately grasp the subtle balance of the complicated correlations between different sources of energy terms (Summa and Levitt 2007). With the carefully parameterized knowledge-based potential terms aided by various advances in the conformational search methods, the accuracy of ab initio modeling for proteins up to 100–150 residues has been significantly improved in the last decade. With the help of co-evolution based contact map predictions, an exciting examples has been recently reported on a free-modeling target (T0806) up to 258 residues in the most recent CASP experiment (Ovchinnikov et al. 2015). However, such performance is only possible when sufficient number of homologous sequences can be obtained to ensure the accuracy of contact predictions: this situation is rare for ab initio modeling target proteins that have no homologues in the PDB.

For further improvement, parallel developments of accurate potential energy functions and efficient optimization methods are both necessary. That is, separate examination/development of potential energy functions is important; meanwhile, systematic benchmarking of various conformational search methods should be performed, so that the advantages as well as the limitations of available search methods can be explored separately. Currently, the ab initio modeling methods solely based on the physicochemical principles of interaction are still far behind, in terms of their modeling speed and accuracy, compared with the methods utilizing bioinformatics and knowledge-based information. However, the physics-based atomic potentials have recently demonstrated their potential in refining the detailed packing of side-chain atoms and peptide backbones (Zhang et al. 2011; Mirjalili and Feig 2013). Development of composite methods using both knowledge-based and physics-based energy terms should represent a promising approach to the problem of ab initio modeling.

It is important to acknowledge that with the progress in structure genomics and structural biology, the number of experimental structures in the PDB has been rapidly increasing, significantly extending the scope of the template-based protein structure predictions. Nevertheless, the traditional comparative modeling approaches can only yield model predictions with the accuracy of the templates,

whereas the efficiency of template structure refinements is highly correlated with our ability in ab initio protein folding, because structure refinements often involve reconstruction of part of the side-chain and local backbone structures, and sometime the global topology for the low-resolution templates. Meanwhile, for most templates available in the PDB, a considerable portion of the sequence is either disordered or unaligned in the query-template alignments; the structures of these portions must be constructed using ab initio modeling. Finally, a very important bottleneck drawback in template-based modeling is that the alignment accuracy dramatically decreases with the sequence identity between query and template becomes low (e.g. <30%). Most recently, it has been demonstrated that the structural models built by free modeling can be used to help identify analogous templates that are of low sequence similarity but high structural similarity to the native, by matching the low-resolution ab initio models to experimentally solved structures in the PDB and thereby improve the success rate of distant-homologous structure predictions (Zhang 2014). Thus, the development of efficient ab initio folding algorithms will remain a major theme in the field and should have important impacts on all aspects of protein structure prediction.

**Acknowledgements** Authors want to thank Drs. Sitao Wu and Haiyou Deng for their contribution to the article. The project is supported in part by the National Institute of General Medical Sciences (GM083107, GM116960, and GM097033).

## References

Bairoch A, Apweiler R, Wu CH et al (2005). The universal protein resource (UniProt). Nucleic Acids Res 33(Database issue): D154–159

Battey JN, Kopp J, Bordoli L et al (2007) Automated server predictions in CASP7. Proteins 69 (S8):68–82

Berendsen HJC, Postma JPM, van Gunsteren WF et al (1981) Interaction models for water in relation to protein hydration. Intermolecular forces, Reidel, The Netherlands

Berg BA, Neuhaus T (1992) Multicanonical ensemble: a new approach to simulate first-order phase transitions. Physical Review Letters 68(1):9–12

Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Research 28(1):235–242

Berrera M, Molinari H, Fogolari F (2003) Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. BMC Bioinform 4:8

Best RB, Buchete NV, Hummer G (2008) Are current molecular dynamics force fields too helical? Biophysical Journal 95(1):L07–09

Best RB, Hummer G (2009) Optimized molecular dynamics force fields applied to the helix-coil transition of polypeptides. J Phys Chem B 113(26):9004–9015

Bowie JU, Eisenberg D (1994) An evolutionary approach to folding small alpha-helical proteins that uses sequence information and an empirical guiding fitness function. Proc Natl Acad Sci U S A 91(10):4436–4440

Bradley P, Malmstrom L, Qian B et al (2005a) Free modeling with Rosetta in CASP6. Proteins 61(Suppl 7):128–134

Bradley P, Misura KM, Baker D (2005b) Toward high-resolution de novo structure prediction for small proteins. Science 309(5742):1868–1871

Brooks BR, Bruccoleri RE, Olafson BD et al (1983) CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. Journal of Computational Chemistry 4(2): 187–217

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16(1):92–112

Cao R, Bhattacharya D, Adhikari B et al (2015). Massive integration of diverse protein quality assessment methods to improve template based modeling in CASP11. Proteins 84:247–259

Case DA, Pearlman DA, Caldwell JA et al (1997). AMBER 5.0, University of California, San Francisco

Chen J, Brooks CL 3rd (2007) Can molecular dynamics simulations provide high-resolution refinement of protein structure? Proteins 67(4):922–930

Chowdhury S, Lee MC, Xiong GM et al (2003) Ab initio folding simulation of the Trp-cage mini-protein approaches NMR resolution. Journal of Molecular Biology 327(3):711–717

Colovos C, Yeates TO (1993) Verification of protein structures: patterns of nonbonded atomic interactions. Protein Science 2(9):1511–1519

Cornell WD, Cieplak P, Bayly CI et al (1995) A Second generation force field for the simulation of proteins, nucleic acids, and organic molecules. Journal of the American Chemical Society 117:5179–5197

Cozzetto D, Kryshtafovych A, Fidelis K et al (2009) Evaluation of template-based models in CASP8 with standard measures. Proteins 77(Suppl 9):18–28

Das R, Qian B, Raman S et al (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69(S8):118–128

Deng H, Jia Y, Zhang Y (2016) 3DRobot: automated generation of diverse and well-packed protein structure decoys. Bioinformatics 32(3):378–387

Deng HY, Jia Y, Wei YY et al (2012) What is the best reference state for designing statistical atomic potentials in protein structure prediction? Proteins-Structure Function and Bioinformatics 80(9):2311–2322

Dominy BN, Brooks CL (2002) Identifying native-like protein structures using physics-based potentials. Journal of Computational Chemistry 23(1):147–160

Duan Y, Kollman PA (1998) Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 282(5389):740–744

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods in Enzymology 277:396–404

Ensign DL, Kasson PM, Pande VS (2007) Heterogeneity even at the speed limit of folding: large-scale molecular dynamics study of a fast-folding variant of the villin headpiece. Journal of Molecular Biology 374(3):806–816

Ezkurdia I, Grana O, Izarzugaza JM et al (2009) Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. Proteins 77(Suppl 9):196–209

Fan H, Mark AE (2004) Refinement of homology-based protein structures by molecular dynamics simulation techniques. Protein Science 13(1):211–220

Feig M, Brooks CL 3rd (2002) Evaluating CASP4 predictions with physical energy functions. Proteins 49(2):232–245

Feig M, Mirjalili V (2015). Protein structure refinement via molecular-dynamics simulations: what works and what does not? Proteins 84:282–292

Felts AK, Gallicchio E, Wallqvist A et al (2002) Distinguishing native conformations of proteins from decoys with an effective free energy estimator based on the OPLS all-atom force field and the Surface Generalized Born solvent model. Proteins 48(2):404–422

Fischer D (2006) Servers for protein structure prediction. Current Opinion in Structural Biology 16(2):178–182

Freddolino PL, Harrison CB, Liu Y et al (2010) Challenges in protein folding simulations: timescale, representation, and analysis. Nature Physics 6(10):751–758

Freddolino PL, Liu F, Gruebele M et al (2008) Ten-microsecond molecular dynamics simulation of a fast-folding WW domain. Biophysical Journal 94(10):L75–77

Freddolino PL, Park S, Roux B et al (2009) Force field bias in protein folding simulations. Biophysical Journal 96(9):3772–3780

Freddolino PL, Schulten K (2009) Common structural transitions in explicit-solvent simulations of villin headpiece folding. Biophysical Journal 97(8):2338–2347

Fujitsuka Y, Chikenji G, Takada S (2006) SimFold energy function for de novo protein structure prediction: consensus with Rosetta. Proteins 62(2):381–398

Ginalski K, Elofsson A, Fischer D et al (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19(8):1015–1018

Hagler A, Euler E, Lifson S (1974) Energy functions for peptides and proteins i. derivation of a consistent force field including the hydrogen bond from amide crystals. Journal of the American Chemical Society 96:5319–5327

Hamelberg D, Mongan J, McCammon JA (2004) Enhanced sampling of conformational transitions in proteins using full atomistic accelerated molecular dynamics simulations. Protein Science 13:76

Helles G (2008) A comparative study of the reported performance of ab initio protein structure prediction algorithms. Journal of the Royal Society, Interface 5(21):387–396

Hendlich M, Lackner P, Weitckus S et al (1990) Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. Journal of Molecular Biology 216(1):167–180

Hills RD Jr, Brooks CL 3rd (2009) Insights from coarse-grained go models for protein folding and dynamics. International Journal of Molecular Sciences 10(3):889–905

Hsieh MJ, Luo R (2004) Physical scoring function based on AMBER force field and poisson-boltzmann implicit solvent for protein structure prediction. Proteins 56(3):475–486

Im W, Lee MS, Brooks CL 3rd (2003) Generalized born model with a simple smoothing function. Journal of Computational Chemistry 24(14):1691–1702

Jagielska A, Wroblewska L, Skolnick J (2008) Protein model refinement using an optimized physics-based all-atom force field. Proceedings of the National Academy of Sciences of the United States of America 105(24):8268–8273

Jauch R, Yeo HC, Kolatkar PR et al (2007) Assessment of CASP7 structure predictions for template free targets. Proteins 69(Suppl 8):57–67

Jonassen I, Klose D, Taylor WR (2006) Protein model refinement using structural fragment tessellation. Computational Biology and Chemistry 30(5):360–366

Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. Journal of Molecular Biology 287(4):797–815

Jones DT, Buchan DW, Cozzetto D et al (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190

Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. Journal of the American Chemical Society 118:11225–11236

Jorgensen WL, Tirado-Rives J (1988) The OPLS potential functions for proteins. energy minimizations for crystals of cyclic peptides and crambin. Journal of the American Chemical Society 110:1657–1666

Kaminski GA, Friesner RA, Tirado-Rives J et al (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105:6474–6487

Karplus K, Barrett C, Hughey R (1998) Hidden markov models for detecting remote protein homologies. Bioinformatics 14:846–856

Kihara D, Lu H, Kolinski A et al (2001) TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci U S A 98(18): 10125–10130

Kinch L, Yong Shi S, Cong Q et al (2011) CASP9 assessment of free modeling target predictions. Proteins 79(Suppl 10):59–73

Kinch LN, Li W, Monastyrskyy B, et al. (2015). Evaluation of free modeling targets in CASP11 and ROLL. Proteins 84: 51–66

Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. Science 220 (4598):671–680

Klepeis JL, Floudas CA (2003) ASTRO-FOLD: a combinatorial and global optimization framework for Ab initio prediction of three-dimensional structures of proteins from the amino acid sequence. Biophysical Journal 85(4):2119–2146

Klepeis JL, Wei Y, Hecht MH et al (2005) Ab initio prediction of the three-dimensional structure of a de novo designed protein: a double-blind case study. Proteins 58(3):560–570

Kocher JP, Rooman MJ, Wodak SJ (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. Journal of Molecular Biology 235(5): 1598–1613

Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE 9(3):e92197

Kryshtafovych A, Barbato A, Fidelis K et al (2014) Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 82(Suppl 2):112–126

Kryshtafovych A, Barbato A, Monastyrskyy B, et al (2015) Methods of model accuracy estimation can help selecting the best models from decoy sets: Assessment of model accuracy estimations in CASP11. Proteins 84: 349–369

Kryshtafovych A, Fidelis K, Tramontano A (2011) Evaluation of model quality predictions in CASP9. Proteins 79(Suppl 10):91–106

Larsson P, Skwark MJ, Wallner B et al (2009) Assessment of global and local model quality in CASP8 using Pcons and ProQ. Proteins 77(Suppl 9):167–172

Lazaridis T, Karplus M (1999a) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. Journal of Molecular Biology 288(3):477–487

Lazaridis T, Karplus M (1999b) Effective energy function for proteins in solution. Proteins 35(2): 133–152

Lee J (1993) New monte carlo algorithm: entropic sampling. Physical Review Letters 71(2): 211–214

Lee J, Kim SY, Joo K et al (2004) Prediction of protein tertiary structure using PROFESY, a novel method based on fragment assembly and conformational space annealing. Proteins 56 4):704–714

Lee J, Scheraga HA, Rackovsky S (1998) Conformational analysis of the 20-residue membrane-bound portion of melittin by conformational space annealing. Biopolymers 46 (2):103–116

Lee MC, Duan Y (2004) Distinguish protein decoys by using a scoring function based on a new AMBER force field, short molecular dynamics simulations, and the generalized born solvent model. Proteins 55(3):620–634

Lee MR, Tsai J, Baker D et al (2001) Molecular dynamics in the endgame of protein structure prediction. Journal of Molecular Biology 313(2):417–430

Lei HX, Wu C, Liu HG et al (2007) Folding free-energy landscape of villin headpiece subdomain from molecular dynamics simulations. Proceedings of the National Academy of Sciences of the United States of America 104(12):4925–4930

Levitt M, Hirshberg M, Sharon R et al (1995) Potential-energy function and parameters for simulations of the molecular-dynamics of proteins and nucleic-acids in solution. Computer Physics Communications 91(1–3):215–231

Li Z, Scheraga HA (1987) Monte carlo-minimization approach to the multiple-minima problem in protein folding. Proc Natl Acad Sci U S A 84(19):6611–6615

Lindahl E, Hess B, van der Spoel D (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Modeling 7:306–317

Lindorff-Larsen K, Maragakis P, Piana S et al (2012) Systematic validation of protein force fields against experimental data. PLoS ONE 7(2):e32131

Lindorff-Larsen K, Piana S, Dror RO et al (2011) How fast-folding proteins fold. Science 334(6055):517–520
Lindorff-Larsen K, Piana S, Palmo K et al (2010) Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins 78(8):1950–1958
Liwo A, Khalili M, Scheraga HA (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci U S A 102(7):2362–2367
Liwo A, Lee J, Ripoll DR et al (1999) Protein structure prediction by global optimization of a potential energy function. Proc Natl Acad Sci U S A 96(10):5482–5485
Liwo A, Pincus MR, Wawak RJ et al (1993) Calculation of protein backbone geometry from alpha-carbon coordinates based on peptide-group dipole alignment. Protein Science 2(10): 1697–1714
Lu H, Skolnick J (2001) A distance-dependent atomic knowledge-based potential for improved protein structure selection. Proteins 44(3):223–232
Luthy R, Bowie JU, Eisenberg D (1992) Assessment of protein models with three-dimensional profiles. Nature 356(6364):83–85
MacKerell AD Jr, Bashford D, Bellott M et al (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102(18):3586–3616
Mariani V, Kiefer F, Schmidt T et al (2011) Assessment of template based protein structure predictions in CASP9. Proteins 79(Suppl 10):37–58
Marks DS, Colwell LJ, Sheridan R et al (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766
Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nature Biotechnology 30(11):1072–1080
McGuffin LJ (2007) Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics 8:345
Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Science 11(2): 430–448
Metropolis N, Rosenbluth AW, Rosenbluth MN et al (1953) Equation of state calculations by fast computing machines. J Chem Phys 21:1087–1092
Miao YL, Feixas F, Eun CS et al (2015) Accelerated molecular dynamics simulations of protein folding. Journal of Computational Chemistry 36(20):1536–1549
Mirjalili V, Feig M (2013) Protein structure refinement through structure selection and averaging from molecular dynamics ensembles. Journal of Chemical Theory and Computation 9(2): 1294–1303
Mitchell M (1996). An Introduction to Genetic Algorithms. Cambridge, MIT Press
Mittal J, Best RB (2010) Tackling force-field bias in protein folding simulations: folding of Villin HP35 and Pin WW domains in explicit water. Biophysical Journal 99(3):L26–28
Montelione GT (2012). Template based modeling assessment in CASP10. 10th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction. Gaeta, Italy
Moult J, Fidelis K, Zemla A et al (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins Suppl 5:2–7
Nemethy G, Gibson KD, Palmer KA et al (1992) Energy parameters in polypeptides. 10. Improved geometric parameters and nonbonded interactions for use in the ECEPP/3 algorithm, with application to proline-containing peptides. J Phys Chem B 96:6472–6484
Neria E, Fischer S, Karplus M (1996) Simulation of activation free energies in molecular systems. J Chem Phys 105(5):1902–1921
Nguyen H, Maier J, Huang H et al (2014) Folding simulations for proteins with diverse topologies are accessible in days with a physics-based force field and implicit solvent. Journal of the American Chemical Society 136(40):13959–13962
Nilges M, Brunger AT (1991) Automated modeling of coiled coils: application to the GCN4 dimerization region. Protein Engineering 4(6):649–659

Oldziej S, Czaplewski C, Liwo A et al (2005) Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: assessment in two blind tests. Proceedings of the National Academy of Sciences of the United States of America 102(21): 7547–7552

Ovchinnikov S, Kim DE, Wang RY, et al (2015) Improved de novo structure prediction in CASP11 by incorporating Co-evolution information into rosetta. Proteins 84:67–75

Park B, Levitt M (1996) Energy functions that discriminate X-ray and near native folds from well-constructed decoys. Journal of Molecular Biology 258(2):367–392

Petrey D, Honig B (2000) Free energy determinants of tertiary structure and the evaluation of protein models. Protein Science 9(11):2181–2191

Pettitt CS, McGuffin LJ, Jones DT (2005) Improving sequence-based fold recognition by using 3D model quality assessment. Bioinformatics 21(17):3509–3515

Piana S, Klepeis JL, Shaw DE (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Current Opinion in Structural Biology 24:98–105

Piana S, Lindorff-Larsen K, Shaw DE (2011) How robust are protein folding simulations with respect to force field parameterization? Biophysical Journal 100(9):L47–49

Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. Proceedings of the National Academy of Sciences of the United States of America 109(44):17845–17850

Piana S, Lindorff-Larsen K, Shaw DE (2013a) Atomic-level description of ubiquitin folding. Proc Natl Acad Sci U S A 110(15):5915–5920

Piana S, Lindorff-Larsen K, Shaw DE (2013b) Atomistic description of the folding of a dimeric protein. J Phys Chem B 117(42):12935–12942

Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nature Protocols 5(4):725–738

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. Journal of Molecular Biology 234(3):779–815

Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. Journal of Molecular Biology 275(5):895–916

Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Science 15(11):2507–2524

Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A 95(19):11158–11162

Simons KT, Kooperberg C, Huang E et al (1997) Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and bayesian scoring functions. Journal of Molecular Biology 268(1):209–225

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. an approach to the knowledge-based prediction of local structures in globular proteins. Journal of Molecular Biology 213(4):859–883

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17(4): 355–362

Skolnick J (2006) In quest of an empirical potential for protein structure prediction. Current Opinion in Structural Biology 16(2):166–171

Skolnick J, Jaroszewski L, Kolinski A et al (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Science 6:676–688

Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR 3.0 threading algorithm. Protein 56:502–518

Skolnick J, Zhang Y, Arakaki AK et al (2003) TOUCHSTONE: A unified approach to protein structure prediction. Proteins 53(Suppl 6):469–479

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21(7): 951–960

Sorin EJ, Pande VS (2005) Exploring the helix-coil transition via all-atom equilibrium ensemble simulations. Biophysical Journal 88(4):2472–2493

Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chemical Physics Letters 314(1–2):141–151
Summa CM, Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci U S A 104(9):3177–3182
Swendsen RH, Wang JS (1986) Replica Monte Carlo simulation of spin glasses. Physical Review Letters 57(21):2607–2609
Tai CH, Bai H, Taylor TJ et al (2014) Assessment of template-free modeling in CASP10 and ROLL. Proteins 82(Suppl 2):57–83
Taylor WR, Bartlett GJ, Chelliah V et al (2008) Prediction of protein structure from ideal forms. Proteins 70(4):1610–1619
Thomas PD, Dill KA (1996) Statistical potentials extracted from protein structures: how accurate are they? Journal of Molecular Biology 257(2):457–469
Tosatto SC (2005) The victor/FRST function for model quality estimation. Journal of Computational Biology 12(10):1316–1327
Tozzini V (2005) Coarse-grained models for proteins. Current Opinion in Structural Biology 15(2):144–150
Tsai J, Bonneau R, Morozov AV et al (2003) An improved protein decoy set for testing energy functions for protein structure prediction. Proteins 53(1):76–87
van Gunsteren WF, Billeter SR, Eising AA et al (1996). Biomolecular simulation: The GROMOS96 Manual and User Guide Univ Publ House, Zurich
Vieth M, Kolinski A, Brooks CL et al (1994) Prediction of the folding pathways and structure of the GCN4 leucine zipper. Journal of Molecular Biology 237(4):361–367
Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Science 12(5): 1073–1086
Wallner B, Elofsson A (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 69(S8):184–193
Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? Journal of Computational Chemistry 21(12):1049–1074
Wang K, Fain B, Levit M et al (2004). Improved protein structure selection using decoy-dependent discriminatory functions. BMC Structural Biology 4(8)
Weiner SJ, Kollman PA, Case DA et al (1984) A new force field for molecular mechanical simulation of nucleic acids and proteins. Journal of the American Chemical Society 106:765–784
Wiederstein M, Sippl MJ (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35(Web Server issue): W407–410
Wroblewska L, Skolnick J (2007) Can a physics-based, all-atom potential find a protein's native structure among misfolded structures? i. large scale AMBER benchmarking. Journal of Computational Chemistry 28(12):2059–2066
Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biology 5:17
Wu S, Szilagyi A, Zhang Y (2011) Improving protein structure prediction using multiple sequence-based contact predictions. Structure 19(8):1182–1191
Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Research 35(10):3375–3382
Wu S, Zhang Y (2008a) A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 24(7):924–931
Wu S, Zhang Y (2008b) MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2):547–556
Wu S, Zhang Y (2010) Recognizing protein substructure similarity using segmental threading. Structure 18(7):858–867

Xu D, Zhang J, Roy A et al (2011) Automated protein structure modeling in CASP9 by I-TASSER pipeline combined with QUARK-based ab initio folding and FG-MD-based structure refinement. Proteins 79(Suppl 10):147–160

Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735

Xu D, Zhang Y (2013) Toward optimal fragment generations for ab initio protein structure assembly. Proteins 81(2):229–239

Yang J, Yan R, Roy A et al (2015a) The I-TASSER Suite: protein structure and function prediction. Nature Methods 12(1):7–8

Yang J, Zhang W, He B, et al (2015) Template-based protein structure prediction in CASP11 and retrospect of I-TASSER in the last decade. Proteins 84: 233–246

Yang Y, Faraggi E, Zhao H et al (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27(15):2076–2082

Zagrovic B, Snow CD, Shirts MR et al (2002) Simulation of folding of a small alpha-helical protein in atomistic detail using worldwide-distributed computing. Journal of Molecular Biology 323(5):927–937

Zhang C, Kim SH (2000) Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci U S A 97(6):2550–2555

Zhang C, Liu S, Zhou H et al (2004) An accurate, residue-level, pair potential of mean force for folding and binding based on the distance-scaled, ideal-gas reference state. Protein Science 13(2):400–411

Zhang J, Liang Y, Zhang Y (2011) Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 19(12):1784–1795

Zhang J, Zhang Y (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS ONE 5 (10):e15386

Zhang W, Yang J, He B et al (2015). Integration of QUARK and I-TASSER for Ab initio protein structure prediction in CASP11. Proteins 84: 76–86

Zhang Y (2008). Progress and Challenges in protein structure prediction. Curr Opin Struct Biol: In press

Zhang Y (2009) I-TASSER: Fully automated protein structure prediction in CASP8. Proteins 77(S9):100–113

Zhang Y (2014) Interplay of I-TASSER and QUARK for template-based and ab initio protein structure prediction in CASP10. Proteins 82(Suppl 2):175–187

Zhang Y, Hubner I, Arakaki A et al (2006) On the origin and completeness of highly likely single domain protein structures. Proc Natl Acad Sci U S A 103:2605–2610

Zhang Y, Kihara D, Skolnick J (2002) Local energy landscape flattening: parallel hyperbolic monte carlo sampling of protein folding. Proteins-Struct Func Genet 48(2):192–201

Zhang Y, Kolinski A, Skolnick J (2003) TOUCHSTONE II: A new approach to ab initio protein structure prediction. Biophysical Journal 85(2):1145–1164

Zhang Y, Skolnick J (2004a) Automated structure prediction of weakly homologous proteins on a genomic scale. Proceedings of the National Academy of Sciences of the United States of America 101:7594–7599

Zhang Y, Skolnick J (2004b) SPICKER: a clustering approach to identify near-native protein folds. Journal of Computational Chemistry 25(6):865–871

Zhang Y, Skolnick J (2005a) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102:1029–1034

Zhang Y, Skolnick J (2005b) TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Research 33(7):2302–2309

Zhang Y, Skolnick J (2013) Segment assembly, structure alignment and iterative simulation in protein structure prediction. BMC Biology 11:44

Zhou H, Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. Biophysical Journal 93(5):1510–1518

Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical Journal 101(8):2043–2052

Zhou H, Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11(11):2714–2726

Zhou R (2003) Free energy landscape of protein folding in water: explicit vs. implicit solvent. Proteins 53(2):148–161

# Chapter 2
# Protein Structures, Interactions and Function from Evolutionary Couplings

**Thomas A. Hopf and Debora S. Marks**

**Abstract** The sequences of biomolecules such as proteins and RNA genes contain information about their three-dimensional states and functions. For over 40 years biologists have used the evolutionary conservation of this information to detect homology and predict important subsets of residues. Recent work has substantially extended this view of conservation by including the detection of evolutionary couplings, interactions, between residues, resulting in a paradigm shift in our ability to compute three-dimensional structures from sequences alone. In addition to three-dimensional structure of single proteins and RNA, this statistical analysis of evolutionary constraints can identify functional residues involved in ligand binding, biomolecule-interactions, alternative ensembles of conformations, "invisible" tertiary states of disordered proteins and allows quantitative prediction of effects of mutations. In this chapter we present an overview of the statistical inference methodologies, a survey of the resulting applications and challenges facing the field.

**Keywords** Sequence coevolution · Covariation · Evolutionary couplings · 3D structure prediction · Function prediction · Protein interactions · Disorder · Conformational changes · Mutation effects · Maximum entropy model

Parts of this chapter have been adapted from (Hopf 2016).

T.A. Hopf (✉) · D.S. Marks
Department of Systems Biology, Harvard Medical School, Boston, MA, USA
e-mail: thomas_hopf@hms.harvard.edu

D.S. Marks
e-mail: debbie@hms.harvard.edu

T.A. Hopf
Department of Cell Biology, Harvard Medical School, Boston, MA, USA

T.A. Hopf
Department of Informatics, Technische Universität München, Garching, Germany



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_2

## 2.1 Introduction

Three-dimensional structure information is missing for a large fraction of known proteins and protein interactions, as experimental structure determination remains low-throughput whilst sequence databases grow exponentially. For instance, only about 50% of Pfam families have a solved structure for any of the family members (Finn et al. 2016) while structural coverage outside of conserved domains is even lower (Perdigao et al. 2015). Similarly, 60–80% of the approx. 10,000 and 40,000 heteromultimeric interactions in *E. coli* and human, respectively, have not yet been characterized structurally (Rajagopala et al. 2014; Mosca et al. 2014). The sustained effort to discover computational methods that have the potential to bypass the need for one-by-one experimental approaches is therefore motivated by this large experimental bottleneck. Comparative modelling transfers the coordinates from a solved protein to a target with similar sequence, based on the observation that the 3D folds of proteins remain conserved even as their amino acid sequences diverge (Webb and Sali 2014) (see also Chap. 4). In cases where no sequence-similar structural template can be identified, de novo fragment assembly methods (Qian et al. 2007) or even ab initio approaches using molecular force fields (Lindorff-Larsen et al. 2011) are an alternative for small proteins (<150 residues) (see also Chap 1). The applicability of these methods is however limited by the enormous size of conformational space that has to be searched as well as the accuracy of the available empirical force fields.

A conceptually different way of approaching the protein structure prediction problem is to mine the information contained in sequences. The evolutionary constraint to maintain residue interactions required for stable and functional proteins causes the coevolution of contacting amino acids. The idea therefore seems simple—find covarying positions in aligned protein sequences to identify residue pairs that correspond to physical contacts in the 3D structure, by analogy to the successful use of this approach in determining RNA secondary structure (Gutell et al. 1992). If correct, and if sufficient, these covarying residues could be transformed into distance constraints to construct 3D models, in a similar way to distances used in NMR structure determination.

However local covariation models applied to protein sequences did not consistently detect residues close in 3D (Shindyalov et al. 1994; Neher 1994; Gobel et al. 1994) despite some successful applications that showed enrichment of interacting residues (Skerker et al. (2008), Pazos et al. (1997)) or identification of contacts across proteins using additional biological information (Skerker et al. 2008). The apparent inability of these early covariance models to systematically identify contacting residues was attributed to a number of different reasons, including a loss of signal due to phylogenetic dependencies, the limited availability of sequence data and even the idea that we should not expect that truly coevolved residues are (mostly) close (Lapedes et al. 2012, 1997). Rather surprisingly, it turned out that changing the underlying model used to compute the couplings was the key innovation needed. This is because raw covariation frequencies or mutual

information between pairs of positions are dominated by 'indirect' transitive correlations, i.e. non-causal correlations between residues positions can be induced by a chaining of causal correlations between intervening residues positions. In a heterogeneous network, such as residues in a protein, these non-causal correlations can appear stronger than causal direct correlations, a well-understood feature of the Ising model in statistical physics where true correlations produce apparent long-range correlation at a distance (Giraud et al. 1999). The solution to this is to use a class of *global probability models* known as Potts model (a maximum entropy model) in statistical physics (Giraud et al. 1999; Lapedes et al. 1997; Ben-Naim and Lapedes 1999; Lapedes et al. 2012) and Markov Random Fields (an undirected graphical model) in computer science (Koller and Friedman 2009). Using these models the dependencies of types of amino acids in pairs of positions are computed simultaneously and consistently, rather than analysing pairs of positions independently of each other.

Application of these global statistical models was the key innovation in the identification of *evolutionary couplings* between pairs of positions in multiple sequence alignments that corresponded to contacting residues (Hopf et al. 2012; Marks et al. 2011, 2012; Morcos et al. 2011; Jones et al. 2012; Balakrishnan et al. 2011; Ekeberg et al. 2013; Lapedes et al. 2012; Michel et al. 2014). A retrospective analysis showed that even sequence data from 1999 PFAM family alignments was sufficient to infer large number of accurate residue contacts with the maximum entropy model for a few protein families (Marks et al. 2011). A pioneering Bayesian approach (Burger and van Nimwegen 2010, 2008) had some success but predictions were not as accurate with respect to residue proximity (Marks et al. 2011) and the use of belief propagation for parameter inference (Weigt et al. 2009) was computationally intractable for all but the smallest proteins. Although the methods required a sufficient number of sequences that diverged under functional selection, global statistical probability approaches such as those in Tables 2.1 and 2.2 provided a chance to obtain detailed structural and functional information for unsolved proteins of biological interest that was unprecedented.

Predicted contacts derived from evolutionary couplings have allowed the de novo prediction of protein 3D structures even for large molecules beyond the scope of previous approaches (Hopf et al. 2012; Marks et al. 2011, 2012; Hopf et al. 2015b; Ovchinnikov et al. 2014, 2015; Michel et al. 2014; Kosciolek and Jones 2014; Sulkowska et al. 2012) their complexes (Ovchinnikov et al. 2014; Hopf et al. 2014), multimeric contacts (Hopf et al. 2012; dos Santos et al. 2015), alternative conformations (Hopf et al. 2012; Toth-Petroczy et al. 2016; Morcos et al. 2013), and even the ability to predict structured states of apparently-disordered proteins (Toth-Petroczy et al. 2016). Many of these reports show, at least anecdotally, that evolutionary couplings models are able to identify functionally constrained residues over and above single column conservation and, most recently, the model has been used to make quantitative prediction of mutational changes in proteins (Hopf et al. 2017; Mann et al. 2014; Figliuzzi et al. 2016). In this chapter, we briefly describe the theoretical approach that underlies the methods, survey the most impactful

**Table 2.1** Webservers for evolutionary couplings (ECs) methods

| Method name | URL | Outputs | Inference method | Generates alignment | Refs. |
|---|---|---|---|---|---|
| EVfold | evfold.org | Alignments, EC pairs, 3D structures, functional residues | PLM | Yes | Marks et al. (2011), Toth-Petroczy et al. (2016) |
| EVcomplex | evcomplex.org | Protein complex alignments, Complex EC pairs | PLM | Yes | Hopf et al. (2014) |
| GREMLIN | gremlin.bakerlab.org | Alignments, EC pairs (incl. complexes), precomputed ECs and 3D structures | PLM | Yes | Kamisetty et al. (2013), Ovchinnikov et al. (2014) |
| DCA | dca.rice.edu | EC pairs | Mean-field | No | Morcos et al. (2011) |
| MetaPSICOV | bioinf.cs.ucl.ac.uk/MetaPSICOV | EC pairs | Sparse inv. cov., PLM, machine learning | Yes | Jones et al. (2012), (2015) |
| PconsC | c2.pcons.net | EC pairs | Sparse inv. cov., PLM, machine learning | Yes | Michel et al. (2014) |

**Table 2.2** Standalone evolutionary couplings inference software

| Method name | Inference algorithm | URL | Special features | Restrictions | Ref. |
|---|---|---|---|---|---|
| plmc | PLM | github.com/debbiemarkslab/plmc | Arbitrary sequences (incl. RNA), probabilistic treatment of gaps | – | (Weinreb et al. 2016; Toth-Petroczy et al. 2016) |
| CCMpred | PLM | github.com/soedinglab/CCMpred | Can be used on GPUs | – | (Seemayer et al. 2014) |
| plmDCA | PLM | plmdca.csc.kth.se | – | Matlab required | (Ekeberg et al. 2013) |
| GREMLIN | PLM | gremlin.bakerlab.org | – | Matlab required | (Balakrishnan et al. 2011; Kamisetty et al. 2013; Ovchinnikov et al. 2014) |
| DCA | Mean-field | dca.rice.edu | – | Matlab required | (Morcos et al. 2011) |
| PSICOV | Sparse inverse covariance | bioinfadmin.cs.ucl.ac.uk/downloads/PSICOV | – | – | (Jones et al. 2012) |
| FreeContact | Mean-field, sparse inverse covariance | rostlab.org/owiki/index.php/FreeContact | Implementation of both DCA and PSICOV algorithms | – | (Kajan et al. 2014) |
| MetaPSICOV | Sparse inverse covariance, PLM, machine learning | bioinfadmin.cs.ucl.ac.uk/downloads/MetaPSICOV | Meta-predictor | – | (Jones et al. 2012; 2015) |
| PconsC | Sparse inverse covariance, PLM, machine learning | c2.pcons.net | Meta-predictor | – | (Michel et al. 2014) |

applications and finally suggest challenges for the future, some of which might be solved by the time you read this!

## 2.2 Evolutionary Couplings from Sequence Alignments

The basis of coevolution-based structure and function prediction methods is the quantification of evolutionary couplings between all amino acid types in all pairs of sites derived from a multiple sequence alignment of the protein family (Fig. 2.1). These evolutionary couplings open up a wide variety of applications (Fig. 2.2).

### *2.2.1 The Global Model*

To avoid indirect correlations of residues pairs (as described above), global methods infer a probabilistic description of the sequence alignment that explains the observed correlations using underlying causative couplings between positions. These couplings are inferred by maximising the likelihood of observing the sequences in the alignment under the maximum entropy/Markov random field probability model.

Pairwise couplings are computed between amino acids to limit the number of model parameters to $O(N^2)$, but models of higher order (e.g. triples) are in principle possible given large enough protein families.

Under the pairwise graphical model the probability of any amino acid sequence $\sigma = (\sigma_1, \ldots, \sigma_n)$ of length $N$ is defined as

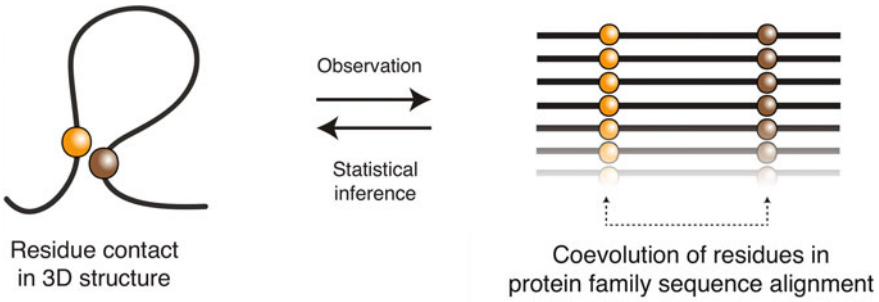



**Fig. 2.1** Residue interactions leave a coevolutionary record in protein sequences. The evolutionary constraint to maintain residue interactions, e.g. required for stable protein structures or complex formation with other molecules, creates a record of amino acid covariation in protein family sequence alignments. Mining this sequence record for residue pairs with strong evolutionary couplings using global statistical models opens a window to protein structure and function prediction (adapted from Hopf 2016, 2015b)

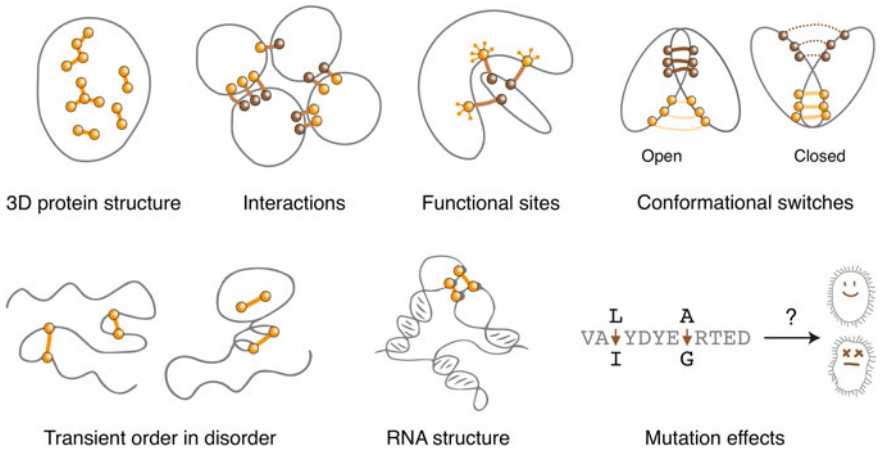



**Fig. 2.2** Applications of evolutionary couplings to predict protein structure and function. Evolutionary couplings allow to predict diverse aspects of protein structure and function that are defined by evolutionarily constrained interactions between residues, including the structures of monomers and complexes and changes in conformation. The approach can also be readily applied to other types of biomolecules, such as RNA, and used to quantify the phenotypic consequences of mutations with explicit modeling of epistatic interactions to the rest of the sequence (adapted from Marks et al. 2012)

$$P(\sigma) = \frac{1}{Z}\exp\left(\sum_{i=1}^{N}\mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{J}_{ij}(\sigma_i,\sigma_j)\right)$$

The model has two types of parameters that describe the constraint on acceptable amino acid configurations $\sigma_i$ and $\sigma_j$ at sites $i$ and $j$: bias terms $h_i$ (single-site conservation) and pair couplings $J_{ij}$ (co-conservation between pairs of sites i, j). Each variable $\sigma_i$ can assume one of the 20 amino acids as a value (most existing approaches treat gaps in the alignment as an additional 21st character, unless modelled as missing data). The *partition function* $Z$ is defined as

$$Z = \sum_{\sigma}\exp\left(\sum_{i=1}^{N}\mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1}\sum_{j=i+1}^{N}\mathbf{J}_{ij}(\sigma_i,\sigma_j)\right)$$

It sums over all possible $21^N$ sequences $\sigma = (\sigma_1, \ldots, \sigma_N)$ of length $N$ and ensures that $P(\sigma)$ is a valid probability distribution. Due to the exponential number of summations, calculating $Z$ is intractable for our application domain and we use a method that approximates Z using a factorization (see below).

To identify evolutionary constraints from an alignment, the inverse problem of inferring the model parameters from sequences has to be solved. Once the parameters are inferred, the pair couplings $J_{ij}$ can be used to quantify the strength of evolutionary coupling between pairs of sites $i$ and $j$.

***Parameter inference***. All the widely used current methods use an approximation to maximum likelihood estimation, which finds the set of parameters that maximizes the probability of observing the data. For the pairwise probability model defined above and a sequence alignment $\Sigma$ with sequences $\sigma$, the likelihood function $\mathcal{L}(\mathbf{h}, \mathbf{J})$ of the model parameters $h$ and $J$ is given by

$$\begin{aligned}\mathcal{L}(\mathbf{h}, \mathbf{J}) =& P(\Sigma|\mathbf{h}, \mathbf{J}) = \prod_{\sigma \in \Sigma} P(\sigma|\mathbf{h}, \mathbf{J}) \\ =& \prod_{\sigma \in \Sigma} \frac{1}{Z(\mathbf{h}, \mathbf{J})} \exp\left(\sum_{i=1}^{N} \mathbf{h}_i(\sigma_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{J}_{ij}(\sigma_i, \sigma_j)\right)\end{aligned}$$

However, since straightforward calculation of the likelihood function is prohibited by the intractability of *Z(h,J)*, several approaches have been taken to approximate parameter inference. These include gradient ascent with Monte Carlo sampling (Lapedes et al. 2012), message passing (Weigt et al. 2009) and mean-field (Marks et al. 2011; Morcos et al. 2011; Michel et al. 2014; Jones et al. 2012; Stein et al. 2015), but most current applications use pseudo-likelihood approximations to the full likelihood (Besag 1975; Balakrishnan et al. 2011; Ekeberg et al. 2013; Kamisetty et al. 2013; Michel et al. 2014; Hopf et al. 2015a, b; 2014; Toth-Petroczy et al. 2016; Weinreb et al. 2016; Ovchinnikov et al. 2014, 2015).

When adopting the pseudo-likelihood maximization (PLM) approach, the full likelihood for each sequence $\sigma = (\sigma_1, \ldots, \sigma_n)$ is approximated by a product of conditional likelihoods for each site $i$, i.e.

$$P(\sigma_1, \ldots, \sigma_N|\mathbf{h}, \mathbf{J}) \approx \prod_{i=1}^{N} P(\sigma_i|\sigma \backslash \sigma_i, \mathbf{h}, \mathbf{J})$$

The conditioning of the probability to observe a selected amino acid $\sigma_i$ in site $i$ on the rest of the sequence $(\sigma \setminus \sigma_i)$ leads to the cancellation of the global partition function *Z(h,J)*. Instead, the pseudo-likelihood normalizes locally over all possible 21 amino acid configurations at each site $i$. This factorization of the full likelihood function reduces the computational complexity of the parameter inference from $O(21^N)$ to $O(|\Sigma|N^2)$. The set of parameters minimizing the pseudo-likelihood is identified using standard iterative optimization algorithms.

***Regularization***. In addition, all published methods use some form of regularization to avoid overfitting to the data, as there are orders of magnitude more parameters in the model than there are effectively-independent samples (Number of parameters = $N$ ($N$-1)/2 (q-1)$^2$ + $N$($q$-1) for protein length $N$ and $q$ = 21 amino acid states). For example, the model has approximately $2*10^6$ parameters for a protein of length $N$ = 100 whereas most protein families only contain $10^2$ to $10^5$ effective (i.e. redundancy-reduced) sequences. This gap increases quadratically as the protein length $N$ increases. The EVcouplings method and others (Kamisetty et al. 2013) typically employ parameter type-specific $l_2$-regularization (equivalent to a Gaussian prior) while the mean-field methods uses pseudocounts (Marks et al. 2011; Morcos

et al. 2011) and sparse inverse covariance method uses $l_1$ (Jones et al. 2012). Finally, since the phylogenetic relationships between sequences mean that they are not independent and identically distributed, most methods for computing evolutionary couplings methods address the issue by sequence reweighting schemes (Weigt et al. 2009; Marks et al. 2011; Morcos et al. 2011; Ekeberg et al. 2013) and we expect this approach to be improved in the future to account more quantitatively for phylogenetic tree structure.

***Positional constraints from evolutionary couplings***

After inference, the coupling parameter matrices $J_{ij}$ contain the family-specific constraints on all 20 × 20 amino acid pair configurations $\sigma_i$ and $\sigma_j$ for each possible combination of positions $i$ and $j$. The last remaining step in the calculation of positional constraints from the evolutionary couplings between pairs of sites is to summarize the $20^2$ numbers in each $J_{ij}$ matrix into a single number that quantifies the total coupling for pair *(i, j)*. The preferred method for this summary statistic is the Frobenius norm.

Of each coupling matrix $J_{ij}$ (after first centring the means of rows and columns around zero, $J'_{ij}$)

$$\mathbf{J}'_{ij}(k,l) = \mathbf{J}_{ij}(k,l) - \mathbf{J}_{ij}(\cdot,l) - \mathbf{J}_{ij}(k,\cdot) + \mathbf{J}_{ij}(\cdot,\cdot)$$

where · means average across these entries,

$$FN(i,j) = \left\|\mathbf{J}_{ij}\right\|_2 = \sqrt{\sum_k \sum_l \mathbf{J}'_{ij}(k,l)^2}$$

which sums across all $21^2$ amino acid combinations *k, l*.

Since the $J_{ij}$ parameters summarized in the *FN* matrix are confounded by factors such as finite sampling and phylogenetic relationships between samples, the empirically derived *average product correction* (APC) is applied to the *FN* matrix to remove background coupling that arises due to noise (Dunn et al. 2008; Jones et al. 2012; Ekeberg et al. 2013). The correction assumes that, on average, each site should only have couplings to a limited subset of all sites. For each site pair *(i, j)*, the APC therefore approximates the noise (background coupling of both sites) with the product of the row and column averages of the FN score matrix (•) and subtracts these from the raw pair scores *FN(i, j)*:

$$EC(i,j) = FN(i,j) - \frac{FN(i,\cdot)FN(\cdot,j)}{FN(\cdot,\cdot)}$$

The final result after applying the correction is the symmetric *N* x *N* evolutionary coupling score matrix (*N* = length of protein). Each entry *EC(i, j)* estimates the strength of evolutionary coupling between a pair of sites *(i, j)*: larger positive values indicate strong evolutionary co-constraints; values around zero indicate that the model could not detect any coupling. The most significant evolutionary couplings can then be selected based on the shape of the score distribution by estimating the

degree to which each pair score is an outlier (Hopf et al. 2014; Ovchinnikov et al. 2014; Toth-Petroczy et al. 2016).

## 2.3 Three-Dimensional Protein Structures from Evolutionary Couplings

Starting from evolutionary couplings inferred from sequence alignments of protein families, one could then test if the couplings provide sufficient information to predict the 3D structure of proteins (Fig. 2.3a). The first publication on proteins folded with evolutionary couplings was using the EVfold method in 2011, and included a diverse set of proteins from 15 families (Marks et al. 2011). The resulting computed 3D structures were typically within 3–5 Å Cα-RMSD from the known experimental structures of these proteins. To our knowledge, this was the first time longer proteins, including some with more than 200 residues, had been folded without comparative modelling, fragments or known long-range contacts to anywhere near this degree of accuracy. Initially, the approach for computing couplings from the sequence alignment was based on a mean field approximation to find the parameters of the maximum entropy model, which was later updated to the more accurate PLM method described above. 3D structures were generated from evolutionary couplings using standard NMR distance geometry and simulated annealing software that use only little compute time, as the number of generated candidate models was only approx. 200–400 per protein. Simple geometric rules were then used to rank the prediction candidates and choose the most favoured models.

Many other groups have since used this or similar approaches to predict accurate long-range contacts from sequences, benchmarking against known contacts in observed 3D structures; such accurate predictions are typically available for thousands of families (Hopf et al. 2012; Michel et al. 2014; Kosciolek and Jones 2014; Ovchinnikov et al. 2015; Toth-Petroczy et al. 2016). The available methods choose different ways of thresholding the number of predicted couplings they display in contact maps and number of couplings used for structure prediction, but overall their strategies and outputs are very similar. Many of the observed differences are just as likely due to different input alignments as they are to do with the algorithms for inferring the couplings. The webservers of EVfold, PSICOV and GREMLIN provide downloads of coupling files that can be used to define restraints for the folding software of your choice. Of the available methods, to date only EVfold will fold on demand for particular sequences of interest, though other methods offer precomputed structures for a limited set of protein families (see Table 2.1 for an overview of available webservers and Table 2.2 for standalone evolutionary couplings software).

To assess the utility of evolutionary couplings for structure prediction, it is important to distinguish between predicting residue contacts and folding the protein. It is possible to have quite accurate residue contact predictions when

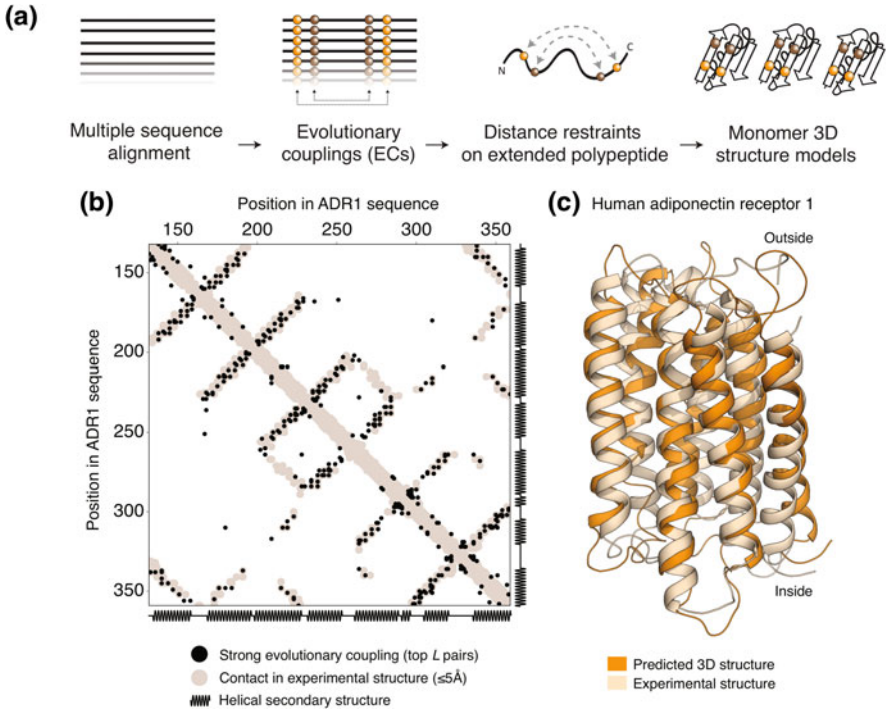



**Fig. 2.3** Protein 3D structure predicted from evolutionary sequences. **a** The 3D structure of a protein can be predicted from a multiple sequence alignment of the protein family by calculating evolutionary couplings between pairs of sites using a global probability model of the sequences. Assuming that residue pairs with strong couplings are close in 3D, the structure can then be computed by restraining the distances of these pairs in an extended polypeptide (adapted from Hopf 2016, 2015b; Marks et al. 2012) **b** Evolutionary couplings (*black dots*) for the human adiponectin receptor 1 (ADR1) largely correspond to residue contacts in the experimental 3D structure (*light brown dots*, precisions of 0.49 (5Å distance cutoff) and 0.77 (8Å cutoff), PDB 3wxv). **c** Models generated by EC-based 3D structure prediction (*dark orange* cartoon, best model) show good agreement with the experimental structure of ADR1 (*pale orange* cartoon, 2.4 Å Cα-RMSD over 192 residues, PDB 3wxv)

comparing evolutionary couplings to experimental structures, and still one may not be able to successfully fold the protein. For instance, predicted contacts may be clustered in one area of the protein, or only local along the chain and therefore missing key long-range contacts that define the overall topology of the molecule, such as contacts connecting the N- and C-termini. Only folding is therefore a definitive test if the computed evolutionary couplings contain sufficient information about the 3D structure of the protein.

While evolutionary couplings give valuable information about the 3D conformation of proteins, they also provide information over and above structure, such as functional residues that are particularly enriched for couplings with other residues (Fig. 2.2). Examples for strong coupling in functional sites include the active site of

trypsin, or the ligand binding pocket of the GPCR rhodopsin, where Lys-296 binds the retinal cofactor and has several strong couplings to other residues (Marks et al. 2011; Hopf et al. 2012). While it may be possible to identify some of these residues by single-site conservation alone, others may appear less conserved, and couplings offer the advantage of identifying the relevant interaction partners.

### 2.3.1 Transmembrane Proteins

Transmembrane proteins are of special biological interest as they mediate information transfer and molecule exchange across the cellular membranes in all forms of life, but are especially challenging to investigate experimentally when compared to globular proteins (see also Chap. 5). Given the resulting lack of experimental structures for the majority of membrane proteins, the most natural leverage of the evolutionary couplings approach was to predict their 3D structures, especially for large multipass proteins of high biomedical interest.

The first work to do so predicted evolutionary couplings and 3D structure for over 40 large membrane proteins, 25 of which were from families that had members with known structures and 18 of which were de novo predictions for families without any structure (Hopf et al. 2012). The blindly predicted structures on the test set of 25 proteins could be compared to known 3D coordinates and resulted in 3–6Å Cα-RMSD over at least 80% of the membrane domain. In similar work, the prediction of a test set of 28 proteins resulted in TM scores of at least 0.5 for most proteins (Jones et al. 2012). More recently, we updated various components of the EVfold prediction pipeline, including sequence alignment generation and inference of evolutionary couplings using PLM. Together with the increased number of sequences since the original publication in 2012, this leads to significant increases in prediction accuracy compared to the original method (average TM score increase of 0.08 on set of 25 proteins, highest TM score 0.82). We expect prediction accuracy to continue improving in the future as more sequences become available and better methods for folding are implemented.

For several examples from our set of de novo predictions, experimental structures have been published since. In general, our predictions show reasonable agreement with the experiment and have identified the correct overall 3D topology (TM score $\geq 0.5$) (Hopf 2016). Amongst these examples, the experimental structures confirmed that we correctly predicted the structural similarity of the unsolved complex 1 subunit 1 (MT-ND1) to the other subunits of the complex despite no detectable homology on the sequence level (Baradaran et al. 2013). We also correctly predicted the fold of the human adiponectin receptor 1 (Fig. 2.3) (TM score 0.69 from model in 2012, TM score = 0.79 in 2016), and successfully identified the cluster of activate site residues on the cytoplasmic side of the membrane (Tanabe et al. 2015). Both cases highlight the predictive power of evolutionary couplings to study the structure and function of proteins with limited experimental data.

### 2.3.2 Protein Interactions and Complexes

The coevolution of interacting residues is not only necessary to maintain the 3D structures of individual proteins, but also to maintain protein interactions and complexes. Based on this premise, others and we developed a general method for computing evolutionary couplings between proteins. The largest scale results identified interacting residues for over 50 protein interactions and the resulting 3D structure for a subset (Hopf et al. 2014; Ovchinnikov et al. 2014) (Fig. 2.4a, Tables 2.1 and 2.2) and many others have now computed a more limited number of interactions that often concentrate on disentangling paralog pairs of histidine kinase and response regulators (Cheng et al. 2016; Boyd et al. 2016; Feinauer et al. 2016; Bitbol et al. 2016; Gueudre et al. 2016).

For those methods with general applicability, the approaches are very similar. First, one must pair the sequences of putatively interacting proteins within each species to create a concatenated sequence alignment of the complex. Second, one computes both the couplings within (intra-protein evolutionary couplings) and between (inter-protein evolutionary couplings) the subunits simultaneously. This way, both the individual proteins can be predicted as well as the complex, using the inter-protein couplings as restraints in a docking protocol. Both EVcomplex and GREMLIN compute the couplings using pseudo-likelihood maximization, and differ only in their alignments, ranking and docking protocols.

However, the scope of both methods is currently limited by the generation of correctly paired sequence alignments that have sufficient sequence diversity. Correctly pairing the sequences when there are paralogs in a species depends on being able to identify the correct interacting proteins. Both EVcomplex and GREMLIN use the observation that interacting proteins are often encoded on the same operon. We have estimated that this excludes 80% of interacting proteins from EC-based prediction, even in *E. coli*. More recent approaches are being developed that aim to solve this issue, but their general applicability outside of a couple of systems still has to be demonstrated.

A second more pernicious assumption of this approach is that that the interactions, as well as the proteins themselves are conserved across evolution. While this may be a reasonable assumption for the components of ATP synthase, how conserved interactions are may be unknown for a large number of protein pairs. We expect to see significant algorithmic developments in this area so that the models can be used to ask the question rather than assume the answer.

Nevertheless, evolutionary couplings from sequence variation allow to predict protein interactions at residue level resolution not possible before (Fig. 2.4b), including the 3D structures of complexes that had not been solved experimentally at the time but whose subsequent characterisation confirmed the accuracy of the approach (e.g. DinJ-YafQ toxin-antitoxin interaction) (Hopf et al. 2014).

Both EVcomplex and GREMLIN also show that one can predict whether or not two proteins in a subunit interact physically, given sufficient sequence diversity and confidence in the matched alignment. In the case of the ATP synthase complex, we

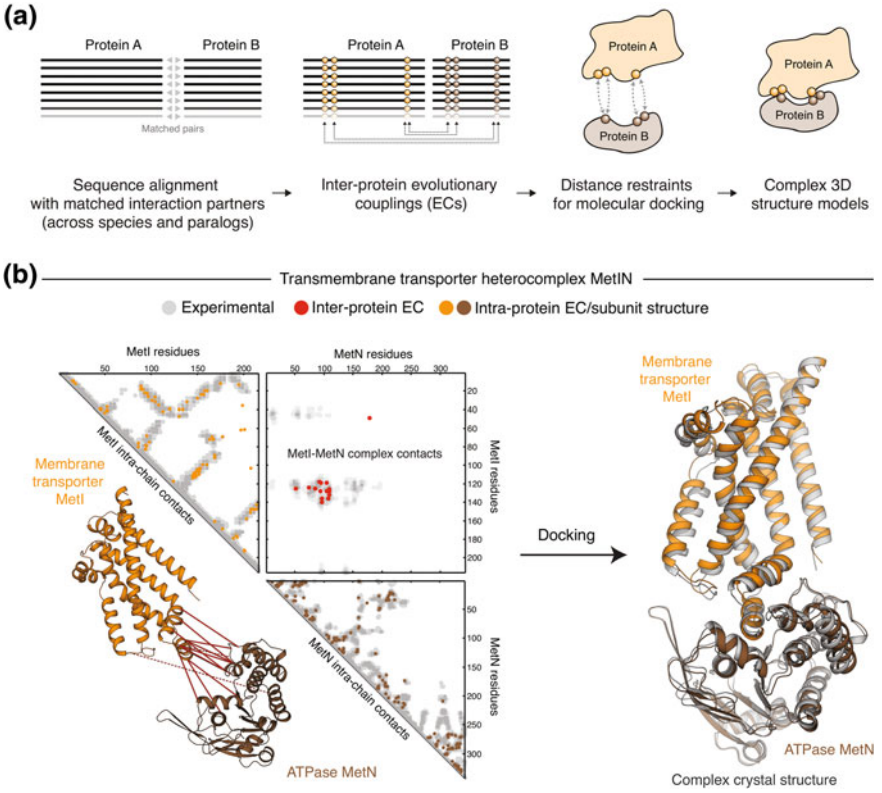



**Fig. 2.4** Protein interactions at residue level detail from evolutionary couplings**. a** Evolutionary couplings across interacting proteins can be calculated by generating a concatenated sequence alignment, where putatively interacting sequences within each species are matched with each other. Assuming coevolution due to structural proximity, the 3D structure of the complex can then be predicted from the monomer structures by docking with distance restraints on the strongly coupled pairs. **b** *Left* Evolutionary couplings (*coloured dots*) in the ABC transmembrane transporter MetIN correspond to structurally proximal residue pairs (*dark/medium/light grey dots* at 5/8/12Å distance cutoffs, PDB 3tui) both in the monomer structures (intra-protein ECs, *triangle* contact maps) as well as between the interacting subunits (inter-protein ECs, *square* contact map). The inter-protein ECs define the structural interaction between both subunits (*red lines* between *orange* and *brown* cartoons). *Right* Docking of the monomer structures (*orange/brown* cartoons) using significant inter-protein ECs leads to an accurate model of the complex (*grey cartoon*, 1.5Å interface-RMSD, PDB 3tui). (Figure adapted from Hopf 2016, 2014)

correctly identified 24 of 28 interactions with only 2 false positives and two interactions that are experimentally ambiguous. Similarly, GREMLIN correctly identified 12/23 interacting protein pairs in the ribosomal 50S subunit. The missing predictions (false negatives) may arise because the models are wrong, or, just as plausibly, the interactions could be weaker and a consequence of constraints between other subunits in the complex. Finally, recent work has also highlighted that evolutionary couplings can be applied to accurately predict the 3D structure of

RNA as well as protein-RNA interactions, in ribosomal complexes and RNaseP, from sequences alone (Weinreb et al. 2016).

### 2.3.3 Conformational Plasticity and Disordered Proteins

Many, if not most proteins may be structurally flexible, with conformational plasticity ranging from simple hinge movements or open-closed conformational switching to ordered stable structures that occur only upon binding or in the appropriate environment. Indeed, it may be the case that even protein segments that are considered highly flexible, such as histone tails, may take on a defined 3D structure in some functional states. Around half of human proteins contain substantially sized regions whose amino acid sequence is considered to indicate structural 'disorder', sometimes called 'intrinsic disorder'(van der Lee et al. 2014; Oates et al. 2013) (see also Chap. 6). These regions can range from 30 amino acid long insertions to longer regions of many hundreds of amino acids that are often present on transcription and translation factors.

Early work on evolutionary couplings showed that these methods will capture contacts from alternative 3D conformation, as demonstrated by the identification of couplings corresponding to open and closed conformations of the glycerol-3-phosphate transporter GlpT (Hopf et al. 2012) and the L-leucine binding protein (Morcos et al. 2013). More recently, this has been explored systematically with another 38 proteins known to have alternative conformations and differential contacts, demonstrating not only fold rearrangement but also, sometimes, secondary structure switching (Toth-Petroczy et al. 2016).

This recent work has extended the exploration of conformational states to proteins considered disordered. Since a small number of disordered proteins are known to become ordered in specific environments and have been captured experimentally, this gave the opportunity to investigate whether evolutionary couplings methods can detect these 3D states. After a number of methodological improvements, including iterative testing for alignment robustness, evolutionary couplings were computed to determine the potential of these proteins forming long-range contacts and secondary structure. In 40 of the 45 cases contacts were successfully predicted for known "order upon binding', including the well-known cyclin inhibitor p27 when it binds the Cyclin A-Cdk2 complex. Importantly, the method also found very little evidence of structural constraints for proteins such as the C-terminal tail of Histone H1 that had multiple lines of evidence for lack of structure (Toth-Petroczy et al. 2016). Hence, the true positive predictions for proteins with ordered conformations do not seem to be at the expense of false positives in proteins without ordered conformations.

To explore the structural potential of apparently disordered regions for which there is currently no experimental information on a proteome-wide scale, Toth-Petroczy et al. systematically surveyed all regions in the human proteome of more than 100 amino acids in length where alignments could be constructed (about

25% of all regions). This analysis resulted in predictions for ~1000 protein regions, of which 40% showed signal for some long-range structure and another 40% secondary structure. The predicted contact maps revealed that some of these disordered domains resembled zinc finger and RNA-binding domains, which could not be identified from their primary sequence (the data from this analysis is available from http://marks.hms.harvard.edu/disorder).

## 2.4 Predicting the Effect of Mutations

A major challenge in biology is being able to predict the functional effects of mutations on phenotype or fitness. New work has shown that the global statistical models of sequences can also be used to predict the effects of mutations by quantifying the change of probabilities between the mutated protein and the wild type sequence ($\Delta E = \log P(\text{mutant}) / P(\text{wild type})$) (Hopf et al. 2017; Figliuzzi et al. 2016; Mann et al. 2014). This quantity $\Delta E$, called the statistical energy difference of a mutant, is computed by summing the changes in couplings and site amino acid preferences between all pairs of positions, to give a total score that describes the effect of any single or higher-order mutation. For instance, as illustrated in the cartoon protein in Fig. 2.5a, the substitution W3L leads to a change in 4 couplings and one single site bias term. Through the evaluation of couplings to other sites, the computation explicitly models the context dependence (or epistasis) of mutations. These interactions are typically neglected by approaches using single-site conservation to quantify the effects of mutations. It is important to note that this approach uses precisely the same statistical model (e.g. PLM or DCA) as one uses to compute residue contacts from the sequence alignment, but *does not depend on computing the structure*. This allows to infer epistatic mutational landscapes for any protein with enough sequence information (Figs. 2.5b, c).

To test the applicability of our implementation of the method, EVmutation, the predicted effects of mutations have been compared against thousands of variants assayed in high-throughput multiplexed mutational scans that have emerged over the last few years, providing a large pool of ground truth for evaluation (Deng et al. 2012; Jacquier et al. 2013; Stiffler et al. 2015; Melamed et al. 2013; 2015; Rockah-Shmuel et al. 2015; Starita et al. 2015; Roscoe and Bolon 2014; Starita et al. 2013; Li et al. 2016; Melnikov et al. 2014). Whilst the exact interpretation of $\Delta E$ effects is not clear a priori, one would expect them to be related to the 'fitness' of the protein sequence. For instance, $\Delta E$ of all single mutations to a bacterial DNA methylase correlated well with an experimental scan testing their effect on bacterial fitness (Spearman's rank correlation $\rho = 0.69$) (Rockah-Shmuel et al. 2015). Similarly, EVmutation effects showed significant correlations across a wide range of 34 experimental datasets for 21 proteins and a tRNA molecule (Hopf et al. 2017). The approach generalizes to any type of biological sequence, and could also be used to predict effects for protein-RNA complexes. Using epistatic interactions with other sites particularly contributed to improved prediction accuracy in functional

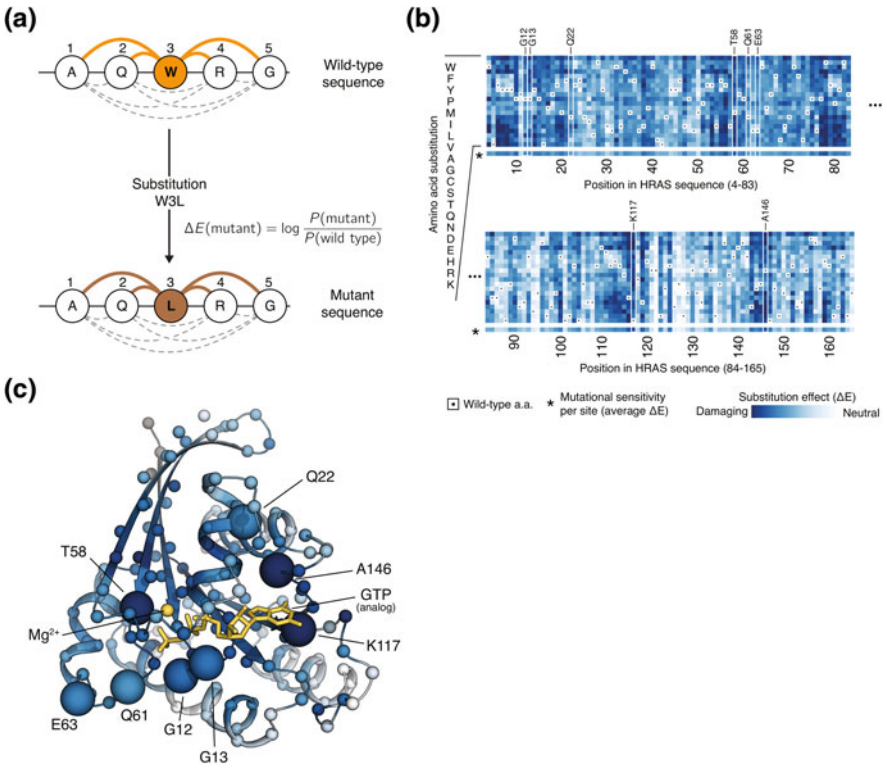



**Fig. 2.5** Prediction of mutation effects using an epistatic model of evolutionary sequences. **a** The global probability model of a protein family can be used to predict the effects of mutations by comparing the probabilities of the wild-type and mutant sequences. The calculation sums the differences in all couplings to mutated positions as well as the change in the single-site amino acid preference terms of the changed sites. Thereby, epistatic interactions with the sequence background are incorporated in the calculation (adapted from (Hopf 2016)) **b** Computed ΔE mutational landscape of the human disease gene HRAS (x-axis: position in HRAS sequence, y-axis: amino acid substitutions, *white boxes*: positions with known disease mutations). **c** Residues (small spheres) around the active site of RASH (GTP ligand analog, *yellow sticks*), including positions with known disease mutations (large spheres), are predicted as sensitive to mutation (colour scale as in (a) from *blue*/damaging to *white*/neutral)

sites, such as ligand binding and protein interaction interfaces, when compared to a model that only uses single-site conservation. When tested on human disease variants, ΔE separated them from neutral variants with similar or higher accuracy than state-of-the-art methods for variant effect classification without, however, being specifically trained on known variants for this problem (Hopf et al. 2017). This suggests that established machine learning methods could benefit from the inclusion of evolutionary statistical energies instead of positional sequence conservation.

## 2.5 Summary and Future Challenges

Over the last 5 years, approaches based on evolutionary couplings from sequence alignments have already shown their power in predicting structural constraints and 3D structures for proteins, RNA, their interactions, the potential structured states of disordered regions, as well as the effects of mutations on protein function. Readers would do well to use this chapter as a basis, but since the field will change rapidly in the next few years, they should be encouraged to search for more recent work than the snapshot presented here.

We expect to see an increase in hybrid approaches that combine evolutionary couplings with experimental methods to accelerate structure determination in such fields as cryoEM, NMR, crystallography or mass spectrometry. First promising work that demonstrates the power of this type of approach has already been published (Tang et al. 2015). Where refined 3D models are desired, there is a still a clear need for improving the structure prediction protocols, although some advances have been made here recently.

Notwithstanding the impressive impact these methods have already had, there are many challenges to be solved, not least with the probability model itself. First, an implicit assumption in the underlying model is that all sequences have been tried by evolution and the ones that we see now are the only possible functional ones, leading to many issues associated with inferring models from undersampled data. Whilst regularization during inference and heuristics for post hoc corrections address this problem somewhat, we expect advances in this area would be beneficial for more accurate models.

A second challenge for the emerging field is to develop improved criteria for assessing the quality of alignments, and the choice of alignment depth that is critically dependent on the research question being asked. If we did not know what the 3D structure of GPCRs looked like, then any family alignment however large and non-specific may be useful; on the other hand, if we want to explore the different ligand-binding pockets of the subfamilies we would need alignments that reflected that specificity. Similarly, for complexes and protein interactions the challenge is to assess the likelihood of interaction with the ambiguity that the interaction may not be conserved in all alignable sequences.

A third challenge is to blindly disambiguate evolutionary couplings that arise due to different aspects of protein function, including the blind assignment of couplings to different conformational states, or the distinction between intra- and inter-protein interactions in homomultimeric complexes in the absence of an experimental structure of the monomer.

All of these challenges are exciting questions for future research, and will help to further increase the usefulness of evolutionary couplings as a tool in exploring diverse aspects of protein structure and function.

## References

Balakrishnan S, Kamisetty H, Carbonell JG, Lee SI, Langmead CJ (2011) Learning generative models for protein fold families. Proteins 79(4):1061–1078. doi:10.1002/prot.22934

Baradaran R, Berrisford JM, Minhas GS, Sazanov LA (2013) Crystal structure of the entire respiratory complex I. Nature 494(7438):443–448. doi:10.1038/nature11871

Ben-Naim E, Lapedes AS (1999) Genetic correlations in mutation processes. Phys Rev E Stat Phys Plasmas Fluids 59(6):7000–7007

Besag J (1975) Statistical analysis of non-lattice data. Statistician 179–195

Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS (2016) Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA 113(43):12180–12185. doi:10.1073/pnas.1606762113

Boyd JS, Cheng RR, Paddock ML, Sancar C, Morcos F, Golden SS (2016) A combined computational and genetic approach uncovers network interactions of the cyanobacterial circadian clock. J Bacteriol 198(18):2439–2447. doi:10.1128/JB.00235-16

Burger L, van Nimwegen E (2008) Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method. Molecular Syst biology 4:165. doi:10.1038/msb4100203

Burger L, van Nimwegen E (2010) Disentangling direct from indirect co-evolution of residues in protein alignments. PLoS Comput Biol 6(1):e1000633. doi:10.1371/journal.pcbi.1000633

Cheng RR, Nordesjo O, Hayes RL, Levine H, Flores SC, Onuchic JN, Morcos F (2016) Connecting the sequence-space of bacterial signaling proteins to phenotypes using coevolutionary landscapes. Mol Biol Evol. doi:10.1093/molbev/msw188

Deng Z, Huang W, Bakkalbasi E, Brown NG, Adamski CJ, Rice K, Muzny D, Gibbs RA, Palzkill T (2012) Deep sequencing of systematic combinatorial libraries reveals beta-lactamase sequence constraints at high resolution. J Mol Biol 424(3–4):150–167. doi:10.1016/j.jmb.2012.09.014

dos Santos RN, Morcos F, Jana B, Andricopulo AD, Onuchic JN (2015) Dimeric interactions and complex formation using direct coevolutionary couplings. Sci Rep 5:13652. doi:10.1038/srep13652

Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E Stat Nonlin Soft Matter Phys 87(1):012707

Feinauer C, Szurmant H, Weigt M, Pagnani A (2016) Inter-protein sequence co-evolution predicts known physical interactions in Bacterial Ribosomes and the Trp Operon. PLoS ONE 11(2):e0149166. doi:10.1371/journal.pone.0149166

Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M (2016) Coevolutionary landscape inference and the context-dependence of mutations in Beta-Lactamase TEM-1. Mol Biol Evol 33(1):268–280. doi:10.1093/molbev/msv211

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res 44(D1):279–285. doi:10.1093/nar/gkv1344

Giraud BG, Heumann JM, Lapedes AS (1999) Superadditive correlation. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics 59 (5 Pt A):4983–4991

Gobel U, Sander C, Schneider R, Valencia A (1994) Correlated mutations and residue contacts in proteins. Proteins 18(4):309–317. doi:10.1002/prot.340180402

Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A (2016) Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci USA 113(43):12186–12191. doi:10.1073/pnas.1607570113

Gutell RR, Power A, Hertz GZ, Putz EJ, Stormo GD (1992) Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. Nucleic Acids Res 20(21):5785–5795

Hopf T (2016) Phenotype prediction from evolutionary sequence covariation. München, Technische Universität München, Diss 2016

Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149(7):1607–1621. doi:10.1016/j.cell.2012.04.012

Hopf TA, Ingraham JB, Poelwijk FJ, Springer M, Sander C, Marks DS (2015a) Quantification of the effect of mutations using a global probability model of natural sequence variation. arXiv preprint arXiv:151004612

Hopf TA, Ingraham JI, Poelwijk FJ, Schärfe CPI, Springer M, Sander C, Marks DS (2017) Mutational effects captured by epistatic models of evolutionary sequence variation. Nat Biotech 35:128–135. doi:10.1038/nbt.3769

Hopf TA, Morinaga S, Ihara S, Touhara K, Marks DS, Benton R (2015b) Amino acid coevolution reveals three-dimensional structure and functional domains of insect odorant receptors. Nat Commun 6:6077. doi:10.1038/ncomms7077

Hopf TA, Schärfe CP, Rodrigues JP, Green AG, Kohlbacher O, Sander C, Bonvin AM, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. eLife 3. doi:10.7554/eLife.03430

Jacquier H, Birgy A, Le Nagard H, Mechulam Y, Schmitt E, Glodt J, Bercot B, Petit E, Poulain J, Barnaud G, Gros PA, Tenaillon O (2013) Capturing the mutational landscape of the beta-lactamase TEM-1. Proc Natl Acad Sci USA 110(32):13067–13072. doi:10.1073/pnas.1215206110

Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190. doi:10.1093/bioinformatics/btr638

Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31(7):999–1006

Kajan L, Hopf TA, Kalas M, Marks DS, Rost B (2014) FreeContact: fast and free software for protein contact prediction from residue co-evolution. BMC Bioinformatics 15:85. doi:10.1186/1471-2105-15-85

Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci USA 110(39):15674–15679. doi:10.1073/pnas.1314045110

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press

Kosciolek T, Jones DT (2014) De novo structure prediction of globular proteins aided by sequence variation-derived contacts. PLoS ONE 9(3):e92197. doi:10.1371/journal.pone.0092197

Lapedes A, Giraud B, Jarzynski C (2012) Using sequence alignments to predict protein structure and stability with high accuracy. arXiv preprint arXiv:12072484

Lapedes AS, Giraud BG, Liu LC, Stormo GD (1997) Correlated Mutations in Protein Sequences: Phylogenetic and Structural Effects. Santa Fe Institute

Li C, Qian W, Maclean CJ, Zhang J (2016) The fitness landscape of a tRNA gene. Science. doi:10.1126/science.aae0568

Lindorff-Larsen K, Piana S, Dror RO, Shaw DE (2011) How fast-folding proteins fold. Science 334(6055):517–520. doi:10.1126/science.1208351

Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, Ndung'u T (2014) The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol 10(8):e1003776. doi:10.1371/journal.pcbi.1003776

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766. doi:10.1371/journal.pone.0028766

Marks DS, Hopf TA, Sander C (2012) Protein structure prediction from sequence variation. Nat Biotechnol 30(11):1072–1080. doi:10.1038/nbt.2419

Melamed D, Young DL, Gamble CE, Miller CR, Fields S (2013) Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. RNA 19(11):1537–1551. doi:10.1261/rna.040709.113

Melamed D, Young DL, Miller CR, Fields S (2015) Combining natural sequence variation with high throughput mutational data to reveal protein interaction sites. PLoS Genet 11(2): e1004918. doi:10.1371/journal.pgen.1004918

Melnikov A, Rogov P, Wang L, Gnirke A, Mikkelsen TS (2014) Comprehensive mutational scanning of a kinase in vivo reveals substrate-dependent fitness landscapes. Nucleic Acids Res 42(14):e112. doi:10.1093/nar/gku511

Michel M, Hayat S, Skwark MJ, Sander C, Marks DS, Elofsson A (2014) PconsFold: improved contact predictions improve protein models. Bioinformatics 30(17):482–488. doi:10.1093/bioinformatics/btu458

Morcos F, Jana B, Hwa T, Onuchic JN (2013) Coevolutionary signals across protein lineages help capture multiple protein conformations. Proc Natl Acad Sci USA 110(51):20533–20538. doi:10.1073/pnas.1315625110

Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA 108(49):1293–1301. doi:10.1073/pnas.1111471108

Mosca R, Ceol A, Stein A, Olivella R, Aloy P (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. Nucleic acids research 42 (Database issue): 374–379. doi:10.1093/nar/gkt887

Neher E (1994) How frequent are correlated changes in families of protein sequences? Proc Natl Acad Sci USA 91(1):98–102

Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2) P(2): database of disordered protein predictions. Nucleic acids research 41 (Database issue): 508–516. doi:10.1093/nar/gks1226

Ovchinnikov S, Kamisetty H, Baker D (2014) Robust and accurate prediction of residue-residue interactions across protein interfaces using evolutionary information. eLife 3: 02030. doi:10.7554/eLife.02030

Ovchinnikov S, Kinch L, Park H, Liao Y, Pei J, Kim DE, Kamisetty H, Grishin NV, Baker D (2015) Large-scale determination of previously unsolved protein structures using evolutionary information. eLife 4: 09248. doi:10.7554/eLife.09248

Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271(4):511–523. doi:10.1006/jmbi.1997.1198

Perdigao N, Heinrich J, Stolte C, Sabir KS, Buckley MJ, Tabor B, Signal B, Gloss BS, Hammang CJ, Rost B, Schafferhans A, O'Donoghue SI (2015) Unexpected features of the dark proteome. Proc Natl Acad Sci USA 112(52):15898–15903. doi:10.1073/pnas.1508380112

Qian B, Raman S, Das R, Bradley P, McCoy AJ, Read RJ, Baker D (2007) High-resolution structure prediction and the crystallographic phase problem. Nature 450(7167):259–264. doi:10.1038/nature06249

Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Hauser R, Siszler G, Wuchty S, Emili A, Babu M, Aloy P, Pieper R, Uetz P (2014) The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol 32 (3):285–290. doi:10.1038/nbt.2831

Rockah-Shmuel L, Toth-Petroczy A, Tawfik DS (2015) Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. PLoS Comput Biol 11(8):e1004421. doi:10.1371/journal.pcbi.1004421

Roscoe BP, Bolon DN (2014) Systematic exploration of ubiquitin sequence, E1 activation efficiency, and experimental fitness in yeast. J Mol Biol 426(15):2854–2870. doi:10.1016/j.jmb.2014.05.019

Seemayer S, Gruber M, Soding J (2014) CCMpred–fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 30(21):3128–3130. doi:10.1093/bioinformatics/btu500

Shindyalov IN, Kolchanov NA, Sander C (1994) Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7(3):349–358

Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT (2008) Rewiring the specificity of two-component signal transduction systems. Cell 133(6):1043–1054. doi:10.1016/j.cell.2008.04.040

Starita LM, Pruneda JN, Lo RS, Fowler DM, Kim HJ, Hiatt JB, Shendure J, Brzovic PS, Fields S, Klevit RE (2013) Activity-enhancing mutations in an E3 ubiquitin ligase identified by high-throughput mutagenesis. Proc Natl Acad Sci USA 110(14):1263–1272. doi:10.1073/pnas.1303309110

Starita LM, Young DL, Islam M, Kitzman JO, Gullingsrud J, Hause RJ, Fowler DM, Parvin JD, Shendure J, Fields S (2015) Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics. doi:10.1534/genetics.115.175802

Stein RR, Marks DS, Sander C (2015) Inferring pairwise interactions from biological data using maximum-entropy probability models. PLoS Comput Biol 11(7):e1004182. doi:10.1371/journal.pcbi.1004182

Stiffler MA, Hekstra DR, Ranganathan R (2015) Evolvability as a function of purifying selection in TEM-1 beta-Lactamase. Cell 160(5):882–892. doi:10.1016/j.cell.2015.01.035

Sulkowska JI, Morcos F, Weigt M, Hwa T, Onuchic JN (2012) Genomics-aided structure prediction. Proc Natl Acad Sci USA 109(26):10340–10345. doi:10.1073/pnas.1207864109

Tanabe H, Fujii Y, Okada-Iwabu M, Iwabu M, Nakamura Y, Hosaka T, Motoyama K, Ikeda M, Wakiyama M, Terada T, Ohsawa N, Hato M, Ogasawara S, Hino T, Murata T, Iwata S, Hirata K, Kawano Y, Yamamoto M, Kimura-Someya T, Shirouzu M, Yamauchi T, Kadowaki T, Yokoyama S (2015) Crystal structures of the human adiponectin receptors. Nature 520(7547):312–316. doi:10.1038/nature14301

Tang Y, Huang YJ, Hopf TA, Sander C, Marks DS, Montelione GT (2015) Protein structure determination by combining sparse NMR data with evolutionary couplings. Nat Methods 12(8):751–754. doi:10.1038/nmeth.3455

Toth-Petroczy A, Palmedo P, Ingraham J, Hopf TA, Berger B, Sander C, Marks DS (2016) Structured states of disordered proteins from genomic sequences. cell 167 (1):158–170 e112. doi:10.1016/j.cell.2016.09.010

van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114(13):6589–6631. doi:10.1021/cr400525m

Webb B, Sali A (2014) Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics 47:5 6 1–32. doi:10.1002/0471250953.bi0506s47

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106(1):67–72. doi:10.1073/pnas.0805923106

Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS (2016) 3D RNA and Functional Interactions from Evolutionary Couplings. Cell 165(4):963–975. doi:10.1016/j.cell.2016.03.030

# Chapter 3
# Fold Recognition

**Lawrence A. Kelley**

**Abstract** Fold recognition is concerned with the prediction of protein three-dimensional structure from amino acid sequence by the detection of extremely remote homologous or analogous relationships to known structures. As such it lies midway between ab initio protein folding and close homology modelling. This chapter surveys both the history of the field and the current state-of-the art, focussing on approaches recently shown to be successful in international blind trials.

**Keywords** CASP · Threading · Pair-potential · PSI-Blast · Profile · Hidden Markov model · Contact map

## 3.1 Introduction

The amino acid sequence of a protein determines its structure, which in turn determines its biological function and mechanism of action. Protein folding is the bridge between the instructions for living things and the living thing itself. This key paradigm in biochemistry accounts for nearly one in four Nobel Prizes in Chemistry since 1956 (Seringhaus and Gerstein 2007) In 2005 *Science* named the protein folding problem one of the 125 biggest unsolved problems in science (Science Editorial, 1st July 2005) and in 2013 the Nobel prize for Chemistry was awarded to Karplus, Levitt and Warshel for their work on computational simulation of proteins.

When the previous 2008 version of this chapter was written there were 5.8 million protein sequences experimentally determined by genome sequencing. There are now 50 million. This number has been exponentially growing for over two decades and this growth is set to continue. The new meta-genomics projects involving shotgun-sequencing random samples of seawater around the globe every 200 miles are finding 1.3 million new genes and as many as 50,000 new species in each barrel of seawater. In 2008, sequencing machines could sequence 100 million

L.A. Kelley (✉)
Structural Bioinformatics Group, Imperial College London, London, UK
e-mail: l.a.kelley@imperial.ac.uk



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_3

base pairs in 24 h. They can now process 1 million base pairs a second and the price of sequencing a human genome has dropped to $1000.

In 2008, 50,000 protein three-dimensional structures had been solved. In 2015, this number is 100,000. So despite the progress of the high-throughput structural genomics initiatives and the large arrays of NMR and crystallography robots working to determine protein structure, we have observed a doubling time of 8 years in structure determination, while the number of sequences has doubled 3 times in the same period.

### 3.1.1 The Importance of Blind Trials: The CASP Competition

Over the past 30 years a bewildering variety of techniques have been developed to attack the problem of protein structure prediction in general and fold recognition in particular. As in any scientific endeavour, it is critical that any new technique is fully tested "experimentally". It is for this reason that the Critical Assessment of Structure Prediction or "CASP" meeting was devised (http://predictioncenter.llnl.gov/; (Moult et al. 2014). The purpose of the CASP meeting or competition (held every two years) is to mimic the real-world situation of being presented with an amino acid sequence for which we do not know the structure. However, there is a critical difference - the organisers of the meeting **do** know the structure. These proteins have had their structures newly solved by experimentalists, but this data has not yet been released to the scientific community. As a result, the assessors of the CASP meeting are in the rare position of knowing the 3-dimensional structures of a set of proteins unknown to the predictors.

CASP acts as a true blind experimental assessment of the viability of techniques for structure prediction in the real world. Therefore, the CASP competition has been my guide in deciding what methodologies to describe in this chapter. This is not to say that other methodologies may not indeed be powerful predictors, which for whatever reason did not perform well at CASP. There are literally hundreds of different techniques that have been developed over the years, and to avoid burdening the reader, I have chosen to use the results of CASP as a filter. For a review of the most recent CASP11 meeting see the CASP11 supplement (Moult et al. 2014).

### 3.1.2 Ab Initio Structure Prediction Versus Homology Modelling

If we are to have any hope of structurally characterising any significant fraction of the proteins in nature, barring the discovery of some revolutionary experimental

technique, then we will require a method to predict structure from sequence computationally. After Anfinsen showed in 1961 that ribonuclease could be refolded after denaturation while preserving enzyme activity, we have been beguiled by the idea that all the information required by a protein to adopt its final conformation is encoded in its sequence. As a result 'pure' methods using only the sequence itself as input and the laws of physics (or approximations to them) have been pursued for decades and are showing some progress. These are covered in Chap. 1 of this book.

However, in general, these methods are either computationally intractable or demonstrate poor performance on everything but the smallest proteins (<100–150 amino acids). Although a physics-based approach may seem like the only true solution to the folding problem, the practical importance of protein structure prediction has meant we have to accept our current limitations and move, if only temporarily, to a more pragmatic solution now. This has led the search for a protein structure prediction technique away from physics and towards a more data-mining approach.

It has long been clear that similar protein sequences fold to similar structures. Thus, given a novel protein sequence whose structure we require, henceforth known as the '*target'* we simply have to check if any other similar sequence with a **known** structure has already been solved. If the sequences are highly similar then this detection process is quite straightforward using basic alignment techniques. Using a simple measure of the similarity of amino acid types, such as the BLOSUM scoring matrix coupled with a dynamic programming algorithm such as Smith-Waterman one can rapidly and optimally (according to the scoring function) align two sequences.

Given an alignment between a sequence and a known structure, henceforth known as the '*template*', one can then build a crude model by simply copying the corresponding three-dimensional coordinates of the template and re-labelling the amino acids in accordance with the equivalent residues from the alignment (Fig. 3.1). The model can be further refined using a slew of techniques described in the comparative modelling chapter of this book (Chap. 4). The advantages of this approach are clear; it is computationally quick, and the accuracy of the resulting model will be very high *given a high sequence similarity between target and template*. This immediately points to the method's limitations. If no similar sequence has yet had its structure solved, we can make no progress at all.

So, we have two lines of attack in the search for a solution to the protein structure prediction problem. One approach, based on general physics principles, aims at providing a well-understood, universal technique to predict structure from sequence, with the added benefit of enabling protein design, a study of dynamics and much more. However, it is extremely difficult and will probably remain computationally intractable for years to come. At the other extreme, we have a straightforward but highly limited heuristic technique, homology modelling, which can give high accuracy models, but only in a very limited number of cases. It is against this backdrop that the term 'fold recognition' was coined, to act as a bridge between these two extremes.

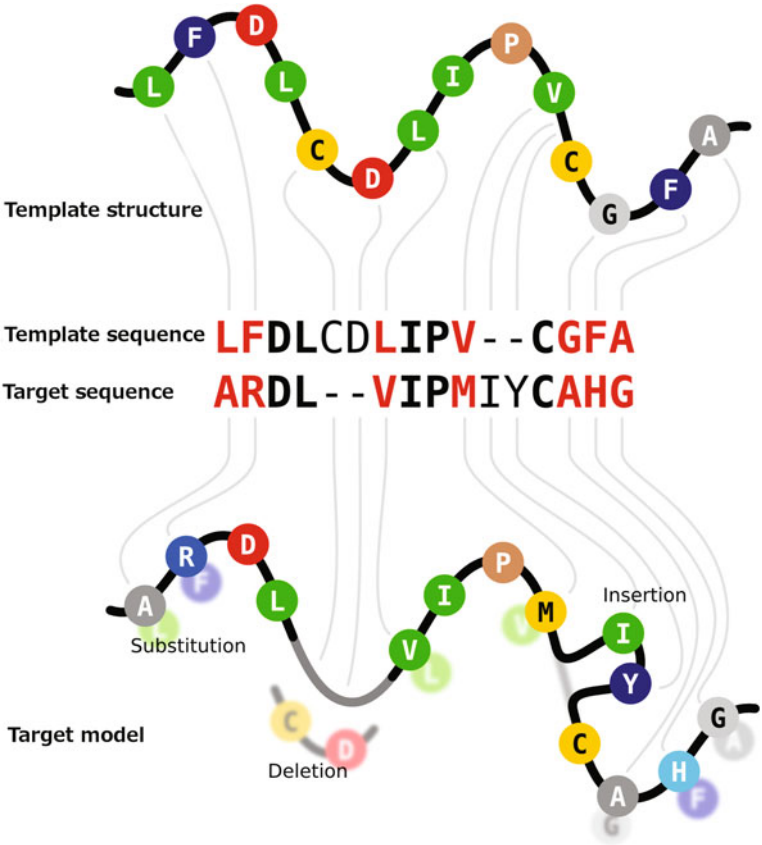



**Fig. 3.1** Cartoon representation of simple model building by target-template alignment. The sequence of the known structure ('Known sequence') is shown aligned to the target sequence. *Dashes* represent insertions and deletions. *Red letters* indicate residue substitutions. Residue type are coloured according to biophysical properties. *Thin wavy lines* connect equivalent positions in the query and template

It should be noted that recent breakthroughs have opened a third path to the structure prediction problem: contact prediction. Its full implications for practical protein structure prediction from sequence alone are still …unfolding… and it is yet to be seen what its full impact will be on the field. Contact prediction is covered fully in Chap. 2, but I will return to the topic briefly towards the end of this chapter.

### *3.1.3 The Limits of Fold Space*

Several key observations about the nature of proteins are in order. According to protein structure classification schemes such as CATH (Sillitoe et al. 2015), the

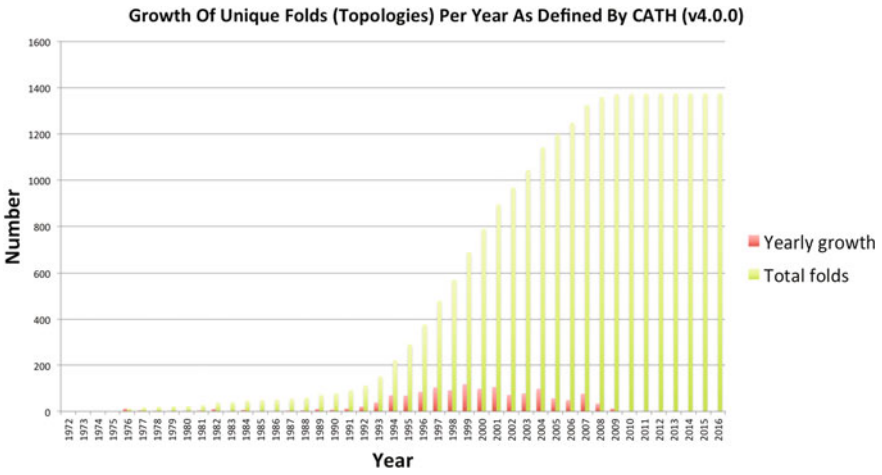



**Fig. 3.2** Graph showing the number of experimentally-determined protein structures included in the CATH database together with the number of topologies as a function of year. It can be seen that although the number of structures added to CATH is increasing rapidly, the number of new folds has remained static since about 2009

approximately 100,000 experimentally determined protein structures in the protein data bank (Berman et al. 2000), can be grouped into about 1200 unique structural folds (unique topologies). As more and more structures are solved experimentally, the number of new folds discovered increases very slowly. In fact, according to data from the RCSB, there have been almost no new folds discovered since the last edition of this book in 2008, despite a doubling of the size of the PDB in that time (Fig. 3.2).

These findings have led to the broad acceptance of the view that there are **a finite and relatively small number of folds found in nature** (Marsden et al. 2006). There are hundreds if not thousands of examples in the structure database demonstrating that highly similar structures may have radically different sequences. So although it is true that highly similar sequences adopt highly similar structures, so too do highly *dissimilar* sequences sometimes adopt *similar* structures.

Thus, it appears that any sequence we choose from the database of sequenced genomes has a high probability of adopting a structure we have already seen. The big question is how to determine which of the 100,000 structures is the right template and how to align our sequence to that structure. **Fold recognition is concerned with the search for scoring functions that can reliably detect the compatibility of a sequence with a known structure and align them accurately when simple sequence similarity cannot be seen**.

Despite the size of sequence space, i.e. the space of all possible protein sequences, the space of protein structures appears considerably smaller. Whether this is related to thermodynamics, the kinetics of folding or to evolutionary selection is difficult to say and beyond the scope of this chapter. However, Magner and

coworkers (Magner et al. 2015) have recently proposed an explanation that suggests thermodynamic stability may be the primary driver for this observation. Regardless of the cause, the restricted nature of fold space is a highly fortuitous fact that has been of great benefit in the field of protein structure prediction.

## 3.2 Pushing Sequence Similarity to the Limits: The Power of Evolutionary Information

The early days of searching a database of sequences for potential homologues was dominated by BLAST and other similar approaches. These were based on use of a generic scoring function such as the BLOSUM or PAM matrices which provide a probability of a mutational transition between one amino acid type and another based on a set of confidently aligned blocks of similar protein sequences. These scoring functions were simple 20 × 20 lookup tables that gave a score for a match between any pair of amino acid types in an alignment. Thus, in general, good scores would be awarded for aligning a hydrophobic residue to another hydrophobic residue (leucine aligned to valine for example) and poor scores were awarded for matching dissimilar residues (glutamate and tryptophan for example). Combining this scoring function with a standard dynamic programming algorithm permitted modest performance in detecting homologous relationships. If one were to search a database of sequences with known structures, and subsequently build a model based on the returned alignment then one would have one of the simplest protein structure prediction techniques.

The obvious shortcoming of this approach is the limited ability of the simple 20 × 20 scoring functions to detect anything but close (>30% sequence identity) homology. Given that we know sequences can diverge well below this threshold of sequence identity whilst maintaining highly similar structures, it was clear that there would be many homologous relationships being missed with this approach, which, if detectable, would permit a substantial increase in our ability to predict structure.

As the sequence databases were rapidly growing in size due to worldwide efforts at genome sequencing, technological developments geared towards using this information efficiently were underway. A simple approach by Park et al. (1997) illustrated how two homologous sequences, which have diverged beyond the point where their homology can be recognised by a simple direct comparison, can be related through a third sequence that is suitably intermediate between the two. Known as ‘intermediate sequence search’, this ‘hopping’ through sequence space showed clear promise, and a more refined approach was developed in Position Specific Iterated BLAST [PSI-Blast; (Altschul et al. 1997)]. Instead of using a fixed 20 × 20 scoring matrix for every protein, and for every position in a protein, one could use a matrix that scores each position in a protein differently. One could construct an $n \times 20$ scoring matrix where $n$ is the length of the protein. This matrix captures the specific propensities of each position in a specific protein sequence to

mutate to one of the 20 possible amino acids. For this reason such a matrix is often called a *position specific scoring matrix* (PSSM) or sometimes just a *profile*.

After an initial standard BLAST scan to collect relatively close homologues, the (pseudo) multiple sequence alignment of these homologues to the target sequence permits one to calculate statistics based on the observed mutations at each position in the target sequence. These statistics form the basis of a new scoring matrix, which can be used for a subsequent round of searching. This process of collecting homologues, building a new scoring function and searching again with this new scoring function can be iterated many (usually between 5 and 10) times, hence the name Position Specific Iterated Blast (PSI-Blast). Coupling this powerful iterative approach with the growing sequence database permitted a substantial improvement in the detection of extremely remote homology. Until very recently, PSI-Blast lay at the core of almost every modern successful structure prediction algorithm. PSI-Blast and BLAST have over 50,000 citations each in the literature making them two of the highest cited scientific papers of all time [combined they are rank 4 of all time (Van Noorden et al. 2014)].

The key to the success of the PSI-Blast approach lies in a realisation that every position in a protein sequence will be under different evolutionary pressures. For example, a glycine in one position may be highly conserved as it is required for a particularly tight turn of the protein chain to maintain its topology. Any mutation in this glycine may be lethal as the protein would fail to fold correctly. A different glycine elsewhere in the sequence may be in a highly variable loop region under minimal selection pressure. Thus when aligning a target sequence against this structure, the first glycine must be present, but the second one may vary. It is this position-specific mutational propensity that permits far more sensitive remote homology detection.

A typical use for PSI-Blast-generated profiles is where the *profile* for a target sequence is scanned against a database of sequences from the PDB, or conversely, a target sequence is scanned against a library of *template profiles*. More generally, a profile of a protein can be considered an ‘evolutionary fingerprint’. This fingerprint captures the evolutionary history of every residue in the protein. These individual histories indirectly reflect the structural environment of a residue and its structural or function role in the protein. It was therefore not long before researchers went beyond matching sequences to fingerprints and tried to match fingerprints to fingerprints, so-called profile-profile matching.

Thus instead of using profiles for only the target sequence or template sequence in isolation, they are used for both and compared to one another (Fig. 3.3). Each position in a sequence can be considered as a vector of probabilities. In the case of simple profiles, one has a 20 dimensional probability vector (1 dimension per amino acid type). A position in the target sequence is similar to a position in a template structure if they are under similar evolutionary pressures, which would be reflected in them having similar probability vectors. Many different techniques have been devised to compare such vectors (the simplest being a dot product), almost all of which surpass the simpler sequence-profile scoring approaches (Soding 2005; Bennett-Lovsey et al. 2008; Rychlewski et al. 2000).

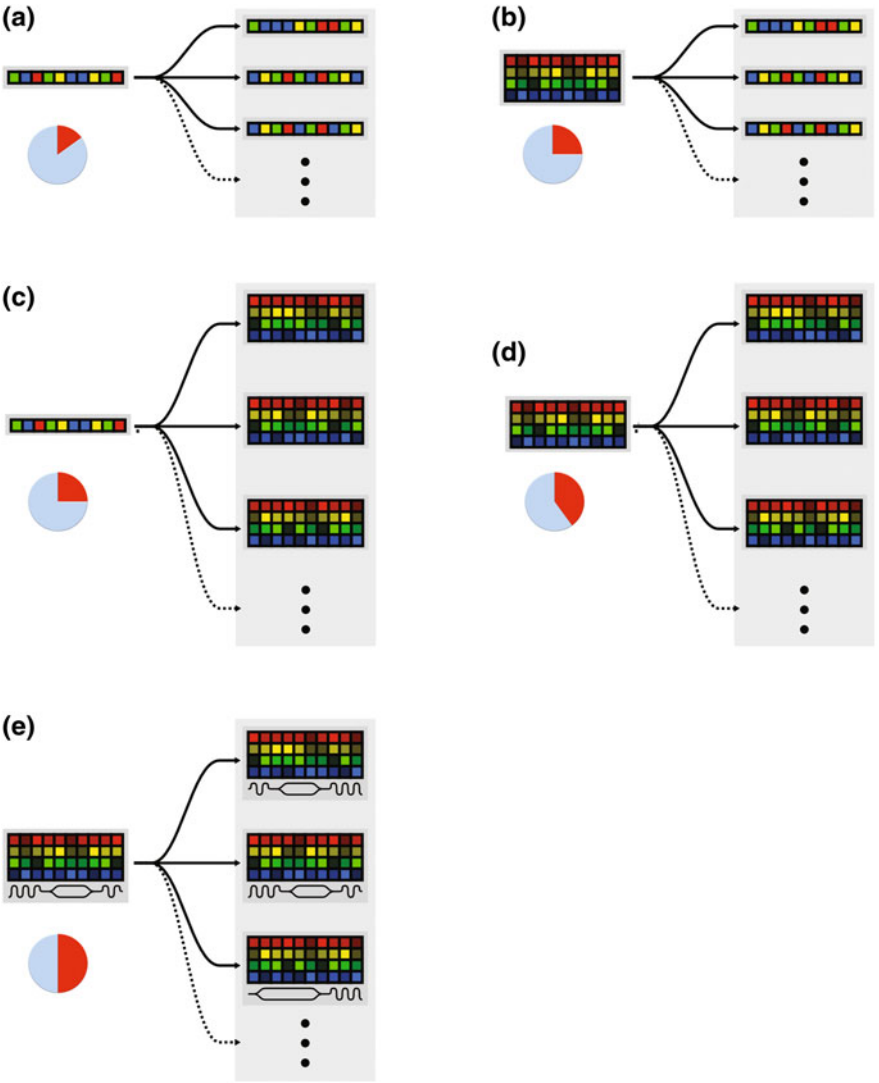



**Fig. 3.3** Schematic representation of the progress of sequence-based fold recognition techniques over time. The *leftmost part* of each figure represents the target sequence. The *grey box to the right* of each figure indicates a database of templates of known structure. The *arrows* indicate a comparison between the query and a particular template. The *pie chart* in each section indicates, in *red*, the approximate proportion of a typical genome that can be confidently modelled by the approach. **a** Simple comparison of an amino acid sequence against a database of sequences. **b** Here the target is represented by a profile of multiple sequences, a PSSM or a hidden Markov model (*the coloured grid*). Each row of the grid represents a different homologous sequence, each column represents a different position in the sequence. **c** The inverse of **b** where now a target sequence is searched against a library of profiles. **d** Profile-profile comparison. **e** profile-profile comparison plus predicted structural features. *Wavy lines* indicate alpha-helices and lozenges indicate beta-strands

### 3.2.1 The Rise of Hidden Markov Models

While profiles were demonstrating strong performance in remote homology recognition, researchers realised that there already existed a well-established, statistically more sound and more powerful approach to capturing the 'evolutionary fingerprint': hidden Markov models (HMMs). HMMs had been used for some time in a range a fields, most notably speech recognition. Their application to proteins would create a new standard for state-of-the-art fold recognition.

Profile hidden Markov models (HMMs) have several advantages over standard profiles. Profile HMMs have a formal probabilistic basis and have a consistent theory behind gap and insertion scores, in contrast to standard profile methods that use heuristic methods. HMMs apply a statistical method to estimate the true frequency of a residue at a given position in the alignment from its observed frequency while standard profiles use the observed frequency itself to assign the score for that residue. This means that a profile HMM derived from only 10 to 20 aligned sequences can be of equivalent quality to a standard profile created from 40 to 50 aligned sequences. The details of the inner workings of HMMs are not appropriate for inclusion in this chapter, and there exist many clear introductions to the method elsewhere.

HMMs had been used for some years in the context of sequence-HMM or HMM-sequence matching. But the work of Soding (Soding 2005) and others successfully applied the idea of profile-profile matching to HMMs in the program HHsearch. HHsearch became one of the leading methods in CASP and was eagerly incorporated into many of the leading predictive systems as we shall later discover.

Thus, to capture a statistically well-behaved 'evolutionary fingerprint' of a protein, one needs to take a target sequence and search the large and ever-increasing sequence database to gather a diverse yet confident array of aligned homologues. From this alignment one constructs a hidden Markov model. This process is repeated for all known structures to create a database of structure-based HMMs. Finally one searches the target HMM against this database using HMM-HMM matching.

It is clear from this description that the raw information that drives the power of this approach comes from the homologous sequences used to derive the HMM. As discussed earlier, this was typically done using PSI-Blast: iteratively search a large sequence database with a fast heuristic approach (BLAST) and build and refine a standard profile at each stage of the iteration. Yet we know that HMMs are more powerful than profiles and that HMM-HMM matching is more powerful still. So if the HMM-HMM approach could be iteratively used during the initial gathering of homologous sequences, the resulting HMM of a sequence will be a superior 'evolutionary fingerprint'. But here a problem arises—computational burden.

Searching a large sequence database (50 million + entries) can only be done in a reasonable time using a range of heuristics like those applied in BLAST. Constructing an HMM for a sequence requires iterative searches of this large database. If one required an HMM for every entry in the database one requires

50 million iterative searches to create the database. Finally HMM-HMM matching is far more computationally intensive than BLAST, so matching an HMM against a database of 50 million HMMs would take a CPU year. These apparent limitations were largely overcome in the program HHblits (Remmert et al. 2012).

HHblits reduces the unmanageable scale of the problem using two approaches: (1) Clustering of the huge sequence database into a representative set that is an order of magnitude smaller ($\sim$3 million). (2) Effectively reducing profile-profile comparison to sequence-to-profile comparison by discretizing the vectors of 20 amino acid probabilities in each HMM column into an alphabet of 219 letters. Each letter represents a typical profile column. The result of these clever heuristics is an approach that is faster than PSI-Blast, has 50–100% higher sensitivity, and generates more accurate alignments.

This brings us to the current summit of purely sequence-based remote homology recognition: (1) Use a target protein sequence to search and gather homologues from the immense sequence database using iterative HMM-HMM matching (HHblits). (2) Use the resulting target HMM to perform another round of HMM-HMM matching against a database of HMMs of known structures. The resulting high scoring targetHMM-templateHMM alignments can then be used for model building. This approach now permits the high confidence structural annotation of approximately half of the human genome (Lewis et al. 2013).

It would be unwise for me to predict that we have reached the limits of extracting all the possible information from pure sequence signal but the current signs are at least that we have reached a plateau (Chubb et al. 2010); how long this remains we will have to wait and see. But even if we have reached the limits of sequence, a vast amount of untapped information remains in the 100,000 experimentally determined 3D protein structures. We shall now see how researchers have successfully harnessed that information to push structure prediction well into the 'twilight zone' of homology.

### *3.2.2 Using Predicted Structural Features*

One of the earliest attempts at extending homology recognition beyond sequence information was developed by Bowie et al. (1991). The idea is based on the fact that certain structural features of a protein sequence can be predicted in the absence of an explicit template. Most notably, the secondary structure, i.e. the locations of alpha-helices and beta-strands, can now be predicted with an accuracy approaching 80% using programs such as PSIPRED (Jones 1999). Given that structure is more conserved than sequence, a pair of remotely homologous proteins will contain similar patterns of secondary structure elements even in the absence of any obvious sequence similarity. In addition, the solvent exposure of a residue can be predicted with relatively high accuracy (e.g. Kim and Park (2004)), as can the presence of tight beta-hairpin turns (e.g. Kumar et al. (2005)). It is worth noting that leading methods to predict these structural features rely on the evolutionary profiles

discussed above and often on machine learning approaches (e.g. neural networks or support vector machines) trained on such profiles.

These predicted structural features provide us with further information that can be used together with sequence matching. When aligning two amino acids from the target and template one can calculate a compatibility score based on a sequence term such as HMM-HMM similarity plus terms involving secondary structure matching and solvent exposure:

$$S_{ij} = Seq_{ij} + SS_{ij} + Solv_{ij}$$

where $S_{ij}$ is the overall score for matching residue $i$ in the target sequence with residue $j$ in the template sequence, $Seq_{ij}$ is the score from sequence similarity (BLOSUM or HMM-HMM comparison) for matching $i$ and $j$, $SS_{ij}$ is the score for matching the predicted secondary structure type at residue $i$ with the known secondary structure at residue $j$, and $Solv_{ij}$ is the score for matching the burial state predicted for residue $i$ with the known burial state at residue $j$. Simple versions of such scoring functions award a fixed value to identical states (helix matched to helix for example) and penalise all other combinations. Often the functions will be more elaborate and be based on empirical observation of the frequency with which the different states tend to be aligned in known homologues, or be weighted by the confidence of the predicted state from the prediction program (e.g. PSIPRED). This is analogous to the progression from a simple identity-based sequence matching matrix towards the more sensitive BLOSUM-style matrix.

Nearly all successful approaches combine secondary structure prediction (and to a lesser extent solvent accessibility) in some form with pure sequence methods and this has repeatedly demonstrated a systematic improvement in fold recognition. Features such as secondary structure and solvent exposure are relatively easy to predict because they are largely determined by features local to the residue in question; a hydrophobic stretch of residues are likely to be buried; a stretch of residues with helical dihedral angle preferences are likely to be a helix. Thus in some sense, these predicted features are a result of *sequence context.* In new work by Meier and Soding (2015b), a more generic context-dependent score is integrated into their HMM-HMM matching program HHsearch that attempts to capture conserved patterns in 13-residue windows of sequence. This method is agnostic about what the conserved patterns 'mean' physically. Some patterns may be correlated to secondary structure but many might not be. This is a data-driven/mining approach. If a contextual pattern is conserved in protein A and in protein B, then A and B may share similar structure in the region of that pattern. The method results in an accuracy improvement in otherwise difficult, remote alignments. The use of context is also being investigated using a more complex approach called *conditional random fields* (CRFs), of which more later.

But the main source of complexity of protein folding is its non-local nature. Residues far apart in sequence can come close together in space. It is this non-locality that contributes significantly to the intractable nature of folding by

computation. Capturing this aspect of the problem requires a different set of tools and new sources of data as we shall now discover.

### 3.2.3 Harnessing 3D Structure to Enhance Recognition

What are we to make of the observation that many highly *dissimilar* protein sequences adopt highly *similar* three-dimensional structures? We have a large body of evidence that suggests that the naturally occurring (native) state of a protein lies in a broad and deep energy well. The protein folds to its (usually, but not always, unique) structure driven by energetically favourable geometry, residue-residue and residue-solvent interactions.

If one were able to understand what geometric, spatial and solvent interactions stabilise a given structure, then one could both detect compatible sequences given a structure and design sequences that fit that structure. This is the concept of **threading**. Given a sequence whose structure we wish to predict, one aligns or 'drapes' this sequence over each of the known structures in our database. In each case one calculates a score to represent how favourable our sequence is with each structure. A structure with a highly favourable score will be our prediction. But what are these favourable interactions and how do we calculate their magnitude? Fortunately, thanks to the diligent work of many experimentalists around the world, we have a database of native protein structures; a database of favourable interactions.

By careful statistical analysis of the distribution of the different amino acid types throughout known protein structures, powerful sequence-structure relationships can be inferred, and used to tackle prediction problems. These empirically-derived or 'knowledge-based' force fields are widely used across the entire spectrum of computational protein structure analysis and their key role in ab initio modelling means many of the details may be found in Chap. 1. Nevertheless, a brief summary will be useful.

### 3.2.4 Knowledge-Based Potentials

To empirically derive rules relating protein sequence to three-dimensional structure requires (1) a large number of examples of sequences and their corresponding structures and (2) a structural feature of proteins one wishes to analyse. A simple illustration of the technique is the generation of a solvation potential. Understanding this simple example will open the way to understanding how almost any structural feature can be encoded and subsequently used to enhance predictive success.

Any globular protein in its folded native state has some residues buried in the (largely hydrophobic) interior and some residues (largely hydrophilic) on the surface exposed to the surrounding solvent. It is straightforward to calculate to what

extent a given residue $r$ is exposed or buried in a protein of known structure. One method, albeit crude, is simply to measure how many other residues are within a certain distance of the residue $r$ (more sophisticated methods are usually used; (Richmond 1984; Kabsch and Sander 1983). So it is possible to compile a list of every residue in every known protein structure together with its associated level of solvent accessibility (in terms of neighbours). With these data it is now possible to use a variety of statistical techniques to attempt to discover any relationship between amino acid type and its propensity to be on the interior or exterior of the protein. The most common methods used are based on statistical mechanics or Bayesian statistics.

First proposed by Tanaka and Scheraga (1976) and later refined by Sippl (1990) and (Miyazawa and Jernigan 1996), we will describe here the method based on Boltzmann statistics. The Bayesian approach is not altogether dissimilar and can be found elsewhere.

First one assumes that protein structures in the database constitute a kind of ensemble and that the levels of solvent exposure of a residue type within proteins distribute themselves according to a Boltzmann distribution. Second, one can calculate the potential of mean force responsible for the observed statistics via the Boltzmann equation. The 'energy' associated with a given property $p$ is:

$$E(p) = -\log\left[\frac{n_{obs}(p)}{n_{exp}(p)}\right]$$

where $n_{obs}(p)$ is the observed value of p and $n_{exp}(p)$ is the 'expected' value of $p$ in a reference state that assumes there are no specific interactions or preferences.

Implementing this usually means discretizing distances and producing a look-up table of force-field values rather than the continuous differentiable functions used in molecular mechanics. From a threading perspective, this look-up table permits one to assign an 'energy' to aligning a target amino acid to a structural position in the template. Each amino acid in the template will have some degree of exposure/burial. Depending on the amino acid type in question, one can reference the look-up table for a value for having, say, a valine that is 30% exposed.

This 'energy' essentially consists of an addition term in the scoring matrix within a standard dynamic programming algorithm. Thus, in addition to maximising the sequence similarity at each position, the structural (in this case solvent) term also contributes to whether a particular pair of amino acids are aligned during the dynamic programming.

There are many sources of variation in the detail of how such potentials are calculated. For example, a force-field may simply be based on the distances between alpha carbons of the backbone which may suffice for relatively crude recognition of the gross topology of a structure. One could add more atom-based interaction sites, possibly to better account for hydrogen-bonding. The framework of the Boltzmann relation is not limited to distances. One may add in angular dependence, or the packing angle between beta-strands. A force-field may have different contributions from residues separated by different distances along the

sequence: i.e. one may use different functions for residues close in sequence ($i$, $i + 3$) and those further apart ($i$, $i + n$; $n > 10$) as mentioned above.

One of the most common uses of this approach is to examine residue-residue contacts distant in sequence. Often termed a 'pair-potential', the idea is to use statistics from the PDB to determine the likelihood of observing amino acid type A in close proximity to amino acid type B. Using such a potential allows one to thread a sequence onto a structure and assess numerically the degree to which the observed pairwise interactions of residues is 'favourable'.

Clearly the power of a threading approach is essentially encapsulated in the power of the energy function. As a result much past and current research focuses on the development of ever more elaborate, and hopefully more powerful, empirical potentials.

### *3.2.5 Summary*

In this section we have covered the central principles of fold recognition. The most naïve approach is to directly compare a target sequence with a template sequence using a generic scoring function to measure residue similarity such as BLOSUM. Instead we see how we can represent a sequence by a range of information at three different levels of description: primary, secondary and tertiary. Primary information comes from sequence: large multiple sequence alignments, capturing mutational preferences expressed as a profile or HMM. Secondary information comes from predicted structural features that are context-dependent but still somewhat locally determined and which may be predicted from local context: secondary structure, solvent exposure. Tertiary information comes from analysis of known 3D structures and incorporates information about contact preferences in 3D space, perhaps residue-solvent, or residue-residue interactions that may be distant in sequence.

So both a target and a template can be represented at each position by these three rich sources of information. One may then try to match this information between target and template optimally, typically by dynamic programming, to produce an alignment from which a model may be generated. Given this background, we shall now turn to how these techniques are successfully combined in real-world systems.

## 3.3 CASP: The Great Filter

As stated in the introduction, CASP provides a much needed filter for selecting which of the vast range of algorithms in the literature should be described in a chapter like this. At the most recent CASP11 meeting (now in its 20th year) there are a handful of leading approaches and one method that has clearly and repeatedly taken the top position in fold recognition. Aside: CASP does not actually use the term fold recognition but instead divides their prediction problems into 3 categories:

template-based modelling (TBM), hard template-based modelling (TBM-hard) and free modelling (FM). For our purposes TBM-hard is the closest to what we would call fold recognition.

### 3.3.1 The Leaders

The recent top groups at CASP11 are representative of the leading methods for the last 4-6 years but of course don't constitute an exhaustive list. I have also chosen to limit my assessment to fully automated approaches as they are both of highest relevance to end-users and are relatively free from unknown variables, such as ad hoc choices made by researchers during the competition. They are Zhang-server [I-TASSER (Yang et al. 2015)], Robetta (Chivian et al. 2005), HHpred (Soding et al. 2005), MULTICOM (Cao et al. 2014), RaptorX (Kallberg et al. 2014) and Phyre (Kelley et al. 2015). Of these, the Zhang-server is the undisputed leader.

One complexity in assessing these systems in terms of their fold recognition ability is that their final accuracy in CASP is a result of many additional factors beyond template identification and alignment. All of these systems perform many complex post-processing steps after template identification that typically increase accuracy substantially. These post-processing steps will be discussed in Sect. 3.4. Another complexity in assessment is the trend for modern successful methods to integrate techniques from the fields of homology modelling, fold recognition and ab initio/template-free modelling into a single system that handles the full range of modelling problems. This can make teasing out the components relevant only to fold recognition, the middle ground of modelling difficulty, problematic. But for now, let's look at how their core fold recognition engines work under the hood/bonnet.

The methods listed can be divided into two general classes: single algorithm and consensus methods. I shall first discuss single algorithms that either succeed on their own or are incorporated into larger consensus systems.

### 3.3.2 Individual Algorithms

HHsearch (HHpred server) (Soding et al. 2005) is a standalone powerful algorithm ranking highly over several CASP competitions. As discussed earlier it combines HMM-HMM and secondary structure matching. In addition it forms one of the modules in four of the six leading techniques: Zhang-server, MULTICOM, Robetta and Phyre.

Sparks (Zhou and Zhou 2004) is a method that does not rank highly by itself, but clearly adds value to consensus methods as it is used in both Robetta and

Zhang-server. Sparks uses standard profile-profile and secondary structure alignment plus two knowledge-based potentials: torsion angles and solvent accessibility. We have already seen how solvent potentials can be derived and it should come as no surprise that torsion angle potentials are derived in an almost identical fashion; protein backbone torsion (phi/psi) angles are first discretized into bins. One then applies the Boltzmann formalism to a large set of observed torsion angles for each residue type in solved structures to create a form of log-odds energy look-up table for each amino acid type and each torsion angle bin.

RaptorX (Kallberg et al. 2014) uses profile-profile, secondary structure and solvent accessibility. But unusually, it uses a rather new approach to alignment called *conditional random fields* (CRF). Most protein threading methods use a scoring function linearly combining sequence and structure features (see equation in Sect. 3.2.2) to measure the quality of a sequence-template alignment so that a dynamic programming algorithm can be used to optimize the scoring function.

However, there are two problems with this central idea: Firstly there are well known correlations between scoring terms, for example secondary structure states and solvent accessibility states. Secondly, the relative importance of different features will vary depending on the level of sequence similarity in a region or the richness of a profile in a region. A standard linear combination of scoring terms cannot fully exploit such interdependency among features and thus, limits alignment accuracy.

Hence it could be argued that for a truly optimal alignment one would require a function that varies from position to position along a protein, varying the contribution of different sequence and structural factors in the alignment. So just as we have historically moved from position-agnostic (BLOSUM) to position-specific (profiles/HMMs) scoring, the CRF moves from a position-agnostic *algorithm* to a position-specific algorithm: different weights for the components of the scoring at each position dependent on the data available. Determining the optimal contribution of these terms at each position is the purpose of a conditional random field and there exist algorithms to estimate the parameters for such models. In the case of RaptorX, the authors applied the technique of regression trees. The details of this approach are beyond the scope of this chapter. Further information is available in (Peng and Xu 2010).

Finally there is the Phyre server (Kelley et al. 2015). Phyre uses HHsearch as its core fold recognition technique. The models generated are then combined using a novel post-processing method that combines multiple models with virtual synthesis as described in Sect. 3.4.

N.B. It should be noted that I have omitted the LEE group (developer of the *nns* server at CASP11) from my list of high performing groups. This is simply because their core methodology is not yet published. However in their abstracts they indicate that they too use CRFs along lines similar to the RaptorX approach.

### 3.3.3 Consensus Methods

Consensus methods have been successful for many years in CASP. The reasons for this are twofold: (1) Every method has its strengths and weaknesses so different methods may succeed when others fail. (2) Spurious individual results tend to be overpowered by the majority: akin to the power of crowds. When designing such systems there tend to be two philosophies: (1) minimise the overlap between methods (maximise their mutual orthogonality) or (2) throw everything into the mix and let the crowd sort it out.

The Robetta group appears to have chosen route (1) In their server they incorporate the three individually strong methods above into their server: HHsearch, Sparks and RaptorX. Robetta uses an iterative recognition procedure whereby models produced by the above three methods are clustered. Remaining regions of the input target sequence that are not covered by templates are excised and re-enter the process for another round. This is done to detect domain boundaries in the protein and repeated until no further regions can be modelled with templates. The templates found by these methods are then processed by the advanced Rosetta suite of modelling tools from the Baker lab of which more in the post-processing section.

MULTICOM and Zhang-server on the other hand, appear to have taken route (2) The MULTICOM server (Cao et al. 2014) uses at least 9 different profile alignment programs (PSI-Blast, HMMER, CS-BLAST, COMPASS, PRC, SAM, HHSearch, MUSTER, RaptorX) to generate models that enter a pool for later processing. The different algorithms are run on different fold databases generated by different criteria. This results in a large pool of potential models and the different databases used contribute to the relative independence of the results from one another. The power of this approach largely stems from the size and variation of the pool coupled with the way these potential models are selected and combined. I will briefly discuss this in the post-processing section near the end.

And finally we come to the Zhang-server (also known as the I-TASSER server, open to the public), which has held the top spot at CASP for at least 6 years. The Zhang-server can be roughly divided into two major components: (1) Template detection and (2) Model assembly. The template detection component of the Zhang-server is called LOMETS (Wu and Zhang 2007). The LOMETS system uses the HHsearch and Sparks methods previously described. In addition it uses 7 other available methods (FFAS-3D, pGenThreader, PRC, PROSPECT2, SP3 and MUSTER and PPAS). Of these, the PPAS method has 7 variants itself, all of which are used within LOMETS. So in total, this amounts to 15 different profile-profile/threading methods. I will not go into the minutiae of these 15 components. Suffice it to say they are all some variant of the [profile-profile + secondary structure + optional potential] paradigm that we have seen throughout.

## 3.4 Post-processing

For many of the leading systems, recognising a template and generating an alignment is just the beginning of an elaborate and complex process of model building and refinement. Both single and consensus methods produce ranked lists of candidate templates. These will be associated with various imperfect confidence scores and may cover different regions of a target protein. In earlier CASP experiments, it was typical to simply build a model based on the single highest scoring alignment from a single ranked list from a single method. This simple approach has been superseded.

Today, a range of techniques, as diverse as those used for fold recognition itself, are applied to the problem of optimally choosing and recombining candidate models. Broadly these methods fall into the categories of (1) clustering (2) model quality assessment, and (3) multiple-template modelling. The additional problem of modelling regions for which a template cannot be found, so-called template-free or ab initio modelling, is often inextricably bound up in with these techniques. However, this is a problem in its own right and is discussed in detail in Chap. 1.

These three approaches can be used multiple times and at different stages of a structure prediction protocol. We will see examples of their use in the leading CASP methods in Sects. 3.4.2. First we will discuss the principles underlying them.

### *3.4.1 Choosing and Combining Candidate Models*

A typical prediction pipeline begins with producing a set of candidate models for different regions of a target protein, using the techniques described above. Often this initial set comes from different core fold recognition algorithms and from different high scoring templates for each algorithm. These models will later be combined using multi-template modelling into a single final result. To maximise the accuracy of this combined model, the quality of the input models should be as high as possible. Thus some method is desirable that can filter out candidate models from the initial pool that are either particularly non-native-like or that disagree substantially with the consensus.

Similarly, the subsequent process of multi-template modelling, as we shall see later, often involves producing tens, hundreds or thousands of candidate solutions. Once again it is desirable to have some method to select form this pool the model most likely to be closest to the true structure.

Many methods are capable of producing high quality models that are otherwise hidden in a sea of low-quality models. Hence the ability to choose the best model(s) from a pool is of critical importance at multiple stages in structure prediction. We will start with probably one of the intuitively simplest and still powerful methods for performing this model selection: clustering.

#### 3.4.1.1 Clustering

In an analogous way to how consensus fold recognition techniques can improve accuracy, clustering models allows the detection of models or regions of models where there is agreement across templates and alignments (confident regions) and where there is significant disagreement (non-confident regions). An intuitive explanation for why this is effective comes from a consideration of the size of conformational space. There are an enormous number of possible conformations for a protein. Hence the likelihood that two relatively independent methods generate similar structures by chance is extremely low. Thus, when a region of similar structure is observed in a pool of models generated by either a range of techniques or a range of templates, there is good cause to be confident in the structure of that region.

One of the simplest methods for predicting model accuracy via clustering is also one of the most powerful: 3D-Jury (Ginalski et al. 2003). 3D-Jury takes as input a pool of models of the same sequence (with possible insertions and deletions). These models are superposed in a sequence-dependent (as opposed to structure-dependent) manner. A standard method for superposition is the MaxSub method (Siew et al. 2000). This algorithm aims to find the maximum subset of atoms in two models that are superposable within some distance threshold, typically 3.5 Å. The size of this subset is the MaxSub score between a pair of models. Given a pool of models, an all-versus-all MaxSub calculation is performed to determine the similarity of every model to one another. The model with the highest similarity to all other models is the 3D-Jury ('best') model. There are variants to this approach to cope with very large pools (thousands or tens of thousands) and to generate multiple top solutions (hierarchical clustering followed by centroid detection), but all share a common principle: recurring features are more likely to be correct than rarely observed features. Hence models with the largest numbers of recurring features across the pool are likely to be closest to the native structure.

Following the success of clustering, further research efforts focussed on the more general problem of model selection, creating a sub-field of model quality assessment.

#### 3.4.1.2 Model Quality Assessment Programs (MQAPs)

Before explaining the general principle of MQAPs, an aside is appropriate:

The protein structure prediction problem is about generating a structure for a sequence and we have seen how difficult that is because of our lack of fundamental understanding of the mechanisms of folding. Although MQAPs may appear at first to be a technical addendum to a structure prediction protocol, they are much more than that. MQAPs are the protein structure prediction problem in another guise. Hypothetically, if one had a perfect method for assigning quality to a protein model,

then the structure prediction problem would be largely solved. All that would be required would be to generate models blindly and keep the ones with the best quality score. Although model generation is not trivial, having a system that could tell you how right or wrong a model was would solve the chief problem. Unsurprisingly, just like folding, model quality assessment is a notoriously difficult problem and many techniques have been employed to tackle it.

The sub-field of 'model quality assessment' is itself assessed at CASP as a separate category of problem (Kryshtafovych et al. 2014) and is a subject in its own right and so only a brief treatment of the subject will be given here. There are two broad categories of MQAP: single-model and multi-model. Single-model MQAP systems look only at the 3D coordinates and amino acid types of an individual model and produce a confidence score. 'Multi-model' systems look at a pool of models generated for the same sequence often from multiple templates and alignment algorithms. Multi-model systems typically outperform single-model systems as they have access to a distribution of models. We have seen already regarding clustering how this distribution contains valuable information on which regions are likely to be correct or incorrect in the model. This is the principle reason multi-model MQAPs surpass single-model methods. This in turn explains the relative success of methods that use a range of fold recognition algorithms (I-TASSER, MULTICOM) to generate a pool of candidate models. This pool contains rich information beyond that which a single alignment technique can produce and this information can be teased out by clustering or multi-model MQAPs.

The underlying methodology of an MQAP usually falls into one of two categories: (1) empirical potential-based or (2) empirical potentials + sequence/structure features + machine learning. Earlier we have seen how empirical potentials can be used to assign a 'score' to a model. A wide range of empirical potentials have been derived for this problem each with their own strengths and weaknesses. It should come as no surprise then that often a consensus of such energy functions is used in practice. Commonly used potentials include DOPE (Shen and Sali 2006) and DFIRE (Zhang et al. 2004). For category (2), a wide range of structural 'features' are calculated from a model or set of models. Features may include: solvent accessibility, residue type, packing angles, torsion angles, secondary structure types, frequency of observation across models etc. In addition, features may be position-specific (i.e. per-residue) or global features of the model (e.g. globularity or percentage of exposed hydrophobics). These features are combined with empirical energy scores to train a machine learning system (often a support vector machine or neural network) to discriminate between good and bad models. The trained machine learning model is then used to assign a score to a given model or pool of models. Examples of widely used MQAPs include: Modfold (McGuffin 2009), Pcons (Larsson et al. 2011), ProQ2 (Wallner et al. 2003) and QMEAN (Benkert et al. 2009).

#### 3.4.1.3 Combining Models Optimally—Multiple Template Modelling

Clustering and model selection by MQAPs can reduce a large pool of candidates into a manageable subset containing an enriched level of high quality models. But how best to combine the information from these models to create a final prediction that is both protein-like and reflects the conformational distribution of the input models?

Multiple template modelling has been around for many years beginning in the area of comparative or close-homology modelling. Here a set of highly similar, typically high sequence similarity templates are superposed and this superposition used to guide the model building. This was first established as a technique in programs such as Comparer (Sutcliffe et al. 1987) and Modeller (Sali and Blundell 1993) and is described in detail in Chap. 4. The guiding principle here is similar to that of clustering. A pool of candidate solutions (input models) describes a probability distribution for the potential positions of atoms in space. The challenge is to use information from that distribution to generate a self-consistent 3D model that simultaneously adheres as closely as possible to that distribution whilst also adhering to basic principles of protein structure. We will see in 3.4.2 how the de facto standard tool Modeller is still widely employed, but so are unique methods such as that of Phyre, I-TASSER and Robetta.

It is at the stage of combining multiple models that the problem of template-free modelling begins to impinge on the problem of fold recognition. Fold recognition lies in the intermediately difficult regime of structure prediction between homology modelling and template-free prediction. In the fold recognition regime there will thus often be regions, sometimes of substantial length, that need to be modelled in the absence of a template. This problem is usually tackled within the multiple-template modelling component of the prediction pipeline. As a result, although template-free modelling is given a thorough treatment in Chap. 1, it will be unavoidable to touch upon it in the next section. Now let us turn to how all three techniques, clustering, MQAPs and multiple-template modelling are applied in practice in various combinations in the leading systems.

### *3.4.2 Post-processing in Practice*

This section delves into some of the gory details of approaches used by the leading groups. As such it is not for the faint of heart but persevering with it I hope will be rewarding and at least give a somewhat accurate flavour of the lengths (or depths) to which this field has striven for the sake of model accuracy.

In Phyre, multiple models are chosen so as to simultaneously maximise the confidence in the models chosen and the coverage of the target protein. Confidence values in this case are taken directly from HHsearch without the use of an explicit MQAP. The resultant subset of high confidence, high coverage models provides a set of confidence-weighted distance constraints. These distances are treated like

linear elastic springs in the Phyre module called *Poing* (Jefferys et al. 2010). Poing is a simplified, fast folding simulator that reduces the protein representation to a 'beads-on-a-string' model where a residue is described by two spheres: a virtual alpha-carbon and a virtual sidechain. The protein is 'synthesised' one residue at a time from a virtual 'ribosome' (a large heavy sphere) in the context of springs representing the distance constraints taken from the input models. This synthesis model is not expected to reflect reality but instead is a useful computational technique to prevent the system from becoming tangled as the distance constraints are slowly introduced. The aim of this approach is to generate a protein-like model that reflects as closely as possible the distances observed in the input models whilst preventing clashes and non-native like local conformations. Unconstrained, template-free regions are encouraged to adopt their predicted secondary structure and to be buried if hydrophobic using a solvent bombardment model. More details are available in (Jefferys et al. 2010).

RaptorX (Kallberg et al. 2014) uses Modeller to construct models for high ranking target-template alignments and re-ranks these models using a neural network MQAP to predict model quality. The MQAP takes as input the contexts of two sequence residues and yields their distance probability distribution. The context of one residue includes sequence profile, predicted secondary structure and amino acid chemical properties in a local window centered at the residue of interest.

In HHpred (Soding et al. 2005), a similar approach to trying to maximise target coverage with a minimum number of confident templates is used. However, in this case the first template chosen is selected by a regression neural network (an MQAP) that predicts the model quality based on four input features from the HHsearch results: HHsearch raw score, secondary structure score, template resolution, and length-normalized sum of posterior probabilities over all aligned residues. Subsequent models are chosen that have high confidence from HHsearch and that cover extra regions of the target whilst trying to minimise the total number of templates used.

These chosen templates are then input to a modified Modeller protocol. Modeller by default treats all input constraints from models as of equal merit. In HHpred, this is modified so that distance constraints reflect position specific alignment confidence. Hence lower quality constraints from weak HMM-HMM alignments are treated as weaker within Modeller, placing greater strength on constraints from confident templates (Meier and Soding 2015a).

The MULTICOM suite of tools and servers constitute a large array of individual methods. A key feature is the heavy use of MQAPs. In MULTICOM, a large pool of candidate models is available from many core alignment algorithms. Models generated are assessed by up to 14 different MQAPs (ModFOLDclust, ProQ2, Pcons, and many more) Each method produces a ranking of models and a consensus of this ranking is used as a final selection criteria. In cases of particularly difficult targets where templates were undetectable, this pool is supplemented by hundreds of template-free models generated by Rosetta (Rohl et al. 2004) Model combination is again done using Modeller but in its default setting. Here it seems

the large input pool from many fold recognition techniques plus careful selection of models using MQAPs is a key to performance.

In Robetta, the central technique for creating a final model is a combination of a highly refined empirical energy function and a conformational sampling technique that uses both template-based models and an approach called fragment assembly. Fragment assembly forms the core of the Baker group's approach to template-free modelling and will be briefly described here. A more complete description is available in Chap. 1. Robetta is named after the Baker group's suite of tools called Rosetta and stands loosely for 'Robot Rosetta' to indicate how it is an automated server in CASP.

To perform fragment assembly, a protein sequence is first divided into short (typically 3- and 9-residue) overlapping fragments of sequence. These fragments are modelled using similar approaches to those for fold recognition, creating for each fragment, a pool of candidate local conformations. In a free modelling context where no templates are available, these fragment pools are used as a source of potential conformations: the aim being to combine these fragments in such a way as to build a protein model with high score according to a wide range of empirical potentials/energy functions. The principle operating here is that most or all viable local conformations of proteins are available in the current PDB. Again this is an empirical approach: using the known structures to indirectly tell us what conformations are acceptable. By sampling fragments at different positions along the target chain and inserting them into the nascent model, the model in its entirety explores conformational space. However it does this only by visiting 'allowed' local regions because of the use of native structural fragments. This thus drastically reduces the search space required for exploration.

In the fold recognition regime, where one or more candidate templates are available, a mixed approach is used. The template-based models are clustered, and distance restraints from these templates are calculated. These models are then fragmented at secondary structure boundaries, generating large fragments. These larger chunks supplement the small 9-residue fragments and form two pools of likely conformations. During modelling these different pools are sampled with small template-free fragments applied to those regions not covered by a template; to 'fill in the gaps' using a free modelling approach where fold recognition fails. Fragments are selected or rejected during modelling using a combination of the sophisticated Rosetta empirical energy function and distance-based restraints from templates.

This process is repeated often thousands of times and final models are selected by clustering the best scoring 100 models from each topologically distinct alignment cluster, and then averaging the models within each cluster. Finally, these averaged structures are refined using another Rosetta protocol tuned to make minor adjustments to ensure local protein-like geometry whilst preserving overall topology. This three-step approach of clustering, averaging and refinement was pioneered by I-TASSER as described below.

In Zhang_server (I-TASSER), the result of its many fold recognition algorithms is a set of models ranked by confidence. Each model is segregated at gap points into

contiguous template-based (sub-)models. These are the large 'fragments'. The conformation of these regions is fixed. However they are allowed to rotate and translate freely, as rigid bodies, in the I-TASSER simulation. The regions without such fragments are modelled on a lattice to reduce search space. Ab initio modelled residues connect the larger, more confident template based 'chunks'. The simulation allows these uncertain regions to move within the lattice, which in turn alters the packing of the larger more confident chunks. Throughout, the varying model is assessed by a battery of energy functions: the aim being to be able to pack these chunks of confident structure together in such a way as to resemble known protein structures.

A large number of these simulations are performed and the resulting models clustered by their program Spicker (Zhang and Skolnick 2004). The clusters with the most members are expected to be closest to the native. For these selected clusters an averaged structure is produced. This average is searched structurally against the Protein Data Bank of known structures to find similar native structures. Restraints from these known structures are pooled with those from the average as well as the original set of input models and all of this information sent back into the simulation for another round. The output is clustered again, representatives selected and then refined using a fragment-based technique called Modrefiner (Xu and Zhang 2011). The models output are finally ranked by several MQAPs [DOPE (Shen and Sali 2006), Rwplus (Zhang and Zhang 2010), GOAP (Zhou and Skolnick 2011)], a consensus of which is used to make a final selection. If you the reader have made it this far, then congratulations are in order!

This system, which remember is also the undisputed leader in the field for some years, illustrates a number of successful principles. We see extensive use of consensus as a predictor of structure—15 different recognition algorithms used to find and align templates. Not only that, but we see a delicate balance of simulation, clustering, and selection, and how entire modules of the protocol are iterated. All of this is geared towards finding a consensus in noisy data: Use all the tools available to make a set of structural guesses, combine those guesses, cluster them, select subsets, assess them, combine again etc. (Fig. 3.4). It is in this area I have termed 'post-processing' on which much of the development in leading methods has been focussed. In light of the complexity of approaches such as I-TASSER and Robetta, I personally hope we will, through better understanding of why these approaches succeed, be able to streamline and simplify this stage of modelling in the future; if for no other reason than to spare the sanity of future researchers.

### *3.4.3 Use of Contacts*

In Sect. 3.1.2 I mentioned a recent breakthrough in our ability to predict which residues are in contact from sequence information alone. I deliberately didn't include their use in the above pipelines for fear of making an already complicated description completely incomprehensible. However, in the last 2 years we have

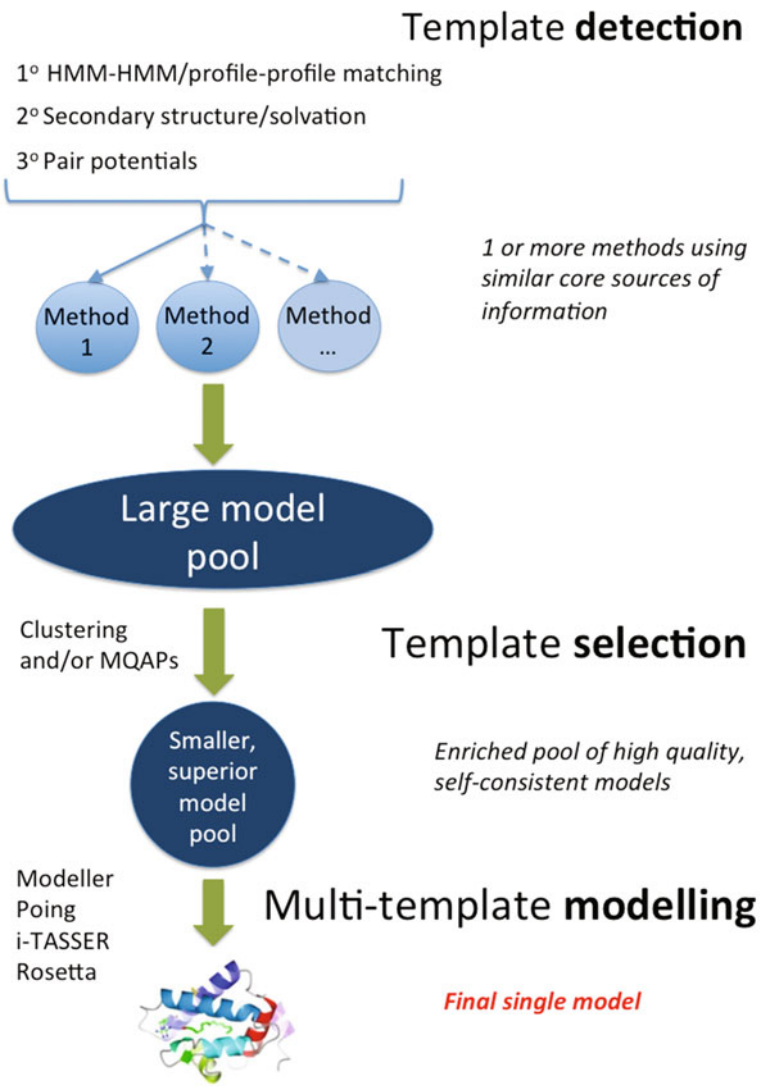



**Fig. 3.4** General flow of structure prediction for a typical CASP predictor. Initial models are produced, sometimes using a range of methods and often integrating 3 levels of information from primary, secondary and tertiary sources, as described in Sects. 3.3 and 3.4. This large model pool is often refined using clustering and/or model quality assessment programs (MQAPs). Finally this refined pool of models are often used as input to multiple template modelling programs such as Modeller

begun to see the first attempts to slot in this new and potentially valuable source of information into existing prediction pipelines.

In I-TASSER, for 'hard' or 'very hard' targets (their internal classification) predicted contacts from several programs are used to supplement distance constraints from templates. In Robetta, the group's own GREMLIN (Kamisetty et al. 2013) program is used to predict contacts, the clusters are re-ranked using this information, and the spatial restraints are supplemented with the predicted contacts. In Phyre, predicted contacts are also used to augment the constraint springs in the sub-module Poing.

It was expected by some at the last CASP11, this author included, that this new information would lead to a dramatic improvement in prediction accuracy. However, with the exception of one notable case in the free modelling section of CASP, it was far from clear whether this contact information had any significant impact. The reasons for this are interesting.

To predict contacts with a useful level of accuracy requires a substantial number of sequence homologues: typically >500 as a minimum and >1000 ideally. In addition, these homologues must demonstrate significant diversity—1000 virtually identical homologues provide little useful information. This large number is required to garner a sufficient statistical signal of mutational correlation between positions.

However, a large number of sequence homologues for a target indicates a protein that is widespread across organisms. This in turn indicates a protein that is likely to both (a) have a powerful sequence profile or HMM and (b) a protein likely to have had at least one of its homologues already structurally determined experimentally as it is likely to be well studied due to its presence in many organisms. In short, proteins with large number of homologues in the sequence database are likely to already be relatively easy to model by existing template-based approaches. Thus, in such cases the addition of a relatively noisy signal from predicted contacts is unlikely to have any significant impact on the final predicted, already high, model accuracy.

The hardest modelling targets in fold recognition tend to be those for which no template can be confidently detected. This is often caused by poorly characterised profiles or HMMs, in turn caused by few sequence homologues. These also of course are the proteins for which contact prediction fails. In this light, the current modest impact of contact prediction on structure prediction is not altogether surprising. However, new approaches that attempt to use contact information in a different manner show promise. One avenue is described below.

#### 3.4.3.1 From Sequence to Profiles to Contact Maps

Recent work in our own lab has investigated whether predicted contacts can be used in an approach directly analogous to classic fold recognition. Instead of comparing a protein HMM to a library of HMMs of known structure, we substitute HMMs for predicted and known contact maps. This involves developing techniques that can

accurately align a (often noisy) predicted contact map with a template contact map. The technical challenges of overcoming prediction noise in the contact map and aligning two-dimensional (contact maps) rather than 1-dimensional (sequences) objects are non-trivial. However, methods have been developed [e.g. Al-eigen (Di Lena et al. 2010)] based on the noise-tolerant approach of eigendecomposition which show promise.

We are at an early stage of development, but if successful, this new approach could open a doorway to structure prediction where sequence similarity or even remote homology is unimportant. Powerful approaches such as HHsearch can tackle roughly 50% of a typical genome. A sizeable proportion of the remaining 50% is likely to consist of proteins of known folds but with sequence similarity so remote that templates cannot be detected.

The contact map approach we are pursuing, called PhyrePower, is completely agnostic regarding sequence similarity. It operates solely in the space of contacts/structure. Preliminary results suggest that in almost 50% of cases, this method can detect appropriate templates completely missed by state-of-the-art sequence approaches such as HHsearch. Even when there are few homologous sequences and thus a poor contact prediction, the new method is able to correctly detect templates, although challenges of alignment accuracy remain. We will have to wait and see how this area develops.

## 3.5 Tools for Fold Recognition on the Web

All of the foregoing discussion is centred on methods that perform well in CASP. This has been helpful to focus attention on the developments in the field and how the boundaries of accuracy are being consistently pushed. However, it is critical to understand that for most biologists and most modelling tasks, the differences in accuracy reflected in CASP between the very best methods and the majority of methods are completely unimportant. Almost all of the methods in the top half of CASP assessed servers will produce excellent models, except in the hardest cases, whose minor differences are largely inconsequential for what many biologists want a model for. In the hardest cases, the top CASP methods may produce a useable model. But in such cases the confidence estimates are usually very low. In CASP, confidence estimates are largely ignored. In the real world, they matter enormously in determining whether a researcher trusts a prediction enough to pursue a subsequent experiment that may consume substantial time, money and effort. Although the top CASP methods can sometimes produce superior models where other methods fail, typically they can't tell you when this is so.

Biologists need tools that are largely in step with state-of-the-art, but more crucially, they need tools that are easy to use, understand, navigate, and that return results in a reasonable timeframe. In addition these tools need to be kept up-to-date with the ever-increasing databases of structure and sequence, and be maintained as a viable service. Very often we find servers in CASP that perform well only to

**Table 3.1** Selected Fold Recognition (FR) servers

| Server name | Web address | Consensus/single | FR/ab initio |
|---|---|---|---|
| HHpred | toolkit.tuebingen.mpg.de/hhpred | Single | FR |
| I-TASSER | zhanglab.ccmb.med.umich.edu/I-TASSER | Consensus | FR + ab initio |
| PCONS | pcons.net | Consensus | FR |
| pGenThreader | bioinf.cs.ucl.ac.uk/psipred | Single | FR |
| Phyre2 | www.imperial.ac.uk/phyre2 | Single | FR + ab initio |
| RaptorX | raptorx.uchicago.edu | Single | FR |
| Robetta | robetta.bakerlab.org | Consensus | FR + ab initio |

discover that as web services, they are transient and disappear as soon as CASP is over or funding has run out. In this light I have chosen a handful of publicly available, maintained, easy to use servers (Table 3.1).

We have seen how consensus in all its forms drives much of the success of the techniques described. This applies to users as well. In general, from a biologist/user perspective, it is unwise to place all one's trust in a single tool/server. For any modelling problem where there is any significant doubt about the result, it is safest to employ a range of leading methods and assess their mutual agreement before drawing conclusions.

## 3.6 The Future

In this chapter we have seen how a diverse set of ideas has been used to tackle the problem of protein structure prediction in the absence of clear sequence similarity. Enormous research efforts have gone into detecting homology between divergent proteins whose common ancestor has vanished millions of years previous. We only have available to us those proteins that exist now. Yet the variety we observe in the proteins of today's organisms gives us insight into the past. We have access only to the leaves on the tree of life. But by analysing sets of nearby leaves we can create a statistical model of what existed closer to the trunk. By taking two apparently unrelated leaves from the tree, we can show they are connected by a common branch. This is HMM-HMM matching.

As genomic sequencing continues, our picture of the current leaves on the tree of life becomes ever more detailed. As this detail increases, we improve in our ability to move up the tree and discover the connections between different branches. Current methods such as HHsearch are able to detect homology to known structures for about 50% of a typical proteome. This fraction will continue to rise for a number of reasons.

As sequencing continues, the profiles or HMMs we can construct for a protein will improve in accuracy and thus recognition power. As we refine our techniques to capture contextual links both close and far in a sequence, we improve our

'evolutionary fingerprint' of a protein. HMMs and simpler profiles capture information for each position in a protein sequence. But there appears to be important contextual information as well. A single residue position does not exist in isolation. Secondary structure for example, is determined by sequence context and we have seen how widespread the use of secondary structure matching is in leading methods. Tools are now being developed that harness this contextual information explicitly and it shall be interesting to see how this approach develops.

Beyond local context, positions distant in sequence can leave traces on one another in the form of correlations in mutations aka evolutionary covariance. And this in turn reveals signals indicating contacts in 3D space. The work in this area has exploded in recent years and holds out much hope for a leap in structure prediction accuracy.

Although growing far more slowly than sequence data, structure data also continues to increase. As more proteins are structurally solved by experiment, our database of 'answers' grows and our likelihood of finding a known homologous structure improves as a result.

Finally we see the amount of effort that has been placed on selecting and combining models, using the power of consensus and variation to refine our picture of the probability distribution for a protein model. Navigating through this distribution to construct a model that best reflects what we know from templates and empirically derived energies is an active area.

However, there may be a limit to how far this general approach will succeed. We are discovering a large number of proteins exhibit a considerable degree of intrinsic disorder (see Chap. 6). Such proteins cannot be meaningfully structurally modelled by any of the approaches described in this chapter. In addition, many proteins are rare and occur only in a relatively small number of organisms. For such cases our knowledge of the mutational preferences in the protein are limited by few homologous sequences. Here neither profiles/HMMs nor contact prediction is of much value. The proportion of proteins in a typical genome that fall into this category of protein 'dark matter' is uncertain, but work in this area is beginning to shed some light (Perdigao et al. 2015). But it is clear that the approaches described in this chapter will never be able to produce accurate models for all proteins nor be able to design new protein folds. For that we require a deeper understanding of folding and a method that does not rely on copying structure based on homology.

The desire to 'solve' the protein folding problem is alive and well as one of the holy grails of molecular biology. To understand protein folding is to understand how the 'software' of DNA becomes the 'hardware' of functional proteins. It is to understand, at a fundamental level, the nature of living things. However, there may be no elegant solution to the protein folding problem. Nature does not necessarily find an elegant solution; simply one that works. Reluctantly, we may have to be satisfied with a complex predictive framework. Nevertheless the hope for a simple, computationally tractable and hitherto undiscovered explanation for protein folding remains strong.

## References


Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Benkert P, Tosatto SC, Schwede T (2009) Global and local model quality estimation at CASP8 using the scoring functions QMEAN and QMEANclust. Proteins 77(Suppl 9):173–180. doi:10.1002/prot.22532

Bennett-Lovsey RM, Herbert AD, Sternberg MJ et al (2008) Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins 70 (3):611–625. doi:10.1002/prot.21688

Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28 (1):235–242

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170

Cao R, Wang Z, Cheng J (2014) Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. BMC Struct Biol 14:13. doi:10.1186/1472-6807-14-13

Chivian D, Kim DE, Malmstrom L et al (2005) Prediction of CASP6 structures using automated Robetta protocols. Proteins 61(Suppl 7):157–166. doi:10.1002/prot.20733

Chubb D, Jefferys BR, Sternberg MJ et al (2010) Sequencing delivers diminishing returns for homology detection: implications for mapping the protein universe. Bioinformatics 26 (21):2664–2671. doi:10.1093/bioinformatics/btq527

Di Lena P, Fariselli P, Margara L et al (2010) Fast overlapping of protein contact maps by alignment of eigenvectors. Bioinformatics 26(18):2250–2258. doi:10.1093/bioinformatics/btq402

Ginalski K, Elofsson A, Fischer D et al (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19(8):1015–1018

Jefferys BR, Kelley LA, Sternberg MJ (2010) Protein folding requires crowd control in a simulated cell. J Mol Biol 397(5):1329–1338. doi:10.1016/j.jmb.2010.01.074

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195–202. doi:10.1006/jmbi.1999.3091

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637. doi:10.1002/bip.360221211

Kallberg M, Margaryan G, Wang S et al (2014) RaptorX server: a resource for template-based protein structure modeling. Methods Mol Biol 1137:17–27. doi:10.1007/978-1-4939-0366-5_2

Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A 110(39):15674–15679. doi:10.1073/pnas.1314045110

Kelley LA, Mezulis S, Yates CM et al (2015) The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 10(6):845–858. doi:10.1038/nprot.2015.053

Kim H, Park H (2004) Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 54(3):557–562. doi:10.1002/prot.10602

Kryshtafovych A, Barbato A, Fidelis K et al (2014) Assessment of the assessment: evaluation of the model quality estimates in CASP10. Proteins 82(Suppl 2):112–126. doi:10.1002/prot.24347

Kumar M, Bhasin M, Natt NK et al (2005) BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. Nucleic Acids Res 33 (Web Server issue):W154–W159

Larsson P, Skwark MJ, Wallner B et al (2011) Improved predictions by Pcons.net using multiple templates. Bioinformatics 27 (3):426–427. doi:10.1093/bioinformatics/btq664

Lewis TE, Sillitoe I, Andreeva A et al (2013) Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic Acids Res 41 (Database issue):D499–D507. doi:10.1093/nar/gks1266

Magner A, Szpankowski W, Kihara D (2015) On the origin of protein superfamilies and superfolds. Sci Rep 5:8166. doi:10.1038/srep08166

Marsden RL, Lee D, Maibaum M et al (2006) Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. Nucleic Acids Res 34 (3):1066–1080. doi:10.1093/nar/gkj494

McGuffin LJ (2009) Prediction of global and local model quality in CASP8 using the ModFOLD server. Proteins 77(Suppl 9):185–190. doi:10.1002/prot.22491

Meier A, Soding J (2015a) Automatic prediction of protein 3D structures by probabilistic multi-template homology modeling. PLoS Comput Biol 11(10):e1004343. doi:10.1371/journal.pcbi.1004343

Meier A, Soding J (2015b) Context similarity scoring improves protein sequence alignments in the midnight zone. Bioinformatics 31(5):674–681. doi:10.1093/bioinformatics/btu697

Miyazawa S, Jernigan RL (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 256 (3):623–644. doi:10.1006/jmbi.1996.0114

Moult J, Fidelis K, Kryshtafovych A et al (2014) Critical assessment of methods of protein structure prediction (CASP)–round x. Proteins 82(Suppl 2):1–6. doi:10.1002/prot.24452

Park J, Teichmann SA, Hubbard T et al (1997) Intermediate sequences increase the detection of homology between sequences. J Mol Biol 273(1):349–354. doi:10.1006/jmbi.1997.1288

Peng J, Xu J (2010) Low-homology protein threading. Bioinformatics 26(12):i294–i300. doi:10.1093/bioinformatics/btq192

Perdigao N, Heinrich J, Stolte C et al (2015) Unexpected features of the dark proteome. Proc Natl Acad Sci U S A 112(52):15898–15903. doi:10.1073/pnas.1508380112

Remmert M, Biegert A, Hauser A et al (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175. doi:10.1038/nmeth.1818

Richmond TJ (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. J Mol Biol 178 (1):63–89

Rohl CA, Strauss CE, Misura KM et al (2004) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93. doi:10.1016/S0076-6879(04)83004-0

Rychlewski L, Jaroszewski L, Li W et al (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9(2):232–241. doi:10.1110/ps.9.2.232

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815. doi:10.1006/jmbi.1993.1626

Seringhaus M, Gerstein M (2007) Chemistry nobel rich in structure. Science 315(5808):40–41. doi:10.1126/science.315.5808.40

Shen MY, Sali A (2006) Statistical potential for assessment and prediction of protein structures. Protein Sci 15(11):2507–2524. doi:10.1110/ps.062416606

Siew N, Elofsson A, Rychlewski L et al (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. Bioinformatics 16(9):776–785

Sillitoe I, Dawson N, Thornton J et al (2015) The history of the CATH structural classification of protein domains. Biochimie 119:209–217. doi:10.1016/j.biochi.2015.08.004

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213(4):859–883

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951–960. doi:10.1093/bioinformatics/bti125

Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33 (Web Server issue):W244–W248. doi:10.1093/nar/gki408

Sutcliffe MJ, Haneef I, Carney D et al (1987) Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1(5):377–384

Tanaka S, Scheraga HA (1976) Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 9(6):945–950

Van Noorden R, Maher B, Nuzzo R (2014) The top 100 papers. Nature 514(7524):550–553. doi:10.1038/514550a

Wallner B, Fang H, Elofsson A (2003) Automatic consensus-based fold recognition using Pcons, ProQ, and Pmodeller. Proteins 53(Suppl 6):534–541. doi:10.1002/prot.10536

Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35(10):3375–3382. doi:10.1093/nar/gkm251

Xu D, Zhang Y (2011) Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys J 101(10):2525–2534. doi:10.1016/j.bpj.2011.10.024

Yang J, Yan R, Roy A et al (2015) The I-TASSER suite: protein structure and function prediction. Nat Methods 12(1):7–8. doi:10.1038/nmeth.3213

Zhang C, Liu S, Zhou Y (2004) Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci 13(2):391–399. doi:10.1110/ps.03411904

Zhang J, Zhang Y (2010) A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. PLoS ONE 5 (10):e15386. doi:10.1371/journal.pone.0015386

Zhang Y, Skolnick J (2004) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25(6):865–871. doi:10.1002/jcc.20011

Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 101(8):2043–2052. doi:10.1016/j.bpj.2011.09.012

Zhou H, Zhou Y (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55 (4):1005–1013. doi:10.1002/prot.20007

# Chapter 4
# Comparative Protein Structure Modelling

**András Fiser**

**Abstract** A prerequisite to understand cell functioning on the system level is the knowledge of three-dimensional protein structures that mediate biochemical interactions. The explosion in the number of available gene sequences set the stage for the next step in genome scale projects, to obtain three dimensional structures for each protein. To achieve this ambitious goal, the costly and slow structure determination experiments are boosted with theoretical approaches. The current state and recent advances in structure modelling approaches are reviewed here, with special emphasis on comparative structure modelling techniques.

**Keywords** Comparative protein structure modelling · Homology modelling · Template-based modelling · Loop modelling

## 4.1 Introduction

### *4.1.1 Structure Determines Function*

Functional characterization of proteins is one of the most frequent problems in biology. While sequences provide valuable information, their high plasticity makes it frequently impossible to identify functionally relevant residues (Todd et al. 2002). For example in case of enzymes, a similar function can be assumed between two proteins if their sequence identity is above 40%, but if the sequence identity drops in between 30 and 40% only the first three Enzyme Commission (EC) numbers can be predicted reliably, and only at 90% accuracy level. Below 30% sequence identity, structural information is necessary to essential for functional annotation.

A. Fiser (✉)
Department of Systems and Computational Biology,
Department of Biochemistry, Albert Einstein College of Medicine,
1300 Morris Park Ave, Bronx, NY 10461, USA
e-mail: andras.fiser@einstein.yu.edu
URL: http://www.fiserlab.org



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_4

Meanwhile it is estimated that 75% of homologous enzymes share less than 30% identical positions (Todd et al. 2001). Another quantitative study on sequence and function divergence was based on the Gene Ontology classification of function in 6828 protein families (Sangar et al. 2007). It was confirmed that among homologous proteins, the proportion of divergent functions decreases dramatically if a threshold of sequence identity is 50% or higher. However, even for proteins with more than 50% sequence identity, transfer of annotation between homologs leads to an erroneous attribution with a totally dissimilar function in 6% of cases. Where the function of a protein is specific binding to another, the sequence similarity is even less informative guide to function. For instance, the systematic functional clustering of all cell surface-expressed Immunoglobulin SuperFamily proteins (IgSF) revealed examples where proteins with unrelated binding specificity shared more similarity than the ones with identical binding specificity (Yap et al. 2014). The ectodomains of CD80 and the functionally related CD86 ligands (both are cognate ligands of CTLA4 receptor) share only 27% sequence identity, whereas CD80 shares greater than 27% sequence identity with other, functionally unrelated IgSF proteins such as IgSF DCC subclass member 4 (IGDC4) and neural cell adhesion molecule L1 (L1CAM) (Hlavin and Lemmon 1991).

Functional characterization of a protein is often facilitated by its three-dimensional (3D) structure. The insight that one may gain from a 3D model ranges from such low level functional descriptions as confirming the fold (Wu et al. 2000) and inferring a general functional role (Fajardo and Fiser 2013), to such high resolution descriptions that allow understanding ligand specificities (Xu et al. 1996) and designing inhibitors in the context of structure based drug discovery (Evers et al. 2003; Becker et al. 2006; Norin and Sundstrom 2001; Wlodawer 2002; Schwede et al. 2009). Finally, structures aid progress towards a higher level understanding of proteomes through analysis of macromolecular assemblies (Stein et al. 2011).

### *4.1.2 Sequences, Structures, Structural Genomics*

Genome scale sequencing projects have already produced around 60 million unique sequences to date (February, 2015) (Apweiler et al. 2004), substantially boosted by metagenomic data, that originally were obtained from Craig Venter's Global Ocean Survey (Rusch et al. 2007; Yooseph et al. 2007; Venter et al. 2004) but are now widely collected from a number of sources. The non-redundant sequence database has increased a staggering 100 fold between 2000 and 2015 (Khafizov et al. 2014) and is doubling every $\sim$18 months (Levitt 2009; Khafizov et al. 2014). Meanwhile only $\sim$120,000 of these proteins have their three-dimensional structures solved experimentally using X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy (Berman et al. 2007). Because of the inherently time-consuming and complicated nature of structure determination techniques, and the less predictable outcomes of these experiments the fraction of known 3D

models is expected to further shrink from the current level of less than 0.25%. Statistics available from large scale efforts of Structural Genomics (SG) centres show that the average success rate of obtaining a structure for a target sequence is only 3–5% (Service 2005), which suggest that even with unlimited resources the structural annotation of all sequences is not feasible with current technologies.

Over the past decade, all structural biology efforts, including structural genomics (Nair et al. 2009), have led to an overall increase in the structural coverage of existing proteins from ~30 to 40% at the residue level, despite the huge growth of the underlying sequence database. However, when redundancy is removed by clustering the entries at 50% sequence identity, the structural coverage exhibits only a modest increase, from 13.3% in 2001 to ~18% by 2013. With existing technologies and strategies, we project (Khafizov et al. 2014) that it would take 15 years to reach a level of ~55% coverage, the level shown to provide considerable utility for defining large-scale functional characterization of organism-specific properties [e.g., the full metabolic network in *T. maritima* (Zhang et al. 2009)]. However, these efforts are now predicted to take twice as long due to the current winding down of US-based SG efforts after 15 years of operation: SG centres contributed 50–60% of novel coverage despite accounting for less than 10% of all structure depositions (Khafizov et al. 2014).

The prospects for increasing structural coverage are tied to the applicability of homology modeling, which provides more than 99.5% of the currently observed ~40% structural coverage of protein sequences (Khafizov et al. 2014). Conservation of protein structure is much higher than that of sequence (Chothia and Lesk 1986; Illergard et al. 2009), which results in a comparatively small number of distinct structural families (Grant et al. 2004). The size distribution of protein fold families is very uneven and the most frequently occurring folds (e.g., Immunoglobulin, TIM barrel, Rossmann fold) have likely already been identified (Andreeva et al. 2008; Zhang and Skolnick 2005). In a typical genome the 10 most populous superfolds cover a third of the protein sequences (Cuff et al. 2011). Therefore, homology modelling can provide structural models for thousands of proteins in a typical genome using only a few dozen popular folds as templates, and it is currently the vastly predominant source of three-dimensional models (Pieper et al. 2006; Kopp and Schwede 2006). However, the usefulness of homology modelling is expected to exponentially decrease in the future as smaller and smaller protein families or "singletons" need to be modelled. These latter proteins either require a targeted experimental exploration, which is often cost prohibitive, or must be modelled by ab initio or "template-free" style approaches (see Chap. 1). However, these approaches are currently suitable to model only relatively small proteins and have a limited success rate (Kryshtafovych et al. 2014; Tai et al. 2014).

### 4.1.3 Approaches to Protein Structure Prediction

The study of principles that dictate the three-dimensional structure of natural proteins can be approached either through the *laws of physics* or *the theory of evolution*. Each of these approaches provides foundation for a class of protein structure prediction methods (Fiser et al. 2002).

The first approach, ab initio or template-free modelling methods, discussed in Chap. 1, predicts the structure from sequence alone (Pillardy et al. 2001). The ab initio methods assume that the native structure corresponds to the global free energy minimum accessible during the lifespan of the protein, and attempt to find this minimum by an exploration of many conceivable protein conformations (Sali et al. 1994; Dill and Chan 1997; Bonneau and Baker 2001).

The second class of methods, called template-based modelling, includes both those threading techniques that return a full three dimensional description for the target (Xu et al. 2007)—see also Chap. 3—and comparative modelling (Fiser 2004). This class relies on detectable similarity spanning most of the modelled sequence and at least one known structure. Comparative modelling refers to those template-based modelling cases when not only the fold is determined from a possible set of available templates, but a full atom model is built (Marti-Renom et al. 2000). When the structure of at least one protein in the family has been determined by experimentation, the other members of the family can be modelled based on their alignment to the known structure. Comparative modelling approach to protein structure prediction is possible because a small change in the protein sequence usually results in a small change in its 3D structure (Chothia and Lesk 1986). It is also facilitated by the fact that 3D structure of proteins from the same family is more conserved than their amino-acid sequences (Lesk and Chothia 1980). Therefore, if similarity between two proteins is detectable at the sequence level, structural similarity can usually be assumed. The increasing applicability of comparative or template-based modelling is due to the observation that the number of different folds that proteins adopt is rather limited (Andreeva et al. 2008; Chothia et al. 2003; Greene et al. 2007).

Both of these approaches to structure prediction have their advantages and limitations. In principle, ab initio approach can be applied to model any sequence. However, due to the complexity and our limited understanding of the protein folding problem, ab initio methods usually result in relatively low resolution models. Despite significant progress in ab initio protein structure prediction (Das et al. 2007), it remains applicable to a limited number of sequences of approximately 100 residues. Benchmarks at recent Critical Assessment of Techniques for Structure Prediction (CASP) experiments indicate that ab initio techniques still cannot get the overall fold correct for the majority of targets (Kryshtafovych et al. 2014; Tai et al. 2014). Our increasing understanding about the accuracy and performance of currently available force fields and sampling techniques should be acknowledged as being due, in substantial part, to the stunning improvement in computational capacity (Piana et al. 2012, 2014; Shaw et al. 2010). To further exploit this resource several “largest ever”

studies took off recently that expected to provide further critical insights into the folding process. These involve among others the Rosetta@home (http://boinc.bakerlab.org/rosetta/), Folding@home (http://folding.stanford.edu/) and the IBM supported Blue Gene projects. In the Rosetta@home and Folding@home projects the process of protein folding or modelling is studied by running simulations on voluntarily contributing private computers, connecting up to a million CPUs worldwide. IBM established a similar scientific target by building Blue Gene, a computer farm of processors with an estimated 20,000 teraflops peak performance (Sequoia). Currently various flavours of Blue Gene computers occupy a total of 4 of the top 10 positions in the TOP500 supercomputer list (http://www.research.ibm.com/bluegene/ and http://www.top500.org/).

In contrast to ab initio techniques comparative protein structure modelling usually provides models that are comparable to low resolution X-ray crystallography or medium resolution NMR solution structures. However, its applicability is limited to those sequences that can be confidently mapped to known structures. Currently, the probability of finding related proteins of known structure for a sequence picked randomly from a genome ranges approximately from 30 to 80%, depending on the genome. Approximately 70% of all known sequences have at least one domain that is detectably related to at least one protein of known structure (Pieper et al. 2006). This fraction is more than two order of magnitude larger than the number of experimentally determined protein structures deposited in the Protein Data Bank (Berman et al. 2007). The applicability of comparative modelling is steadily increasing because the increasing number of experimentally determined novel structures.

As we will see, in practice, template based modelling always includes information that is independent from the template, in form of various force restraints from general statistical observations or molecular mechanical force fields. As a consequence of improving force fields and search algorithms the most successful approaches are more and more often explore template independent conformational space (Zhang 2007; Das et al. 2007). Similarly, the most successful ab initio approaches, in fact, are using fragments of known structures to build up models (Rohl et al. 2004b; Yang et al. 2015; Lee et al. 2011; Zhou and Skolnick 2007) and as such they should rather be referred as template-free approaches and distinguished from methods that employ first principles only. While it makes sense to discuss the two fundamental principles behind the techniques employed in structure modelling separately, the current trends are pointing to approaches that extensively combine both. While truly ab initio approaches can shed light on the dynamics of the actual folding process, in practice, effective structure modelling almost always involves a certain flavour of template-based modelling.

While template based modeling techniques assumed to be close to their peak performance (errors in template based models built on 40% or higher sequence identity templates are comparable to errors observed in experimental models), template-free modeling methods certainly have room for considerable improvement (Tai et al. 2014). One trivial way to do that is to obtain additional spatial restrains

that can aid template-free approaches either at the sampling or at the scoring evaluation step. These additional restraints can be obtained either computationally or experimentally. An effective computational way to add restraints is to predict possible three dimensional contacts from sequence variations. This can be done efficiently for target sequences for which a large number (i.e. hundreds) of similar sequences (orthologs and paralogs) are available. In this case the extensive sequence profile allows to detect co-evolving residues, which often from spatial contacts in the corresponding structure (Marks et al. 2011; Morcos et al. 2011) (a detailed overview of these techniques is provided in Chap. 2). Another possible source to obtain additional spatial restraints is to use fast, semi-high throughput experiments that can yield indirect spatial restraints. Several such hybrid methods have been developed recently that primarily utilize NMR experiments, and using restraints through chemical shift, dipolar coupling or limited NOE information (Menon et al. 2013; Rohl and Baker 2002; Lange et al. 2012; Bowers et al. 2000).

The benefit of using additional restraints from either limited experimentation or from co-evolutionary conservation data is that that accuracy of template free modelling approaches can sometimes dramatically improve and become competitive with those of template-based techniques.

## 4.2 Steps in Comparative Protein Structure Modelling

Comparative or homology (template-based) protein structure modelling builds a three-dimensional model for a protein of unknown structure (the target) based on one or more related proteins of known structure (the templates) (Greer 1981; Blundell et al. 1987; Marti-Renom et al. 2000; Fiser 2004; Ginalski 2006; Petrey and Honig 2005). The necessary conditions for getting a useful model are (i) detectable similarity between the target sequence and the sequence of the template structure and (ii) availability of a correct alignment between them.

All current comparative modelling methods consist of five sequential steps. The first step is to search for proteins with known 3D structures that are related to the target sequence. The second step is to pick those structures that will be used as templates. The third step is to align their sequences with the target sequence. The fourth step is to build the model for the target sequence given its alignment with the template structures. The last step is to evaluate the model, using a variety of criteria.

There are several computer programs and web servers that automate the comparative modelling process (Table 4.1). While the web servers are convenient and useful (Battey et al. 2007; Fernandez-Fuentes et al. 2007a; Rai et al. 2006; Zhang 2007), the best results are still obtained by non-automated, expert use of the various modelling tools (Kopp et al. 2007). Complex decisions for selecting the structurally and biologically most relevant templates, optimally combining multiple template information, refining alignments in non trivial cases, selecting segments for loop modelling, including cofactors and ligands in the model or specifying external restraints require an expert knowledge that is difficult to fully automate (Fiser and

**Table 4.1** Names and WWW addresses of some online tools useful for various aspects of comparative modeling

| | |
|---|---|
| Fold recognition by database searches | |
| PSI- and DELTA-BLAST | www.ncbi.nlm.nih.gov/BLAST/ |
| FastA/SSEARCH | www.ebi.ac.uk/fasta33 |
| FFAS03 | ffas.sanfordburnham.org |
| HHblits | toolkit.tuebingen.mpg.de/hhblits |
| Fold recognition by threading | |
| PHYRE2 | www.sbg.bio.ic.ac.uk/~phyre2/ |
| RaptorX | raptorx.uchicago.edu/ |
| LOOPP | clsb.ices.utexas.edu/loopp/web/ |
| MUSTER | zhanglab.ccmb.med.umich.edu/MUSTER/ |
| SAM-T06 | www.soe.ucsc.edu/research/compbio/SAM_T06/T06-query.html |
| pGenTHREADER | bioinf.cs.ucl.ac.uk/psipred |
| Sparks | sparks-lab.org |
| FUGUE | mizuguchilab.org/fugue/ |
| LOMETS | zhanglab.ccmb.med.umich.edu/LOMETS/ |
| Sequence alignment tools | |
| Smith-Waterman | jaligner.sourceforge.net/ |
| ClustalW | www.clustal.org/clustal2/ |
| MUSCLE | www.drive5.com/muscle/ |
| T-COFFEE | tcoffee.vital-it.ch |
| PROMALS | prodata.swmed.edu/promals/promals.php |
| PROBCONS | probcons.stanford.edu |
| SALIGN | salilab.org/salign |
| *Comparative modeling, loop and side chain modeling* | |
| MMM | www.fiserlab.org/servers/MMM |
| M4T | www.fiserlab.org/servers/M4T |
| MODELLER | www.salilab.org/modeller/ |
| MODWEB | modbase.compbio.ucsf.edu/modweb/ |
| I-TASSER | zhanglab.ccmb.med.umich.edu/I-TASSER/ |
| HHPRED | toolkit.tuebingen.mpg.de/hhpred |
| 3D-JIGSAW | bmm.crick.ac.uk/~3djigsaw/ |
| CPH-MODELS | www.cbs.dtu.dk/services/CPHmodels/ |
| IntFOLD | www.reading.ac.uk/bioinf/IntFOLD/ |
| SWISSMODEL | swissmodel.expasy.org/workspace |
| FAMS | www.pharm.kitasato-u.ac.jp/fams |
| PRISM | honig.c2b2.columbia.edu/prism |
| RAPPER | mordred.bioc.cam.ac.uk/~ rapper |
| ESYPRED3D | www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/ |
| PCONS | pcons.net |

(continued)

**Table 4.1** (continued)

| | |
|---|---|
| Loop modeling | |
| ARCHPRED | fiserlab.org/servers/archpred |
| MODLOOP | salilab.org/modloop |
| FALC-LOOP | falc-loop.seoklab.org/ |
| FREAD | opig.stats.ox.ac.uk/webapps/fread/php/ |
| SUPERLOOPER | bioinf-applied.charite.de/superlooper/ |
| Side chain modeling | |
| SCWRL4 | dunbrack.fccc.edu/scwrl4/ |
| IRECS | irecs.bioinf.mpi-inf.mpg.de/index.php |
| *Model evaluation* | |
| PROCHECK | www.ebi.ac.uk/thornton-srv/software/PROCHECK/ |
| Prosa-web | prosa.services.came.sbg.ac.at/prosa.php |
| WHATCHECK | swift.cmbi.ru.nl/gv/whatcheck |
| VERIFY3D | services.mbi.ucla.edu/Verify_3D/ |
| ANOLEA | melolab.org/anolea/ |
| PROQ | www.sbc.su.se/~bjornw/ProQ/ProQ.cgi |
| ModEVAL | modbase.compbio.ucsf.edu/evaluation/ |
| Qmean | swissmodel.expasy.org/qmean/cgi/index.cgi |

Sali 2003a) although more and more efforts on automation point to this direction (Fernandez-Fuentes et al. 2007b; Contreras-Moreira et al. 2003).

### 4.2.1 *Searching for Structures Related to the Target Sequence*

Comparative modelling usually starts by searching the Protein Data Bank (PDB) (Berman et al. 2007) of known protein structures using the target sequence as the query. This search is generally done by comparing the target sequence with the sequence of each of the structures in the database.

There are two main classes of protein comparison methods that are useful in fold identification. The first class compares the sequences of the target with each of the database templates independently. This can be done by using pairwise sequence-sequence comparison (Apostolico and Giancarlo 1998). The performance of these methods in sequence searching (Pearson 2000; Sauder et al. 2000) and fold assignments (Brenner et al. 1998) has been evaluated exhaustively. The most popular programs in the class include FASTA (Pearson 2000) and BLAST (Schaffer et al. 2001). To improve the sensitivity of the sequence based searches evolutionary information can be incorporated in form of multiple sequence alignment (Rychlewski et al. 2000; Krogh et al. 1994; Henikoff et al. 2000; Marti-Renom et al. 2004; Altschul et al. 1997). These approaches begin by finding

all sequences in a sequence database that are clearly related to the target and easily aligned with it. The multiple alignment of these sequences is the target sequence profile, which implicitly carries additional information about the location and pattern of evolutionary conserved positions of the protein. The most well known program in this class is PSI-BLAST (Altschul et al. 1997) which implements a heuristic search algorithm for short motifs and its newer generation version, delta-BLAST that in addition uses domain specific information (Boratyn et al. 2012). A further step to increase the sensitivity of this approach is to pre-calculate sequence profiles for all the known structures and then use pairwise dynamic programming algorithm to compare the two profiles. This has been implemented, among other programs, in COACH (Edgar and Sjolander 2004) and in FFAS03 (Jaroszewski et al. 1998, 2005). The construction of profile-based Hidden Markov Models (HMM) is another sensitive way to locate universally conserved motifs among sequences (Karplus et al. 1998). A substantial improvement in HMM approaches was achieved by incorporating information about predicted secondary structural elements (Karchin et al. 2003; Karplus et al. 2005). Another development in this group of methods is the phylogenetic tree-driven HMM, which selects a different subset of sequences for profile HMM analysis at each node in the evolutionary tree (Edgar and Sjolander 2003). Important development was the HHblits sequence search method (Remmert et al. 2012) to compile sequence profiles by quick and sensitive search of large databases, which profiles then can be used to perform HMM-HMM alignments against a precompiled database of profiles of known structures to identify remotely related templates for homology modelling (Soding 2005). Locating sequence intermediates that are homologous to both sequences may also enhance the template searches (Sauder et al. 2000; John and Sali 2004; Rubinstein et al. 2013). These more sensitive fold identification techniques are especially useful for finding significant structural relationships when sequence identity between the target and the template drops below 25%. More accurate sequence profiles and structural alignments can be constructed with consistency-based approaches such as T-Coffee (Moretti et al. 2007) PROMAL (and PROMAL3D for structures) (Pei et al. 2008; Pei and Grishin 2007), ProbCons (Do et al. 2005) etc. For reviews of multiple sequence alignments see (Notredame 2007; Edgar and Batzoglou 2006).

The second class of methods relies on pairwise comparison of a protein sequence and a protein structure; the target sequence is matched against a library of 3D profiles or threaded through a library of 3D folds. These methods are also called fold assignment, threading or 3D template matching (Bowie et al. 1991; Jones et al. 1992; Finkelstein and Reva 1991). These methods, discussed in detail in Chap. 2, are especially useful when sequence profiles are not possible to construct because there are not enough known sequences that are clearly related to the target or potential templates.

Template search methods "outperform" the needs of comparative modelling in the sense that they are able to locate sequences that are so remotely related as to render construction of a reliable comparative model impossible. The reason for this is that sequence relationships are often established on short conserved segments,

while a successful comparative modelling exercise requires an overall correct alignment for the entire modelled part of the protein. This is an important distinction between fold recognition and comparative modelling: while both are template based and deliver a 3D description of the target as a result, fold recognition aims at identifying the general 3D shape of the target sequence or at least the class of shapes where it belongs to, while comparative modelling aims at generating an all atom model for the entire target sequence.

### *4.2.2 Selecting Templates*

Once a list of potential templates is obtained using searching methods, it is necessary to select one or more templates that are appropriate for the particular modelling problem. Several factors need to be taken into account when selecting a template.

**Considerations in Template Selection**
The simplest template selection rule is to select the structure with the highest sequence similarity to the modelled sequence. The family of proteins that includes the target and the templates can frequently be organized into sub-families. The construction of a multiple alignment and a phylogenetic tree (Felsenstein 1981) can help in selecting the template from the subfamily that is closest to the target sequence. The similarity between the "environment" of the template and the environment in which the target needs to be modelled should also be considered. The term "environment" is used here in a broad sense, including everything that is not the protein itself (e.g., solvent, pH, ligands, quaternary interactions). If possible, a template bound to the same or similar ligands as the modelled sequence should generally be used. The quality of the experimentally determined structure is another important factor in template selection. Resolution and R-factor of a crystal structure and the number of restraints per residue for an NMR structure are indicative of their accuracy. For instance, if two templates have comparable sequence similarity to the target, the one determined at the highest resolution should generally be used. The criteria for selecting templates also depend on the purpose of a comparative model. For example, if a protein-ligand model is to be constructed, the choice of the template that contains a similar ligand is probably more important than the resolution of the template.

**Advantage of Using Multiple Templates**
It is not necessary to select only one template. In fact, the optimal use of several templates increases the model accuracy (Venclovas and Margelevicius 2005; Sanchez and Sali 1997; Fernandez-Fuentes et al. 2007a, b); however, not all modelling programs are designed to accept more than one template. The benefit of combining multiple template structures can be twofold. First, multiple template structures may be aligned with different domains of the target, with little overlap between them, in which case, the modelling procedure can construct a

homology-based model of the whole target sequence. Second, the template structures may be aligned with the same part of the target and build the model on the locally best template.

An elaborate way to select suitable templates is to generate and evaluate models for each candidate template structure and/or their combinations. The optimized all-atom models can then be evaluated by an energy or scoring function, such as the Z-score of PROSA (Sippl 1995) or VERIFY3D (Eisenberg et al. 1997). These scoring methods are often sufficiently accurate to allow selection of the most accurate of the generated models (Wu et al. 2000). This trial-and-error approach can be viewed as limited threading (i.e., the target sequence is threaded through similar template structures). However these approaches are good only at selecting various templates on a global level.

A recently developed method M4T (Multiple Mapping Method with Multiple Templates) selects and combines multiple template structures through an iterative clustering approach that takes into account the "unique" contribution of each template, their sequence similarity among themselves and to the target sequence, and their experimental resolution (Fernandez-Fuentes et al. 2007a, b). The resulting models systematically outperformed models that were based on the single best template.

Another important observation from this study was that below 40% sequence identity, models built using multiple templates are more accurate than those built using a single template only and this trend is accentuated as one moves into more remote target-template pair cases. Meanwhile the advantage of using multiple templates gradually disappears above 40% target-template sequence identity cases (Fig. 4.1). This suggests that in this range the average differences between the template and target structures are smaller than the average differences among

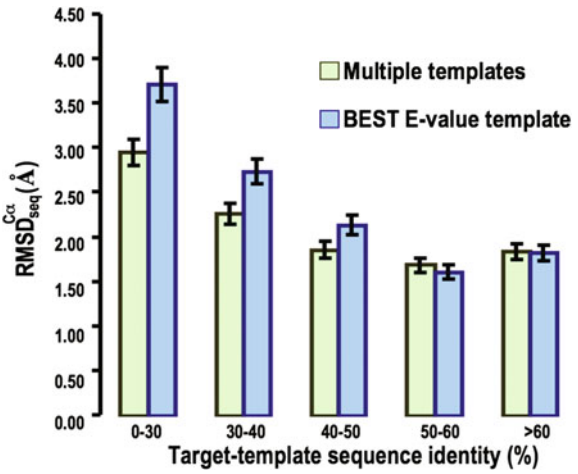



**Fig. 4.1** Comparing accuracy (y-axis) of models built for the same set of 765 protein target sequences using either one template (best E-value hit only; *blue bars*), or multiple templates (*green bars*). The percentage of sequence identity (x-axis) is calculated between the hit with the highest E-value and the query sequence. *Error bars* indicate standard errors of the mean

alternative template structures that are all highly similar to the target (Fernandez-Fuentes et al. 2007b).

### 4.2.3 Sequence to Structure Alignment

To build a model, all comparative modelling programs depend on a list of assumed structural equivalences between the target and template residues. This list is defined by the alignment of the target and template sequences. Many template search methods will produce such an alignment and these sometimes can directly be used as the input for modelling. Often, however, especially in the difficult cases, this initial alignment is not the optimal target-template alignment e.g., at less than 30% sequence identity (where sequence identity is defined as the number of identical positions in the alignment normalized by the length of the target sequence). Search methods tend to be tuned for detection of remote relationships, which is often realized based on a local motif and not for a full length, optimal alignment. Therefore, once the templates are selected, an alignment method should be used to align them with the target sequence. The alignment is relatively simple to obtain when the target-template sequence identity is above 40%. If the target-template sequence identity is lower than 40%, the alignment accuracy becomes the most important factor affecting the quality of the resulting model. A misalignment by only one residue position will result in an error of approximately 4 Å in the model.

**Taking Advantage of Structural Information in Alignments**

Alignments in comparative modelling represent a unique class, because on one side of the alignment there is always a 3D structure, the template. Therefore alignments can be improved by including structural information from the template. For example, gaps should be avoided in secondary structure elements, in buried regions, or between two residues that are far in space. Some alignment methods take such criteria into account (Jennings et al. 2001; Shi et al. 2001; Blake and Cohen 2001).

When multiple template structures are available, a good strategy is to superpose them with each other first, to obtain a multiple structure-based alignment highlighting structurally conserved residues (Petrey et al. 2003; Reddy et al. 2001; Al-Lazikani et al. 2001). In the next step, the target sequence is aligned with this multiple structure-based alignment. The benefits of using of multiple structures and multiple sequences derive from the evolutionary and structural information about the templates as well as evolutionary information about the target sequence, and often produces a better alignment for modelling than the pairwise sequence alignment methods (Sauder et al. 2000; Rychlewski et al. 2000).

Multiple Mapping Method (MMM) directly relies on information from the 3D structure (Rai and Fiser 2006; Rai et al. 2006). MMM minimizes alignment errors by selecting and optimally splicing differently aligned fragments from a set of alternative input alignments. This selection is guided by a scoring function that determines the preference of each alternatively aligned fragment of the target

sequence in the structural environment of the template. The scoring function has four terms, which are used to assess the compatibility of alternative variable segments in the protein environment: (a) environment specific substitution matrices from FUGUE (Shi et al. 2001); (b) residue substitution matrix, Blosum (Henikoff and Henikoff 1992) (c) A 3D-1D substitution matrix, H3P2, that scores the matches of predicted secondary structure of the target sequence to the observed secondary structures and accessibility types of the template residues (Luthy et al. 1991); (d) a statistically derived residue-residue contact energy term (Rykunov and Fiser 2007). MMM essentially performs a limited and inverse threading of short fragments: in this exercise the actual question is not the identification of a right fold, but identification of the correct alignment mapping, among many alternatives, for sequence segments that are threaded on the same fold. These local mappings are evaluated in the context of the rest of the model, where alignments provide a consistent solution and framework for the evaluation.

### *4.2.4 Model Building*

When discussing the model building step within comparative protein structure modelling it is useful to distinguish two parts: *template dependent* and *template independent* modelling. This distinction is necessary because certain parts of the target must be built without the aid of any template. These parts correspond to gaps in the template sequence within the target-template alignment. Modelling of these regions is commonly referred to as loop modelling problem. It is evident, that these loops are responsible for the most characteristic differences between the template and target, and therefore are chiefly responsible for structural and consequently functional differences. In contrast to these loops, the rest of the target, and in particular the conserved core of the fold of the target, is built using information from the template structure. First, we will review a few major approaches of this latter part, the template dependent modelling. This is also the logical first step during the building of a model, since the template dependent modelling step provides a structure for most of the target protein, which then serves as a starting structural framework for any subsequent loop modelling exercise.

#### 4.2.4.1 Template Dependent Modelling

**Modelling by Assembly of Rigid Bodies**

The first and still widely used approach in comparative modelling is to assemble a model from a framework of small number of rigid bodies obtained from the aligned template protein structures (Greer 1990; Blundell et al. 1987; Browne et al. 1969). The approach is based on the natural dissection of the protein structure into conserved core regions, variable loops that connect them, and side chains that decorate the backbone (Topham et al. 1993). A widely used program in this class is

COMPOSER (Sutcliffe et al. 1987). The accuracy of a model can be somewhat increased when more than one template structure is used to construct the framework and when the templates are averaged into the framework using weights corresponding to their sequence similarities to the target sequence (Srinivasan and Blundell 1993).

**Modelling by Segment Matching or Coordinate Reconstruction**

The basis of modelling by coordinate reconstruction is the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (Unger et al. 1989). Thus, comparative models can be constructed by using a subset of atomic positions from template structures as "guiding" positions, and by identifying and assembling short, all-atom segments that fit these guiding positions. The guiding positions usually correspond to the Cα atoms of the segments that are conserved in the alignment between the template structure and the target sequence. The all-atom segments that fit the guiding positions can be obtained either by scanning all the known protein structures, including those that are not related to the sequence being modelled (Claessens et al. 1989; Holm and Sander 1991), or by a conformational search restrained by an energy function (van Gelder et al. 1994; Bruccoleri and Karplus 1990). For example, a general method for modelling by segment matching (SEGMOD) (Levitt 1992) is guided by the positions of some atoms (usually Cα atoms) to find the matching segments in a representative database of all known protein structures. This method can construct both main chain and side chain atoms, and can also model gaps. Even some side chain modelling methods (Chinea et al. 1995) and the class of loop construction methods based on finding suitable fragments in the database of known structures (Jones and Thirup 1986) can be seen as segment matching or coordinate reconstruction methods.

**Modelling by Satisfaction of Spatial Restraints**

The methods in this class begin by generating many constraints or restraints on the structure of the target sequence, using its alignment to related protein structures as a guide. The procedure is conceptually similar to that used in determination of protein structures from NMR-derived restraints. The restraints are generally obtained by assuming that the corresponding distances between aligned residues in the template and the target structures are similar. These homology-derived restraints are usually supplemented by stereochemical restraints on bond lengths, bond angles, dihedral angles, and non-bonded atom-atom contacts that are obtained from a molecular mechanics force field (Brooks et al. 2009). The model is then derived by minimizing the violations of all the restraints. This can be achieved either by distance geometry or real-space optimization. For example, an elegant distance geometry approach constructs all-atom models from lower and upper bounds on distances and dihedral angles (Havel and Snow 1991). Although further efforts were made to apply distance geometry for comparative modelling, e.g. (Aszodi and Taylor 1996), more successful but also more conservative, real space modelling approaches dominate the field, perhaps because evolution also proved to be surprisingly conservative in preserving structural features in various proteins (Kihara and Skolnick 2003).

Comparative modelling by satisfaction of spatial restraints is implemented in the computer program MODELLER (Fiser and Sali 2003a; Sali and Blundell 1993), currently the most popular protein modelling program. In the first step of model building, distance and dihedral angle restraints on the target sequence are derived from its alignment with template 3D structures. The form of these restraints was obtained from a statistical analysis of the relationships between similar protein structures. By scanning the database of alignments, tables quantifying various correlations were obtained, such as the correlations between two equivalent Cα-Cα distances, or between equivalent main chain dihedral angles from two related proteins (Sali and Blundell 1993). These relationships are expressed as conditional probability density functions (pdf's) and can be used directly as spatial restraints. For example, probabilities for different values of the main chain dihedral angles are calculated from the type of residue considered, from the main chain conformation of an equivalent template residue, and from sequence similarity between the two proteins. An important feature of the method is that the forms of spatial restraints were obtained empirically, from a database of protein structure alignments, without any user imposed subjective assumption. Finally, the model is obtained by optimizing the objective function in Cartesian space. The optimization is carried out by the use of the variable target function method (Braun and Go 1985) employing methods of conjugate gradients and molecular dynamics with simulated annealing (Clore et al. 1986).

A similar comprehensive package is NEST that can build a homology model based on single sequence-template alignment or from multiple templates. It can also consider different structures for different parts of the target (Petrey et al. 2003).

Benchmarks of comparative modelling programs have shown similar performance of major approaches but with Modeller usually outperforming the rest (Builder, Nest, SegMod, Swiss-Model, 3D-jigsaw) (Dalton and Jackson 2007; Wallner and Elofsson 2005a)

**Combining Alignments, Combining Structures**

It is frequently difficult to select the best templates or calculate a good alignment. One way of improving a comparative model in such cases is to proceed with an iteration of template selection, alignment, and model building, guided by model assessment. This iteration can be repeated until no improvement in the model is detected (Guenther et al. 1997; Fiser and Sali 2003a). More recently these anecdotal and manual approaches were automated (Petrey et al. 2003). For instance, an automated method was introduced that optimizes both the alignment and the model implied by it (John and Sali 2003). This task is achieved by a genetic algorithm protocol that starts with a set of initial alignments and then iterates through re-alignment, model building and model assessment to optimize a model assessment score. During this iterative process new alignments are constructed by application of a number of operators, such as alignment mutations and cross-overs; comparative models corresponding to these alignments are built and assessed by a variety of criteria, partly depending on an atomic statistical potential. In another approach, a genetic algorithm was applied to automatically combine templates and

alignments. A relatively simple structure dependent scoring function was used to evaluate the sampled combinations. Despite some limitations, the procedure is shown to be robust to alignment errors, while simplifying the task of selecting templates (Contreras-Moreira et al. 2003).

Other attempts to optimize target-template alignments include the Robetta server, where alignments are generated by dynamic programming using a scoring function that combines information on many protein features, including a novel measure of how obligate a sequence region is to the protein fold. By systematically varying the weights on the different features that contribute to the alignment score, very large ensembles of diverse alignments are generated. A variety of approaches to select the best models from the ensemble, including consensus of the alignments, a hydrophobic burial measure, low- and high-resolution energy functions, and combinations of these evaluation methods were explored (Chivian and Baker 2006).

Those meta-server approaches that do not simply score and rank alternative models obtained from a variety of methods but further combine them could also be perceived as approaches that explore the alignment and conformational space for a given target sequence (Kolinski and Bujnicki 2005).

Another alternative for combined servers is provided by M4T. The M4T program automatically identifies the best templates and explores and optimally splices alternative alignments according to its internal scoring function that focuses on the features of the structural environment of each template (Fernandez-Fuentes et al. 2007b).

**Meta-servers**

Meta-server approaches have been developed to take advantage of the variety of other existing programs. Meta-servers collect models from alternative methods and either use them for inputs to make new models or look for consensus solutions within them. For instance FAMS-ACE (Terashi et al. 2007) takes inputs from other servers as starting points for refinement and remodelling after which Verify3D (Eisenberg et al. 1997) is used to select the most accurate solution. Other consensus approaches include PCONS, a neural network approach that identifies a consensus model by combining information on reliability scores and structural similarity of models obtained from other techniques (Wallner et al. 2007). 3D-JURY operates along the same idea, its selection is mainly based on the consensus of model structure similarity (Ginalski et al. 2003).

#### 4.2.4.2 Template Independent Modelling: Modelling Loops, Insertions

In comparative modelling, target sequences often have inserted residues relative to the template structures or have regions that are structurally different from the corresponding regions in the templates. Thus, no structural information about these inserted segments can be extracted from the template structures. These regions frequently correspond to surface loops. Loops often play an important role in defining the functional specificity of a given protein framework, forming the

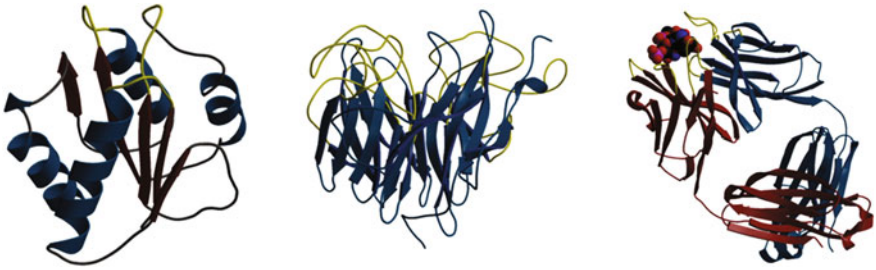


**Fig. 4.2** Examples of loops (rendered in *yellow*) that are responsible for functional specificity within protein superfamilies. From *left* to *right*: Flavodoxin, Immunoglobulin, Neuraminidase from, respectively, the α + β barrel, Ig and antiparallel β-barrel protein fold families

functional sites such as antibody complementary determining regions (Rudolph et al. 2006), ligand binding sites (for ATP (Saraste et al. 1990), calcium (Grabarek 2006), and NAD(P) (Lesk 1995), for example), DNA binding sites (Tainer et al. 1995) or enzyme active sites [e.g. Ser-Thr kinases (Johnson et al. 1998) or Asp proteases (Wlodawer et al. 1989)]. The accuracy of loop modelling is a major factor determining the usefulness of comparative models in applications such as ligand docking or functional annotation (Fig. 4.2). Loop modelling can be seen as a mini protein folding problem because the correct conformation of a given segment of a polypeptide chain has to be calculated mainly from the sequence of the segment itself. However, loops are generally too short to provide sufficient information about their local fold—unless a very substantial part of the fragments match sequentially and a known conformation—and on the other hand, the environment of each loop is uniquely defined by the solvent and the protein that cradles it. In a few rare cases it was shown that even identical decapeptides in different proteins do not always have the same conformation (Mezei 1998; Fernandez-Fuentes and Fiser 2006).

There are two main classes of loop modelling methods: (i) the database search approaches, where a segment that fits on the anchor core regions is found in a database of all known protein structures (Jones and Thirup 1986; Chothia and Lesk 1987) and (ii) the conformational search approaches (Shenkin et al. 1987; Moult and James 1986). There are also methods that combine these two approaches (Deane and Blundell 2001; van Vlijmen and Karplus 1997; de Bakker et al. 2003).

**Fragment Based Approach to Loop Modelling**

The database or fragment search approach to loop modelling is accurate and efficient when a database of specific loops is created to address the modelling of the same class of loops, such as β-hairpins (Sibanda et al. 1989), or loops on a specific fold, such as the hypervariable regions in the immunoglobulin fold (Chothia et al. 1989). Earlier it was predicted that it is unlikely that structure databanks will ever reach a point when fragment based approaches become efficient to model loops (Fidelis et al. 1994), which resulted in a boost in the development of conformational search approaches from around 2000. However, many details of the fold universe has been explored during the last decade due to the large number of new folds

solved experimentally, which had a profound effect on the extent of known structural fragments. Recent analyses showed that loop fragments are not only well represented in current structure databanks but shorter segments are possibly completely explored already (Du et al. 2003). It was reported that sequence segments up to 10 residues had a related (i.e. at least 50% identical segment) in PDB with a known conformation, and despite the six fold increase in sequence databank size and the doubling of PDB since 2002 there was not a single unique loop conformation entered in the PDB or sequence segment observed that shares less than 50% sequence identity to a PDB fragment, which indicates that newly sequenced proteins keep recycling the same set of already known short structural segments. All sequence segments up to 10–12 residues have at least one corresponding structural segment that shares at least 50% identity thus ensuring structural similarity, except a very few notable exceptions mentioned above (Fernandez-Fuentes and Fiser 2006). Consequently more recent efforts have tried to classify loop conformations into more general categories, thus extending the applicability of the database search approach for more cases (Fernandez-Fuentes et al. 2006a; Michalsky et al. 2003). A recent work described the advantage of using HMM sequence profiles in classifying and predicting loops (Espadaler et al. 2004). An another recently published loop prediction approach first predicts conformation for a query loop sequence and then structurally aligns the predicted structural fragments to a set of non-redundant loop structural templates. These sequence-template loop alignments are then quantitatively evaluated with an artificial neural network model trained on a set of predictions with known outcomes (Peng and Yang 2007).

ArchPred, perhaps the most accurate database loop modelling approach is briefly described here (Fernandez-Fuentes et al. 2006a, b). ArchPred exploits a hierarchical and multidimensional database that has been set up to classify about 300,000 loop fragments and loop flanking secondary structures. Besides the length of the loops and types of bracing secondary structures the database is organized along four internal coordinates, a distance and three types of angles characterizing the geometry of stem regions (Oliva et al. 1997). Candidate fragments are selected from this library by matching the length, the types of bracing secondary structures of the query and satisfying the geometrical restraints of the stems and subsequently inserted in the query protein framework where their fit is assessed by the root mean squared deviation (RMSD) of stem regions and by the number of rigid body clashes with the environment. In the final step, remaining candidate loops are ranked by a Z-score that combines information on sequence similarity and fit of predicted and observed $\phi/\psi$ main chain dihedral angle propensities. Confidence Z-score cut-offs were determined for each loop length that identify those predicted fragments that outperform a competitive ab initio method. A web server implements the method, regularly updates the fragment library and performs predictions. Predicted segments are returned or, optionally, these can be completed with side chain reconstruction and subsequently annealed in the environment of the query protein by conjugate gradient minimization.

In summary, the recent reports about the more favourable coverage of loop conformations in the PDB suggest that database approaches are now limited by their ability to recognize suitable fragments, and not by the lack of these segments (i.e. sampling), as earlier thought.

**Ab Initio Modelling of Loops**

To overcome the limitations of the database search methods, conformational search methods were developed. There are many such methods, exploiting different protein representations, objective function terms, and optimization or enumeration algorithms. The search strategies include the minimum perturbation method (Fine et al. 1986), molecular dynamics simulations (Bruccoleri and Karplus 1987), genetic algorithms (Ring and Cohen 1993), Monte Carlo and simulated annealing (Collura et al. 1993; Abagyan and Totrov 1994), multiple-copy simultaneous search (Zheng et al. 1993), self-consistent field optimization (Koehl and Delarue 1995), and an enumeration based on the graph theory (Samudrala and Moult 1998). Loop prediction by optimization is applicable to both simultaneous modelling of several loops and to those loops interacting with ligands, neither of which is straightforward for the database search approaches, where fragments are collected from unrelated structures with different environments.

The MODLOOP module in MODELLER implements the optimization-based approach (Fiser and Sali 2003b; Fiser et al. 2000). Loop optimization in MODLOOP relies on conjugate gradients and molecular dynamics with simulated annealing. The pseudo energy function is a sum of many terms, including some terms from the CHARMM-22 molecular mechanics force field (Brooks et al. 2009) and spatial restraints based on distributions of distances (Sippl 1990; Melo and Feytmans 1997) and dihedral angles in known protein structures. To simulate comparative modelling problems, the loop modelling procedure was optimized and evaluated on a large number of loops of known structure both in native and in only approximately correct environments. The performance of the approach later was further improved by using CHARMM molecular mechanic forcefield with Generalized Born (GB) solvation potential to rank final conformations (Fiser et al. 2002). Incorporation of solvation terms in the scoring function was a central theme in several other subsequent studies (Das and Meirovitch 2003; Forrest and Woolf 2003; DePristo et al. 2003; de Bakker et al. 2003). Improved loop prediction accuracy resulted from the incorporation of an entropy like term to the scoring function, the “colony energy”, derived from geometrical comparisons and clustering of sampled loop conformations (Xiang et al. 2002; Fogolari and Tosatto 2005). The continuous improvement of scoring functions delivers improving loop modelling methods. Two recent loop modelling procedures have been introduced that are utilizing the effective statistical pair potential that is encoded in DFIRE (Soto et al. 2008; Zhang et al. 2004). Very long loops are predicted either using the Rosetta approach, essentially performing a mini folding exercise for the loop segments (Rohl et al. 2004a) or, more recently, by the InsEnds method that use pivot movements of torsion angles to capture the conformation of very long loops or long terminal segments (Adhikari et al. 2012). In the Prime program large

numbers of loops are generated by using a dihedral angle-based building procedure followed by iterative cycles of clustering, side-chain optimization, and complete energy minimization of selected loop structures using a full atom molecular mechanic force field (OPLS) with implicit solvation model (Jacobson et al. 2004). Modeling loops in proteins remains an active topic in the field (Tang et al. 2014).

#### 4.2.4.3 Refining Models

Comparative models are constructed with the best possible set of restraints available, which is a usually a combination of various template structure dependent distance and angle restraints combined with molecular mechanic force field terms and restraints imposed by a variety of statistical potential functions. Because of the large number of available restraints the problem is overdefined. The model building step is relatively straightforward and primarily focuses on resolving the conflicting restraints. In case of MODELLER this is achieved by a combination of conjugate gradient minimization and molecular dynamics simulation, and concludes a model typically just within a few minutes. Because of the dominance of template dependent restraints it is often difficult to generate a model that is more similar on the backbone accuracy level to the target protein than to the actual template (if one assumes no alignment errors). It is a difficult task to further refine models because of the fact that the most accurate restraints and forcefield terms were already used in model building. It essentially poses the same task as an ab initio modelling problem, since any novel refinement should take place in a template independent style. Various studies and a recent survey suggested that most refinements decrease the accuracy of models (Summa and Levitt 2007), although promising newer studies suggest that knowledge-based potential of mean force was able to systematically improve model by a modest 1% GDT_TS (Rodrigues et al. 2012; Chopra et al. 2010). Very recently molecular mechanic energy function was able to improve the initial model but by a small margin. At a recent CASP meeting it was reported that a restrained molecular mechanics optimization that employs model averaging has resulted in a systematic improvement in model quality, albeit only a very small one (an average of $\sim$0.06 Å RMSD improvement) (Mirjalili et al. 2014).

Other promising refinement approaches try to intelligently restrict the conformational search space around the high quality initial model. This can be achieved by simply defining a certain maximum deviation that is allowed for the backbone movements during sampling (Kolinski et al. 2001). A more recent promising approach identifies Evolutionary and Vibrational Armonics subspace, a reduced sampling subspace that consists of a combination of evolutionarily favored directions, defined by the principal components of the structural variation within a homologous family, plus topologically favored directions, derived from the low frequency normal modes of the vibrational dynamics, up to 50 dimensions. This subspace is accurate enough so that the cores of most proteins can be represented within 1 Å accuracy, and reduced enough so that effective optimization

approaches, such as the Replica Exchange Monte Carlo simulation can be applied (Han et al. 2008; Qian et al. 2004).

#### 4.2.4.4 Hybrid Modelling of Proteins and Complexes with Experimental Restraints

Some comparative modelling techniques are able to incorporate constraints or restraints derived from a number of different sources other than the homologous template structure. For example, restraints could be provided by rules for secondary structure packing (Cohen et al. 1989), analyses of hydrophobicity (Aszodi and Taylor 1994) and correlated mutations (Taylor and Hatrick 1994; Marks et al. 2011), empirical potentials of mean force (Sippl 1995), nuclear magnetic resonance experiments (Sutcliffe et al. 1992), or from experiments on chemical cross-linking, spin and photoaffinity labelling (Orr et al. 1998), hydrogen/deuterium exchange coupled with mass spectrometry (Xiao et al. 2006), hydroxyl radical footprinting (Kiselar et al. 2003), fluorescence spectroscopy, image reconstruction in electron microscopy (Topf et al. 2008), site-directed mutagenesis (Boissel et al. 1993) etc. In this way, a comparative model, especially in the difficult cases, could be improved by making it consistent with available experimental data and with more general knowledge about protein structure.

In the past, comparative modelling relied mostly on template information and statistically-derived restraints from known protein structures and sequences. But it is expected that with the advances of large scale genetic and proteomics techniques more and more experimentally derived restraints will be available for automatic incorporation in the modelling process.

A particularly active topic within “hybrid modeling”, i.e. to employ limited, easily obtainable indirect experimental information to improve modeling, is focusing on NMR restraints. This is because during the last decade Structural Genomics centres successfully automated many steps of protein production, but the successful crystallization of proteins remains a major bottleneck. According to large-scale statistics, about 63% of purified proteins will result in crystals but only 10% will be of diffracting quality that is suitable for X-ray crystallography. This means that, considering only the four high throughput SG centres, about 5000 purified proteins per centre are produced that will never get solved. Structural Genomics delivers about 10% of all structures solved so, even according to a conservative estimate, tens of thousands of purified proteins are produced each year for the purpose of structure solution but will not end up in a structural model. These proteins are accessible from the PSI Materials Repository (Cormier et al. 2011) and are suitable for NMR studies unless their structure exceeds about 200 residues, where the resonance assignment of NMR spectra becomes difficult. Like crystallography, NMR studies can also be time consuming and ultimately unsuccessful unless all chemical shifts and NOESY peak lists are assigned. However, there are a number of NMR data types (often insufficient to produce structural models on their own) that can be collected within days in a semi-automated manner. This sparse

NMR data can subsequently be combined with computational modeling to deliver molecular models.

Technological advances facilitate the collection and analysis of a variety of experimental data in a high-throughput fashion. These include the use of automated programs that can speed up assignment of resonances in NMR spectra (Raman et al. 2010a), or the use of robotics for protein production (Graslund et al. 2008) and labeling (Crublet et al. 2014). This is in contrast to conventional NMR structure determination, which requires a nearly complete assignment of chemical shifts and cross peaks in a NOESY spectrum, calling for an iterative, manual approach.

Following this trend, a growing number of methods incorporate a variety of easily obtainable NMR data as restraints to guide protein structure modelling or simulation. Many of these methods focus on backbone NMR chemical shift (CS) assignments. Obtaining CS is a necessary first step in the classical NMR structure determination process. Backbone CS data are the easier to obtain in comparison to assigning side chain resonances or determining large numbers of interproton distances (NOEs). Residual dipolar coupling data is another possible source of structural restraints for molecular modelling (Rohl and Baker 2002), although this type of data is available for far fewer proteins.

A number of programs use NMR CS data to predict secondary structure conformations (Hung and Samudrala 2003; Shen et al. 2009a; Wishart and Sykes 1994). Within the framework of developing the TALOS program, it was shown that CS data can guide the selection of tripeptide segments with similar conformations and provide preferences/restraints for mainchain dihedral angles (Cornilescu et al. 1999; Shen et al. 2009a). Recently, TALOS was extended to specifically address CS-based dihedral angle predictions in loop segments (Shen and Bax 2012). The Rosetta ab initio fragment assembly program (Bonneau et al. 2002) was combined with chemical shift data and sparse NOE restraints ($\sim$1 per residue) to steer the selection and filtering of three and nine residue fragments, besides taking into account sequence similarity measures of these fragments (Bowers et al. 2000). The method explored a range of structures between 52 and 152 residues and delivered models as good as 1.5 Å RMSD from the experimental solution, although the results became weaker with larger proteins. In a similar approach by Rose et al. (Gong et al. 2007), experimentally determined CS and sequence patterns were used to search the protein database for consecutively overlapping six residue long backbone fragments, which then were "stitched" together using Monte Carlo simulation. In more recent applications, CS-Rosetta was shown to be successful in delivering high quality models (below 2.5 Å RMSD from the experimental solution structure) when using CS data in combination with sequence information (Shen et al. 2008, 2009b). Once CS restraints were added to the scoring function the selection of low energy models consistently improved.

The applicability of CS-Rosetta was recently extended for larger molecules (>12 kDa) through the incorporation of NMR residual dipolar coupling data (Raman et al. 2010b). The combined approach uses sequence information of short three and nine residue segments, NMR CS and residual dipolar coupling data together were shown to produce homology modelling quality models (with

GDT_TS (Zemla 2003) values above 41%, but with an average in the high 70% for the superposable parts of the proteins) for molecules up to 266 residues. Similar ideas are implemented in the CHESHIRE method, which first predicts secondary structures of three and nine residue fragments using CS data and then combines these fragments into larger ones by matching sequence information, secondary structures and CS patterns (Cavalli et al. 2007). In an elegant approach from the same group, NMR CS data were converted into forces in molecular dynamics simulations and were successfully used to fold short polypeptide chains or to refine partially unfolded structures (Robustelli et al. 2009, 2010). An important advance for that work was the development of CamShift method (Kohlhoff et al. 2009) that quickly predicts CS values from structures, approximating CS with a polynomial function of interatomic distances. This results in a readily differentiable function with respect to the coordinates of atomic positions and therefore is suitable to use as restraints in molecular dynamics simulations. Using these CS imposed restraints, it was possible to properly fold 11 out of 12 partially unfolded test proteins, while without the CS restraints and using only the molecular dynamics force field, only one protein folded properly. The RMSD of unfolded models for the parts of the protein with a translational symmetry were within 3.2–7 Å RMSD away from the solution structures and were refined to below 2.2 Å RMSD after the simulation. Besides CamShift several other approaches are available that calculate theoretical CS values for a given structure, such as SHIFTX2 (Han et al. 2011), SPARTA+ (Shen and Bax 2010) and PROSHIFT (Meiler 2003). GENMR (Berjanskii et al. 2009) is a very fast modelling implementation that combines homology models with CS and/or NOE data. The component of GENMR that relies on structure calculation using CS and sequence information without NOE data is CS23D (Wishart et al. 2008). CS23D incorporates various other methods, such as threading, homology modeling or small fragment assembly using the Rosetta program.

Recently, the limits of applicability of a previously-developed fragment-based loop modeling approach was explored (Fernandez-Fuentes et al. 2006a, b) revealing that the protein structure universe seems to have saturated on the level of super-secondary motifs (Fernandez-Fuentes and Fiser 2006). It was observed that the library of Smotifs with similar internal geometries have saturated and new folds discovered during the last decade did not require the emergence of new Smotifs, but new folds appear to be novel combination of existing Smotifs (Fernandez-Fuentes et al. 2010). This observation presents a hypothesis according to which, it should be possible to build any new or yet to be discovered structure by combining existing Smotifs from already known structures. The library of Smotifs is a backbone-only, geometrically-defined fragment library, which means that for practical modeling applications, a relation needs to be made between the target protein and specific fragments in the library. A method SmotifCS was developed (Menon et al. 2013), that use of NMR CS data to select Smotifs. Even without any input about sequence information, when tested on a set of 102 different fold topologies the method

returned a homology model quality solution for about a 90% of cases and at least a topologically correct fold for almost all of them (Menon et al. 2013).

In addition to deliver more accurate models by hybrid modeling, the idea of using limited experimental restraints should particularly facilitate the modelling of protein complexes and assemblies (Alber et al. 2008).

A systematic approach to tackle the modelling of large protein complexes with the aid of experimental restraints was developed for the modelling of the nuclear pore complex, the largest known protein complex in the cell that consist of 456 proteins (Alber et al. 2008). The approach integrated a wealth of experimental information. For instance, quantitative immunoblotting determined the stoichiometry, while hydrodynamics experiments provided insight about the approximate shape and excluded volume of each nucleoporins; immuno—EM helped in coarse localization of nucleoporins; affinity purification determined the composition of complexes; cryo-EM and bioinformatics analysis uncovered locations of transmembrane segments and overlay experiments gave information on direct binary interactions. All these data inputs were integrated in a hierarchical process that combined comparative modelling, threading, rigid and flexible docking techniques. The ultimate goal of the data integration is to convert all available experimental information into spatial restraints that can guide the generalized modelling procedure. The procedure is flexible to combine entities of various representations and resolutions (for instance atoms, atomistic models of proteins, symmetry units or whole assemblies) and optimization procedures (Alber et al. 2007a, b, 2008). This and similar efforts will leverage benefits simultaneously from efforts of genome sequencing, functional genomics, proteomics systems biology and structural biology.

### *4.2.5 Model Evaluation*

After a model is built, it is important to check it for possible errors. The quality of a model can be approximately predicted from the sequence similarity between the target and the template. Sequence identity above 30% is a relatively good predictor of the expected accuracy of a model. If the target-template sequence identity falls below 30%, the sequence identity becomes significantly less reliable as a measure of the expected accuracy of a single model. It is in such cases that model evaluation methods are most informative.

Two types of evaluation can be carried out. “Internal” evaluation of self-consistency checks whether or not a model satisfies the restraints used to calculate it, including restraints that originate from the template structure or obtained from statistical observations. “External” evaluation relies on information that was not used in the calculation of the model.

Assessment of the stereochemistry of a model (e.g., bonds, bond angles, dihedral angles, and non-bonded atom-atom distances) with programs such as PROCHECK

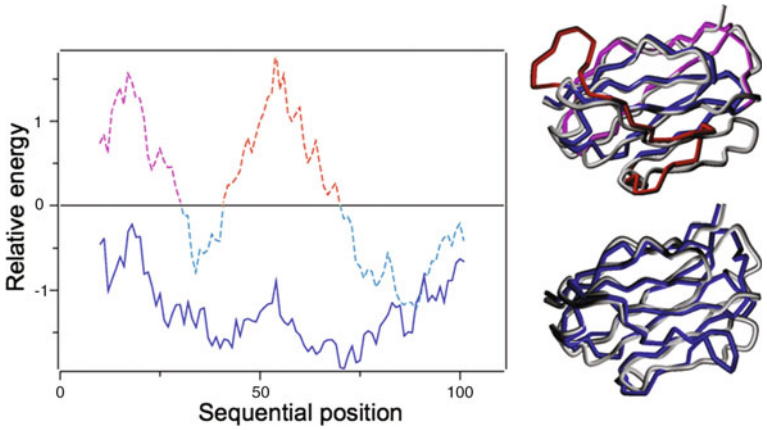



**Fig. 4.3** Residue energy, using a pairwise statistical potential, is plotted as a function of sequential residue positions for two alternative models of the same protein. Negative (*blue color*) and positive (*red color*) energies indicate energetically favorable and unfavorable residue environments, respectively. The energy profiles correspond to the models shown on the *right* with the inaccurate model placed above the more accurate model. Corresponding parts in the models and energy profiles use the same colour coding scheme while a colourless trace represents the actual experimental structure

(Laskowski et al. 1993) and WHATCHECK (Hooft et al. 1996) is an example of internal evaluation. Although errors in stereochemistry are rare and less informative than errors detected by methods for external evaluation, a cluster of stereochemical errors may indicate that the corresponding region also contains other larger errors (e.g., alignment errors).

As a minimum, external evaluations test whether or not a correct template was used. Luckily a wrong template can be detected easily with the currently available scoring functions. A more challenging task for the scoring functions is the prediction of unreliable regions in the model. One way to approach this problem is to calculate a "pseudo energy" profile of a model, such as that produced by PROSA (Sippl 1993) or Verify3D (Eisenberg et al. 1997). The profile reports the energy for each position in the model (Fig. 4.3). Peaks in the profile frequently correspond to errors in the model. There are several pitfalls in the use of energy profiles for local error detection. For example, a region can be identified as unreliable only because it interacts with an incorrectly modelled region (Fiser et al. 2000). The development of accurate model assessment scoring methods remain very active (Rykunov and Fiser 2010; Rykunov et al. 2009; Zhou and Skolnick 2011). Other recent approaches usually combine a variety of inputs to assess the models, either as a whole (Eramian et al. 2006) or locally (Fasnacht et al. 2007). In benchmarks the best quality assessor techniques use a simple consensus approach where reliability of a model is assessed by the agreement among alternative models that are sometimes obtained from a variety of methods (Wallner and Elofsson 2005b, 2007). Model assessment is an important but difficult area, due to a circular argument: scoring

function terms of an effective model assessment approach should be used in the first place to produce accurate models.

## 4.3 Performance of Comparative Modelling

### *4.3.1 Accuracy of Methods*

An informative way to test protein structure modelling methods, including comparative modelling, is provided by the bi-annual meetings on Critical Assessment of Techniques for Protein Structure Prediction (CASP) (Moult 2005). Protein modellers are challenged to model sequences with unknown 3D structure and to submit their models to the organizers before the meeting. At the same time, the 3D structures of the prediction targets are being determined by X-ray crystallography or NMR methods. They only become available after the models are calculated and submitted. Thus, a *bona fide* evaluation of protein structure modelling methods is possible, although in these exercises it is not trivial to separate the contributions from programs and human expert knowledge.

Alternatively a large scale, continuous, and automated prediction benchmarking experiment was implemented in the program EVA—Evaluation of Automatic protein structure prediction (Eyrich et al. 2001). Every week EVA submitted pre-released PDB sequences to participating modelling servers, collected the results and provided detailed statistics on secondary structure prediction, fold recognition, comparative modelling, and prediction on 3D contacts. The LiveBench program had implemented its evaluations in a similar spirit (Bujnicki et al. 2001). After these initial attempts, currently there are two operational continuous evaluation servers that benchmark publicly accessible methods, CASPRoll (http://predictioncenter.org/casproll/) and CAMEO (Continuous Automated Model Evaluation) (Haas et al. 2013).

A rigorous statistical evaluation (Marti-Renom et al. 2002) of a blind prediction experiment illustrated that the accuracies of the various model-building methods, using segment matching, rigid body assembly, satisfaction of spatial restraints or any combinations of these are relatively similar when used optimally (Wallner and Elofsson 2005a; Dalton and Jackson 2007). This also reflects on the fact that such major factors as template selection and alignment accuracy have a large impact on the overall model accuracy, and that the core of protein structures is highly conserved. From a practical point of view models should be evaluated by their usefulness regarding the functional insight they provide. A unique functional role must be connected with unique structural features, which is more often found in variable loop regions than in the conserved core. However, functional site descriptions are not only manually defined but, in an increasing fraction of cases, are missing or incomplete. This particularly applies to the outputs from Structural Genomics projects, which often focus specifically and deliberately on proteins of unknown function. Therefore, while large-scale benchmarking of modelling methods through

the evaluation of the accuracy of functional annotations based on the resulting models is desirable, it is not yet straightforward to carry out in practice (Chakravarty et al. 2005; Chakravarty and Sanchez 2004).

### 4.3.2 Errors in Comparative Models

The overall accuracy of comparative models spans a wide range. At the low end of the spectrum are the low resolution models whose only essentially correct feature is their fold. At the high end of the spectrum are the models with an accuracy comparable to medium resolution crystallographic structures (Baker and Sali 2001). Even low resolution models are often useful to address biological questions, because function can many times be predicted from only coarse structural features of a model, as later chapters of this book illustrate.

The errors in comparative models can be divided into five categories: (1) Errors in side chain packing. (2) Distortions or shifts of a region that is aligned correctly with the template structures. (3) Distortions or shifts of a region that does not have an equivalent segment in any of the template structures. (4) Distortions or shifts of a region that is aligned incorrectly with the template structures. (5) A misfolded structure resulting from using an incorrect template. Significant methodological improvements are needed to address each of these errors.

Errors 3–5 are relatively infrequent when sequences with more than 40% identity to the templates are modelled. For example, in such a case, approximately 90% of the main chain atoms are likely to be modelled with an RMS error of about 1 Å (Sanchez and Sali 1998). In this range of sequence similarity, the alignment is mostly straightforward to construct, there are not many gaps, and the structural differences between the proteins are usually limited to loops and side chains. When sequence identity is between 30 and 40%, the structural differences become larger, and the gaps in the alignment are more frequent and longer, misalignments and insertions in the target sequence become the major problems. As a result, the main chain RMS error rises to about 1.5 Å for about 80% of residues. The rest of the residues are modelled with large errors because the methods generally fail to model structural distortions and rigid body shifts, and are unable to recover from misalignments. When sequence identity drops below 30%, the main problem becomes the identification of related templates and their alignment with the sequence to be modelled. In general, it can be expected that about 20% of residues will be misaligned, and consequently incorrectly modelled with an error larger than 3 Å, at this level of sequence similarity. These misalignments are a serious impediment for comparative modelling because it appears that most structurally related protein pairs share less than 30% sequence identity (Rost 1999).

To put the errors in comparative models into perspective, we list the differences among structures of the same protein that have been determined experimentally.

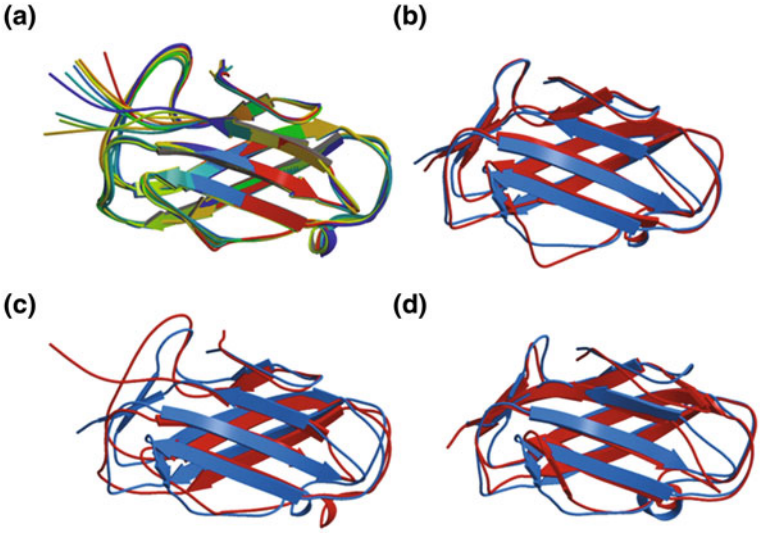



**Fig. 4.4** Illustrating accuracies of structural models obtained from various experimental and computational sources for the same, Der P 2 allergen protein. **a** Superposition of 10 alternative NMR solution structures (PDB code 1A9V) for Der P 2; Average RMSD = 0.97 Å. **b** Superposition of X-ray crystallographic structures of two isoforms of Der P 2 protein (2F08 (2.20 Å resolution) and 1KTJ (2.15 Å resolution) sharing 87% sequence identity). RMSD = 1.33 Å; **c** Superposition of NMR and X-ray solutions of Der P 2 protein (1A9V and 1KTJ). RMSD = 2.2 Å; **d** Superposition of the comparative model built for 1NEP protein using 1KTJ as a template and the X-ray solution structure of 1NEP. 1NEP and 1KTJ share 28% sequence identity representing a typical difficult comparative modeling scenario. RMSD = 1.66 Å. All RMSD values refer to Cα superpositions

A 1 Å accuracy of main chain atom positions corresponds to X-ray structures defined at a low-resolution of about 2.5 Å and with an R-factor of about 25% (Ohlendorf 1994), as well as to medium-resolution NMR structures determined from 10 inter-proton distance restraints per residue (Fig. 4.4). Similarly, differences between the highly refined X-ray and NMR structures of the same protein also tend to be about 1 Å (Clore et al. 1993). Changes in the environment (e.g., oligomeric state, crystal packing, solvent, ligands) can also have a significant effect on the structure (Faber and Matthews 1990). Overall, comparative modelling based on templates with more than 40% identity is almost as good as medium resolution experimental structures, simply because the proteins at this level of similarity are likely to be as similar to each other as are the structures for the same protein determined by different experimental techniques under different conditions. However, the caveat in comparative protein modelling is that some regions, mainly loops and sidechains, may have larger errors.

The performance of comparative modelling may sometimes appear overstated, because what is usually discussed in the literature are the mean values of backbone deviations. However, individual errors in certain residues essential for the protein

function, even in the context of an overall backbone RMSD of less than 1 Å, can still be large enough to prevent reliable conclusions to be drawn regarding mechanism, protein function or drug design.

## 4.4 Applications of Comparative Modelling

### *4.4.1 Modelling of Individual Proteins*

Comparative modelling is often an efficient way to obtain useful information about the proteins of interest. For example, comparative models can be helpful in designing genetic experiments, such as designing mutants to test hypotheses about the function of a protein (Vernal et al. 2002; Wu et al. 1999; Shin et al. 2012), identifying active and binding sites (Sheng et al. 1996). Models are useful for studying protein-protein, protein-nucleic acid (Pujato et al. 2014) and protein-ligand interactions, designing inhibitors, e.g. searching, designing and improving ligands for a given binding site (Ring et al. 1993), modelling substrate specificity (Xu et al. 1996), predicting antigenic epitopes (Sali et al. 1993; Abboud et al. 2009), simulating protein-protein docking (Vakser 1995). Models can reveal physico-chemical features that are not possible to predict from sequence information only, for instance, inferring function from calculated electrostatic potential around the protein (Sali et al. 1993; Fiser and Vertessy 2000) and in general, rationalizing known experimental observations (Fiser et al. 2003). Models are also very useful to enhance structure solutions by facilitating molecular replacement in X-ray structure determination (Schwarzenbacher et al. 2008), refining models based on NMR constraints (Barrientos et al. 2001), confirming a remote structural relationship (Guenther et al. 1997; Wu et al. 1999). Comparative modeling is extensively used in applied research, in the context of structure-based drug discovery (Norin and Sundstrom 2001; Wlodawer 2002; Schwede et al. 2009; Evers et al. 2003) or designing proteins as drugs. For instance, of the 21 antibodies currently on the market, it is estimated (Schwede et al. 2009) that 11 were the result of computational design of humanized constructs via homology modelling [e.g. Zenapax (Carter et al. 1992), Herceptin (Lippow et al. 2007; Presta et al. 1997) and Avastin (Queen et al. 1989)].

### *4.4.2 Comparative Modelling and the Protein Structure Initiative*

The full impact of the genome projects will only be realized once we assign and understand the functions of the new encoded proteins. This understanding will be

facilitated by structural information for all or almost all proteins. Much of the structural information will be provided by Structural Genomics (Burley et al. 2008; Chance et al. 2002), a large-scale determination of protein structures by X-ray crystallography and nuclear magnetic resonance spectroscopy, combined efficiently with accurate, automated and large-scale comparative protein structure modelling techniques. Given the performance of the current modelling techniques, it seems reasonable to require models based on at least 30% sequence identity (Vitkup et al. 2001), corresponding to one experimentally determined structure per sequence family, rather than fold family.

To enable large-scale comparative modelling needed for structural genomics, the steps of comparative modelling are being assembled into a completely automated pipelines such as SWISS-model repository (Biasini et al. 2014) or MODBASE (Pieper et al. 2014), which contain more than 3 and 30 million models, respectively. Statistics of these databases show that domains in approximately 70% of the known protein sequences can be modelled, at least partially. This is due substantially of the almost 7000 structures that were deposited by the structural genomics centres, which focus on new folds or novel structure. These depositions contributed 73% of all novel structural features in the PDB (Burley et al. 2008).

While the current number of at least partially modelled proteins may look impressive, usually only one domain per protein is modelled. On average, in contrast, proteins have two or three domains. For example, the average length of a yeast ORF is 472 amino acids, while the average size of domains in CATH, a database of structural domains, is 175 amino acids. The average model size in MODBASE, a database of comparative models, is only slightly longer at 192 residues. Furthermore, in two thirds of the modelling cases the template shares less than 30% sequence identity to the closest template.

## 4.5 Summary

Comparative modelling has already proven to be a useful tool in many biological applications and its importance among structure prediction methods is expected to be further accentuated because of the many experimental structures emerging from Protein Structure Initiative projects and the continuous improvements in methodologies.

The average sequence identity between structurally related proteins in general is just around 8–9%, and most of them share less than 15% identity (Rost 1997). Comparative modelling is largely restricted to that subset of sequences that share a recognizable sequence similarity to a protein with a known structure; therefore it is safe to assume that this approach is still only scratching the surface of possibilities in terms of recognizing and utilizing useful structural information. Indeed, recent results suggest that there is a fairly limited number of structural building blocks that make up all known protein folds and that the library of these building blocks has

already saturated around year 2000 (Fernandez-Fuentes et al. 2010). Now it is largely up to the ability of current computational methods to relate any known sequences to one of the known folds or at least associate its building blocks with one of the known structural motifs. Fold recognition methods discussed in Chap. 2 will have an important role in extending the possibilities for comparative modelling towards ever remote homologues and even structural analogues. Hybrid methods, where limited, indirect experimental data supplements weak sequence signal will also have an increasing role in structure modeling, when trying to relate local structural motifs to known ones.

Improved and new methods to refining comparative models by adding accurate loops and side chains, refining internal packing of secondary structural elements, setting up scoring functions that can measure model quality, optimally combining fragments from known folds and detecting errors in the 3D models are critical issues. Even a small improvement in these techniques will have a large impact because most of the protein structural relationships are too remote to utilize them in comparative modelling. On the other hand, while improvements in these topics may not have a significant impact on the overall accuracy of already existing protein models, their importance in achieving *functionally* more reliable 3D models i.e. models that can confidently be used for functional annotation, can not be emphasized enough.

The above advances in comparative protein structure modelling techniques are necessary prerequisites to develop new "structural proteomics" modelling methods with the aim of combining the basic building blocks of fold models into physiologically more relevant quaternary structures and assemblies. This will create possibilities for modelling interactions among the many known protein structures.

## References

Abagyan R, Totrov M (1994) Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. J Mol Biol 235(3):983–1002

Abboud N, De Jesus M, Nakouzi A, Cordero RJ, Pujato M, Fiser A, Rivera J, Casadevall A (2009) Identification of linear epitopes in *Bacillus anthracis* protective antigen bound by neutralizing antibodies. J Biol Chem 284(37):25077–25086

Adhikari AN, Peng J, Wilde M, Xu J, Freed KF, Sosnick TR (2012) Modeling large regions in proteins: applications to loops, termini, and folding. Protein Sci 21(1):107–121

Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Rout MP, Sali A (2007a) Determining the architectures of macromolecular assemblies. Nature 450(7170):683–694

Alber F, Dokudovskaya S, Veenhoff LM, Zhang W, Kipper J, Devos D, Suprapto A, Karni-Schmidt O, Williams R, Chait BT, Sali A, Rout MP (2007b) The molecular architecture of the nuclear pore complex. Nature 450(7170):695–701

Alber F, Forster F, Korkin D, Topf M, Sali A (2008) Integrating diverse data for structure determination of macromolecular assemblies. Annu Rev Biochem 77:443–477

Al-Lazikani B, Sheinerman FB, Honig B (2001) Combining multiple structure and sequence alignments to improve sequence detection and alignment: application to the SH2 domains of Janus kinases. Proc Natl Acad Sci U S A 98(26):14796–14801

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36 (Database issue):D419–D425

Apostolico A, Giancarlo R (1998) Sequence alignment in molecular biology. J Comput Biol: J Comput Mol Cell Biol 5(2):173–196

Apweiler R, Bairoch A, Wu CH (2004) Protein sequence databases. Curr Opin Chem Biol 8 (1):76–80

Aszodi A, Taylor WR (1994) Secondary structure formation in model polypeptide chains. Protein Eng 7(5):633–644

Aszodi A, Taylor WR (1996) Homology modelling by distance geometry. Fold Des 1(5):325–334

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294 (5540):93–96

Barrientos LG, Campos-Olivas R, Louis JM, Fiser A, Sali A, Gronenborn AM (2001) 1H, 13C, 15N resonance assignments and fold verification of a circular permuted variant of the potent HIV-inactivating protein cyanovirin-N. J Biomol NMR 19(3):289–290

Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. Proteins 69(Suppl 8):68–82

Becker OM, Dhanoa DS, Marantz Y, Chen D, Shacham S, Cheruku S, Heifetz A, Mohanty P, Fichman M, Sharadendu A, Nudelman R, Kauffman M, Noiman S (2006) An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. J Med Chem 49 (11):3116–3135

Berjanskii M, Tang P, Liang J, Cruz JA, Zhou J, Zhou Y, Bassett E, MacDonell C, Lu P, Lin G, Wishart DS (2009) GeNMR: a web server for rapid NMR-based protein structure determination. Nucleic Acids Res 37 (Web Server issue):W670–W677

Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res 35 (Database issue):D301–D303

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42 (Web Server issue): W252–W258

Blake JD, Cohen FE (2001) Pairwise sequence alignment below the twilight zone. J Mol Biol 307 (2):721–735

Blundell TL, Sibanda BL, Sternberg MJ, Thornton JM (1987) Knowledge-based prediction of protein structures and the design of novel molecules. Nature 326(6111):347–352

Boissel JP, Lee WR, Presnell SR, Cohen FE, Bunn HF (1993) Erythropoietin structure-function relationships. Mutant proteins that test a model of tertiary structure. J Biol Chem 268 (21):15983–15993

Bonneau R, Baker D (2001) Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173–189

Bonneau R, Strauss CE, Rohl CA, Chivian D, Bradley P, Malmstrom L, Robertson T, Baker D (2002) De novo prediction of three-dimensional structures for major protein families. J Mol Biol 322(1):65–78

Boratyn GM, Schaffer AA, Agarwala R, Altschul SF, Lipman DJ, Madden TL (2012) Domain enhanced lookup time accelerated BLAST. Biol Direct 7:12

Bowers PM, Strauss CE, Baker D (2000) De novo protein structure determination using sparse NMR data. J Biomol NMR 18(4):311–318

Bowie JU, Luthy R, Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. Science 253(5016):164–170

Braun W, Go N (1985) Calculation of protein conformations by proton-proton distance constraints. A new efficient algorithm. J Mol Biol 186(3):611–626

Brenner SE, Chothia C, Hubbard TJ (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A 95 (11):6073–6078

Brooks BR, Brooks CL 3rd, Mackerell AD Jr, Nilsson L, Petrella RJ, Roux B, Won Y, Archontis G, Bartels C, Boresch S, Caflisch A, Caves L, Cui Q, Dinner AR, Feig M, Fischer S, Gao J, Hodoscek M, Im W, Kuczera K, Lazaridis T, Ma J, Ovchinnikov V, Paci E, Pastor RW, Post CB, Pu JZ, Schaefer M, Tidor B, Venable RM, Woodcock HL, Wu X, Yang W, York DM, Karplus M (2009) CHARMM: the biomolecular simulation program. J Comput Chem 30(10):1545–1614

Browne WJ, North AC, Phillips DC, Brew K, Vanaman TC, Hill RL (1969) A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. J Mol Biol 42(1):65–86

Bruccoleri RE, Karplus M (1987) Prediction of the folding of short polypeptide segments by uniform conformational sampling. Biopolymers 26(1):137–168

Bruccoleri RE, Karplus M (1990) Conformational sampling using high-temperature molecular dynamics. Biopolymers 29(14):1847–1862

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 10(2):352–361

Burley SK, Joachimiak A, Montelione GT, Wilson IA (2008) Contributions to the NIH-NIGMS protein structure initiative from the PSI production centers. Structure 16(1):5–11

Carter P, Presta L, Gorman CM, Ridgway JB, Henner D, Wong WL, Rowland AM, Kotts C, Carver ME, Shepard HM (1992) Humanization of an anti-p185HER2 antibody for human cancer therapy. Proc Natl Acad Sci U S A 89(10):4285–4289

Cavalli A, Salvatella X, Dobson CM, Vendruscolo M (2007) Protein structure determination from NMR chemical shifts. Proc Natl Acad Sci U S A 104(23):9615–9620

Chakravarty S, Sanchez R (2004) Systematic analysis of added-value in simple comparative models of protein structure. Structure 12(8):1461–1470

Chakravarty S, Wang L, Sanchez R (2005) Accuracy of structure-derived properties in simple comparative models of protein structures. Nucleic Acids Res 33(1):244–259

Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK (2002) Structural genomics: a pipeline for providing structures for the biologist. Protein Sci 11(4):723–738

Chinea G, Padron G, Hooft RW, Sander C, Vriend G (1995) The use of position-specific rotamers in model building by homology. Proteins 23(3):415–421

Chivian D, Baker D (2006) Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res 34(17):e112

Chopra G, Kalisman N, Levitt M (2010) Consistent refinement of submitted models at CASP using a knowledge-based potential. Proteins 78(12):2668–2678

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826

Chothia C, Lesk AM (1987) Canonical structures for the hypervariable regions of immunoglobulins. J Mol Biol 196(4):901–917

Chothia C, Lesk AM, Tramontano A, Levitt M, Smith-Gill SJ, Air G, Sheriff S, Padlan EA, Davies D, Tulip WR (1989) Conformations of immunoglobulin hypervariable regions. Nature 342(6252):877–883

Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300(5626):1701–1703

Claessens M, Van Cutsem E, Lasters I, Wodak S (1989) Modelling the polypeptide backbone with 'spare parts' from known protein structures. Protein Eng 2(5):335–345

Clore GM, Brunger AT, Karplus M, Gronenborn AM (1986) Application of molecular dynamics with interproton distance restraints to three-dimensional protein structure determination. A model study of crambin. J Mol Biol 191(3):523–551

Clore GM, Robien MA, Gronenborn AM (1993) Exploring the limits of precision and accuracy of protein structures determined by nuclear magnetic resonance spectroscopy. J Mol Biol 231 (1):82–102

Cohen FE, Kuntz ID, Fasman GD (1989) Tertiary structure prediction. In: Fasman GD (ed) Prediction of protein structure and the principles of protein conformations. Plenum, New York, pp 647–705

Collura V, Higo J, Garnier J (1993) Modeling of protein loops by simulated annealing. Protein Sci 2(9):1502–1510

Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA (2003) Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 53(Suppl 6):424–429

Cormier C, Steel J, Fiacco M, Park J, Kramer J, LaBaer J (2011) PSI: biology-materials repository: developing a public resource for structural biology plasmids. Biophys J 100(3):52

Cornilescu G, Delaglio F, Bax A (1999) Protein backbone angle restraints from searching a database for chemical shift and sequence homology. J Biomol NMR 13(3):289–302

Crublet E, Kerfah R, Mas G, Noirclerc-Savoye M, Lantez V, Vernet T, Boisbouvier J (2014) A cost-effective protocol for the parallel production of libraries of 13CH3-specifically labeled mutants for NMR studies of high molecular weight proteins. Methods Mol Biol 1091:229–244

Cuff AL, Sillitoe I, Lewis T, Clegg AB, Rentzsch R, Furnham N, Pellegrini-Calace M, Jones D, Thornton J, Orengo CA (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. Nucleic Acids Res 39 (Database issue): D420–D426

Dalton JA, Jackson RM (2007) An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics 23(15):1901–1908

Das B, Meirovitch H (2003) Solvation parameters for predicting the structure of surface loops in proteins: transferability and entropic effects. Proteins 51(3):470–483

Das R, Qian B, Raman S, Vernon R, Thompson J, Bradley P, Khare S, Tyka MD, Bhat D, Chivian D, Kim DE, Sheffler WH, Malmstrom L, Wollacott AM, Wang C, Andre I, Baker D (2007) Structure prediction for CASP7 targets using extensive all-atom refinement with Rosetta@home. Proteins 69(Suppl 8):118–128

de Bakker PI, DePristo MA, Burke DF, Blundell TL (2003) Ab initio construction of polypeptide fragments: accuracy of loop decoy discrimination by an all-atom statistical potential and the AMBER force field with the generalized born solvation model. Proteins 51(1):21–40

Deane CM, Blundell TL (2001) CODA: a combined algorithm for predicting the structurally variable regions of protein models. Protein Sci 10(3):599–612

DePristo MA, de Bakker PI, Lovell SC, Blundell TL (2003) Ab initio construction of polypeptide fragments: efficient generation of accurate, representative ensembles. Proteins 51(1):41–55

Dill KA, Chan HS (1997) From Levinthal to pathways to funnels. Nat Struct Biol 4(1):10–19

Do CB, Mahabhashyam MS, Brudno M, Batzoglou S (2005) ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res 15(2):330–340

Du P, Andrec M, Levy RM (2003) Have we seen all structures corresponding to short protein fragments in the Protein Data Bank? An update. Protein Eng 16(6):407–414

Edgar RC, Batzoglou S (2006) Multiple sequence alignment. Curr Opin Struct Biol 16(3):368–373

Edgar RC, Sjolander K (2003) Simultaneous sequence alignment and tree construction using hidden Markov models. Pac Symp Biocomput 180–191

Edgar RC, Sjolander K (2004) COACH: profile-profile alignment of protein families using hidden Markov models. Bioinformatics 20(8):1309–1318

Eisenberg D, Luthy R, Bowie JU (1997) VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 277:396–404

Eramian D, Shen MY, Devos D, Melo F, Sali A, Marti-Renom MA (2006) A composite score for predicting errors in protein structure models. Protein Sci 15(7):1653–1666

Espadaler J, Fernandez-Fuentes N, Hermoso A, Querol E, Aviles FX, Sternberg MJ, Oliva B (2004) ArchDB: automated protein loop classification as a tool for structural genomics. Nucleic Acids Res 32 (Database issue):D185

Evers A, Gohlke H, Klebe G (2003) Ligand-supported homology modelling of protein binding-sites using knowledge-based potentials. J Mol Biol 334(2):327–345

Eyrich VA, Marti-Renom MA, Przybylski D, Madhusudhan MS, Fiser A, Pazos F, Valencia A, Sali A, Rost B (2001) EVA: continuous automatic evaluation of protein structure prediction servers. Bioinformatics 17(12):1242–1243

Faber HR, Matthews BW (1990) A mutant T4 lysozyme displays five different crystal conformations. Nature 348(6298):263–266

Fajardo JE, Fiser A (2013) Protein structure based prediction of catalytic residues. BMC Bioinform 14:63

Fasnacht M, Zhu J, Honig B (2007) Local quality assessment in homology models using statistical potentials and support vector machines. Protein Sci 16(8):1557–1568

Felsenstein J (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17(6):368–376

Fernandez-Fuentes N, Fiser A (2006) Saturating representation of loop conformational fragments in structure databanks. BMC Struct Biol 6:15

Fernandez-Fuentes N, Oliva B, Fiser A (2006a) A supersecondary structure library and search algorithm for modeling loops in protein structures. Nucleic Acids Res 34(7):2085–2097

Fernandez-Fuentes N, Zhai J, Fiser A (2006b) ArchPRED: a template based loop structure prediction server. Nucleic Acids Res 34 (Web Server issue):W173–W176

Fernandez-Fuentes N, Madrid-Aliste CJ, Rai BK, Fajardo JE, Fiser A (2007a) M4T: a comparative protein structure modeling server. Nucleic Acids Res 35 (Web Server issue):W363–W368

Fernandez-Fuentes N, Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A (2007b) Comparative protein structure modeling by combining multiple templates and optimizing sequence-to-structure alignments. Bioinformatics 23(19):2558–2565

Fernandez-Fuentes N, Dybas JM, Fiser A (2010) Structural characteristics of novel protein folds. PLoS Comput Biol 6 (4):e1000750

Fidelis K, Stern PS, Bacon D, Moult J (1994) Comparison of systematic search and database methods for constructing segments of protein structure. Protein Eng 7(8):953–960

Fine RM, Wang H, Shenkin PS, Yarmush DL, Levinthal C (1986) Predicting antibody hypervariable loop conformations. II: Minimization and molecular dynamics studies of MCPC603 from many randomly generated loop conformations. Proteins 1(4):342–362

Finkelstein AV, Reva BA (1991) A search for the most stable folds of protein chains. Nature 351 (6326):497–499

Fiser A (2004) Protein structure modeling in the proteomics era. Expert Rev Proteomics 1 (1):97–110

Fiser A, Sali A (2003a) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374:461–491

Fiser A, Sali A (2003b) ModLoop: automated modeling of loops in protein structures. Bioinformatics 19(18):2500–2501

Fiser A, Vertessy BG (2000) Altered subunit communication in subfamilies of trimeric dUTPases. Biochem Biophys Res Commun 279(2):534–542

Fiser A, Do RK, Sali A (2000) Modeling of loops in protein structures. Protein Sci 9 (9):1753–1773

Fiser A, Feig M, Brooks CL III, Sali A (2002) Evolution and physics in comparative protein structure modeling. Acc Chem Res 35(6):413–421

Fiser A, Filipe SR, Tomasz A (2003) Cell wall branches, penicillin resistance and the secrets of the MurM protein. Trends Microbiol 11(12):547–553

Fogolari F, Tosatto SC (2005) Application of MM/PBSA colony free energy to loop decoy discrimination: toward correlation between energy and root mean square deviation. Protein Sci 14(4):889–901

Forrest LR, Woolf TB (2003) Discrimination of native loop conformations in membrane proteins: decoy library design and evaluation of effective energy scoring functions. Proteins 52 (4):492–509

Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16 (2):172–177

Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 19(8):1015–1018

Gong H, Shen Y, Rose GD (2007) Building native protein conformation from NMR backbone chemical shifts using Monte Carlo fragment assembly. Protein Sci 16(8):1515–1521

Grabarek Z (2006) Structural basis for diversity of the EF-hand calcium-binding proteins. J Mol Biol 359(3):509–525

Grant A, Lee D, Orengo C (2004) Progress towards mapping the universe of protein folds. Genome Biol 5(5):107

Graslund S, Nordlund P, Weigelt J, Hallberg BM, Bray J, Gileadi O, Knapp S, Oppermann U, Arrowsmith C, Hui R, Ming J, dhe-Paganon S, Park HW, Savchenko A, Yee A, Edwards A, Vincentelli R, Cambillau C, Kim R, Kim SH, Rao Z, Shi Y, Terwilliger TC, Kim CY, Hung LW, Waldo GS, Peleg Y, Albeck S, Unger T, Dym O, Prilusky J, Sussman JL, Stevens RC, Lesley SA, Wilson IA, Joachimiak A, Collart F, Dementieva I, Donnelly MI, Eschenfeldt WH, Kim Y, Stols L, Wu R, Zhou M, Burley SK, Emtage JS, Sauder JM, Thompson D, Bain K, Luz J, Gheyi T, Zhang F, Atwell S, Almo SC, Bonanno JB, Fiser A, Swaminathan S, Studier FW, Chance MR, Sali A, Acton TB, Xiao R, Zhao L, Ma LC, Hunt JF, Tong L, Cunningham K, Inouye M, Anderson S, Janjua H, Shastry R, Ho CK, Wang D, Wang H, Jiang M, Montelione GT, Stuart DI, Owens RJ, Daenke S, Schutz A, Heinemann U, Yokoyama S, Bussow K, Gunsalus KC (2008) Protein production and purification. Nat Methods 5(2):135–146

Greene LH, Lewis TE, Addou S, Cuff A, Dallman T, Dibley M, Redfern O, Pearl F, Nambudiry R, Reid A, Sillitoe I, Yeats C, Thornton JM, Orengo CA (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35 (Database issue):D291–D297

Greer J (1981) Comparative model-building of the mammalian serine proteases. J Mol Biol 153 (4):1027–1042

Greer J (1990) Comparative modeling methods: application to the family of the mammalian serine proteases. Proteins 7(4):317–334

Guenther B, Onrust R, Sali A, O'Donnell M, Kuriyan J (1997) Crystal structure of the ë-subunit of the clamp-loader complex of *E. coli* DNA polymerase III. Cell 91(3):335–345

Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, Bordoli L, Schwede T (2013) The protein model portal—a comprehensive resource for protein structure and model information. Database: J Biol Databases Curation 2013:bat031

Han R, Leo-Macias A, Zerbino D, Bastolla U, Contreras-Moreira B, Ortiz AR (2008) An efficient conformational sampling method for homology modeling. Proteins 71(1):175–188

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50(1):43–57

Havel TF, Snow ME (1991) A new method for building protein conformations from sequence alignments with homologues of known structure. J Mol Biol 217(1):1–7

Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22):10915–10919

Henikoff JG, Pietrokovski S, McCallum CM, Henikoff S (2000) Blocks-based methods for detecting protein homology. Electrophoresis 21(9):1700–1706

Hlavin ML, Lemmon V (1991) Molecular structure and functional testing of human L1CAM: an interspecies comparison. Genomics 11(2):416–423

Holm L, Sander C (1991) Database algorithm for generating protein backbone and side-chain co-ordinates from a C alpha trace application to model building and detection of co-ordinate errors. J Mol Biol 218(1):183–194

Hooft RW, Vriend G, Sander C, Abola EE (1996) Errors in protein structures. Nature 381 (6580):272

Hung LH, Samudrala R (2003) Accurate and automated classification of protein secondary structure with PsiCSI. Protein Sci 12(2):288–295

Illergard K, Ardell DH, Elofsson A (2009) Structure is three to ten times more conserved than sequence—a study of structural response in protein cores. Proteins 77(3):499–508

Jacobson MP, Pincus DL, Rapp CS, Day TJ, Honig B, Shaw DE, Friesner RA (2004) A hierarchical approach to all-atom protein loop prediction. Proteins 55(2):351–367

Jaroszewski L, Rychlewski L, Zhang B, Godzik A (1998) Fold prediction by a hierarchy of sequence, threading, and modeling methods. Protein Sci 7(6):1431–1440

Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) FFAS03: a server for profile–profile sequence alignments. Nucleic Acids Res 33 (Web Server issue):W284–W288

Jennings AJ, Edge CM, Sternberg MJ (2001) An approach to improving multiple alignments of protein sequences using predicted secondary structure. Protein Eng 14(4):227–231

John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model building and model assessment. Nucleic Acids Res 31(14):3982–3992

John B, Sali A (2004) Detection of homologous proteins by an intermediate sequence search. Protein Sci 13(1):54–62

Johnson LN, Lowe ED, Noble ME, Owen DJ (1998) The Eleventh Datta Lecture. The structural basis for substrate recognition and control by protein kinases. FEBS Lett 430(1–2):1–11

Jones TA, Thirup S (1986) Using known substructures in protein model building and crystallography. EMBO J 5(4):819–822

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358(6381):86–89

Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51 (4):504–514

Karplus K, Barrett C, Hughey R (1998) Hidden Markov models for detecting remote protein homologies. Bioinformatics 14(10):846–856

Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R (2005) SAM-T04: what is new in protein-structure prediction for CASP6. Proteins 61 (Suppl 7):135–142

Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative (vol 111, pg 3733, 2014). Proc Natl Acad Sci U S A 111(13):5060

Kihara D, Skolnick J (2003) The PDB is a covering set of small protein structures. J Mol Biol 334 (4):793–802

Kiselar JG, Janmey PA, Almo SC, Chance MR (2003) Structural analysis of gelsolin using synchrotron protein footprinting. Mol Cell Proteomics 2(10):1120–1132

Koehl P, Delarue M (1995) A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. Nat Struct Biol 2(2):163–170

Kohlhoff KJ, Robustelli P, Cavalli A, Salvatella X, Vendruscolo M (2009) Fast and accurate predictions of protein NMR chemical shifts from interatomic distances. J Am Chem Soc 131 (39):13894–13895

Kolinski A, Bujnicki JM (2005) Generalized protein structure prediction based on combination of fold-recognition with de novo folding and evaluation of models. Proteins 61(Suppl 7):84–90

Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 44(2):133–149

Kopp J, Schwede T (2006) The SWISS-MODEL repository: new features and functionalities. Nucleic Acids Res 34 (Database issue):D315–D318

Kopp J, Bordoli L, Battey JN, Kiefer F, Schwede T (2007) Assessment of CASP7 predictions for template-based modeling targets. Proteins 69(Suppl 8):38–56
Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 235(5):1501–1531
Kryshtafovych A, Fidelis K, Moult J (2014) CASP10 results compared to those of previous CASP experiments. Proteins 82(Suppl 2):164–174
Lange OF, Rossi P, Sgourakis NG, Song YF, Lee HW, Aramini JM, Ertekin A, Xiao R, Acton TB, Montelione GT, Baker D (2012) Determination of solution structures of proteins up to 40 kDa using CS-Rosetta with sparse NMR data from deuterated samples. Proc Natl Acad Sci U S A 109(27):10873–10878
Laskowski RA, Moss DS, Thornton JM (1993) Main-chain bond lengths and bond angles in protein structures. J Mol Biol 231(4):1049–1067
Lee J, Lee J, Sasaki TN, Sasai M, Seok C, Lee J (2011) De novo protein structure prediction by dynamic fragment assembly and conformational space annealing. Proteins 79(8):2403–2417
Lesk AM (1995) NAD-binding domains of dehydrogenases. Curr Opin Struct Biol 5(6):775–783
Lesk AM, Chothia C (1980) How different amino acid sequences determine similar protein structures: the structure and evolutionary dynamics of the globins. J Mol Biol 136(3):225–270
Levitt M (1992) Accurate modeling of protein conformation by automatic segment matching. J Mol Biol 226(2):507–533
Levitt M (2009) Nature of the protein universe. Proc Natl Acad Sci U S A 106(27):11079–11084
Lippow SM, Wittrup KD, Tidor B (2007) Computational design of antibody-affinity improvement beyond in vivo maturation. Nat Biotechnol 25(10):1171–1176
Luthy R, McLachlan AD, Eisenberg D (1991) Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239
Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766
Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325
Marti-Renom MA, Madhusudhan MS, Fiser A, Rost B, Sali A (2002) Reliability of assessment of protein structure prediction methods. Structure 10(3):435–440
Marti-Renom MA, Madhusudhan MS, Sali A (2004) Alignment of protein sequences by their profiles. Protein Sci 13(4):1071–1087
Meiler J (2003) PROSHIFT: protein chemical shift prediction using artificial neural networks. J Biomol NMR 26(1):25–37
Melo F, Feytmans E (1997) Novel knowledge-based mean force potential at atomic level. J Mol Biol 267(1):207–222
Menon V, Vallat BK, Dybas JM, Fiser A (2013) Modeling proteins using a super-secondary structure library and NMR chemical shift information. Structure 21(6):891–899
Mezei M (1998) Chameleon sequences in the PDB. Protein Eng 11(6):411–414
Michalsky E, Goede A, Preissner R (2003) Loops in proteins (LIP)—a comprehensive loop database for homology modelling. Protein Eng 16(12):979–985
Mirjalili V, Noyes K, Feig M (2014) Physics-based protein structure refinement through multiple molecular dynamics trajectories and structure averaging. Proteins 82(Suppl 2):196–207
Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci U S A 108(49):E1293–E1301
Moretti S, Armougom F, Wallace IM, Higgins DG, Jongeneel CV, Notredame C (2007) The M-Coffee web server: a meta-method for computing multiple sequence alignments by combining alternative alignment methods. Nucleic Acids Res 35 (Web Server issue):W645–W648
Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15(3):285–289

Moult J, James MN (1986) An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. Proteins 1(2):146–163
Nair R, Liu J, Soong TT, Acton TB, Everett JK, Kouranov A, Fiser A, Godzik A, Jaroszewski L, Orengo C, Montelione GT, Rost B (2009) Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genomics 10(2):181–191
Norin M, Sundstrom M (2001) Protein models in drug discovery. Curr Opin Drug Discov Devel 4 (3):284–290
Notredame C (2007) Recent evolutions of multiple sequence alignment algorithms. PLoS Comput Biol 3(8):e123
Ohlendorf DH (1994) Accuracy of refined protein structures. Comparison of four independently refined models of human interleukin 1 beta. Acta Crystallogr D Biol Crystallogr D50:808–812
Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ (1997) An automated classification of the structure of protein loops. J Mol Biol 266(4):814–830
Orr GA, Rao S, Swindell CS, Kingston DG, Horwitz SB (1998) Photoaffinity labeling approach to map the Taxol-binding site on the microtubule. Methods Enzymol 298:238–252
Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. Methods Mol Biol 132:185–219
Pei J, Grishin NV (2007) PROMALS: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23(7):802–808
Pei J, Kim BH, Grishin NV (2008) PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36(7):2295–2300
Peng HP, Yang AS (2007) Modeling protein loops with knowledge-based prediction of sequence-structure alignment. Bioinformatics 23(21):2836–2842
Petrey D, Honig B (2005) Protein structure prediction: inroads to biology. Mol Cell 20(6):811–819
Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A, Koh IY, Alexov E, Honig B (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. Proteins 53(Suppl 6):430
Piana S, Lindorff-Larsen K, Shaw DE (2012) Protein folding kinetics and thermodynamics from atomistic simulation. Proc Natl Acad Sci U S A 109(44):17845–17850
Piana S, Klepeis JL, Shaw DE (2014) Assessing the accuracy of physical models used in protein-folding simulations: quantitative evidence from long molecular dynamics simulations. Curr Opin Struct Biol 24:98–105
Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, Marti-Renom M, Karchin R, Webb BM, Eramian D, Shen MY, Kelly L, Melo F, Sali A (2006) MODBASE: a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 34 (Database issue):D291–D295
Pieper U, Webb BM, Dong GQ, Schneidman-Duhovny D, Fan H, Kim SJ, Khuri N, Spill YG, Weinkam P, Hammel M, Tainer JA, Nilges M, Sali A (2014) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42 (Database issue):D336–D346
Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kazmierkiewicz R, Oldziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA (2001) Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci U S A 98(5):2329–2333
Presta LG, Chen H, O'Connor SJ, Chisholm V, Meng YG, Krummen L, Winkler M, Ferrara N (1997) Humanization of an anti-vascular endothelial growth factor monoclonal antibody for the therapy of solid tumors and other disorders. Cancer Res 57(20):4593–4599
Pujato M, Kieken F, Skiles AA, Tapinos N, Fiser A (2014) Prediction of DNA binding motifs from 3D models of transcription factors; identifying TLX3 regulated genes. Nucleic Acids Res 42(22):13500–13512
Qian B, Ortiz AR, Baker D (2004) Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. Proc Natl Acad Sci U S A 101(43):15346–15351

Queen C, Schneider WP, Selick HE, Payne PW, Landolfi NF, Duncan JF, Avdalovic NM, Levitt M, Junghans RP, Waldmann TA (1989) A humanized antibody that binds to the interleukin 2 receptor. Proc Natl Acad Sci U S A 86(24):10029–10033

Rai BK, Fiser A (2006) Multiple mapping method: a novel approach to the sequence-to-structure alignment problem in comparative protein structure modeling. Proteins 63(3):644–661

Rai BK, Madrid-Aliste CJ, Fajardo JE, Fiser A (2006) MMM: a sequence-to-structure alignment protocol. Bioinformatics 22(21):2691–2692

Raman S, Huang YJP, Mao BC, Rossi P, Aramini JM, Liu GH, Montelione GT, Baker D (2010a) Accurate automated protein NMR structure determination using unassigned NOESY data. J Am Chem Soc 132(1):202–207

Raman S, Lange OF, Rossi P, Tyka M, Wang X, Aramini J, Liu G, Ramelot TA, Eletsky A, Szyperski T, Kennedy MA, Prestegard J, Montelione GT, Baker D (2010b) NMR structure determination for larger proteins using backbone-only data. Science 327(5968):1014–1018

Reddy BV, Li WW, Shindyalov IN, Bourne PE (2001) Conserved key amino acid positions (CKAAPs) derived from the analysis of common substructures in proteins. Proteins 42 (2):148–163

Remmert M, Biegert A, Hauser A, Soding J (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175

Ring CS, Cohen FE (1993) Modeling protein structures: construction and their applications. FASEB J 7(9):783–790

Ring CS, Sun E, McKerrow JH, Lee GK, Rosenthal PJ, Kuntz ID, Cohen FE (1993) Structure-based inhibitor design by using protein models for the development of antiparasitic agents. Proc Natl Acad Sci U S A 90(8):3583–3587

Robustelli P, Cavalli A, Dobson CM, Vendruscolo M, Salvatella X (2009) Folding of small proteins by Monte Carlo simulations with chemical shift restraints without the use of molecular fragment replacement or structural homology. J Phys Chem B 113(22):7890–7896

Robustelli P, Kohlhoff K, Cavalli A, Vendruscolo M (2010) Using NMR chemical shifts as structural restraints in molecular dynamics simulations of proteins. Structure 18(8):923–933

Rodrigues JP, Levitt M, Chopra G (2012) KoBaMIN: a knowledge-based minimization web server for protein structure refinement. Nucleic Acids Res 40 (Web Server issue):W323–W328

Rohl CA, Baker D (2002) De novo determination of protein backbone structure from residual dipolar couplings using Rosetta. J Am Chem Soc 124(11):2723–2729

Rohl CA, Strauss CE, Chivian D, Baker D (2004a) Modeling structurally variable regions in homologous proteins with rosetta. Proteins 55(3):656–677

Rohl CA, Strauss CE, Misura KM, Baker D (2004b) Protein structure prediction using Rosetta. Methods Enzymol 383:66–93

Rost B (1997) Protein structures sustain evolutionary drift. Fold Des 2(3):S19–S24

Rost B (1999) Twilight zone of protein sequence alignments. Protein Eng 12(2):85–94

Rubinstein R, Ramagopal UA, Nathenson SG, Almo SC, Fiser A (2013) Functional classification of immune regulatory proteins. Structure 21(5):766–776

Rudolph MG, Stanfield RL, Wilson IA (2006) How TCRs bind MHCs, peptides, and coreceptors. Annu Rev Immunol 24:419–466

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcon LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol 5 (3):e77

Rychlewski L, Jaroszewski L, Li W, Godzik A (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. Protein Sci 9(2):232–241

Rykunov D, Fiser A (2007) Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. Proteins 67(3):559–568

Rykunov D, Fiser A (2010) New statistical potential for quality assessment of protein models and a survey of energy functions. BMC Bioinform 11(1):128

Rykunov D, Steinberger E, Madrid-Aliste CJ, Fiser A (2009) Improved scoring function for comparative modeling using the M4T method. J Struct Funct Genomics 10(1):95–99

Sali A, Blundell TL (1993) Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 234(3):779–815

Sali A, Matsumoto R, McNeil HP, Karplus M, Stevens RL (1993) Three-dimensional models of four mouse mast cell chymases. Identification of proteoglycan binding regions and protease-specific antigenic epitopes. J Biol Chem 268(12):9023–9034

Sali A, Shakhnovich E, Karplus M (1994) How does a protein fold? Nature 369(6477):248–251

Samudrala R, Moult J (1998) A graph-theoretic algorithm for comparative modeling of protein structure. J Mol Biol 279(1):287–302

Sanchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1:50–58

Sanchez R, Sali A (1998) Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. Proc Natl Acad Sci U S A 95(23):13597–13602

Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinform 8:294

Saraste M, Sibbald PR, Wittinghofer A (1990) The P-loop–a common motif in ATP- and GTP-binding proteins. Trends Biochem Sci 15(11):430–434

Sauder JM, Arthur JW, Dunbrack RL Jr (2000) Large-scale comparison of protein sequence alignment algorithms with structure alignments. Proteins 40(1):6–22

Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29(14):2994–3005

Schwarzenbacher R, Godzik A, Jaroszewski L (2008) The JCSG MR pipeline: optimized alignments, multiple models and parallel searches. Acta Crystallogr D Biol Crystallogr 64(Pt 1):133–140

Schwede T, Sali A, Honig B, Levitt M, Berman HM, Jones D, Brenner SE, Burley SK, Das R, Dokholyan NV, Dunbrack RL Jr, Fidelis K, Fiser A, Godzik A, Huang YJ, Humblet C, Jacobson MP, Joachimiak A, Krystek SR Jr, Kortemme T, Kryshtafovych A, Montelione GT, Moult J, Murray D, Sanchez R, Sosnick TR, Standley DM, Stouch T, Vajda S, Vasquez M, Westbrook JD, Wilson IA (2009) Outcome of a workshop on applications of protein models in biomedical research. Structure 17(2):151–159

Service R (2005) Structural biology. Structural genomics, round 2. Science 307(5715):1554–1558

Shaw DE, Maragakis P, Lindorff-Larsen K, Piana S, Dror RO, Eastwood MP, Bank JA, Jumper JM, Salmon JK, Shan Y, Wriggers W (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330(6002):341–346

Shen Y, Bax A (2010) SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network. J Biomol NMR 48 (1):13–22

Shen Y, Bax A (2012) Identification of helix capping and beta-turn motifs from NMR chemical shifts. J Biomol NMR 52: 211–232

Shen Y, Lange O, Delaglio F, Rossi P, Aramini JM, Liu G, Eletsky A, Wu Y, Singarapu KK, Lemak A, Ignatchenko A, Arrowsmith CH, Szyperski T, Montelione GT, Baker D, Bax A (2008) Consistent blind protein structure generation from NMR chemical shift data. Proc Natl Acad Sci U S A 105(12):4685–4690

Shen Y, Delaglio F, Cornilescu G, Bax A (2009a) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4):213–223

Shen Y, Vernon R, Baker D, Bax A (2009b) De novo protein structure generation from incomplete chemical shift assignments. J Biomol NMR 43(2):63–78

Sheng Y, Sali A, Herzog H, Lahnstein J, Krilis SA (1996) Site-directed mutagenesis of recombinant human beta 2-glycoprotein I identifies a cluster of lysine residues that are critical for phospholipid binding and anti-cardiolipin antibody activity. J Immunol 157(8):3744–3751

Shenkin PS, Yarmush DL, Fine RM, Wang HJ, Levinthal C (1987) Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. Biopolymers 26(12):2053–2085

Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310 (1):243–257

Shin DS, Zhao RB, Yap EH, Fiser A, Goldman ID (2012) A P425R mutation of the proton-coupled folate transporter causing hereditary folate malabsorption produces a highly selective alteration in folate binding. Am J Physiol-Cell Ph 302(9):C1405–C1412

Sibanda BL, Blundell TL, Thornton JM (1989) Conformation of beta-hairpins in protein structures. A systematic classification with applications to modelling by homology, electron density fitting and protein engineering. J Mol Biol 206(4):759–777

Sippl MJ (1990) Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 213(4):859–883

Sippl MJ (1993) Recognition of errors in three-dimensional structures of proteins. Proteins 17 (4):355–362

Sippl MJ (1995) Knowledge-based potentials for proteins. Curr Opin Struct Biol 5(2):229–235

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951–960

Soto CS, Fasnacht M, Zhu J, Forrest L, Honig B (2008) Loop modeling: sampling, filtering, and scoring. Proteins 70(3):834–843

Srinivasan N, Blundell TL (1993) An evaluation of the performance of an automated procedure for comparative modelling of protein tertiary structure. Protein Eng 6(5):501–512

Stein A, Ceol A, Aloy P (2011) 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res 39 (Database issue):D718–D723

Summa CM, Levitt M (2007) Near-native structure refinement using in vacuo energy minimization. Proc Natl Acad Sci U S A 104(9):3177–3182

Sutcliffe MJ, Haneef I, Carney D, Blundell TL (1987) Knowledge based modelling of homologous proteins, part I: three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1(5):377–384

Sutcliffe MJ, Dobson CM, Oswald RE (1992) Solution structure of neuronal bungarotoxin determined by two-dimensional NMR spectroscopy: calculation of tertiary structure using systematic homologous model building, dynamical simulated annealing, and restrained molecular dynamics. Biochemistry 31(11):2962–2970

Tai CH, Bai H, Taylor TJ, Lee B (2014) Assessment of template-free modeling in CASP10 and ROLL. Proteins 82(Suppl 2):57–83

Tainer JA, Thayer MM, Cunningham RP (1995) DNA repair proteins. Curr Opin Struct Biol 5 (1):20–26

Tang K, Zhang J, Liang J (2014) Fast protein loop sampling and structure prediction using distance-guided sequential chain-growth Monte Carlo method. PLoS Comput Biol 10(4): e1003539

Taylor WR, Hatrick K (1994) Compensating changes in protein multiple sequence alignments. Protein Eng 7(3):341–348

Terashi G, Takeda-Shitaka M, Kanou K, Iwadate M, Takaya D, Hosoi A, Ohta K, Umeyama H (2007) Fams-ace: a combined method to select the best model after remodeling all server models. Proteins 69(Suppl 8):98–107

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143

Todd AE, Orengo CA, Thornton JM (2002) Plasticity of enzyme active sites. Trends Biochem Sci 27(8):419–426

Topf M, Lasker K, Webb B, Wolfson H, Chiu W, Sali A (2008) Protein structure fitting and refinement guided by cryo-EM density. Structure 16(2):295–307

Topham CM, McLeod A, Eisenmenger F, Overington JP, Johnson MS, Blundell TL (1993) Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. J Mol Biol 229(1):194–220

Unger R, Harel D, Wherland S, Sussman JL (1989) A 3D building blocks approach to analyzing and predicting structure of proteins. Proteins 5(4):355–373

Vakser IA (1995) Protein docking for low-resolution structures. Protein Eng 8(4):371–377

van Gelder CW, Leusen FJ, Leunissen JA, Noordik JH (1994) A molecular dynamics approach for the generation of complete protein structures from limited coordinate data. Proteins 18 (2):174–185

van Vlijmen HW, Karplus M (1997) PDB-based protein loop prediction: parameters for selection and methods for optimization. J Mol Biol 267(4):975–1001

Venclovas C, Margelevicius M (2005) Comparative modeling in CASP6 using consensus approach to template selection, sequence-structure alignment, and structure assessment. Proteins 61(Suppl 7):99–105

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304 (5667):66–74

Vernal J, Fiser A, Sali A, Muller M, Jose CJ, Nowicki C (2002) Probing the specificity of a trypanosomal aromatic alpha-hydroxy acid dehydrogenase by site-directed mutagenesis. Biochem Biophys Res Commun 293(1):633–639

Vitkup D, Melamud E, Moult J, Sander C (2001) Completeness in structural genomics. Nat Struct Biol 8(6):559–566

Wallner B, Elofsson A (2005a) All are not equal: a benchmark of different homology modeling programs. Protein Sci 14(5):1315–1327

Wallner B, Elofsson A (2005b) Pcons5: combining consensus, structural evaluation and fold recognition scores. Bioinformatics 21(23):4248–4254

Wallner B, Elofsson A (2007) Prediction of global and local model quality in CASP7 using Pcons and ProQ. Proteins 69(Suppl 8):184–193

Wallner B, Larsson P, Elofsson A (2007) Pcons.net: protein structure prediction meta server. Nucleic Acids Res 35 (Web Server issue):W369–W374

Wishart DS, Sykes BD (1994) The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data. J Biomol NMR 4 (2):171–180

Wishart DS, Arndt D, Berjanskii M, Tang P, Zhou J, Lin G (2008) CS23D: a web server for rapid protein structure generation using NMR chemical shifts and sequence data. Nucleic Acids Res 36 (Web Server issue):W496–W502

Wlodawer A (2002) Rational approach to AIDS drug design through structural biology. Annu Rev Med 53:595–614

Wlodawer A, Miller M, Jaskolski M, Sathyanarayana BK, Baldwin E, Weber IT, Selk LM, Clawson L, Schneider J, Kent SBH (1989) Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. Science 245(4918):616–621

Wu G, Fiser A, ter Kuile B, Sali A, Muller M (1999) Convergent evolution of Trichomonas vaginalis lactate dehydrogenase from malate dehydrogenase. Proc Natl Acad Sci U S A 96 (11):6285–6290

Wu G, McArthur AG, Fiser A, Sali A, Sogin ML, Mllerm M (2000) Core histones of the amitochondriate protist, Giardia lamblia. Mol Biol Evol 17(8):1156–1163

Xiang Z, Soto CS, Honing B (2002) Evaluating conformational free energies: the colony energy and its application to the poblem of loop prediction. Proc Natl Acad Sci U S A 99:7432–7437

Xiao H, Verdier-Pinard P, Fernandez-Fuentes N, Burd B, Angeletti R, Fiser A, Horwitz SB, Orr GA (2006) Insights into the mechanism of microtubule stabilization by Taxol. Proc Natl Acad Sci U S A 103(27):10166–10173

Xu LZ, Sanchez R, Sali A, Heintz N (1996) Ligand specificity of brain lipid-binding protein. J Biol Chem 271(40):24711–24719

Xu J, Jiao F, Yu L (2007) Protein structure prediction using threading. Methods Mol Biol 413:91–122

Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y (2015) The I-TASSER Suite: protein structure and function prediction. Nat Methods 12(1):7–8

Yap EH, Rosche T, Almo S, Fiser A (2014) Functional clustering of immunoglobulin superfamily proteins with protein-protein interaction information calibrated hidden Markov Model sequence profiles. J Mol Biol 426(4):945–961

Yooseph S, Sutton G, Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol 5(3):e16

Zemla A (2003) LGA: A method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

Zhang Y (2007) Template-based modeling and free modeling by I-TASSER in CASP7. Proteins 69(Suppl 8):108–117

Zhang Y, Skolnick J (2005) The protein structure prediction problem could be solved using the current PDB library. Proc Natl Acad Sci U S A 102(4):1029–1034

Zhang C, Liu S, Zhou Y (2004) Accurate and efficient loop selections by the DFIRE-based all-atom statistical potential. Protein Sci 13(2):391–399

Zhang Y, Thiele I, Weekes D, Li ZW, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A (2009) Three-dimensional structural view of the central metabolic network of thermotoga maritima. Science 325(5947):1544–1549

Zheng Q, Rosenfeld R, Vajda S, DeLisi C (1993) Determining protein loop conformation using scaling-relaxation techniques. Protein Sci 2(8):1242–1248

Zhou H, Skolnick J (2007) Ab initio protein structure prediction using chunk-TASSER. Biophys J 93(5):1510–1518

Zhou H, Skolnick J (2011) GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophys J 101 (8):2043–2052

# Chapter 5
# Advances in Computational Methods for Transmembrane Protein Structure Prediction

**Tim Nugent, David Jones and Sikander Hayat**

**Abstract** Transmembrane (TM) proteins fulfill many crucial cellular functions such as substrate transport, biogenesis and signalling, and make up a significant fraction of any given proteome. Estimates suggest that up to 30% of all human genes may encode α-helical TM proteins, while β-barrel TM proteins, which are found in the outer-membrane of gram-negative bacteria, mitochondria and chloroplast, are encoded by 2–3% of genes. However, relatively few high resolution TM protein structures are known, making it all the more important to extract as much structural information as possible from amino acid sequences. In this chapter, we review the existing methods for the identification, topology prediction and three-dimensional modelling of TM proteins, including a discussion of the recent advances in identifying residue-residue contacts from large multiple sequence alignments that have enabled impressive gains to be made in the field of TM protein structure prediction.

**Keywords** Transmembrane proteins · Structure prediction · 3D modelling

T. Nugent (✉)
Thomson Reuters, Corporate Research and Development, 30 South Colonnade, Canary Wharf, EC2A 4EG London, UK
e-mail: tim.nugent@thomsonreuters.com

D. Jones
Bioinformatics Group, Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
e-mail: d.jones@cs.ucl.ac.uk

S. Hayat
Computational Biology Program, Memorial Sloan Kettering Cancer Center, New York City, USA
e-mail: hayats@mskcc.org



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_5

## 5.1 Introduction

Transmembrane (TM) proteins are involved in a wide range of essential biological processes including cell signalling, transport of membrane-impermeable molecules, cell-cell communication, cell recognition, cell adhesion and biogenesis of the bacterial outer membrane. Many are also prime drug targets, with approximately 60% of all drugs currently on the market targeting membrane proteins (Hopkins and Groom 2002). Despite recent progress in TM protein structure determination, the experimental difficulties associated with obtaining crystals that diffract to high resolution mean that TM protein are severely under-represented in structural databases, making up only 1% of known structures in the PDB (White 2004) of which only about 500 are unique. TM proteins, which have both hydrophobic and hydrophilic regions on their surfaces, are much more difficult to isolate than water-soluble proteins as the native membrane surrounding the protein must be disrupted and replaced with detergent molecules without causing any denaturation. Given the biological and pharmacological importance of TM proteins, an understanding of their structure and topology—the total number of TM helices, their boundaries and in/out orientation relative to the membrane—is essential for functional analysis and directing further experimental work. In the absence of vital structural data, bioinformatics strategies thus turn to sequence-based prediction methods.

## 5.2 Membrane Protein Structural Classes

TM proteins can be classified into two basic types: α-helical and β-barrel proteins. α-helical membrane proteins form the major category of TM proteins and are present in all type of biological membranes, including bacterial outer membranes. They consist of one or more α-helices, each of which contains a stretch of hydrophobic amino acids, embedded in the membrane and linked to subsequent helices by extra-membranous loop regions. It is thought such proteins may have up to 20 TM helices allowing a diverse range of differing topologies. Loop regions are known to contain substructures including re-entrant loops—short α-helices that enter and exit the membrane on the same side—as well as amphipathic helices that lie parallel to the membrane plane, and globular domains. β-barrel TM proteins (TMBs) mainly consist of transmembrane β-strands that form a closed barrel in the membrane. Analysis of solved β-barrel 3D structures show that these proteins can consist of 8–26 β-strands arranged in an anti-parallel manner in the bacterial outer-membrane. Some TMBs also have large plug-domains and outer loops that can interact with the barrel region to control substrate transport.

### 5.2.1 α-Helical Bundles

α-helical TM proteins can be further divided into a number of subtypes based on their topology. Type I and II membrane proteins consist of a single TM α helix, type III have multiple membrane-spanning helices while type IV membrane proteins have multiple domains which form an assembly that spans the membrane multiple times. Type I membrane proteins are attached to the membrane with an anchor sequence targeting their amino terminus to the endoplasmic reticulum lumen and the carboxy terminus exposed to the cytoplasmic side. These proteins are further divided into two subtypes. Type Ia—which constitutes most eukaryotic membrane proteins—contain cleavable signal sequences, while type Ib do not. Type II membrane proteins are similar to type I in that they span the membrane only once but their orientation is reversed; they have their amino terminus on the cytoplasmic side of the cell and the carboxy terminus on the exterior. Type III membrane proteins, which include G protein coupled receptors (e.g. PDB code 1gzm) consist of multiple TM helices and are also divided into two subtypes. Type IIIa have cleavable signal sequence while type IIIb do not, but do have their amino terminus exposed to the extracellular side of the membrane. Type IV membrane proteins have multiple domains which form an assembly that spans the membrane multiple times. Domains may reside on a single polypeptide chain but are often composed of more than one. Examples include Photosystem I, which comprises nine unique chains (1jb0).

### 5.2.2 Transmembrane β-Barrels

TMBs can be divided into two main categories depending on whether the barrel pore is formed from a single-chain, or via a homo-oligomeric complex, with each chain contributing 2–4 strands. All known bacterial transmembrane β-barrels consist of anti-parallel β-strands that traverse the outer-membrane in a regular manner (Fig. 5.1). Residues on a transmembrane β-strand follow a strict-dyad repeat such that alternate side-chain face the lipids and barrel pore, respectively. The lipid-facing residues are mostly hydrophobic, but the pore-facing residues can be a mixture of both polar and hydrophobic amino acids. Moreover, transmembrane β-strands generally have fewer residues than transmembrane α-helices and have a less prominent hydrophobic profile. Residues on adjacent β-strands are hydrogen bonded to each other such that alternate residues on strand S1 form a N–O and O–N bond with residues in-register on strand S2, where S1 and S2 are adjacent strands. Solved 3D structures of bacterial TMBs have 8 to 26 β-strands, while the only known Eukaryotic TMB structure - mitochondrial voltage dependent anion channel (VDAC) has 19 strands, where the first and the last strand are parallel to each other. TMBs have long extra-cellular loops that generally protrude away from the barrel pore region but can interact with the barrel domain and short inner loops.

Additionally, a few TMBs have plug domains (Fig. 5.1) that sit inside the barrel and participate in gating and signaling (Ferguson et al. 2002). It is generally estimated that TMBs account for 2–3% of the genes in bacteria, but there is scope for improvement in accurately determining the number of yet unknown TMB families.

Multi-chain TMBs mainly fall into one of four known superfamilies—(a) the pore-forming toxins (PDB codes 3w9t, 3o44, 4h56, 3b07, 7ahl) that are secreted by pathogenic bacteria such as *Staphylococcus aureus*, *Clostridium perfringens* and *Vibrio cholerae*, (b) outer membrane efflux proteins (PDB codes 4mt4, 4mt0, 2xmn, 3pik, 1wp1, 1yc9, 1ek9) that are used by bacteria to expel a wide range of molecules including antibacterial drugs thereby increasing multi-drug resistance, (c) mycobacterial porins (PDB code 1uun) in Mycobacteria that can be used to transport drugs through an otherwise low-permeability outer membrane environment that renders them resistant to many antibiotics, and (d) trimeric autotransporters (PDB codes 2lme, 2gr7) such as the Hia autotransporter of *Haemophilus*

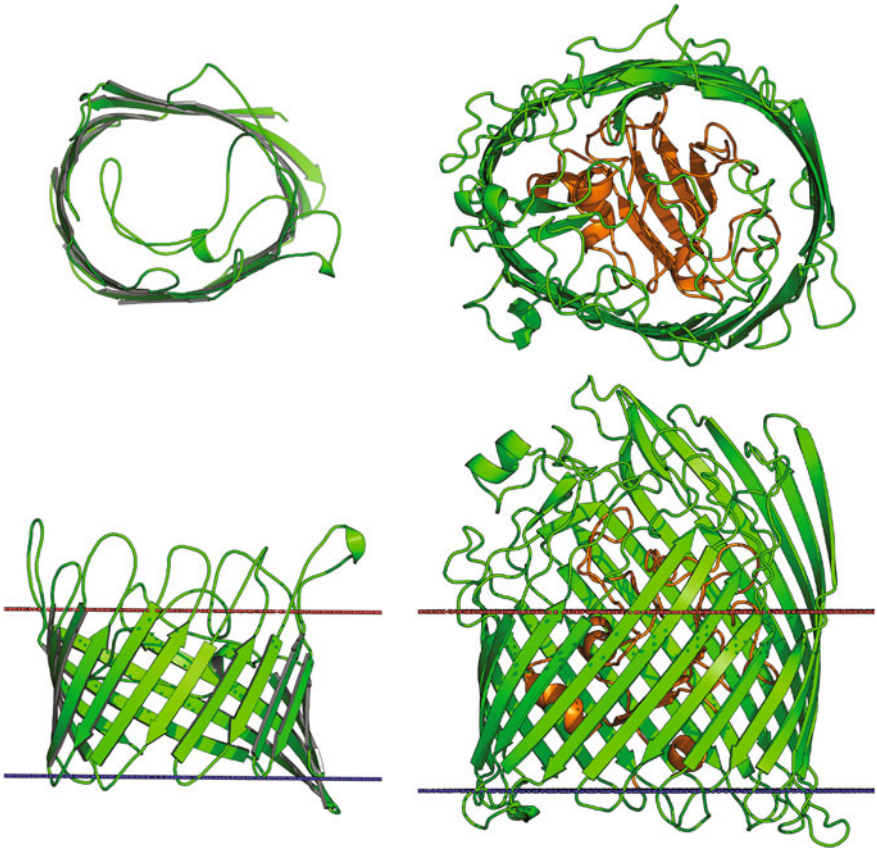


**Fig. 5.1** Top and front views of a diffusion porin (PDB code 3prn) and outer membrane iron transporter FecA (PDB code 1kmp). Both proteins have long outer-loops. The large plug domain of FecA (*orange*) sits in the barrel and facilitates substrate transport and allosteric transitions

*influenzae* that belongs to the largest family of virulence proteins mediating bacterial adhesion, invasion and spread to host cells. Sequence-based analysis methods to identify protein sequences that belong to those families, and therefore estimate the number of multi-chain TMB families, are currently lacking. Additionally, better computational methods for their topology prediction and 3D assembly need to be developed to increase our understanding of their assembly mechanism and function.

## 5.3 Databases

There now exist a number of databases that serve as repositories for the sequences and structures of both α-helical and β-barrel TM proteins (Table 5.1). OPM (Lomize et al. 2006b, 2011), PDBTM (Tusnady et al. 2004, 2005a; Kozma et al. 2013), CGDB (Chetwynd et al. 2008) and the mpstruc database (http://blanco.biomol.uci.edu/mpstruc/) all contain TM proteins of known structure determined using X-ray and electron diffraction, nuclear magnetic resonance and cryo-electron microscopy. OPM, PDBTM and CGDB additionally contain orientation predictions of the protein relative to the membrane based on water-lipid transfer energy minimisation (Lomize et al. 2006a), hydrophobicity/structural feature analysis (Tusnady et al. 2005b) and coarse grained molecular dynamic simulations (Sansom et al. 2008), while MemProtMD (http://sbcb.bioch.ox.ac.uk/memprotmd/) contains orientations calculated using a knowledge-based statistical potential (Nugent and Jones 2013). TOPDB (Tusnady et al. 2008; Dobson et al. 2015a) and MPtopo (Jayasinghe et al. 2001) include topology data that has been experimentally validated using low-resolution techniques such as gene fusion, antibody and mutagenesis studies. Other TM protein databases tend to focus on specific families such as

**Table 5.1** Transmembrane protein databases

| Method | URL | Features |
|---|---|---|
| OPM | http://opm.phar.umich.edu/ | Known structures |
| PDB_TM | http://pdbtm.enzim.hu/ | Known structures |
| CGDB | http://sbcb.bioch.ox.ac.uk/cgdb/ | Coarse grained simulations |
| MemProtMD | http://sbcb.bioch.ox.ac.uk/memprotmd/ | Coarse grained simulations |
| TOPDB | http://topdb.enzim.hu/ | Experimental validation |
| Mptopo | http://blanco.biomol.uci.edu/mptopo/ | Experimental validation |
| VKCDB | http://vkcdb.biology.ualberta.ca/ | Potassium channels |
| KDB | http://sbcb.bioch.ox.ac.uk/kdb/ | Potassium channels |
| TCDB | http://www.tcdb.org/ | Transporters |
| TMBB-DB | http://beta-barrel.tulane.edu/ | Predicted TMBs |
| TMBETA-GENOME | http://tmbeta-genome.cbrc.jp/annotation | Predicted TMBs |
| OMPdb | http://bioinformatics.biol.uoa.gr/OMPdb | Predicted TMB families |
| HHomp | http://toolkit.tuebingen.mpg.de/hhomp | TMB remote homology detection |

voltage-gated potassium channels, including VKCDB (Li and Gallin 2004; Gallin and Boutet 2011) and KDB (http://sbcb.bioch.ox.ac.uk/kdb/), while others such as the Transporter Classification Database (Saier et al. 2006, 2009, 2014) focus on particular structural or functional classes.

For TMBs, TMBB-DB (Freeman and Wimley 2012), TMBETA-GENOME (Gromiha et al. 2007) and OMPdb (Tsirigos et al. 2011) provide an exhaustive list of putative TMBs predicted using computational methods. In addition, HHomp (Remmert et al. 2009) provides a list of putative TMBs found by comprehensive, transitive homology search. As with all bioinformatics databases, care should be taken to ensure that a given resource is frequently updated. The rate at which new sequences and structures are deposited in GenBank and the PDB [and occasionally retracted e.g. (Chang et al. 2006)] results in significant manual annotation for database administrators, and much evidence suggests that this workload often exceeds the amount of time an administrator is willing to commit.

## 5.4 Multiple Sequence Alignments

As with globular proteins, multiple sequence alignments play an important role in TM protein structure prediction. Homologous sequences identified via database searches can be used to construct sequence profiles which can significantly enhance TM topology prediction accuracy (Henricson et al. 2005; Jones 2007), while recent co-evolution-based approaches (Jones et al. 2012, 2015) are dependent on high-quality alignments to infer residue-residue contacts which can be used for de novo modelling (Nugent and Jones 2012).

Conventional pair-wise alignment methods return possible matches based on a scoring function that relies on amino acid substitution matrices such as PAM (Dayhoff and Schwartz 1978) or BLOSUM (Henikoff and Henikoff 1992). Such matrices are derived from globular protein alignments, and as amino acid composition, hydrophobicity and conservation patterns differ between globular and TM proteins (Jones et al. 1994a), they are in principle unsuitable for TM protein alignment. A number of TM-specific substitution matrices have therefore been developed, which take into account such differences. For example, the JTT TM matrix (Jones et al. 1994b) was based on the observation that polar residues in TM proteins are highly conserved, while hydrophobic residues are more interchangeable. Other matrices such as SLIM (Muller et al. 2001), were reported to have the highest accuracy for detecting remote homologues in a manually curated GPCR dataset, while PHAT (Ng et al. 2000) has been shown to outperform JTT, especially on database searching.

More recently, a number of methods have been developed to improve actual TM protein alignment. HMAP (Tang et al. 2003) showed that alignment accuracy could be improved significantly using a profile-profile based approach incorporating structural information. STAM (Shafrir and Guy 2004) implemented higher penalties for insertion/deletions in TM segments compared to loop regions, with combinations of different substitution matrices to produce alignments resulting in more accurate

homology models. PRALINETM (Pirovano et al. 2008), which integrates state-of-the-art sequence prediction techniques with membrane-specific substitution matrices, was shown to outperform standard multiple alignment techniques such as ClustalW (Thompson et al. 1994) and MUSCLE (Edgar 2004) when tested on the TM alignment benchmark set within BAliBASE (Bahr et al. 2001). AlignMe (Stamm et al. 2014, 2013; Khafizov et al. 2010), which uses secondary structure matching combined with evolutionary information, also demonstrated high quality alignments when tested on BAliBASE, although it was noted that accuracy was generally lower when transmembrane topology predictions were also included, although the inclusion of this information may still be useful in cases of extremely distantly related proteins for which sequence information is less informative. PSI-Coffee—a modification of the T-Coffee method (Chang et al. 2012; Notredame et al. 2000)—employs a homology extension technique that can be used to reveal and use specific conservation patterns found within transmembrane proteins, such as amphiphilic α-helices, resulting in significant improvements to the accuracy of alignments. Hill and co-workers constructed substitution tables for different environments within membrane proteins, demonstrating that, in the 10–25% sequence identity range, alignments could be improved by an average of 28 correctly aligned residues compared with alignments made using default substitution tables, leading to improved structural models (Hill and Deane 2013; Hill et al. 2011).

For TMBs, Jimenez-Morales and Liang (2011) have estimated the evolutionary pattern of residue substitutions which can be useful for improved sequence alignment of TMBs, while Yan et al. (2011), have shown the utility of secondary structure element alignment for the identification of putative TMBs. Additionally, a structure based alignment method for TMBs that uses TMB-specific topology features has been shown to improve alignment (Wang et al. 2013).

## 5.5 Transmembrane Protein Topology Prediction

The under-representation of TM proteins in structural databases makes their study extremely difficult. As a result, tools to analyse TM proteins have historically focused on sequence-based topology prediction—identifying the total number of TM helices, their boundaries, and in/out orientation relative to the membrane. Experimental approaches for determining TM topology include glycosylation analysis, insertion tags, antibody studies and fusion protein constructs; however, such studies are time consuming, often conflicting (Mao et al. 2003; Kyttala et al. 2004; Ratajczak et al. 2014), and also risk upsetting the natural topology by altering the protein sequence. Theoretical prediction methods therefore provide an important strategy for furthering our understanding of these biological and pharmacological important proteins.

### 5.5.1 Early α-Helical Topology Prediction Approaches

Early topology predictions methods were based on physicochemical observations of TM proteins. Even before the arrival of the first crystal structures, stretches of hydrophobic residues long enough to span the lipid bilayer were identified as TM spanning α-helices. Prediction methods by Kyte and Doolittle (1982) and Engelman et al. (1986), and later by Wimley and White (1996), relied on experimentally determined hydropathy indices to create a hydropathy plot for a protein. This involved taking a sliding window of 19–21 residues and averaging the score with peaks in the plots (regions of high hydrophobicity) corresponding to the locations of TM helices. With more sequences came the discovery that aromatic Trp and Tyr residues tend to cluster near the ends of the transmembrane segments (Wallin et al. 1997), possibly acting as physical buffers to stabilise TM helices within the lipid bilayer. Later, studies identified the appearance of sequence motifs, such as the GxxxG motif (Senes et al. 2000), within TM helices and also periodic patterns implicated in helix-helix packing and 3D structure (Samatey et al. 1995). However, perhaps the most important realisation was that positively-charged residues tend to cluster on cytoplasmic loop—the 'positive-inside' rule of Gunnar von Heijne (von Heijne 1992). Combined with hydrophobicity-based prediction of TM helices, this led to early topology prediction methods such as TopPred (Claros and von Heijne 1994).

### 5.5.2 Machine Learning Approaches for α-Helical Topology Prediction

Despite their early success, these methods based on hydrophobicity analysis combined with the 'positive-inside' rule have since been superseded by machine learning approaches which offer substantially higher prediction accuracy due to their probabilistic formulation (Table 5.2). Hidden Markov models (HMMs) were among the first supervised learning algorithms to be applied to TM topology prediction, with both TMHMM (Krogh et al. 2001) and HMMTOP (Tusnady and Simon 1998) proving highly successful. TMHMM implemented a cyclic model with seven states for a TM helix, while HMMTOP used HMMs to distinguish between five structural states [helix core, inside loop, outside loop, helix caps (C and N) and globular domains]. These states were connected by transition probabilities before dynamic programming was used to match a sequence against a model with the most probable topology. HMMTOP also allowed constrained predictions to be made, where specific residues could be fixed to a topological location based on experimental data, as did other methods such as HMM-TM (Bagos et al. 2006). Later HMM-based predictors include PRODIV-TMHMM and PolyPhobius, both of which made use of evolutionary information from homologs resulting in substantially increased performance (Viklund and Elofsson 2004; Kall et al. 2005).

**Table 5.2** Topology prediction methods for α-helical transmembrane proteins

| Method | Features | URL |
|---|---|---|
| TMHMM (Krogh et al. 2001) | HMM | http://www.cbs.dtu.dk/services/TMHMM/ |
| HMMTOP (Tusnady and Simon 1998) | HMM | http://www.enzim.hu/hmmtop/ |
| HMM-TM (Bagos et al. 2006) | HMM | http://bioinformatics.biol.uoa.gr/HMM-TM/ |
| PRODIV-TMHMM (Viklund and Elofsson 2004) | HMM + Evolutionary information | https://www.pdc.kth.se/hakanv/prodiv-tmhmm |
| Phobius (Kall et al. 2005) | HMM + Evolutionary information + Signal peptide prediction | http://phobius.sbc.su.se/ |
| OCTOPUS (Viklund and Elofsson 2008) | HMM + NN + Evolutionary information | http://octopus.cbr.su.se/ |
| SPOCTOPUS (Viklund and Elofsson 2008) | HMM + NN + Evolutionary information + Signal peptide prediction | http://octopus.cbr.su.se/ |
| PHDhtm (Rost et al. 1996) | NN | https://www.predictprotein.org/ |
| MEMSAT3 (Jones 2007) | NN + Evolutionary information + Signal peptide prediction | http://bioinf.cs.ucl.ac.uk/psipred/ |
| MEMSAT-SVM (Nugent and Jones 2009) | SVM + Evolutionary information + Signal peptide prediction | http://bioinf.cs.ucl.ac.uk/psipred/ |
| Philius (Reynolds et al. 2008) | Dynamic Bayesian networks | http://noble.gs.washington.edu/proj/philius/ |
| WRF-TMH (Hayat and Khan 2013) | Random forests | http://111.68.99.218/WRF-TMH/ |
| TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009) | Consensus | http://topcons.cbr.su.se/ |
| CCTOP (Dobson et al. 2015b) | Consensus | http://cctop.enzim.ttk.mta.hu/ |

Neural networks (NNs) were employed by early methods including PHDhtm (Rost et al. 1996) and MEMSAT3 (Jones 2007). PHDhtm used multiple sequence alignments to perform a consensus prediction of TM helices by combining two NNs. The first created a 'sequence-to-structure' network, which represented the structural propensity of the central residue in a window. A 'structure-to-structure' network then smoothed these propensities to predict TM helices, before the 'positive-inside' rule was applied to produce an overall topology. MEMSAT3 uses a neural network and dynamic programming in order to predict not only TM helices, but also to score the topology and to identify possible signal peptides.

Additional evolutionary information provided by multiple sequence alignments led to prediction accuracies increasing to as much as 80%. OCTOPUS (Viklund and Elofsson 2008) used a novel combination of hidden Markov models and artificial neural networks to further increase performance.

Later, Support Vector Machines (SVMs) gained in popularity and were successfully applied to TM protein topology prediction (Yuan et al. 2004; Lo et al. 2006, 2008). Particularly using non-linear kernel functions, SVMs are capable of learning complex relationships among the amino acids within a given window with which they are trained, particularly when provided with evolutionary information, and are also more resilient to the problem of over-training compared to other machine learning methods. MEMSAT-SVM (Nugent and Jones 2009), an extension of MEMSAT3, used multiple SVM models to classify sequence into one of four states [TM helix, inside or outside loop, re-entrant helix, or signal peptide] before calculating the most likely topologies using dynamic programming, while a further SVM was used to discriminate between globular and TM proteins. Although multiclass SVMs do exist, their performance is typically poorer than binary SVMs since in many cases no single mathematical function exists to separate all classes of data from one another.

More recently, other machine learning algorithms have been applied to TM helix and topology prediction including dynamic Bayesian networks (Reynolds et al. 2008), random forests (Hayat and Khan 2013), self-organizing maps (Deng 2006) and deep learning (Qi et al. 2012). A selection of machine learning-based predictors can be found in Table 5.2.

### *5.5.3 Signal Peptides and Re-entrant Helices*

One significant challenge faced by topology predictors is the discrimination between TM helices and other highly hydrophobic structural features. These include targeting motifs such as signal peptides and signal anchors, amphipathic helices, and re-entrant helices, membrane penetrating helices that enter and exit the membrane on the same side, common in many ion channel families (Fig. 5.2). The similarity between such features and the hydrophobic profile of a TM helix frequently leads to crossover between the different types of predictions. Should these elements be predicted as TM helices, the ensuing topology prediction is likely to be severely disrupted. Some prediction methods, such as SignalP (Petersen et al. 2011; Bendtsen et al. 2004) and TargetP (Emanuelsson et al. 2007), are effective in identifying signal peptides in TM proteins, and may be used as a pre-filter prior to analysis using a TM topology predictor. Phobius (Kall et al. 2004) used a HMM to successfully address the problem of signal peptides in TM protein topology prediction, while PolyPhobius (Kall et al. 2005) further increased accuracy by including homology information. Other methods such as MEMSAT-SVM, OCTOPUS and SPOCTOPUS (Viklund et al. 2008) have also attempted to incorporate identification of re-entrant regions and signal peptides into TM

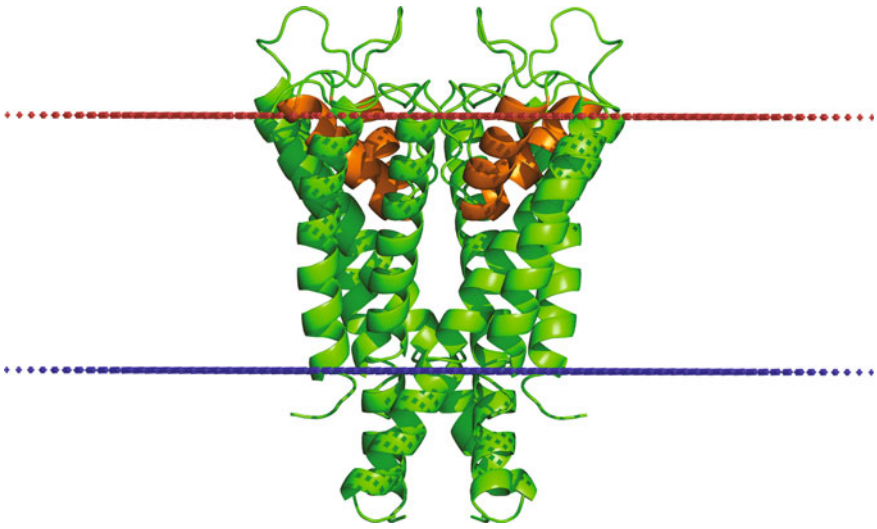


**Fig. 5.2** Potassium channel KcsA (PDB code 1R3J). Each monomer of the homo-tetrameric complex consists of two TM helices and one re-entrant helix (*orange*), which surrounds the central pore and is involved in channel gating

topology prediction but there is significant room for improvement. The problem, particularly regarding re-entrant helices, is the lack of reliable data with which to train machine-learning based methods.

### 5.5.4 *Consensus Approaches for α-Helical Topology Prediction*

While a number of methods successfully combine multiple machine learning approaches, for example ENSEMBLE (Martelli et al. 2003) uses a NN and two HMMs while OCTOPUS uses two sets of four NNs and one HMM, perhaps the best overall methods are those which adopt a consensus approach by combining the results of several predictors to yield more reliable results. Early consensus predictors such as BPROMPT (Taylor et al. 2003) combined the outputs of five different predictors to produce an overall topology using a Bayesian belief network, while Nilsson et al. (2002) used a simple majority-vote approach to return the best topology from their five predictors. The PONGO server (Amico et al. 2006) returns the results of 5 high scoring methods in a graphical format for direct comparison. More recently, MetaTM (Klammer et al. 2009) is based on SVM models and combines the results of six TM topology predictors and two signal peptide predictors. TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009) combines a number of topology predictions into one consensus prediction, while also quantifying the reliability of the prediction based on

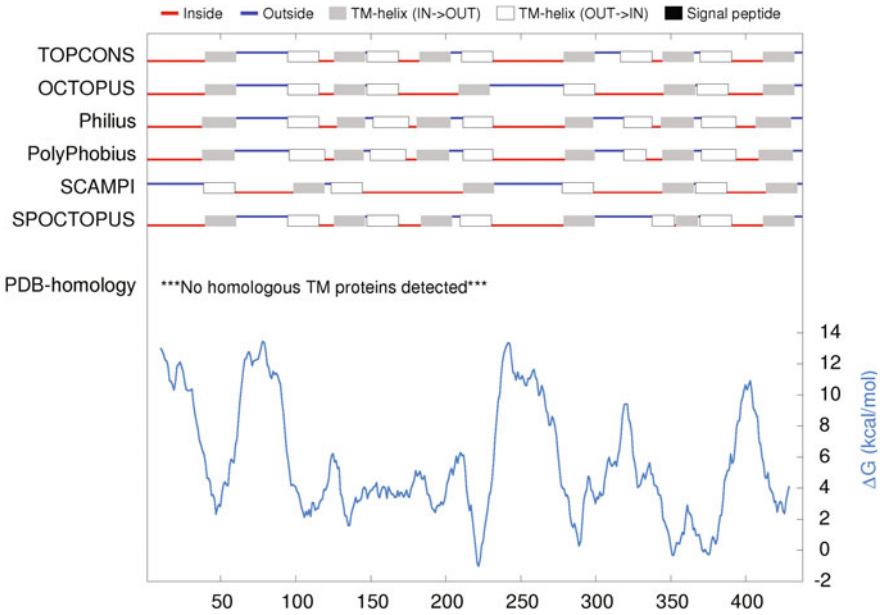



**Fig. 5.3** Consensus topology prediction by TOPCONS (Tsirigos et al. 2015; Bernsel et al. 2009). The results from a number of individual predictors are combined to produce the TOPCONS prediction

the level of agreement between the underlying methods, both at the protein level and at the level of individual TM regions (Fig. 5.3). Results indicate an overall increase in performance by 4% compared to the currently available best-scoring methods. CCTOP (Dobson et al. 2015b) makes use of ten different topology prediction methods, while also incorporating topology information from existing experimental and computational resources such as the PDBTM, TOPDB and TOPDOM databases, using a HMM. In most cases, but particularly proteins whose topology is not straightforward, using a consensus-based method is highly advisable.

### 5.5.5 *Transmembrane β-Barrel Topology Prediction*

Topology prediction of TMBs entails the estimation of the number and the location of TM β-strands. Traditional methods based on a sliding-window hydrophobicity profile are not sufficiently accurate, most likely due to the shorter size and less prominent hydrophobic nature of the TM β-strands. This problem is further complicated by the presence of other β-sheet rich regions in full protein sequences such as the pre-barrel region (seen, for example, in *EstA* Autotransporter protein; PDB code 3kvn) and large plug domains that reside inside the barrel (as seen in *FecA*

**Table 5.3** Computational methods for identifying transmembrane β-barrels

| Method | Features | URL |
|---|---|---|
| boctopus + PSORTb (Imai et al. 2013) | Predicted topology + Subcellular localization | http://boctopus.cbr.su.se/ |
| BETAWARE (Savojardo et al. 2013a) | N-to-1 Extreme Learning Machine | http://betaware.biocomp.unibo.it/BetAware |
| SSEA-OMP (Yan et al. 2011) | Secondary structure element alignment | http://protein.cau.edu.cn/SSEA-OMP/index.html |
| TMB-Hunt (Garrow et al. 2005) | K-nearest neighbor | http://bioinformatics.leeds.ac.uk/betaBarrel/ |
| TMBETA-NET (Gromiha et al. 2005) | Amino acid composition + NN | http://psfs.cbrc.jp/tmbeta-net/ |
| BOMP (Berven et al. 2004) | C-terminal pattern + Integral b-score | http://services.cbu.uib.no/tools/bomp |
| F-W barrel analyzer (Freeman and Wimley 2010) | Empirical Score | http://www.tulane.edu/biochem/WW/apps.html |

protein; PDB code 1fep). Additionally, the absence of long stretches of hydrophobic residues makes it harder to distinguish TM β-strands from β-sheets in globular proteins. One strategy to predict the topology of TMBs relies on first predicting if the query sequence is a TMB or not (Table 5.3) and then using dedicated computational methods to predict the topology of sequences that are predicted to be TMBs. This can potentially improve the accuracy of computational methods that are based on learning from data points available from known 3D structures. Boctopus in combination with PSORTb (Imai et al. 2013), which is a bacterial subcellular localization tool, can be used to identify putative TMBs. The idea here is that proteins for which topology predictor methods predict at least 8 strands with predicted subcellular localization as 'outer-membrane' can be potential TMBs. BETAWARE (Savojardo et al. 2013a) is a machine learning based tool that predicts if a protein is TMB using N-to-1 network encoding and then predicts the topology using a constrained grammar. Other methods employ a combination of secondary structure features, hydrophobicity, amino acid composition and empirical scores to identify putative TMBs. In general, TMB topology prediction methods can be classified as empirical, machine learning and consensus-based. A few of these methods are discussed below (Table 5.4).

### 5.5.6 *Empirical Approaches for β-Barrel Topology Prediction*

Traditionally, features based on knowledge gained from 3D structures, such as the hydrophobicity analyses over a sliding window, amino acid distribution, length of

**Table 5.4** Topology prediction methods for transmembrane β-barrels

| Method | Features | URL |
|---|---|---|
| BETAWARE (Savojardo et al. 2013a) | Conditional Random Fields | http://www.biocomp.unibo.it/ |
| boctopus (Hayat and Elofsson 2012a) | HMM + SVM | http://boctopus.cbr.su.se/ |
| tobmodel (Hayat and Elofsson 2012b) | HMM + SVM | http://tmbmodel.cbr.su.se/ |
| TMBHMM (Singh et al. 2011) | HMM | http://www.zbi.uni-saarland.de/en |
| partiFold (Waldispühl et al. 2008) | http://partifold.csail.mit.edu/ | Inter-strand residue interaction probabilities |
| PROFtmb (Bigelow and Rost 2006) | HMM | https://www.predictprotein.org/ |
| transFold (Waldispühl et al. Waldispühl et al. 2006) | Multi-tape S-attribute grammars | http://bioinformatics.bc.edu/clotelab/transFold/ |
| PRED-TMBB (Bagos et al. 2004) | HMM | http://bioinformatics.biol.uoa.gr/PRED-TMBB/ |
| tbbpred (Natt et al. 2004) | SVM + NN | http://www.imtech.res.in/raghava/tbbpred/ |
| TMBETAPRED-RBF (Ou et al. 2010) | SVM | http://rbf.bioinfo.tw/ |
| TMBETA-NET (Gromiha et al. 2005) | NN | http://psfs.cbrc.jp/tmbeta-net/ |

TM β-strands and outer/inner loops, have been used for the topology prediction of TMBs (Schirmer and Cowan 1993; Gromiha et al. 1997; Gromiha and Ponnuswamy 1993; Diederichs et al. 1998). Wimley et al. (2002) combined features such as hydrophobicity profile, amino acid composition, known variation in the length of inner loops and the abundance of proteins facing the lipids of the barrel pore to formulate a computational score to predict TM stretches and also identify putative TMBs. The distribution of amino acids on a transmembrane β-strand along the membrane normal and the occurrence of the dyad-repeat pattern were employed by Jackups and Liang (2005) to improve the location of predicted strands and estimate the strand-registration such that the maximum number of hydrogen-bonds were satisfied between two adjacent β-strands.

### *5.5.7 Machine Learning Approaches for β-Barrel Topology Prediction*

Machine learning-based methods for the topology prediction of TMBs are typically trained on a dataset of labeled data points extracted from known 3D structures. Rost and Sander (1993) showed early on that the use of information obtained from

multiple sequence alignments yields higher prediction accuracy as compared to using features from a single-sequence alone. SVMs, neural networks and hidden Markov models have all been used for TMB topology prediction (Table 5.4). The use of a sequence profile-based HMM for the identification and topology prediction of TMBs was first introduced by Martelli et al. (2002). PROFtmb (Bigelow and Rost 2006) and PRED-TMBB (Bagos et al. 2004) used a similar approach, where an HMM is used to predict strands, inner-loop and outer-loop states using a sequence profile. The HMM architecture employed in these methods was chosen such that it resembled a pair of strands (up and down), a self-loop representing long outer-loops that connect the two strands on the extracellular side and a self-loop of the inner-membrane side. The number of states representing the β-strand region was chosen to account for the variation in the length of these elements that form TMBs.

Recently, two-stage predictors such as BOCTOPUS (Hayat and Elofsson 2012a) and tobmodel (Hayat and Elofsson 2012b) have been implemented. These methods employ SVMs in the first stage to predict the local preference of each residue to form an outer-loop, inner-loop or membrane strand region. The output of this stage is then fed to an HMM that predicts the overall topology. Another approach called BETAWARE (Savojardo et al. 2013a) consists of two methods, first an N-to-1 Extreme Learning Machine algorithm is used for the identification of TMBs, followed by a Grammatical-Restrained Hidden Conditional Random Field approach to predict the topology. In contrast to other methods, transFold (Waldispühl et al. 2006) does not require a training set but uses a grammar to predict the β-strands and inter-strand residue contacts. Most of these topology prediction methods can also be used for distinguishing TMBs from non-TMBs.

### 5.5.8 *Consensus Approaches for β-Barrel Topology Prediction*

To our knowledge, conBBPRED (Bagos et al. 2005) is the only consensus method available for TMB topology prediction. conBBPRED assigns a per-residue score by averaging over contributions of each individual predictor followed by a dynamic programming step to obtain the overall topology. On a dataset of 20 proteins, conBBPRED increases the accuracy of predicted topologies by 15% (Bagos et al. 2005). With larger datasets and more topology predictors becoming available, it will be interesting to see if consensus topology prediction methods for TMBs show improved accuracy over single methods.

## 5.6 3D Structure Prediction

As with globular proteins, 3D structure prediction of TM proteins can be dealt with via two main approaches, homology modelling and de novo modelling, covered in Chaps. 1 and 4 of this book.

### *5.6.1 Homology Modelling of α-Helical Transmembrane Proteins*

Homology modelling involves the use of a related template structure in order to build a 3D model of a target protein. The method is based on the observation that protein structure is conserved more highly than amino acid sequence, hence even proteins that have diverged significantly in sequence but still share detectable similarity may also share common structural properties, and in particular, the overall fold. When a suitable template is available, predicting TM protein structure by homology modelling can be highly effective, especially when tools specifically designed for modelling TM proteins are used. Compared to globular proteins, lower sequence conservation is required for fold preservation in transmembrane regions, so it may even be possible to generate useful 3D models with templates that share as little as 20% sequence identity to the target, although the paucity of high resolution membrane protein structures will still limit the number of families that such methods are applicable to (Olivella et al. 2013).

A homology modelling protocol can be subdivided into a number of key steps which can each be performed iteratively to improve the quality of the final model: template selection, target-template alignment, model construction, and model quality assessment (Marti-Renom et al. 2000; Sanchez and Sali 1997). Aside from SWISS-MODEL (Peitsch 1996; Biasini et al. 2014) which has a 7TM/GPCR interface, few TM protein-specific homology modelling methods exist. MEDELLER (Kelm et al. 2010) is designed to approach the steps in structure prediction to take into account the differences between the physical environments of globular and TM proteins. The method is optimized to build a highly reliable core structure shared by the template and target proteins by first calculating membrane insertion using iMembrane (Kelm et al. 2009) which is used to guide target-template alignment by MP-T (Hill and Deane 2013). The core is gradually extended using a specialized membrane-specific substitution score, before loops are completed using the loop modelling protocols FREAD (Choi and Deane 2010) and Modeller (Marti-Renom et al. 2000). Results show that MEDELLER produces accurate core models and achieves a core model accuracy of 1.97 Å RMSD versus 2.57 Å for Modeller. The Memoir modeling pipeline now provides a fully automated web server that applies this protocol to both α-helical and β-barrel TM proteins (Ebejer et al. 2013).

Chen and co-workers developed a method specifically to deal with the issue of building homology models from very distantly related homologues exhibiting distinct loop and TM helix conformations (Chen et al. 2014). The approach is based on efficient sampling techniques of alternative TM helix structures, in order to reconstruct both TM core and loop regions from distant structural homologues, resulting in high quality models that were top-ranked when stringently validated in two blind predictions (Kufareva et al. 2011; Michino et al. 2009). Since the method requires only a single distant homolog, they estimate that around 60% of human membrane proteins can be reliably modeled using their approach, allowing the generation of 3D models for a large and diverse fraction of structurally uncharacterized TM proteins.

A number of tools also exist to model specific regions of TM proteins. These include TM loop regions, which have been shown to differ significantly from loop regions in globular proteins. Kelm and co-workers showed that it is possible to accurately predict the structure of TM loops using a database of small TM protein loop fragments (0.8–1.6 Å). Their findings show that while many globular protein fragments have similar shapes to their TM counterparts, their sequences are often very different, although they do not appear to differ in their substitution patterns. Their method is implemented in a modification to FREAD (Kelm et al. 2014). Modelling of TM kinks has also attracted a lot of attention, as they have been observed to provide important functional and structural roles in TM proteins (Yohannan et al. 2004). Tools to model TM kinks include the Monte Carlo method based algorithm, MC-HELAN, which determines helical axes alongside positions and angles of helical kinks (Langelaan et al. 2010), HELANAL-Plus (Kumar and Bansal 2012), a web server for analysis of helix geometry in TM protein structures, and TMKink, a neural network predictor which identifies over two-thirds of all bends with high sensitivity and specificity (Meruelo et al. 2011).

### 5.6.2 *Homology Modelling of Transmembrane β-Barrel Proteins*

For transmembrane β-barrel proteins, HHomp (Remmert et al. 2009) can be used to identify remote homologues with a known 3D structure that can act as template/s for 3D modelling of these proteins. Standard application of MEDELLER or MODELLER can then be used to generate all-atom homology models (Kelm et al. 2010; Marti-Renom et al. 2000). The TMBpro method (Randall et al. 2008) uses a combination of machine-learning to predict the location of β-strands and inter-strand contacts and then selects templates from TMBs with known 3D structure by matching the number of β-strands. However, as stated above, a key limitation of such an approach is that it is only limited to protein sequences for which a reliable template can be found. Additionally, for transmembrane β-barrels,

where identification of novel families is still an open issue, such an approach might miss reliable templates.

### 5.6.3 De Novo Modelling of α-Helical Transmembrane Proteins

De novo modelling, or ab initio modelling, involves the construction of a 3D model in the absence of any tertiary structural data relating to the target protein. As with homology modelling, most methods address globular proteins although recently a number of methods have emerged specifically to deal with TM proteins including FILM (Pellegrini-Calace et al. 2003), RosettaMembrane (Barth et al. 2007, 2009) and BCL::MP-fold (Weiner et al. 2013) (Table 5.5).

FILM (Folding In Lipid Membranes) is a modification of the globular protein structure prediction method FRAGFOLD (Jones and McGuffin 2003; Jones 1997). FRAGFOLD employs simulated annealing in order to perform a conformational search using high-resolution super-secondary structural fragments to assemble the tertiary fold, guided by a statistical function that includes pairwise, solvation, steric and hydrogen bonding energy terms. FILM added a knowledge-based membrane potential term to the FRAGFOLD energy function, derived from the statistical analysis of a data set of 640 transmembrane helices whose topologies had been determined experimentally. The relative frequencies of each amino acid at fixed distances from the membrane centre were assessed, allowing the membrane potential term to be calculated by transforming these values using the inverse Boltzmann equation. Results indicated that it was possible to predict both the topology and conformation of small proteins at a reasonable level of accuracy, although attaining the level of compactness observed in larger TM helix bundles was challenging, since TM helix bundles are usually not optimally compact despite neighboring helices being closely packed together. Further modification to FILM allowed progress to be made in the prediction of larger TM helix bundles by incorporating another term accounting for lipid exposure into the energy function. This allowed models of seven TM helix bacteriorhodopsin and rhodopsin to be

**Table 5.5** 3D modelling tools for α-helical transmembrane proteins

| Method | Features | URL |
|---|---|---|
| RosettaMembrane (Barth et al. 2009, 2007) | Knowledge-based potential | https://www.rosettacommons.org/ |
| Evfold_membrane (Hopf et al. 2012; Sheridan et al. 2015) | Evolutionary couplings | http://evfold.org/transmembrane |
| FILM3 (Nugent and Jones 2012) | Evolutionary couplings | http://bioinfadmin.cs.ucl.ac.uk/downloads/FILM3/ |
| BCL::MP-fold (Weiner et al. 2013) | Knowledge-based potential | http://www.meilerlab.org/index.php/servers |

generated to within 6–7 Å root mean square deviation (rmsd) of the native structure (Hurwitz et al. 2006).

RosettaMembrane is also a modification of a globular protein structure prediction method—Rosetta (Rohl et al. 2004; Simons et al. 1999), which, like FRAGFOLD, assembles folds using fragments of known structures using simulated annealing or parallel tempering—an effective algorithm to overcome the slow convergence in low-temperature protein simulation. RosettaMembrane added terms to the Rosetta energy function that described intra-protein and protein-solvent interactions in the anisotropic membrane environment, treating hydrogen bonds explicitly and membrane protein/lipid interactions implicitly. The method describes interactions between protein residues at atomic detail while applying continuum solvent models to the water, hydrophobic core, and lipid head group regions of the membrane. Results suggest that the model captures the essential physical properties that govern the solvation and stability of membrane proteins, allowing the structures of 12 small TM protein domain (<150 residues) to be predicted successfully to a resolution of <2.5 Å (129), comparing favourably with predictions obtained on small water-soluble protein domains. More recently, the method was extended to incorporate distance constraints into the predictions to direct helix-helix interactions, the constraints being derived from either experimental data or sequence-based predictions (Fuchs et al. 2009; Lo et al. 2009; Nugent et al. 2011; Nugent and Jones 2010). This allowed larger (90–300 residues) structures with more complicated topologies to be successfully modelled to within 4 Å rmsd in the best four cases, with results indicating that only a single constraint was sometimes sufficient to enrich the population of near-native models.

A recent method BCL::MP-fold (Weiner et al. 2013), a modification of BCL::Fold (Karakas et al. 2012), generates models within a static membrane object by evaluating conformations using a knowledge-based energy potential which takes into account the unique properties of the apolar membrane in the amino acid environment potential, as well as an increased radius of gyration along the membrane normal. Three additional terms are introduced first to describe the preferential orientation of secondary structure elements with respect to the membrane, secondly to penalise connection of two neighboring TM helices that would require passage through the membrane, and finally to assess the agreement of residue placement in TM regions with predictions from sequence. Additionally, a symmetry folding mode allows for the prediction of obligate homo-multimeric TM complexes. A benchmark test using 40 TM protein 3D structures demonstrated that the method is able to accurately predict the correct topology in 34 cases, suggesting the approach can successfully predict protein topology without the need for large multiple sequence alignments, homologous template structures, or experimental restraints.

### 5.6.4 De Novo Modelling of Transmembrane β-Barrels

The topological arrangement of β-strands in transmembrane β-barrels is regular and can be exploited to generate 3D models of TMBs based on an idealized geometry (Naveed et al. 2012; Hayat and Elofsson 2012b). Existing methods based on idealized geometry approximate the diameter of a TMB, calculated based on its number of strands. Additionally, 3D coordinates of Cα atoms along β-strands and their placement with respect to the in-register Cα atom can also be determined using a theoretical description (Chou et al. 1990; Murzin et al. 1994a, b). Tobmodel uses these regular structural features to generate idealized Cα atoms of TMBs (Hayat and Elofsson 2012b). Another method, 3d-SpoT, uses an empirical scoring function derived from frequencies of lipid-facing and pore-facing residues in known TMB structures to find the optimal strand-registration and then uses a geometric model of intertwined coils to generate 3D models (Naveed et al. 2012) (Table 5.6).

### 5.6.5 Covariation-Based Approaches

Up until recently, using knowledge-based potentials derived from the statistical analysis of known protein structures has been the standard approach for de novo structure prediction. Over the last five years, the field has seen dramatic progress as new methods have emerged that are capable of accurately inferring residue-residue contacts from large multiple sequence alignments (MSAs), allowing 3D structures to be computed directly from sequence data. Two key factors have led to this revolution; firstly, the rapid growth in the size of sequence databases, which has resulted in the number of sequences available for a typical protein family increasing by orders of magnitude (Sadowski and Taylor 2013), and secondly, the application of advanced statistical methods to this sequence data that allows the detection of true correlated mutations between sites in MSAs. The main idea behind correlated

**Table 5.6** 3D modelling tools for transmembrane β-barrels

| Method | Features | URL |
|---|---|---|
| EVfold_bb (Hayat et al. 2015) | Evolutionary couplings + Strand-registration prediction | http://cbio.mskcc.org/foldingproteins/transmembrane/betabarrels/ |
| tobmodel (Hayat and Elofsson 2012b) | Topology + Strand-registration prediction | http://tmbmodel.cbr.su.se/ |
| 3D-SpoT (Naveed et al. 2012) | Inter-strand pairing + Idealized barrel | http://tanto.bioe.uic.edu/TMBB-Explorer/ |
| TMBpro (Randall et al. 2008) | Machine learning + Templates | http://tmbpro.ics.uci.edu/ |

mutations is that residues that are proximal in 3D space are more likely to impose constraints on each other, which should lead to a correlation in their substitution patterns in the MSA. Mutation of either residue might disrupt the stability of the contact, which is likely to have an impact on the stability of the overall fold. Subsequent mutation of one or both residues to a more physicochemically complementary pairing may increase the likelihood of the contact being maintained; therefore residue pairs that form contacts are often seen to covary. It is this property that modern contact prediction methods seek to exploit.

A number of different methods have been developed for predicting contacts from sequence data based on the recognition of these residue covariation patterns. Up until now, they major obstacle in achieving performance useful for structure prediction has been in dealing with indirect coupling effects: should a direct contact exist at sites A–B and A–C, an apparent interaction may appear between B-C even though no direct contact exists. The approach of Lapedes et al. (1999) dealt with this so-called chaining problem by applying a maximum entropy approach, but at a high computational cost. The Direct Coupling Analysis (DCA) method reduced the problem to one of maximum entropy inference, applying a heuristic message passing approach to determine the solution of the contact weights (Weigt et al. 2009). This allowed the approach of Lapedes et al. to be put to practical use, with prediction accuracy achieving sufficient quality to be useful in structure prediction (Taylor and Sadowski 2011). PSICOV is based on sparse inverse covariance estimation (Jones et al. 2012). It applies the graphical lasso method (Friedman et al. 2008) to estimate the inverse of the covariance matrix, which is calculated from the MSA, whilst also constraining the solution to be sparse. The inverse covariance matrix, also known as the precision matrix, gives the correlation between any two sites in the MSA, conditional on observations at all other sites. This global statistical model was able to predict contacts with an accuracy approaching 80%, even for long-range contacts (those separated by >23 residues in the sequence), which is sufficient to identify to the native fold for medium sized (<200 residue) globular proteins, where sufficient numbers of aligned sequences are available. A more recent method, plmDCA (Ekeberg et al. 2013) uses a pseudo-likelihood approach applied to the Potts models. This has been shown to significantly outperform existing DCA-based approaches, while consensus approaches such as PconsC (Skwark et al. 2013) and MetaPSICOV further improve performance (Jones et al. 2015).

### 5.6.6 *Evolutionary Covariation-Based Methods for De Novo Modelling of α-Helical Membrane Proteins*

The performance of these methods has led to the development of a number of de novo structure prediction methods capable of generating accurate models for even large domains, guided primarily by predicted contacts. Evfold_membrane (Hopf et al. 2012) incorporates predicted transmembrane topology into the EVfold protocol

(Marks et al. 2011), which uses DCA in combination with the CNS molecular dynamics software suite to generate 3D models. A webserver to de novo fold proteins using EVfold protocol with DCA and plmDCA has also been implemented (Sheridan et al. 2015). It was shown to be capable of generating accurate models within the top-10 ranked structures for fifteen targets ranging in size from 50 to 260 residues to within 2.7–4.8 Å rmsd of their native structures over at least two-thirds of the protein length. The latest version of FILM, FILM3, replaces the statistical potential with a single scoring function based on predicted contacts and their estimated probabilities (Nugent and Jones 2012). Using contacts predicted by PSICOV, results indicate that models with TM-scores >0.5 could be generated for 25 out of 28 membrane protein targets with complex topologies and an average length over 300 residues (Fig. 5.4). In the most remarkable case, it was possible to build a model for all 514 residues of cytochrome c oxidase polypeptide I with a TM-score >0.75. As encouraging as these results are, data suggests that even with perfect distance constraints, folding methods are unable to generate models less than 2 Å rmsd of the native structure, suggesting that protein refinement protocols will play an increasingly important role in generating higher accuracy models.

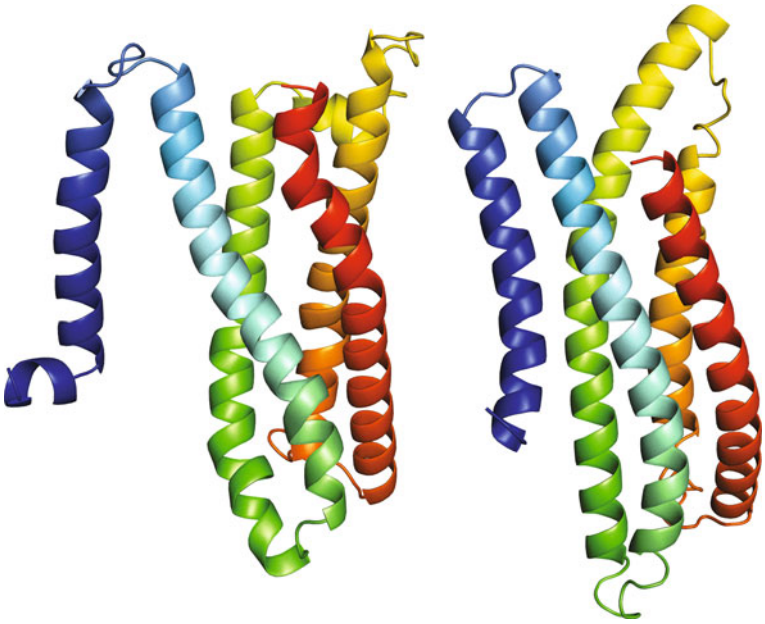


**Fig. 5.4** Model of CASP 11 free modelling target T0836 (*right*)—a 5-helix TM protein. Predicted contacts were generated using MetaPSICOV (Jones et al. 2015) enabling a model to be produced using the FILM3 protocol (Nugent and Jones 2012) resulting in a TM-score of 0.60 (Kosciolek and Jones 2015). The native structure is on the *left*

### 5.6.7 Evolutionary Covariation-Based Methods for Transmembrane β-Barrel Structure Prediction

Transmembrane β-barrels have a uniform β-strand topological pattern, where alternate strands traverse from the inside to the outside and vice versa, and additionally, anti-parallel β-strands have a unique hydrogen-bonding pattern. These structural features can be exploited to enhance the accuracy of predicting residues pairs in contact between two adjacent β-strands. Further, these can also be used to estimate the registration (relative position of two strands with respect to each other) of two adjacent β-strands. This has been shown to be useful for 3D modelling of TMBs (Hayat and Elofsson 2012b; Naveed et al. 2012; Randall et al. 2008). Additionally, Hayat et al. (2015) have implemented a simple strand-shift algorithm, where adjacent strands are shifted up/down relative to each other to ascertain the position that gives the highest sum of evolutionary couplings (ECs) between paired residues to identify the correct registration of TM β-strands in TMBs. This hybrid algorithm that combines empirical knowledge about TM β-strands and evolutionary covariation analysis-based contact prediction improves the prediction accuracy of inter-strand residue contacts. These predicted inter-strands constraints can then used to identify the underlying hydrogen-bonding network and the resulting interactions are used as distance constraints to de novo fold large TMBs using a tool called EVfold_bb (Hayat et al. 2015). EVfold_bb method can correctly predict the 3D structure with an average TM-score of 0.54 for the top-ranking models. EVfold_bb can also identify the correct inter-strand registration with an accuracy of 44% (in generated models), which is an improvement over tobmodel (18%), which does not use ECs to guide optimal strand registration search. Moreover, the generated models are not restricted to idealized geometries and do not require a template. Most interestingly, EVfold_bb can also identify and model 3D interactions between the barrel and the large plug domain in FecA protein (TM-score 0.68). The plug domain sits in the TM barrel domain and participates in gating and signaling (Noinaj et al. 2012).

Furthermore, methods specifically meant for improving prediction of β-sheet contacts in both globular and membrane proteins have also been developed. These methods can be broadly divided into two groups based on the use of ECs. BetaPro (Cheng and Baldi 2005) and MLN-2S (Lippi and Frasconi 2009) use neural networks and Markov logic networks, respectively, to predict β-sheet contacts. Maximum entropy-based correlated mutation measures (CMM) (Burkoff et al. 2013), Bcov (Savojardo et al. 2013b), bbcontacts (Andreani and Söding 2015) and MetaPSICOV (Jones et al. 2015) all use evolutionary covariation. In addition, these methods employ an additional layer of machine-learning techniques such as deep learning or HMMs on predicted evolutionary couplings to increase the accuracy of predicted residue-residue contacts in β-sheets. In future, methods that combine the general principles of anti-parallel β-stands along with machine-learning based methods that employ predicted contacts should be able to improve the applicability of these techniques to TMBs.

## 5.7 Future Directions

Substantial progress has been made in the field of membrane protein structure prediction over recent years. Methods for the detection of remote homologues have drastically improved, making it possible to generate template-based models for a larger number of protein families. Advances in techniques for predicting pairwise residue contacts have made it possible to generate de novo 3D models of large membrane proteins. However, these techniques are only applicable to protein families with large multiple sequence alignments. It is anticipated that as more sequencing data becomes available, 3D models of yet unknown TM protein families will become model-able based on predicted contacts. Future challenges lie in further improving these contact prediction methods by optimizing multiple sequence alignments, generation of fragment libraries, statistical inference methods used and the tools employed to predict 3D models.

**Competing Interests** The authors declare that they have no competing interests.

## References

Amico M, Finelli M, Rossi I, Zauli A, Elofsson A, Viklund H, von Heijne G, Jones D, Krogh A, Fariselli P, Luigi Martelli P, Casadio R (2006) PONGO: a web server for multiple predictions of all-alpha transmembrane proteins. Nucleic Acids Res 34 (Web Server issue):W169–172

Andreani J, Söding J (2015) bbcontacts: prediction of β-strand pairing from direct coupling patterns. Bioinformatics:btv041

Bagos PG, Liakopoulos TD, Hamodrakas SJ (2005) Evaluation of methods for predicting the topology of β-barrel outer membrane proteins and a consensus prediction method. BMC Bioinformatics 6(1):7

Bagos PG, Liakopoulos TD, Hamodrakas SJ (2006) Algorithms for incorporating prior topological information in HMMs: application to transmembrane proteins. BMC Bioinformatics 7:189

Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004) PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins. Nucleic Acids Res 32(suppl 2):W400–W404

Bahr A, Thompson JD, Thierry JC, Poch O (2001) BAliBASE (Benchmark Alignment dataBASE): enhancements for repeats, transmembrane sequences and circular permutations. Nucleic Acids Res 29(1):323–326

Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci USA 104(40):15682–15687

Barth P, Wallner B, Baker D (2009) Prediction of membrane protein structures with complex topologies using limited constraints. Proc Natl Acad Sci USA 106(5):1409–1414

Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783–795

Bernsel A, Viklund H, Hennerdal A, Elofsson A (2009) TOPCONS: consensus prediction of membrane protein topology. Nucleic Acids Res 37(Web Server issue):W465–468

Berven FS, Flikka K, Jensen HB, Eidhammer I (2004) BOMP: a program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 32(suppl 2):W394–W399

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42(Web Server issue): W252–258

Bigelow H, Rost B (2006) PROFtmb: a web server for predicting bacterial transmembrane beta barrel proteins. Nucleic Acids Res 34(suppl 2):W186–W188

Burkoff NS, Várnai C, Wild DL (2013) Predicting protein β-sheet contacts using a maximum entropy based correlated mutation measure. Bioinformatics:btt005

Chang G, Roth CB, Reyes CL, Pornillos O, Chen YJ, Chen AP (2006) Retraction. Science 314 (5807):1875

Chang JM, Di Tommaso P, Taly JF, Notredame C (2012) Accurate multiple sequence alignment of transmembrane proteins with PSI-Coffee. BMC Bioinformatics 13(Suppl 4):S1

Chen KY, Sun J, Salvo JS, Baker D, Barth P (2014) High-resolution modeling of transmembrane helical protein structures from distant homologues. PLoS Comput Biol 10(5):e1003636

Cheng J, Baldi P (2005) Three-stage prediction of protein β-sheets by neural networks, alignments and graph algorithms. Bioinformatics 21(suppl 1):i75–i84

Chetwynd AP, Scott KA, Mokrab Y, Sansom MS (2008) CGDB: a database of membrane protein/lipid interactions by coarse-grained molecular dynamics simulations. Mol Membr Biol 25(8):662–669

Choi Y, Deane CM (2010) FREAD revisited: Accurate loop structure prediction using a database search algorithm. Proteins 78(6):1431–1440

Chou KC, Carlacci L, Maggiora GM (1990) Conformational and geometrical properties of idealized beta-barrels in proteins. J Mol Biol 213(2):315–326

Claros MG, von Heijne G (1994) TopPred II: an improved software for membrane protein structure predictions. Comput Appl Biosci 10(6):685–686

Dayhoff MO, Schwartz RM (1978) Chapter 22: A model of evolutionary change in proteins. In: in Atlas of Protein Sequence and Structure. Silver Spring

Deng Y (2006) TSFSOM: transmembrane segments prediction by fuzzy self-organizing map. In: Advances in neural networks-ISNN 2006. Springer, pp 728–733

Diederichs K, Freigang J, Umhau S, Zeth K, Breed J (1998) Prediction by a neural network of outer membrane β-strand protein topology. Protein Sci 7(11):2413–2420

Dobson L, Lango T, Remenyi I, Tusnady GE (2015a) Expediting topology data gathering for the TOPDB database. Nucleic Acids Res 43(Database issue):D283–289

Dobson L, Remenyi I, Tusnady GE (2015b) CCTOP: a Consensus constrained TOPology prediction web server. Nucleic Acids Res

Ebejer JP, Hill JR, Kelm S, Shi J, Deane CM (2013) Memoir: template-based structure prediction for membrane proteins. Nucleic Acids Res 41(Web Server issue):W379–383

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Ekeberg M, Lovkvist C, Lan Y, Weigt M, Aurell E (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys Rev E: Stat, Nonlin, Soft Matter Phys 87(1):012707

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971

Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biophys Chem 15:321–353

Ferguson AD, Chakraborty R, Smith BS, Esser L, van der Helm D, Deisenhofer J (2002) Structural basis of gating by the outer membrane transporter FecA. Science 295(5560):1715–1719

Freeman TC Jr, Wimley WC (2010) A highly accurate statistical approach for the prediction of transmembrane beta-barrels. Bioinformatics 26(16):1965–1974. doi:10.1093/bioinformatics/btq308

Freeman TC, Wimley WC (2012) TMBB-DB: a transmembrane β-barrel proteome database. Bioinformatics 28(19):2425–2430

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441
Fuchs A, Kirschner A, Frishman D (2009) Prediction of helix-helix contacts and interacting helices in polytopic membrane proteins using neural networks. Proteins 74(4):857–871
Gallin WJ, Boutet PA (2011) VKCDB: voltage-gated K+ channel database updated and upgraded. Nucleic Acids Res 39(Database issue):D362–366
Garrow AG, Agnew A, Westhead DR (2005) TMB-Hunt: a web server to screen sequence sets for transmembrane β-barrel proteins. Nucleic Acids Res 33(suppl 2):W188–W192
Gromiha MM, Ahmad S, Suwa M (2005) TMBETA-NET: discrimination and prediction of membrane spanning β-strands in outer membrane proteins. Nucleic Acids Res 33(suppl 2): W164–W167
Gromiha MM, Majumdar R, Ponnuswamy P (1997) Identification of membrane spanning beta strands in bacterial porins. Protein Eng 10(5):497–500
Gromiha MM, Ponnuswamy P (1993) Prediction of transmembrane β-strands from hydrophobic characteristics of proteins. Int J Pept Protein Res 42(5):420–431
Gromiha MM, Yabuki Y, Kundu S, Suharnan S, Suwa M (2007) TMBETA-GENOME: database for annotated β-barrel membrane proteins in genomic sequences. Nucleic Acids Res 35(suppl 1):D314–D316
Hayat M, Khan A (2013) WRF-TMH: predicting transmembrane helix by fusing composition index and physicochemical properties of amino acids. Amino Acids 44(5):1317–1328
Hayat S, Elofsson A (2012a) BOCTOPUS: improved topology prediction of transmembrane β barrel proteins. Bioinformatics 28(4):516–522
Hayat S, Elofsson A (2012b) Ranking models of transmembrane β-barrel proteins using Z-coordinate predictions. Bioinformatics 28(12):i90–i96
Hayat S, Sander C, Marks DS, Elofsson A (2015) All-atom 3D structure prediction of transmembrane β-barrel proteins from sequences. Proc Natl Acad Sci 112(17):5413–5418
Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 89(22):10915–10919
Henricson A, Kall L, Sonnhammer EL (2005) A novel transmembrane topology of presenilin based on reconciling experimental and computational evidence. FEBS J 272(11):2727–2733
Hill JR, Deane CM (2013) MP-T: improving membrane protein alignment for structure prediction. Bioinformatics 29(1):54–61. doi:10.1093/Bioinformatics/Bts640
Hill JR, Kelm S, Shi J, Deane CM (2011) Environment specific substitution tables improve membrane protein alignment. Bioinformatics 27(13):15–23
Hopf TA, Colwell LJ, Sheridan R, Rost B, Sander C, Marks DS (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149(7):1607–1621. doi:10.1016/j.cell.2012.04.012
Hopkins AL, Groom CR (2002) The druggable genome. Nat Rev Drug Discov 1(9):727–730
Hurwitz N, Pellegrini-Calace M, Jones DT (2006) Towards genome-scale structure prediction for transmembrane proteins. Philos Trans R Soc Lond B Biol Sci 361(1467):465–475
Imai K, Hayat S, Sakiyama N, Fujita N, Tomii K, Elofsson A, Horton P (2013) Localization prediction and structure-based in Silico analysis of bacterial proteins: with emphasis on outer membrane proteins. In: Data mining for systems biology. Springer, Berlin, pp 115–140
Jackups R, Liang J (2005) Interstrand pairing patterns in β-barrel membrane proteins: the positive-outside rule, aromatic rescue, and strand registration prediction. J Mol Biol 354 (4):979–993
Jayasinghe S, Hristova K, White SH (2001) MPtopo: a database of membrane protein topology. Protein Sci 10(2):455–458
Jimenez-Morales D, Liang J (2011) Pattern of amino acid substitutions in transmembrane domains of β-barrel membrane proteins for detecting remote homologs in bacteria and mitochondria. PLoS ONE 6(11):e26400
Jones DT (1997) Successful ab initio prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. Proteins Suppl 29(1):185–191

Jones DT (2007) Improving the accuracy of transmembrane protein topology prediction using evolutionary information. Bioinformatics 23(5):538–544
Jones DT, Buchan DW, Cozzetto D, Pontil M (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 28(2):184–190. doi:10.1093/bioinformatics/btr638
Jones DT, McGuffin LJ (2003) Assembling novel protein folds from super-secondary structural fragments. Proteins 53(suppl 6):480–485
Jones DT, Singh T, Kosciolek T, Tetchner S (2015) MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 31(7):999–1006
Jones DT, Taylor WR, Thornton JM (1994a) A model recognition approach to the prediction of all-helical membrane protein structure and topology. Biochemistry 33(10):3038–3049
Jones DT, Taylor WR, Thornton JM (1994b) A mutation data matrix for transmembrane proteins. FEBS Lett 339(3):269–275
Kall L, Krogh A, Sonnhammer EL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338(5):1027–1036
Kall L, Krogh A, Sonnhammer EL (2005) An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 21(Suppl 1):i251–i257
Karakas M, Woetzel N, Staritzbichler R, Alexander N, Weiner BE, Meiler J (2012) BCL:Fold-de novo prediction of complex and large protein topologies by assembly of secondary structure elements. PLoS ONE 7(11):e49240. doi:10.1371/journal.pone.0049240
Kelm S, Shi J, Deane CM (2009) iMembrane: homology-based membrane-insertion of proteins. Bioinformatics 25(8):1086–1088
Kelm S, Shi J, Deane CM (2010) MEDELLER: homology-based coordinate generation for membrane proteins. Bioinformatics 26(22):2833–2840
Kelm S, Vangone A, Choi Y, Ebejer JP, Shi J, Deane CM (2014) Fragment-based modeling of membrane protein loops: successes, failures, and prospects for the future. Proteins 82(2):175–186. doi:10.1002/prot.24299
Khafizov K, Staritzbichler R, Stamm M, Forrest LR (2010) A study of the evolution of inverted-topology repeats from LeuT-fold transporters using AlignMe. Biochemistry 49 (50):10702–10713
Klammer M, Messina DN, Schmitt T, Sonnhammer EL (2009) MetaTM—a consensus method for transmembrane protein topology prediction. BMC Bioinformatics 10:314
Kosciolek T, Jones DT (2015) Accurate contact predictions using coevolution techniques and machine learning. Proteins. doi:10.1002/prot.24863
Kozma D, Simon I, Tusnady GE (2013) PDBTM: Protein Data Bank of transmembrane proteins after 8 years. Nucleic Acids Res 41(Database issue):D524–529
Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580
Kufareva I, Rueda M, Katritch V, Stevens RC, Abagyan R (2011) Status of GPCR modeling and docking as reflected by community-wide GPCR Dock 2010 assessment. Structure 19(8):1108–1126
Kumar P, Bansal M (2012) HELANAL-Plus: a web server for analysis of helix geometry in protein structures. J Biomol Struct Dyn 30(6):773–783
Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132
Kyttala A, Ihrke G, Vesa J, Schell MJ, Luzio JP (2004) Two motifs target Batten disease protein CLN3 to lysosomes in transfected nonneuronal and neuronal cells. Mol Biol Cell 15(3):1313–1323
Langelaan DN, Wieczorek M, Blouin C, Rainey JK (2010) Improved helix and kink characterization in membrane proteins allows evaluation of kink sequence predictors. J Chem Inf Model 50(12):2213–2220

Lapedes AS, Giraud B, Liu L, Stormo GD (1999) Correlated mutations in models of protein sequences: phylogenetic and structural effects. In: Seillier-Moiseiwitsch F (ed) Statistics in molecular biology and genetics, vol 33. Lecture Notes-Monograph Series. Institute of Mathematical Statistics, Hayward, CA, pp 236–256

Li B, Gallin WJ (2004) VKCDB: voltage-gated potassium channel database. BMC Bioinformatics 5:3

Lippi M, Frasconi P (2009) Prediction of protein β-residue contacts by Markov logic networks with grounding-specific weights. Bioinformatics 25(18):2326–2333

Lo A, Chiu HS, Sung TY, Hsu WL (2006) Transmembrane helix and topology prediction using hierarchical SVM classifiers and an alternating geometric scoring function. Comput Syst Bioinformatics Conf, 31–42

Lo A, Chiu HS, Sung TY, Lyu PC, Hsu WL (2008) Enhanced membrane protein topology prediction using a hierarchical classification method and a new scoring function. J Proteome Res 7(2):487–496

Lo A, Chiu YY, R?dland EA, Lyu PC, Sung TY, Hsu WL (2009) Predicting helix-helix interactions from residue contacts in membrane proteins. Bioinformatics 25(8):996–1003

Lomize AL, Pogozheva ID, Lomize MA, Mosberg HI (2006a) Positioning of proteins in membranes: a computational approach. Protein Sci 15(6):1318–1333

Lomize AL, Pogozheva ID, Mosberg HI (2011) Anisotropic solvent model of the lipid bilayer. 2. Energetics of insertion of small molecules, peptides, and proteins in membranes. J Chem Inf Model 51(4):930–946

Lomize MA, Lomize AL, Pogozheva ID, Mosberg HI (2006b) OPM: orientations of proteins in membranes database. Bioinformatics 22(5):623–625. doi:10.1093/bioinformatics/btk023

Mao Q, Foster BJ, Xia H, Davidson BL (2003) Membrane topology of CLN3, the protein underlying Batten disease. FEBS Lett 541(1–3):40–46

Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, Sander C (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766. doi:10.1371/journal.pone.0028766

Martelli PL, Fariselli P, Casadio R (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. Bioinformatics 19(Suppl 1):i205–i211

Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. Bioinformatics 18(suppl 1):S46–S53

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Meruelo AD, Samish I, Bowie JU (2011) TMKink: a method to predict transmembrane helix kinks. Protein Sci 20(7):1256–1264

Michino M, Abola E, Brooks CL, Dixon JS, Moult J, Stevens RC (2009) Community-wide assessment of GPCR structure modelling and ligand docking: GPCR Dock 2008. Nat Rev Drug Discov 8(6):455–463

Muller T, Rahmann S, Rehmsmeier M (2001) Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics 17(Suppl 1):S182–S189

Murzin AG, Lesk AM, Chothia C (1994a) Principles determining the structure of beta-sheet barrels in proteins. I. A theoretical analysis. J Mol Biol 236(5):1369–1381

Murzin AG, Lesk AM, Chothia C (1994b) Principles determining the structure of beta-sheet barrels in proteins. II. The observed structures. J Mol Biol 236(5):1382–1400

Natt NK, Kaur H, Raghava G (2004) Prediction of transmembrane regions of β-barrel proteins using ANN-and SVM-based methods. Proteins: Struct Funct Bioinf 56(1):11–18

Naveed H, Xu Y, Jackups R Jr, Liang J (2012) Predicting three-dimensional structures of transmembrane domains of β-barrel membrane proteins. J Am Chem Soc 134(3):1775–1781

Ng PC, Henikoff JG, Henikoff S (2000) PHAT: a transmembrane-specific substitution matrix. Predicted hydrophobic and transmembrane. Bioinformatics 16(9):760–766

Nilsson J, Persson B, Von Heijne G (2002) Prediction of partial membrane protein topologies using a consensus approach. Protein Sci 11(12):2974–2980

Noinaj N, Easley NC, Oke M, Mizuno N, Gumbart J, Boura E, Steere AN, Zak O, Aisen P, Tajkhorshid E, others (2012) Structural basis for iron piracy by pathogenic Neisseria. Nature 483(7387):53–58

Notredame C, Higgins DG, Heringa J (2000) T-Coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol 302(1):205–217

Nugent T, Jones DT (2009) Transmembrane protein topology prediction using support vector machines. BMC Bioinformatics 10:159

Nugent T, Jones DT (2010) Predicting transmembrane helix packing arrangements using residue contacts and a force-directed algorithm. PLoS Comput Biol 6(3):e1000714

Nugent T, Jones DT (2012) Accurate de novo structure prediction of large transmembrane protein domains using fragment-assembly and correlated mutation analysis. Proc Natl Acad Sci USA 109(24):E1540–E1547

Nugent T, Jones DT (2013) Membrane protein orientation and refinement using a knowledge-based statistical potential. BMC Bioinformatics 14:276

Nugent T, Ward S, Jones DT (2011) The MEMPACK alpha-helical transmembrane protein structure prediction server. Bioinformatics 27(10):1438–1439

Olivella M, Gonzalez A, Pardo L, Deupi X (2013) Relation between sequence and structure in membrane proteins. Bioinformatics 29(13):1589–1592

Y-y Ou, S-a Chen, Gromiha MM (2010) Prediction of membrane spanning segments and topology in β-barrel membrane proteins at better accuracy. J Comput Chem 31(1):217–223

Peitsch MC (1996) ProMod and Swiss-Model: internet-based tools for automated comparative protein modelling. Biochem Soc Trans 24(1):274–279

Pellegrini-Calace M, Carotti A, Jones DT (2003) Folding in lipid membranes (FILM): a novel method for the prediction of small membrane protein 3D structures. Proteins 50(4):537–545

Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8(10):785–786

Pirovano W, Feenstra KA, Heringa J (2008) PRALINETM: a strategy for improved multiple alignment of transmembrane proteins. Bioinformatics 24(4):492–497

Qi Y, Oja M, Weston J, Noble WS (2012) A unified multitask architecture for predicting local protein properties. PLoS ONE 7(3):e32235

Randall A, Cheng J, Sweredoski M, Baldi P (2008) TMBpro: secondary structure, β-contact and tertiary structure prediction of transmembrane β-barrel proteins. Bioinformatics 24(4):513–520

Ratajczak E, Petcherski A, Ramos-Moreno J, Ruonala MO (2014) FRET-assisted determination of CLN3 membrane topology. PLoS ONE 9(7):e102593

Remmert M, Linke D, Lupas AN, Söding J (2009) HHomp?prediction and classification of outer membrane proteins. Nucleic Acids Res 37(suppl 2):W446–W451

Reynolds SM, Kall L, Riffle ME, Bilmes JA, Noble WS (2008) Transmembrane topology and signal peptide prediction using dynamic bayesian networks. PLoS Comput Biol 4(11): e1000213

Rohl CA, Strauss CE, Misura KM, Baker D (2004) Protein structure prediction using Rosetta. Meth Enzymol 383:66–93

Rost B, Fariselli P, Casadio R (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 5(8):1704–1718

Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci 90(16):7558–7562

Sadowski MI, Taylor WR (2013) Prediction of protein contacts from correlated sequence substitutions. Sci Prog 96(Pt 1):33–42

Saier MH, Reddy VS, Tamang DG, Vastermark A (2014) The transporter classification database. Nucleic Acids Res 42(Database issue):D251–258

Saier MH, Tran CV, Barabote RD (2006) TCDB: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Res 34(Database issue): D181–186

Saier MH, Yen MR, Noto K, Tamang DG, Elkan C (2009) The Transporter Classification Database: recent advances. Nucleic Acids Res 37(Database issue):D274–278

Samatey FA, Xu C, Popot JL (1995) On the distribution of amino acid residues in transmembrane alpha-helix bundles. Proc Natl Acad Sci USA 92(10):4577–4581

Sanchez R, Sali A (1997) Advances in comparative protein-structure modelling. Curr Opin Struct Biol 7(2):206–214

Sansom MS, Scott KA, Bond PJ (2008) Coarse-grained simulation: a high-throughput computational approach to membrane proteins. Biochem Soc Trans 36(Pt 1):27–32

Savojardo C, Fariselli P, Casadio R (2013a) BETAWARE: a machine-learning tool to detect and predict transmembrane beta barrel proteins in Prokaryotes. Bioinformatics:bts728

Savojardo C, Fariselli P, Martelli PL, Casadio R (2013b) BCov: a method for predicting β-sheet topology using sparse inverse covariance estimation and integer programming. Bioinformatics: btt555

Schirmer T, Cowan SW (1993) Prediction of membrane-spanning β-strands and its application to maltoporin. Protein Sci 2(8):1361–1363

Senes A, Gerstein M, Engelman DM (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. J Mol Biol 296(3):921–936

Shafrir Y, Guy HR (2004) STAM: simple transmembrane alignment method. Bioinformatics 20(5):758–769

Sheridan R, Fieldhouse RJ, Hayat S, Sun Y, Antipin Y, Yang L, Hopf T, Marks DS, Sander C (2015) EVfold. org: Evolutionary Couplings and Protein 3D Structure Prediction. bioRxiv:021022

Simons KT, Bonneau R, Ruczinski I, Baker D (1999) Ab initio protein structure prediction of CASP III targets using ROSETTA. Proteins Suppl 3:171–176

Singh NK, Goodman A, Walter P, Helms V, Hayat S (2011) TMBHMM: a frequency profile based HMM for predicting the topology of transmembrane beta barrel proteins and the exposure status of transmembrane residues. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1814(5):664–670

Skwark MJ, Abdel-Rehim A, Elofsson A (2013) PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 29(14):1815–1816. doi:10.1093/bioinformatics/btt259

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2013) Alignment of helical membrane protein sequences using AlignMe. PLoS ONE 8(3):e57731

Stamm M, Staritzbichler R, Khafizov K, Forrest LR (2014) AlignMe–a membrane protein sequence alignment web server. Nucleic Acids Res 42(Web Server issue):W246–251

Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B (2003) On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles. J Mol Biol 334(5):1043–1062

Taylor PD, Attwood TK, Flower DR (2003) BPROMPT: a consensus server for membrane protein prediction. Nucleic Acids Res 31(13):3698–3700

Taylor WR, Sadowski MI (2011) Structural constraints on the covariance matrix derived from multiple aligned protein sequences. PLoS ONE 6(12):e28265

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680

Tsirigos KD, Bagos PG, Hamodrakas SJ (2011) OMPdb: a database of β-barrel outer membrane proteins from Gram-negative bacteria. Nucleic Acids Res 39(suppl 1):D324–D331

Tsirigos KD, Peters C, Shu N, Kall L, Elofsson A (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. Nucleic Acids Res

Tusnady GE, Dosztanyi Z, Simon I (2004) Transmembrane proteins in the Protein Data Bank: identification and classification. Bioinformatics 20(17):2964–2972

Tusnady GE, Dosztanyi Z, Simon I (2005a) PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. Nucleic Acids Res 33(Database issue):D275–278

Tusnady GE, Dosztanyi Z, Simon I (2005b) TMDET: web server for detecting transmembrane regions of proteins by using their 3D coordinates. Bioinformatics 21(7):1276–1277

Tusnady GE, Kalmar L, Simon I (2008) TOPDB: topology data bank of transmembrane proteins. Nucleic Acids Res 36(Database issue):D234–239

Tusnady GE, Simon I (1998) Principles governing amino acid composition of integral membrane proteins: application to topology prediction. J Mol Biol 283(2):489–506

Viklund H, Bernsel A, Skwark M, Elofsson A (2008) SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. Bioinformatics 24(24):2928–2929. doi:10.1093/bioinformatics/btn550

Viklund H, Elofsson A (2004) Best alpha-helical transmembrane protein topology predictions are achieved using hidden Markov models and evolutionary information. Protein Sci 13(7):1908–1917

Viklund H, Elofsson A (2008) OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. Bioinformatics 24(15):1662–1668. doi:10.1093/bioinformatics/btn221

von Heijne G (1992) Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. J Mol Biol 225(2):487–494

Waldispühl J, Berger B, Clote P, Steyaert J-M (2006) transFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res 34(suppl 2): W189–W193

Waldispühl J, O'Donnell CW, Devadas S, Clote P, Berger B (2008) Modeling ensembles of transmembrane β-barrel proteins. Proteins: Structure, Function, Bioinform 71(3):1097–1112

Wallin E, Tsukihara T, Yoshikawa S, von Heijne G, Elofsson A (1997) Architecture of helix bundle membrane proteins: an analysis of cytochrome c oxidase from bovine mitochondria. Protein Sci 6(4):808–815

Wang H, Liu B, Sun P, Ma Z (2013) A topology structure based outer membrane proteins segment alignment method. Mathematical Problems in Engineering 2013

Weigt M, White RA, Szurmant H, Hoch JA, Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA 106(1):67–72

Weiner BE, Woetzel N, Karakas M, Alexander N, Meiler J (2013) BCL:MP-fold: folding membrane proteins through assembly of transmembrane helices. Structure 21(7):1107–1117. doi:10.1016/j.str.2013.04.022

White SH (2004) The progress of membrane protein structure determination. Protein Sci 13(7):1948–1949

Wimley WC (2002) Toward genomic identification of β-barrel membrane proteins: Composition and architecture of known structures. Protein Sci 11(2):301–312

Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3(10):842–848

Yan R-X, Chen Z, Zhang Z (2011) Outer membrane proteins can be simply identified using secondary structure element alignment. BMC Bioinformatics 12(1):76

Yohannan S, Faham S, Yang D, Whitelegge JP, Bowie JU (2004) The evolution of transmembrane helix kinks and the structural diversity of G protein-coupled receptors. Proc Natl Acad Sci USA 101(4):959–963

Yuan Z, Mattick JS, Teasdale RD (2004) SVMtm: support vector machines to predict transmembrane segments. J Comput Chem 25(5):632–636

# Chapter 6
# Bioinformatics Approaches to the Structure and Function of Intrinsically Disordered Proteins

**Zsuzsanna Dosztányi and Peter Tompa**

**Abstract** Intrinsically disordered proteins and protein regions (IDPs/IDRs) exist without a well-defined structure. They carry out their function by relying on their highly flexible conformational states and are mostly involved in signal transduction and regulation. By a battery of biophysical techniques, the structural disorder of about 600 proteins has been demonstrated, and functional studies have provided the basis of classifying their functions into various schemes. Indirect evidence suggests that the occurrence of disorder is widespread, and several thousand proteins with significant disorder exist in the human proteome alone. To narrow the wide gap between known and anticipated IDPs, a range of bioinformatics algorithms have been developed, which can reliably predict the disordered state from the amino acid sequence. Attempts have also been made to predict IDP function. However, due to their fast evolution, and reliance on short motifs for function, capturing sequence clues for IDP functions is a much more challenging task. In this chapter we give a brief survey of the IDP field, with particular focus on their functions and bioinformatics approaches developed for predicting their structure and function.

**Keywords** Intrinsically unstructured proteins · Natively unfolded proteins · Conformational ensembles · Linear motifs · Flexibility · Prediction methods

Z. Dosztányi (✉)
MTA-ELTE Lendület Bioinformatics Research Group,
Department of Biochemistry, Eötvös Loránd University, Budapest, Hungary
e-mail: dosztanyi@caesar.elte.hu

P. Tompa (✉)
Research Center for Natural Sciences, Institute of Enzymology,
Hungarian Academy of Sciences, Budapest, Hungary
e-mail: tompa@enzim.hu

P. Tompa
VIB Center for Structural Biology (CSB), Vrije Universiteit Brussel, Brussels, Belgium



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_6

## 6.1 The Concept of Protein Disorder

The classical paradigm, which equated protein function with a stable 3D structure, had tremendous success in interpreting the function of enzymes, receptors and structural proteins. Decades of structure determination efforts and recent structural genomics programs have yielded over 100,000 well-defined structures deposited in the Protein Data Bank (PDB, www.pdb.org), strongly reinforcing the traditional view. The recent recognition that many proteins or regions of proteins lack a well-defined three-dimensional structure under native, physiological conditions, however, challenged the universality of this paradigm (Dunker et al. 2001; Tompa 2002; Dyson and Wright 2005). The rapid accumulation of data in support of this emerging alternative view of proteins in recent years led to the reassessment and extension of the structure-function paradigm (Wright and Dyson 1999; Tompa 2012).

A range of biophysical techniques, primarily X-ray crystallography, NMR, SAXS and CD, have provided evidence that intrinsically disordered, or unstructured, proteins (IDPs/IUPs) or regions of proteins (IDRs) assume no well-defined conformations, but rather a rapidly fluctuating ensemble of alternative structural states (Tompa 2002, 2005; Dyson and Wright 2005; Uversky et al. 2005). IDPs can occupy conformational states anywhere between the fully disordered (*random coil*) and compact (*molten globule*) states with characteristic distributions of transient secondary and tertiary contacts (Uversky et al. 2000; Uversky 2002) similarly to the denatured states of globular proteins. At variance with globular proteins, IDP functions directly stem from the unfolded states, and are exploited mostly in regulating processes of signal transduction and gene transcription (Iakoucheva et al. 2002; Ward et al. 2004; Tompa et al. 2006).

Not only are IDPs able to function despite their lack of stable structures, structural disorder actually provides functional advantages in regulatory functions, such as the separation of specificity from binding strength (Wright and Dyson 1999), adaptability to various partners (Tompa 2005), increased rate of interaction (Pontius 1993) and frequent involvement in post-translational modifications (Iakoucheva et al. 2004). These advantages enable IDPs to fit into unique functional niches, and explain the advance of protein disorder in evolution, with a critical difference in frequency between eukaryotes and prokaryotes (Iakoucheva et al. 2002; Ward et al. 2004; Tompa et al. 2006). The advantages also explain a high level of disorder in functionally important regulatory proteins, which also play central roles in disease, such as the prion protein (Lopez Garcia et al. 2000), BRCA1 (Mark et al. 2005), tau protein (Schweers et al. 1994), p53 (Bell et al. 2002), and α-synuclein (Weinreb et al. 1996). The current most complete collection of IDPs, the DisProt database (www.disprot.org), contains about 600 disordered proteins, mostly observed serendipitously as such (Sickmeier et al. 2007). The application of predictors based on such collection of proteins, however, suggests that, in the proteomes of metazoa, about 5–15% of proteins are fully disordered, and 30–50% of proteins contain at least one long disordered region (Dunker et al. 2000;

Ward et al. 2004; Tompa et al. 2006). To narrow this apparently wide gap in knowledge, a lot of effort is spent on developing bioinformatics algorithms to predict disorder and function from amino acid sequence. This review focuses on the principles and recent developments in this area of IDP research.

## 6.2 Sequence Features of IDPs

### *6.2.1 The Unusual Amino Acid Composition of IDPs*

It has been observed first by Dunker et al. (2001) that the frequency of amino acids in disordered proteins significantly differs from that of ordered proteins. The difference does not depend on the method used to establish the structural status of the protein, as they are always depleted in hydrophobic amino acids, and are enriched in polar and charged amino acids. The former group (Trp, Cys, Phe, Ile, Tyr, Val, and Leu) is termed order-promoting, whereas the latter (Ala, Arg, Gly, Gln, Ser, Pro, Glu, and Lys) are disorder-promoting (Dunker et al. 2001) amino acids. Similar trends have been found in other studies (Uversky 2002; Tompa 2002), and it is now generally accepted that the main attribute determining disorder is a low overall level of hydrophobicity, which precludes the formation of a stable globular core. This is often accompanied with high net charge, which favors an extended structural state due to electrostatic repulsion (Uversky et al. 2000).

### *6.2.2 Low Sequence Complexity and Disorder*

Another manifestation of the sequential bias of IDPs is the low sequence complexity of their polypeptide chains. Application of an entropy function to amino acid sequences of proteins (Wootton 1994; Wootton and Federhen 1996) has shown that globular proteins appear mostly to be in a high-entropy (complexity) state, whereas in many other proteins long regions apparently of low complexity can be observed. As much as 25% of all amino acids in SwissProt are in low-complexity regions, and 34% of all proteins have at least one such segment (Wootton 1994; Wootton and Federhen 1996). The exact relationship of low complexity and disorder has been addressed in two studies. First, the relation of alphabet size (number of amino acids) and complexity to the capacity of folding was studied (Romero et al. 1999). It was found that SwissProt proteins cover the entire possible range of alphabet size (1–20) and entropy range (K = 0.0–4.5), whereas globular domains only occupied a limited region (alphabet = 10–20, K = 3.0–4.2). Regions corresponding to lower values (down to alphabet size = 3 and K = 1.5) mostly correspond to structured, fibrous proteins, such as coiled coils, collagens and fibroins. It was concluded that a minimal alphabet size of 10 and entropy near 2.9 are

necessary and sufficient to define a sequence that can fold into a globular structure. By extending these studies to IDPs (Romero et al. 2001), it was shown that the complexity distribution of disordered proteins is shifted to lower values, but significantly overlaps with that of ordered proteins. Overall, disordered and low-complexity regions correlate and are abundant in proteomes, but low-complexity and disorder should not be treated as synonyms.

### *6.2.3 Flavours of Disorder*

Despite the pronounced differences between disordered and ordered protein regions, it is pertinent to mention that IDPs are heterogeneous in terms of their structural, and functional properties that is also reflected in their sequential properties. It was suggested that disordered regions characterized by various experimental techniques show different biases in their amino acid compositions (Dunker et al. 2001). One potential category corresponds to segments collected from the PDB database as missing residues in the electron density map. These are typically shorter segments (less than 10 residues long), often corresponding to terminal residues or flexible loop regions attached to globular domains. These regions show significant differences in their sequence properties compared to the typically longer (over 30 residue) segments collected in the DisProt database, that are usually identified by CD, NMR or hydrodynamic radius measurements which capture the global properties of these regions. The observed differences among these groups have important implications from the viewpoint of predictions as well, as disorder predictors trained on one group of proteins often perform poorly on other groups (Le Gall et al. 2007).

The specific sequential biases observed in certain disordered proteins can often be correlated with functional properties. One example for this is the trans-activator domains of transcription factors. These regions have a strong tendency to be disordered (Sigler 1988; Minezaki et al. 2006) and are often classified based on their amino acid composition. Traditionally, transcription factors are distinguished on the basis of the amino acid preferences of their trans-activator domains, such as acidic, Pro-rich and Gln-rich (Triezenberg 1995). Although the statistical foundation of these differences is practically non-existent, this categorization can be justified by that function within one category of transcription factors is rather insensitive to amino acid changes as long as the above character of the domain is maintained (Hope et al. 1988). On the other hand, mutations that change this character impair trans-activation function (Gill and Ptashne 1987). Thus, some features apparent at the level of composition are closely related to function. Whereas the insights gained from these analyses are too limited to establish clear categories, it may be suggestive of important directions of future research.

## 6.3 Prediction of Disorder

The underlying and unifying feature is that IDPs have "unusual" amino acid composition and sequence that distinguishes them from ordered proteins, suggesting that the primary determinant of protein disorder is encoded in the amino acid sequence proteins. Based on the noted compositional bias, about 50 predictors of disorder have been developed (reviewed in He et al. 2009; Dosztanyi et al. 2010) many of them conveniently available as web-servers or program packages (see Table 6.1). The best predictors approach the accuracy of the best secondary structure prediction algorithms, and the principles of comparing their performance have already been laid down.

### *6.3.1 Charge-Hydropathy Plot*

The classical approach to assess the disordered status of a protein is based on the observation of Uversky that a combination of low mean hydrophobicity and high net charge distinguishes IDPs from ordered proteins. This principle can be applied in a simple fashion, by plotting net charge versus net hydrophobicity (Uversky 2002), in a plot termed either the charge-hydropathy (CH) plot or the Uversky plot. On the plot IDPs tend to be positioned in the high net charge—low net hydrophobicity region, and are separated from globular proteins by a linear function of a formula <charge> = 2.743 <hydropathy> − 1.109 (Fig. 6.1), determined at high precision in a later study (Oldfield et al. 2005). A limitation of the CH plot is that it only enables a binary classification of proteins, without providing information at amino acid resolution. To deal with this situation, Sussman and colleagues have extended this principle (Prilusky et al. 2005) by applying a sliding window along a protein sequence to calculate mean hydrophobicity and net charge and thereby predict the disorder of the middle residue.

### *6.3.2 Propensity-Based Predictors*

The simplest approaches for the prediction of protein disorder are based on amino acid propensity scales, and assess if a given disorder-related amino acid feature is enriched or depleted within a pre-defined segment of the protein. Given the specific compositional bias of disordered proteins, the highest discriminatory power are expected from propensities that are related to various hydrophobicity scales, such as flexibility and coordination number (Xie et al. 1998). In a related fashion, GlobPlot (Linding et al. 2003), applies an amino acid propensity scale, which expresses the

**Table 6.1** Computational resources for IDPs

| Name | Web URL | Description |
|---|---|---|
| *Databases* | | |
| DisProt | http://www.disprot.org/ | Experimentally verified database of protein disorder |
| IDEAL | http://www.ideal.force.cs.is.nagoya-u.ac.jp/IDEAL/ | Intrinsically disordered proteins with manually curated annotations with a special focus on functional sites |
| MobiDB | http://mobidb.bio.unipd.it/ | A database of protein disorder and mobility annotations based on DisProt and PDB X-ray structures, several different flavours of disorder predictors |
| $D^2P^2$ | http://d2p2.pro/ | Database of pre-computed disorder predictions on a large library of proteins from completely-sequenced genomes |
| PE-DB | http://pedb.vib.be/ | A database for the deposition of structural information on IDP- and denatured protein ensembles based on NMR and SAXS data |
| ELM | http://elm.eu.org/ | Database of eukaryotic linear motifs |
| *Disorder prediction methods* | | |
| PONDR methods (e.g. VL_XT, VSL2) | http://www.pondr.com/cgi-bin/PONDR/pondr.cgi | Various methods based on machine learning principles |
| DISOPRED (DISOPRED2 and 3) | http://bioinf.cs.ucl.ac.uk/disopred | Machine learning methods (current method is based on NN) |
| IUPred | http://iupred.enzim.hu | Estimated pairwise interaction energy per residue |
| DisEMBL | http://dis.embl.de | Neural network based prediction of residues in loops, in loops with high B-factor and in REMARK 465 lines of PDB files |
| GlobPlot | http://globplot.embl.de | Amino acid propensity, preference for ordered secondary structure |
| FoldUnfold | http://skuld.protres.ru/~mlobanov/ogu/ogu.cgi | Amino acid propensity based on contact numbers |
| FoldIndex | http://bip.weizmann.ac.il/fldbin/findex | Uses the combination of amino acid propensity of net charge and hydrophobicity calculated with a sliding window |
| Predictprotein (UCON, NORSp, MD) | http://ppopen.rostlab.org/ | Various machine learning methods |
| RONN | https://app.strubi.ox.ac.uk/RONN/ | Bio-basis Function Neural Network that recognizes similarity to known disordered segments |

(continued)

**Table 6.1** (continued)

| Name | Web URL | Description |
|---|---|---|
| ESpritz | http://biocomp.bio.unipd.it/espritz/ | Bidirectional recursive neural networks and trained on three different flavours of disorder |
| MFDp | http://biomine-ws.ece.ualberta.ca/MFDp2/index.php | Meta server for prediction of disordered protein |
| OnD-CRF | http://babel.ucmp.umu.se/ond-crf/ | Conditional random fields based prediction using features generated from the amino acids sequence and from secondary structure prediction |
| Genesilico-Metadisorder | http://genesilico.pl/metadisorder/ | Meta server trained on PDB REMARK 465 lines, CASP7 and Disprot datasets; incorporates fold recognition method |
| *Prediction of IDP functional modules* | | |
| ANCHOR | http://anchor.enzim.hu | Prediction of disordered binding regions based on estimated energies |
| DISOPRED3 | http://bioinf.cs.ucl.ac.uk/disopred | Machine learning method for disordered binding regions |
| iELM | http://i.elm.eu.org/search/ | Prediction of matches to the regular expression of known linear motifs |
| MORFPRED | http://biomine-ws.ece.ualberta.ca/MoRFpred/index.html | Machine learning method that combines features provide information about evolutionary profiles, selected physiochemical properties of amino acids, and predicted disorder, solvent accessibility and B-factors |
| DILIMOT | http://dilimot.russelllab.org/ | *De novo* linear motif discovery based on enrichment |
| SLiMSuite | http://www.slimsuite.unsw.edu.au/software.php | Various tools for linear motif discovery and searches |
| *Conservation analysis* | | |
| DisCons | http://pedb.vib.be/discons/ | Tool to classify residues based on the combined conservation scores of the sequence and of the disorder propensity |

The table lists publicly available computational resources for IDPs. It includes databases, prediction methods for disordered regions and their functional modules, their URL addresses, and their descriptions. Further details on the predictors are found in the text, and in references (He et al. 2009; Dosztanyi et al. 2010)

tendency for a given amino acid to be in a region of coil versus a regular secondary structure and delineate ordered and disordered regions based on the calculated measure. A specific amino acid propensity scale was also optimized for discriminating ordered and disordered segments (Campen et al. 2008).

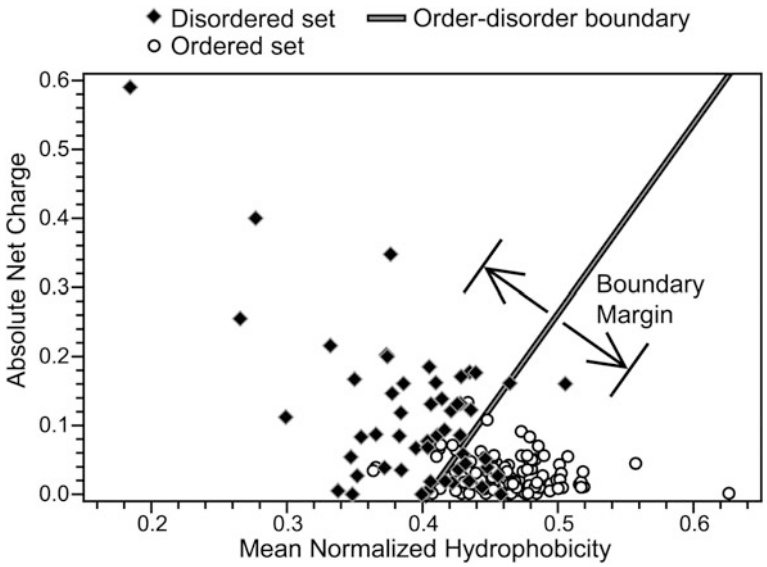



**Fig. 6.1** Charge-hydropathy plot of protein disorder. Net charge versus mean hydrophobicity has been plotted for intrinsically disordered (*full diamond*) and ordered (*empty circle*) proteins. The two are separated by a *straight line* <charge> = 2.743 <hydropathy> − 1.109, with *arrows pointing* to the *lines* delimiting the zone with a prediction accuracy of 95% for disordered proteins and 97% of ordered proteins, at the expense of discarding 50% of all proteins (adapted with permission from Oldfield 2005. Copyright 2005 American Chemical Society)

### *6.3.3 Prediction Based on Simplified Biophysical Models*

Some predictors operate based on the idea that IDPs cannot fold because their amino acids cannot make sufficient inter-residue interactions to overcome the unfavorable decrease in entropy accompanying folding. There are several predictors based on this principle, that apply simple statistical principles [FoldUnfold (Galzitskaya et al. 2006)], based on contact predictions [Ucon (Schlessinger et al. 2007)], or estimate the total inter-residue interaction energy of a chain [IUPred (Dosztanyi et al. 2005a, b)]. This latter is described in some detail.

To estimate the total pair-wise interaction energy realized by a polypeptide chain, IUPred uses low-resolution force fields (statistical potentials) derived from globular proteins. The underlying idea is that the contribution of a residue depends not only on its type, but also on other amino acids, i.e. its potential partners, in the sequence. Because a probabilistic treatment of the potential interactions of all residues with all others is not tractable, the problem is simplified by a quadratic expression in the amino acid composition. The contribution of an amino acid is approximated by an energy predictor matrix, which relates the energy contribution of amino acid $i$ to that of amino acid $j$. The parameters of the matrix are determined by least squares fitting to actual globular proteins. By this approach, the average

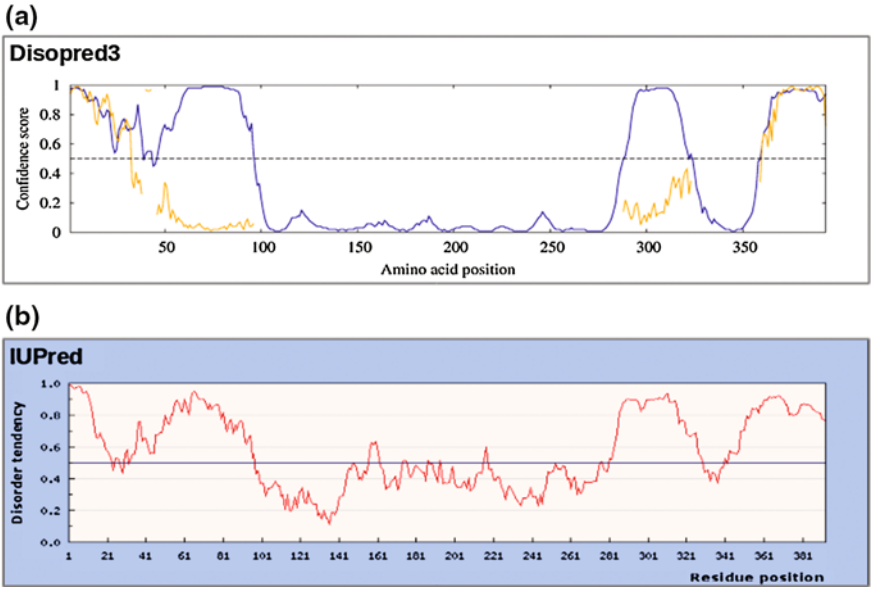



**Fig. 6.2** Plots of predicted disorder for p53. Disorder of the tumor suppressor p53 has been predicted by the **a** DISOPRED3 (Jones and Cozzetto 2015) and **b** IUPred (Dosztanyi et al. 2005a) methods. Residues with scores above 0.5 are predicted disordered, while below 0.5 residues are predicted ordered. The predictions are basically in agreement with biophysical data that suggest disorder within the N-terminal and the C-terminal regions, while the central tetramerization domain is predicted to be more ordered. For DISOPRED3, the *orange line* indicates predicted disordered *binding* regions within the predicted disordered regions plotted with *blue*. In the presented case only the very terminal regions are predicted as disordered *binding* regions, based on their scores above 0.5

energy level of disordered proteins (−0.07 arbitrary units) is significantly more unfavorable than that of globular proteins (−0.81 arbitrary units), which suggests that the approach is informative on the gross structural status of proteins. When only a pre-defined local sequential neighborhood is considered in the calculations, the approach provides sequence-specific information on disorder, forming the basis of IUPred disorder prediction method (Fig. 6.2b).

### *6.3.4 Machine Learning Algorithms*

The prediction of protein disorder is basically a simple binary classification problem that can be approached using standard machine learning (ML) algorithms. Compared to the previous simpler approaches, these methods can incorporate non-trivial amino acid features and deduce hidden sequence properties, which can lead to superior performance. At the same time, their correct prediction often does

not rely on known principles, and thus they do not add to our understanding of what defines disorder. Furthermore, ML approaches are more prone to be biased to our current limited collection of disordered protein examples.

One of the most commonly applied ML techniques is artificial neural network (ANN). ANN is a computational model that was inspired by the learning process of the brain and uses a system of weighted connections. During training, the weights are optimized in such a way that correct relationship between input data (e.g. sequence features) and outputs (e.g. order or disorder category) is recognized. Another commonly applied technique uses the support vector machine (SVM) algorithm. This method searches for a hyperplane in a feature space that separates ordered and disordered proteins. The hyperplane may either be linear on non-linear. It can also take into account unbalanced class frequencies of data, which is the typical case in the prediction of order and disorder. Other types of machine learning approaches have also been exploited for the prediction of protein disorder. These include radial basis function network (Su et al. 2006), biobasis function neural networks (Yang et al. 2005), recurrent function neural networks (Hecker et al. 2008), or conditional random fields (Wang and Sauer 2008). The main reason to use these more advanced ML techniques is to capture some of the hidden, higher order sequence dependences of protein disorder.

The group of Keith Dunker developed a family of predictors termed PONDR (predictor of natural disordered regions), including the first method in the field, VL-XT. The VL-XT method relies on three separate ANNs, one trained specifically for the N-terminal region, one for the C-terminal regions and another for the middle regions of variously characterized long disordered regions. As an input it uses local amino acid composition, flexibility and other sequence features (Li et al. 1999). This method is still in use today, as it was suggested that it can recognize regions that are likely to serve as recognition motifs (Iakoucheva et al. 2002). The VL2 method was aimed at directly capturing various flavours of IDPs (Vucetic et al. 2003). The authors clustered 145 IDPs by setting up competition among increasing numbers of predictors, with the criterion of prediction accuracy used to partition individual proteins. The resulting three groups appeared to have only weakly discernible functional associations. Another predictor, VL3 is also based neural networks but it was trained on a much larger dataset. The VSL2 method represented a significant step forward in the field of disorder prediction methods (Peng et al. 2006) as it aimed to give equally good performance for both short and longer disordered segments using a combination of SVMs trained specifically on short and long disordered regions. Because short disordered regions are context dependent, i.e. their lack of structure depends on their structural environment, whereas disorder of long regions stands on its own, this combined approach resulted in one of the most powerful algorithms of disorder prediction. A more recent method in this series of predictors, PONDR-FIT, is a meta-approach. This is a commonly applied approach that takes the output of several individual prediction methods, and combines them into a single prediction in order to achieve improved performance (Xue et al. 2010a).

Another family of predictors was developed by the group of David Jones. The original version of DISOPRED relied on a neural network based approach that was adopted from the prediction of secondary structure elements (Jones and Ward 2003). It incorporated sequence profiles generated by PSI-BLAST as an input at the expense of increased computational cost. It was found later that using linear SVMs could achieve better performance, as long as the training sets included only high resolution data (Ward et al. 2004). For several years, DISORPED2 was one of the best methods for the prediction of missing residues in X-ray structures. At the same time, it had a clear tendency to under-predict long disordered regions. The latest incarnation in this series, DISOPRED3, was developed to tackle this issue (Jones and Cozzetto 2015). The authors returned to the original neural network based DISOPRED method, but retrained it on data rich in long disordered regions. They also developed a nearest neighbour classifier, and together with the SVM based predictor, the predictions from these components were feed into an additional module. The final prediction of this new method was shown to be more specific than its predecessor and produced more accurate predictions across different IDR lengths and positions along the sequence. DISOPRED3 also predicts disordered binding regions (see Sect. 6.7.2) (Fig. 6.2).

There are several additional disorder prediction methods (He et al. 2009; Dosztanyi et al. 2010). Many of them were specifically trained either on long disordered regions, or missing residues of X-ray structures. More recent methods are predominantly meta-predictors, which integrate the output of independent tools (Kozlowski and Bujnicki 2012; Disfani et al. 2012). A list of publicly available disorder prediction methods is given in Table 6.1.

### *6.3.5 Related Approaches for the Prediction of Protein Disorder*

As shown by the aforementioned studies, low sequence complexity differs from disorder, yet prediction of low complexity regions can be considered as a first reasonable approach to assessing disorder, or at least the lack of globularity. The entropy function of Shannon, adapted to the case of protein sequences (Wootton 1994; Wootton and Federhen 1996) forms the basis of the SEG program routinely used to identify sequentially biased fragments of low compositional complexity measures. This practice has a definite value in delineating non-globular regions of proteins.

A different approach relies on the prediction of regular secondary structural elements (α-helix, β-strand). The underlying assumption is that long regions (>70 consecutive amino acids) devoid of predicted regular secondary structure are structurally disordered (Liu and Rost 2003). Whereas the predictor, NORSp, clearly offers an orthogonal approach to disorder predictors, it should be noted that there are well-ordered proteins composed entirely of non-repetitive local structural

elements [termed loopy proteins (Liu et al. 2002)]. Disordered regions which can undergo disorder-to-order transition are also often predicted to contain regular secondary structure elements, which generally correspond to their structure adopted in the complex, some of them present even in the disordered state as transient local structural elements (Fuxreiter et al. 2004). Therefore, predicted secondary structure elements can be compatible with disordered regions.

Another concept that is strongly related to disorder is flexibility. While disorder is inherent to only a subset of proteins, all proteins possess flexibility and are in constant motion. This structural flexibility can be characterized by B-values derived from experimental data for structures determined by X-ray crystallography. Methods that were specifically trained to recognize flexible residues (i.e., residues with high normalized B-values) can capture some aspects of disorder and represent an orthogonal approach for the identification of structural disorder. This can be exploited in meta-predictors to improve the predictions for experimentally characterized disordered segments (Schlessinger et al. 2009). A more direct relationship between backbone dynamics and protein disorder was suggested recently based on backbone $S^2$ order parameters (Cilia et al. 2013). NMR chemical shifts provide information on local fast dynamics of the backbone up to the microsecond and low millisecond range and are closely linked to $S^2$ order parameters. Such data are available for a diverse collection of proteins from fully structured to disordered in the Biological Magnetic Resonance Data Bank (BMRB, http://www.bmrb.wisc.edu/). The DynaMine method predicts residue level backbone dynamics from the amino acid sequence in the form of backbone $S^2$ order parameters (Cilia et al. 2013, 2014). The method was trained using a linear regression algorithm, and it takes an input sequence fragment of size 51 and provides the prediction for the central element of the fragment. The predicted values can indicate that a given residue is likely to be rigid, flexible or has highly context-dependent dynamics. Rigid and flexible residues showed a strong correlation with order and disorder, further demonstrating a close connection between disorder and dynamics.

### *6.3.6 Comparison of Disorder Prediction Methods*

Disorder prediction methods have been evaluated in the last six rounds of the critical assessment of structure prediction algorithms (CASP), a biannual, community-wide blind experiment that started in 1994 (Melamud and Moult 2003; Monastyrskyy et al. 2014). CASP motivated several groups to develop their own methods and established standard evaluation criteria. The most commonly used measures for evaluation are Matthews correlation coefficient (MCC), balanced accuracy (ACC) and area under ROC curve (AUC) (Monastyrskyy et al. 2011, 2014). These measures usually ensure a balanced assessment of sensitivity and specificity. In the CASP10 dataset, the best performing groups achieved AUC above 0.9, MCC score above 0.5 and ACC score around 0.75 (Monastyrskyy et al. 2014). It should be noted that the CASP target selection procedure is mainly aimed

at those sequences for which the structure is expected to be released by the end of 3 month prediction stage. This criterion inevitably selects against disordered residues, especially longer IDRs, which can hinder structure determination efforts. A different benchmark dataset can also be created using from both PDB and DisProt entries (Mizianty et al. 2010). While a good, unbiased, estimate of the performance of disorder prediction methods remains a challenging task, as a fair assessment, one might state that the predictors mentioned above perform at a level approaching the best secondary-structure prediction algorithms. To arrive at a dependable assessment of disorder, it is recommended that several predictors based on different principles should be used.

## 6.4 Databases of IDPs

There are several resources that collect experimental and computational annotations on disordered regions in proteins. The Database of Protein Disorder (DisProt) database was developed to enable IDP research by collecting and organizing knowledge regarding the experimental characterization and the functional associations of IDPs (Sickmeier et al. 2007). The latest version of DisProt at the time of writing (6.02, May 2013) contained 694 proteins with 1539 disordered regions. The IDEAL database also collects experimentally verified IDPs with an additional focus on regions that undergo coupled folding and binding upon interaction with other proteins (Fukuchi et al. 2014). IDEAL contains manually curated annotations on IDPs in locations, structures, and functional sites such as protein binding regions and posttranslational modification sites together with references and structural domain assignments. The latest release (Oct 2014) contained 557 disordered protein regions and 203 binding regions.

The MobiDB database provides more comprehensive information about disordered segments by combining experimentally verified disorder annotation with computational predictions (Potenza et al. 2015). The database features three levels of annotation: manually curated, indirect and predicted. Manually curated data is extracted from the DisProt database. Indirect data is inferred from PDB structures as missing residues in X-ray structures and mobile regions in NMR structures. Currently the predictions from 10 methods are included (three ESpritz flavours, two IUPred flavours, two DisEMBL flavours, GlobPlot, VSL2b and JRONN) to enable MobiDB to provide disorder annotations for every protein in absence of more reliable data. Its most up-to-date version (July 2014) contains intrinsic disorder annotations for 80,370,243 Uniprot entries. The Database of Disorder Protein Prediction ($D^2P^2$) stores pre-computed disorder predictions made by 9 different methods for proteins from completely sequenced genomes (Oates et al. 2013). Complementing disorder predictions, the database contains information on disordered binding regions, PTM sites and domains. Currently (as of Apr 2015), it holds 10,429,761 sequences in 1765 genomes.

## 6.5 Structural Features of IDPs

After recognizing the basic properties of IDPs that set them apart from globular proteins with well-defined structure, the next big challenge is to characterize the various conformational states of IDPs and to understand how these are related to function. The common property of IDPs is that they fluctuate rapidly over an ensemble of conformations in their native unbound state. At a closer look, however, the conformational states of IDPs show significant heterogeneity. These can be characterized using various experimental techniques, providing detailed information in terms of apparent molecular dimension and shape, presence of transient local structural elements or transient long-range contacts. Based on various observations, proteins have been proposed to exist along a continuum of conformational states that cover the spectrum of tightly folded domains that display either no disorder or only local disorder in loops and tails, compact molten globules containing extensive secondary structure, unfolded states that transiently populate local elements of secondary structure, and highly extended states that resemble statistical coils (Dyson and Wright 2005). In this model, there are no boundaries between the described states and native proteins could appear anywhere within this continuous landscape. In contrast, the protein quartet model distinguishes four types of conformational states with increasing amount of compactness and secondary structure content: random-coil, pre-molten globule, molten globule and folded states (Uversky 2002). These states correspond to the conformational average that is formed by an ensemble of individual conformations which can be located anywhere along the structural continuum.

The dynamic nature of IDPs is best modeled by statistical approaches that describe the probabilities of individual conformations in the ensemble (van der Lee et al. 2014). The focus of several recent studies was to generate a pool of conformations that satisfy various experimental constraints (Fisher and Stultz 2011; Mittag and Forman-Kay 2007). As the number of degrees of freedom is much greater than can be determined with available experimental measurements, the ensemble descriptions of IDPs are highly underdetermined with several ensembles fitting the data equally well. In order to reduce this ambiguity, various experimentally measurable constraints are combined into a single amino acid-specific ensemble description. Measurements typically involve NMR chemical shifts and residual dipolar couplings that predominantly report on local order, and paramagnetic relaxation enhancements and SAXS that mainly report on transient intrachain contacts and sampling of the volume space by the unfolded chain (Jensen et al. 2014). The conformational pool can be generated by a statistical coil generator, or molecular dynamics simulations. This approach was applied to characterize Tau and α-synuclein, two intrinsically disordered amyloidogenic polypeptides involved in human neurodegenerative disease (Schwalbe et al. 2014). The resulting ensembles could predict independent experimental observations and suggested local conformational features potentially involved in function and disease. In order to aid further research in this area, the Protein Ensemble Database was launched

(Varadi et al. 2014). This database collects the best fit conformational ensembles together with their experimental constraints and currently contains 16 proteins with 25 ensembles.

The structural characteristics and populations of individual states in the conformational ensemble of IDPs are determined by the nature of the amino acids and their distribution in the sequence. Disordered regions are characterized by distinct compositional biases and they are in general depleted in canonical hydrophobic residues and enriched in polar and charged residues (Uversky 2013). These biases lead to a weakened hydrophobic effect that makes IDPs unable to fold independently. However, IDRs can be categorized into three further compositional classes that reflect the fraction of charged versus polar residues: polar tracts, polyelectrolytes and polyampholytes (Mao et al. 2013). In polar tracts polar residues are dominant at the expense of hydrophobic, charged and proline residues; polyelectrolytes have an excess of either positively or negatively charged residues; while polyampholytes also contain a large number of charged residues but the number of opposite charges is comparable. The balance between solvent mediated intra-chain attractions versus repulsions determines the types of conformations that make up the ensemble that is thermodynamically accessible to an IDP sequence (Das et al. 2015). Accordingly, polar tracts usually form compact globules that are largely devoid of significant secondary structural elements, while strong polyelectrolytes form expanded coil-like structures. The molecular dimension of the ensemble of IDPs is also influenced by the distribution of charged residues. If oppositely charged residues are segregated in the linear sequence, then the oppositely charged blocks can form hairpins or globular conformations. Sequences with well-mixed oppositely charged residues adopt random coil or globular conformations depending on the total charge. The analysis of curated disordered segments from the DisProt database suggested that a majority of IDPs have amino acid compositions that predispose them to form globules or chimeras of globules and coils (Das et al. 2015). This latter category includes IDPs that can undergo folding upon binding, while more swollen random coil-like conformations can help to improve the solubility. These biophysical principles can help to understand how the detailed properties of the structural ensembles of IDPs are related to their function (van der Lee et al. 2014).

## 6.6 Functional Classification of IDPs

Predicting function of IDPs is even more challenging than predicting their structure for several reasons. First, IDPs evolve very fast, and even though their structural state as such is often preserved, there is very little information on how much their functions change. Another reason is that the functional classification of proteins/genes is usually done at the level of the whole gene, and it is often very obscure in what way and to what extent disorder of a segment (IDR) contributes to

these. In addition, in many cases the functions of IDPs cannot be incorporated into functional classification schemes developed for ordered proteins. The area of functional classification of IDPs witnesses immense activity, which has so far resulted in two fundamentally different approaches to classification. Key aspects of these are reviewed next.

### 6.6.1 *Gene Ontology-Based Functional Classification of IDPs*

In several studies the prevalence of disorder in functional classes of proteins has been addressed (Iakoucheva et al. 2002; Ward et al. 2004; Tompa et al. 2006; Xie et al. 2007). These are usually based on the Gene Ontology (GO) scheme (Ashburner et al. 2000), and have addressed the prevalence of disorder in all three ontologies, namely molecular function (MF), biological process (BP) and cellular localization (CL). In practically complete agreement, different works suggest that the frequency of disorder is higher in eukaryotes than in prokaryotes, and that disorder is common in regulatory and signaling functions. In terms of MF, the highest levels of disorder appear in categories such as transcription regulation, protein kinase, transcription factor, DNA binding, whereas it is lowest in oxidoreductase, catalytic, ligase, structural molecule categories. In terms of BP, categories such as development, protein phosphorylation, regulation of transcription, and signal transduction have the highest level of disorder, whereas it appears infrequently in biosynthesis and energy pathways. With respect to localization, it prevails in nuclear, cytoskeletal and chromosomal proteins, for example, with low levels in mitochondrial, cytoplasmic and membrane proteins.

When prediction is focused on long disordered regions, thought to be functionally significant (Xie et al. 2007), similar observations have been made. When SwissProt BP key-words were analyzed, significant positive (e.g. differentiation, transcription, transcription regulation), and negative (e.g. biosynthesis, transport, electron transport, glycolysis) correlations with disorder were found (Xie et al. 2007). When MF keywords were analyzed, most positively correlated were ribonucleoprotein, ribosomal protein, developmental protein, whereas negatively correlated were oxidoreductase, transferase, lyase, and hydrolase classes. In terms of 710 functional SwissProt key-words, 238 were in strongly positive, whereas 302 in strongly negative, correlation with disorder; 170 keywords were ambiguous.

Taken together, all the pertinent studies agree that proteins of regulatory functions are positively correlated with disorder, whereas proteins with catalytic functions are negatively associated.

### 6.6.2 *Classification of IDPs Based on Their Mechanism of Action*

In another system taking the molecular mechanisms of IDPs into consideration, disordered proteins have been classified into five (Tompa 2002) and later into six categories (Tompa 2005). Further observations suggested the addition of prion proteins as an additional category (Pierce et al. 2005). This classification scheme (Table 6.2) can accommodate all distinct modes of IDP/IDR actions described thus far.

**Table 6.2** Classification scheme of IDPs

| Protein | Partner | Function |
|---|---|---|
| *Entropic chains* | | |
| Nup2p FG repeat region | n.a. | Gating in NPC |
| K channel N-terminal region | n.a. | Timing of gate inactivation |
| *Display sites* | | |
| CREB KID | PKA | Phosphorylation site |
| Cyclin B N-terminal domain | E3 ubiquitin ligase | Ubiquitination site |
| *Chaperones* | | |
| ERD 10/14 | (e.g.) Luciferase | Prevention of aggregation |
| hnRNP A1 | (e.g.) DNA | Strand re-annealing |
| *Effectors* | | |
| p27Kip1 | CycA-Cdk2 | Inhibition of cell-cycle |
| Securin | Separase | Inhibition of anaphase |
| *Assemblers* | | |
| RNAP II CTD | mRNA maturation factors | Regulation of mRNA maturation |
| CREB | p300/CBP | Initiation of transcription |
| *Scavengers* | | |
| Casein | Calcium phosphate | Stabilization of calcium phosphate in milk |
| Salivary PRPs | Tannin | Neutralization of plant tannins |
| *Prions* | | |
| Ure2p | | Utilization of urea under nitrogen |
| Sup35p | NusA, mRNA | Suppression of stop codon, translation readthrough |

Classification of IDPs encompassing seven functional categories based on their molecular modes of action. Two examples within each category are given, specifying the binding partner (if applicable) and the actual cellular function of the protein

#### 6.6.2.1 Entropic Chains

The first functional category, unique to disordered proteins, is that of *entropic chains*, the function of which does not involve partner recognition, but directly results from disorder. Sub-categories within this class are termed entropic springs, bristles/spacers, linkers, and clocks, and the underlying mechanisms can be best described as either influencing the localization of attached domains, or generating force against movements/structural changes (Dunker et al. 2002). The best characterized examples in this category are entropic gating in nuclear pore complex by disordered regions of NUPs (Elbaum 2006), the entropic spacer/bristle function of projection domains of microtubule-associated proteins in the cytoskeleton (Mukhopadhyay and Hoh 2001), and the entropic spring action of the PEVK region of titin, ensuring passive tension in resting muscle due to its elasticity (Trombitas et al. 1998).

#### 6.6.2.2 Function by Transient Binding

In the other six categories, IDPs function via molecular recognition, i.e. they bind other macromolecule(s) or small ligand(s) either transiently or permanently. *Display sites* are primarily targeted for post-translational modifications. For example, enzymatic modifications require flexible and structurally adaptable regions in proteins, as shown by limited proteolysis, which occurs in linker regions in globular proteins (Fontana et al. 1997). Phosphorylation (Iakoucheva et al. 2004), ubiquitination (Cox et al. 2002) and deacetylation (Khan and Lewis 2005) also preferentially occur in locally disordered regions. The general correlation of disorder with such sites has been demonstrated by predicting disorder in proteins that contain short recognition elements [also known as linear motifs (Puntervoll et al. 2003)]. It was found that linear motifs preferentially reside in locally disordered sequential environments within the parent protein (Fuxreiter et al. 2007).

Another category of IDPs functioning by transient binding is *chaperones*, as suggested in a statistical analysis on the level of disorder in protein- and RNA chaperones (Tompa and Csermely 2004). RNA chaperones have a very high proportion of disorder (40% of their residues fall into long disordered regions), and protein chaperones also tend to be among the most disordered proteins (15% of their residues are located within long disordered regions). Because disordered regions are often directly involved in chaperone function, an "entropy transfer" model of structural disorder in chaperone function could be formulated (Tompa and Csermely 2004). Implications of this model were verified by observations of fully disordered chaperone proteins (Kovacs et al. 2008).

#### 6.6.2.3 Functions by Permanent Binding

In the other four categories IDPs/IDRs function by permanent partner binding. Proteins termed *effectors* bind and modify the activity of their partner, primarily an enzyme (Tompa 2002). Several IDPs characterized in great detail, such as p27Kip1, the inhibitor of Cdks (Kriwacki et al. 1996; Lacy et al. 2004), securin, the inhibitor of separase (Waizenegger et al. 2002) and calpastatin, the inhibitor of calpain (Kiss et al. 2008a, b), belong here. Interestingly, such effectors sometimes have the potential to both inhibit and activate their partners, as shown for p27Kip1 (Olashaw et al. 2004), or the C fragment of DHPR II-III loop (Haarmann et al. 2003). These and other observations have led to the concept of the involvement of structural disorder in multiple, sometimes opposing, activities of proteins, i.e. moonlighting (Tompa et al. 2005).

The next category of IDPs functioning by permanent partner binding is that of *assemblers*, which either target the activity of attached domains, or assemble multi-protein complexes (Tompa 2002). A high level of disorder in some scaffolding proteins, such as BRCA1 and Ste5 (Mark et al. 2005; Bhattacharyya et al. 2006), an increased level of disorder in hub proteins of the interactome (Dosztanyi et al. 2006; Haynes et al. 2006; Patil and Nakamura 2006), and the correlation of the average level of disorder with the number of partners in multi-protein complexes (Hegyi et al. 2007) attest to the generality of this relation.

In the third class within this category, *scavengers*, there are disordered proteins which store and/or neutralize small ligand molecules. Milk nutrient casein(s), for example, also function as calcium phosphate stores in milk, enabling a high total calcium phosphate concentration (Holt et al. 1996). They might be representative of all, highly disordered, proteins involved in biomineralization (Kalmar et al. 2012).

The final functional category of IDPs is that of *prions*, not included in previous classification schemes (Tompa 2002, 2005). Prions have been traditionally considered as pathogens, mostly because of their causal association with "mad cow diseases" (Prusiner 1998). Many papers, however, showed that the autocatalytic conformational change underlying the prion phenomenon also occurs in the normal physiological functions of proteins of yeast (Tuite and Koloteva-Levin 2004), or even higher organisms, such as *Drosophila melanogaster* (Si et al. 2003a, b; Fowler et al. 2007). These prion proteins have disordered Q/N-rich prion domains (Pierce et al. 2005), primarily responsible for the autocatalytic conformational transition that has functional consequences on neighbouring domains.

### 6.6.3 *Functional Features of IDPs*

There are several different but interconnected concepts in this area that emphasize different structural or sequential aspects with slightly different implications for function. These include short linear motifs, disordered regions that can undergo disorder-to-order transition upon binding, and disordered domains.

#### 6.6.3.1 Short Linear motifs

The analysis of sequences involved in protein-protein interactions has suggested that in certain proteins the element of recognition is a short motif of discernible conservation, often denoted as a "consensus" sequence, such as those involved in modification by kinases or binding by SH3 domains (Neduva et al. 2005; Van Roey et al. 2014). These functional elements are constructed as a few conserved specificity determinant residues interspersed within largely variable residues, with a typical length between 5 and 25 residues and usually located within locally disordered regions (Fuxreiter et al. 2007). The consensus motif can be captured by a regular expression. These functional modules are often termed linear motifs (LMs), also denoted as eukaryotic linear motifs (ELMs), short linear motifs (SLiMs) or MiniMotifs (Diella et al. 2008; Davey et al. 2012b; Mi et al. 2012). They primarily bind to globular proteins and form small compact binding surfaces that result in low affinity interactions. Due to their small size, LMs enable both high functional diversity and functional density to polypeptide segments that contain them. They can also evolve rapidly, and emerge convergently in unrelated proteins, conferring evolutionary plasticity on the interactome (Diella et al. 2008; Davey et al. 2012b). While LMs play essential roles in the regulation of dynamic cellular processes, they can also be hijacked by pathogenic viruses and bacteria that evolved to mimic these linear motifs (Davey et al. 2011).

Linear motifs can be broadly divided into two major classes: modifications sites, which are recognized and altered by modifying enzymes and binding sites, which mediate interactions with globular domains (Van Roey et al. 2014). Modification sites can be further classified into proteolytic cleavage sites (e.g. caspase sites), structural modification sites (e.g. peptidylprolyl cis-trans isomerase) and post-translational removal or addition sites (e.g. phosphorylation sites). Binding motifs include well-known classes of motifs, such as the C-terminal motifs that bind PDZ domain, or the proline-rich PxxP motifs that interact with SH3 (Src homology 3) domains. The main function of these motifs is to promote complex formation. Another category of ligand binding sites correspond to docking motifs. These sites increase the specificity and efficiency of modification events. Docking motifs are usually distinct from the actual modification sites but are located on the same protein. Example docking motifs are KEN box and D box degrons that act as recognition surfaces for ubiquitin ligases and play a role in the degradation of the given protein. Another important class of ligand binding linear motifs is targeting motifs. Targeting motifs can direct proteins into specific subcellular localization or act as traffic proteins that ensure that their cargo is delivered to the right location (Van Roey et al. 2014).

Linear motifs guide proteins through their life (Tompa et al. 2014). They regulate and coordinate the processing, localization and degradation of almost all proteins. Many proteins contain several distinct motifs, including both binding and modification sites that can form molecular switches (Van Roey et al. 2013). Common mechanisms involve a PTM inside or in the immediate flanking region of a binding motif, or adjacent or overlapping binding motifs that can also function

competitively or cooperatively. The complex interplay between multiple binding and/or modification sites is crucial for the information processing and creation of dynamic signaling networks (Van Roey et al. 2012). Many components of these networks, however, are not yet known. While the ELM database currently catalogues around 2000 motif instances, the estimated number of motifs in the human proteome is around a million (Tompa et al. 2014).

#### 6.6.3.2 Disordered Binding Regions/Molecular Recognition Features

A different type of functional module located within IDPs involved in protein-protein interactions can be identified by their ability to undergo disorder-to-order transition upon binding (Vacic et al. 2007; Meszaros et al. 2007). These functional elements are called disordered binding regions or molecular recognition features (MoRFs). The PDB database contains several examples of segments that are disordered in isolation but adopt a well-defined conformation in complex (Meszaros et al. 2007). The length of these functional binding regions is typically between 10 and 70 residues, shorter than that of globular domains (Vacic et al. 2007). Nevertheless, this criterion in itself is not sufficient to identify disordered binding regions; the disorder status in the unbound form has to be verified. Based on the analysis of examples of disordered regions bound to a globular proteins, they can be categorized into α-MoRFs, β-MoRFs, ι-MoRFs, and mixed MoRFs depending on the dominant secondary structure element in the complexed form (Vacic et al. 2007). It was suggested that the unbound form of these MoRFs is biased towards the conformation they adopt in the complex and this can influence the kinetic and thermodynamic properties of the binding (Fuxreiter et al. 2004). However, the structure of the MoRFs in general is heavily shaped by their partner. An extreme case of this behavior is showcased in the C-terminal region of p53 where a short segment can adopt four different types of conformation depending on the partner (Hsu et al. 2013). MoRFs can not only show significant binding plasticity, they can also retain significant disorder even in their bound form. Fuzziness is a concept of disorder in the bound state of IDPs. One manifestation of this phenomenon is when binding occurs without the acquisition of a single dominant structure, instead involving multiple states, and thus may be considered as polymorphism in the bound state (Tompa and Fuxreiter 2008). This has been observed in the case of T-cell factor 4 (Tcf4) binding to β-catenin (Graham et al. 2001) and nuclear localization signal (NLS) to α-importin (Fontes et al. 2000). This type of fuzziness is another type of mechanism that can help to fine tune functional properties of IDPs (Tompa and Fuxreiter 2008).

#### 6.6.3.3 Intrinsically Disordered Domains

Functional modules are often identified based on their evolutionary conservation. Although IDRs in general are less conserved (Brown et al. 2011), structural

disorder is not completely opposed to conservation, as certain disordered regions appear to be evolutionarily conserved (Chen et al. 2006a, b). In almost 30% of the domain families collected in the Pfam database, there are at least 20 consecutive amino acids predicted to be disordered and with significant conservation (Chen et al. 2006a, b). Furthermore, 14% of Pfam domains have more than 50% of their residues predicted disordered. These intrinsically disordered domains (IDDs) include experimentally verified disordered segments, such as WH2 of actin-binding proteins and the KID domain of CDK inhibitors (Tompa et al. 2009). IDDs are involved in a variety of functions, which usually coincide with the general functional preferences of IDPs, such as DNA/RNA binding, ribosome structure, protein binding (both signalling/regulation and complex formation) (Chen et al. 2006a, b). One feature that could be specific to IDDs is that some of them can also bind other IDRs or IDDs by mutually induced folding (Demarest et al. 2002). IDDs can also co-occur with specific protein domains and particular combinations of domains have been observed in the cases of receptors and ion-channels, and in proteins involved in binding and regulation (Pentony and Jones 2010).

The definition of functional features located within IDPs captures different aspects: the key amino acids in the function in the case of LMs, their ability to undergo a disorder-to-order transition during molecular recognition in the case MoRFs, and sequential conservation in the case of IDDs (van der Lee et al. 2014). Their characteristic length is also different. Despite the differences, these functional modules also share many common features (Meszaros et al. 2012). While MoRFs and IDDs are disordered by definition, 80% of LMs are also located in disordered regions (Fuxreiter et al. 2007). These concepts share functional similarities and show a tendency to be involved in molecular recognition to promote complex formation, primarily in various signaling and regulatory functions (van der Lee et al. 2014). An example for the overlap of the various types of functional modules is showcased for the fully disordered human p27Kip1 that contains a conserved PFAM family, an almost 70 residue long region that can undergo a disorder-to-order transition upon binding to the complex of the phosphorylated cyclin A-cyclin-dependent kinase 2 (Cdk2) (Russo et al. 1996) and a linear motif that is shared by several other protein binding to cyclin protein, according to ELM (Dinkel et al. 2014) (Fig. 6.3).

## 6.7 Prediction of the Function of IDPs

As suggested by the foregoing considerations, reliable all-round prediction of the functions of IDPs is still a long way off, and we have only taken the first steps towards this goal. As discussed in the next section, there are several approaches that may shed some light on the function of an IDP not yet experimentally characterized. Sequence-based prediction of short LMs by a variety of algorithms (Davey et al. 2006; Neduva et al. 2005), prediction of MoRFs or disordered binding regions in IDPs/IDRs (Meszaros et al. 2009; Vacic et al. 2007), and combination of

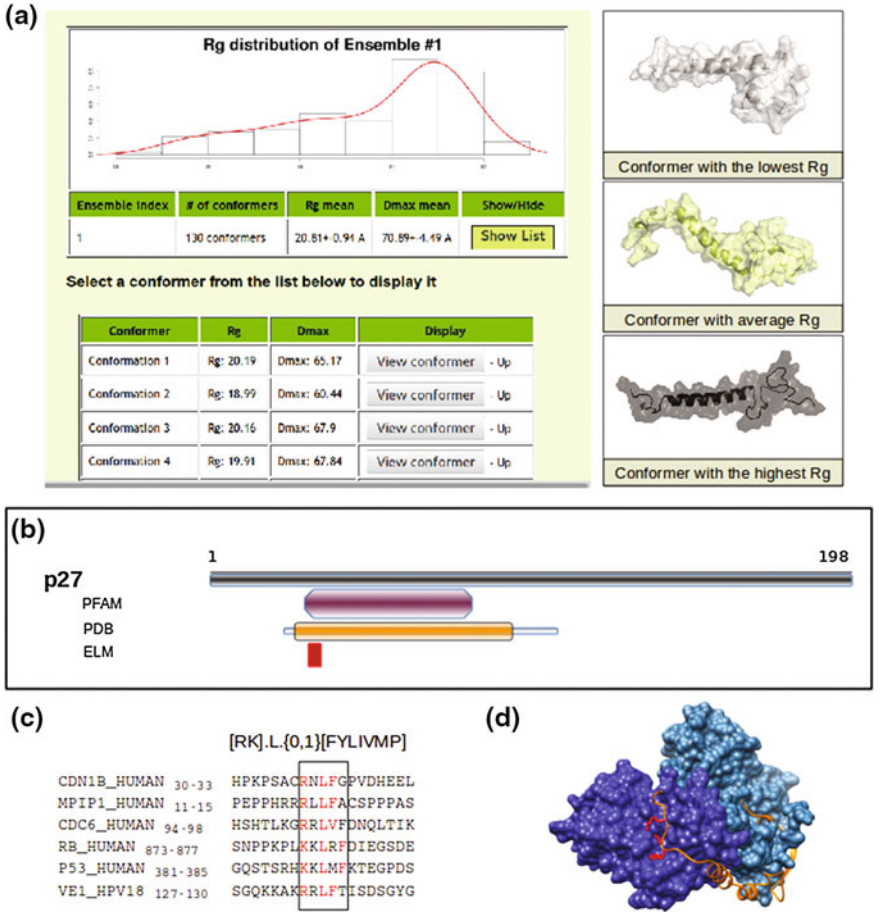



**Fig. 6.3** Various representations of disorder and disordered binding regions of p27Kip1. **a** The distribution of the radius of gyration for the ensemble of conformations and selected conformers for human p27Kip1 obtained using molecular dynamics simulation deposited into the pE-DB database (Varadi et al. 2014). **b** The overlap of the different types of functional modules for the p27Kip1 that contain a conserved PFAM family (Finn et al. 2014) PF02234 (*purple box*), a disordered binding region (*orange box*) and linear motif (*red box*) according to the ELM database (Dinkel et al. 2014). **c** The motif definition and UNIPROT ID and sequence regions of representative instances for the occurrence for the DOC_CYCLIN_1 motif, including p27Kip1 (CDN1B_HUMAN). **d** The complex of p27Kip1 kinase inhibitory domain bound to the phosphorylated cyclin A-cyclin-dependent kinase 2 (Cdk2) (PDB code: 1jsu) (Russo et al. 1996). P27Kip1 is indicated with *orange ribbon* and the amino acids matching the linear motif are represented by *red sticks*. Cdk2 subunits are shown in a space-filling representation in different *shades of blue*

sequence information with disorder (Iakoucheva et al. 2004; Radivojac et al. 2006) and taking advantage of functional correlation of the global pattern of disorder (Lobley et al. 2007) are reasonable approaches to assess the function of an unknown piece of disordered protein.

### 6.7.1 Predicting Short Recognition Motifs in IDRs

Currently, the most comprehensive resource for linear motifs is the ELM database (Dinkel et al. 2014). It contains around 200 motif classes with over 2400 experimentally validated instances with in-depth manual annotation. A similar resource is the MiniMotif database, although its annotations are not publicly available (Mi et al. 2012). The known motifs can be used to map functionality onto regions with unknown functions. This approach relies on the regular expression derived from known motifs, which is then used to search protein sequences to find new matches. The main problem in the computational detection of linear motifs is that such matches can occur with very high false positive rate by pure chance (Meszaros et al. 2012). Therefore, it is difficult to identify functionally relevant instances among the randomly occurring nonfunctional hits. To overcome this issue additional filters are usually employed, such as accessibility based on structural models, prediction of intrinsic protein disorder, evolutionary conservation, annotations based on cellular localization or protein-protein interaction data (Via et al. 2009). While these filters can drastically reduce the number of false positive hits, there can be still a significant number of candidates. To define potential functionality, more precise definition of the binding motif or further biological insights are needed.

The de novo discovery of linear motifs aims at identifying putative uncharacterized motifs in protein sequences. One approach seeks to find short sequence elements that are overrepresented in a set a sequences that share a common interaction partner. From the input sequences, regions unlikely to contain instances of linear motifs (globular domains, signal peptides, trans-membrane and coiled-coil regions) are removed. Motifs are then uncovered in the remaining sequences by a pattern-matching algorithm, and ranked according to measures of over-representation. This approach was implemented in the DILIMOT method [DIscovery of LInear MOTifs (Neduva and Russell 2006)], and applied to high-throughput interaction datasets of yeast, fly, worm and human sequences that resulted in the re-discovery of many previously known ELM instances, and also the recognition of novel motifs. Conceptually closely related to DILIMOT is the SliMDisc (Short Linear Motif Discovery) approach (Davey et al. 2006). This method takes advantage of evolutionarily related sequences, but upweights putative motifs that are present in apparently unrelated sequences. Building on the principle of the SLiMDisc algorithm, SLiMFinder and its more recent version, QSLiMFinder, rely on an improved statistical model and a reduced motif search space that can result in an increased sensitivity and specificity for de novo motif discovery (Davey et al. 2010; Edwards et al. 2007; Palopoli et al. 2015).

Linear motifs can also be discovered based on their specific pattern of conservation as well. By looking at their sequence alignments, linear motif sites often appear as islands of conservation among evolutionarily more flexible positions, reflecting a stronger evolutionary constraint on the functionally important positions compared to their generally disordered sequential neighborhood. Methods such SliMPrints or Phylo-HMM have been shown to be able to identify novel linear motifs (Davey et al. 2012a; Nguyen Ba et al. 2012). Nevertheless, approaches of this type are able to recover only around 30% of known linear motifs, and fail if the aligned sequences are either too similar or too diverse.

### *6.7.2 Prediction of Disordered Binding Regions/MoRFs*

MoRFs are short functional motifs involved in partner binding and so their recognition is of predictive value with respect to the function of the parent protein. Several machine learning methods have been developed for their prediction. Early attempts were based on the observations that the location of MoRFs in the sequence is often indicated by short dips in the disorder prediction profiles (Vacic et al. 2007). MoRFPred uses sequence features that provide information about evolutionary profiles, selected physiochemical properties of amino acids, and predicted disorder, solvent accessibility and B-factors (Disfani et al. 2012). These features are combined by an SVM for the predictions and complemented with annotations generated using sequence alignment. MoRFCHiBi is a fast, novel method that combines the outcomes of two SVM models for the prediction of MoRFs (Malhis and Gsponer 2015). The first, SVMS, is designed to extract information from the general contrast in amino acid compositions between MoRFs, their surrounding regions (Flanks), and the remainders of the sequences. The second, SVMT, is used to identify similarities between regions in a query sequence and MoRFs of the training set. The DISOPRED3 method also has a component that predicts MoRFs located within their predicted disordered segments (Jones and Cozzetto 2015). Prediction of disordered binding regions is based on an SVM-based classifier that uses a 15 amino acid long sliding window, which considers sequence profile data, the length and location of the input IDR relative to the whole protein sequence, and the amino acid composition of the window.

In all these approaches, disordered residues not annotated as protein binding were considered to be part of the negative dataset, without taking into account the possibility that there could be other, not yet characterized binding regions also present within the same protein. This assumption is highly conservative and arguably somewhat unrealistic, given the occurrence of IDRs in protein–protein interaction network hubs. A completely different philosophy is behind the ANCHOR method that aims to identify regions in the amino acid sequence that can undergo a disorder-to-order transition upon binding (Meszaros et al. 2009; Dosztanyi et al. 2009). It seeks to find segments that cannot form enough favorable intrachain interactions to fold on their own and are likely to gain stabilizing energy

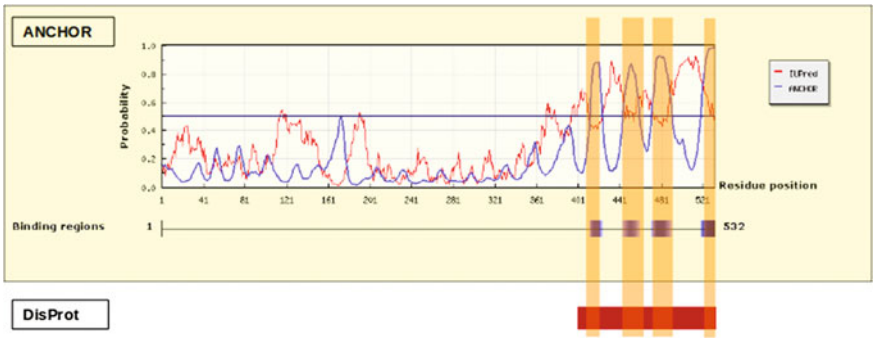



**Fig. 6.4** Prediction of MoRFs/disordered binding regions for the Nucleoprotein from Nipah virus by the ANCHOR method. The output from the ANCHOR server (Dosztanyi et al. 2009) for the nucleoprotein from Nipah virus showing the predicted disorder by the IUPred method (*red line*) and the location predictions of disordered binding regions by ANCHOR (*blue line*) that are also indicated by *blue boxes* underneath the plot. The predictions are in very good agreement with the annotation given in the DISPROT database (DP00697), that assigns disorder status to the region 400–532 indicated by a *red box* below the ANCHOR prediction and MoRFs to four regions: two α-MoRFs at 408–422 and 473–493, an irregular ι-MoRF at 523–532, and a β-MoRF at 444–464 indicated by the *shaded* areas. These regions are based on experimental results using a combination of techniques (Habchi et al. 2010)

by interacting with a globular protein. To evaluate these properties, an energy estimation method similar to IUPred is used. The balance between the various energy terms is determined using a linear regression model optimized on a dataset of known examples. During training, the ratio of the fraction of residues predicted to be in disordered binding regions relative to the number of residues within general disordered segments was minimized and it was not assumed that a given protein contained no other disordered binding region. As a result, ANCHOR has a higher sensitivity, at the expense of lower specificity, compared to the other methods (Disfani et al. 2012). An example for the prediction of ANCHOR for the nucleoprotein from Nipah virus is given in Fig. 6.4.

### *6.7.3 Combination of Information on Sequence and Disorder: Phosphorylation Sites and CaM Binding Motifs*

Prediction of short recognition motifs can be improved by incorporating information on disorder, as demonstrated in the case of phosphorylation sites and calmodulin binding sites (CaMBT) in proteins. Dunker and colleagues have reported (Iakoucheva et al. 2004), by comparing a collection of experimentally determined phosphorylation sites (at Ser, Thr or Tyr) to potential sites that are actually not phosphorylated, that the regions around phosphorylation sites are

significantly enriched in disorder-promoting amino acids, and depleted in order-promoting amino acids (Dunker et al. 2001). By combining the sets of positive examples and the corresponding negative examples and considering local disorder, a predictor of phosphorylation sites could be constructed. DISPHOS (disorder-enhanced phosphorylation predictor) has an improved accuracy over other phosphorylation-site predictor algorithms, such as NetPhos (Blom et al. 1999) and Scansite (Obenauer et al. 2003).

The other thoroughly-studied example is the interaction between calmodulin (CaM) and its binding targets, which involves significant flexibility on both sides. It is known that CaM usually wraps around a helical binding peptide/target (CaMBT) of about 20 amino acids in length (Ikura and Ames 2006). In a comprehensive analysis it has been pointed out that CaM recognition requires disorder of the partner (Radivojac et al. 2006). For example, CaM-dependent enzymes are often stimulated by limited proteolytic digestion (e.g. calcineurin (Manalan and Klee 1983) or cyclic nucleotide phosphodiesterase (Tucker et al. 1981), which suggests local disorder of the binding site. The inclusion of disorder was used for developing a predictor of CaMBTs with an improved performance (Radivojac et al. 2006).

### *6.7.4 Correlation of Disorder Pattern and Function*

Jones and colleagues have taken a direct approach to find association between the global pattern of disorder and the function of a protein (Lobley et al. 2007) described by standard Gene Ontology (GO) categories. It was first found that both location- and length-descriptors of disorder correlate with functional categories associated with signal transduction and transcription regulation. Both molecular function (MF) and biological process (BP) annotations were used. The location descriptors displayed several trends associated with GO categories, such as an elevated level in the middle of the protein in transcription regulator, DNA binding, and RNA pol II transcription factor functions, in the C-terminus in transcription factor activator, transcription factor repressor, and transcription factor or in the N-terminus in potassium channel annotated proteins. Length descriptors showed even more significant associations with function than position descriptors. For example, disordered regions of more than 500 continuous residues are over-represented in transcription-related categories, whereas shorter regions of the order of 50 residues or fewer are over-represented in proteins performing metal ion binding, ion channel, and GTPase regulatory functions. The observed associations could be used to improve prediction of protein function: an SVM predictor applied to 26 GO categories, prediction of 11 BP categories and 12 MF categories showed improvements resulting from the addition of disorder features. In all, disorder adds significantly to the prediction of protein function, with more significant improvements observed in BP than in MF classification.

## 6.8 Evolution of IDPs

IDPs, lacking a well-defined structure, generally have fewer evolutionary constraints and thus tend to evolve faster than globular proteins. The fast evolution of IDPs/IDRs has been directly demonstrated in several protein families by comparing the amino acid replacement rate of disordered and globular regions in protein families, in which both regions are simultaneously present (Brown et al. 2011). Nevertheless, disordered residues display a wide range of evolutionary rates. Using a combination of disorder prediction and multiple sequence alignments, three different scenarios can be discriminated: (i) constrained disorder describes disordered regions that are also highly conserved, (ii) flexible disorder corresponds to regions where disorder tendency is conserved but the actual amino acid sequence is not, and (iii) non-conserved disorder, where not even the property of disorder is conserved among closely-related species (Bellay et al. 2011). A novel tool was recently introduced to provide information on the evolutionary context of a disordered protein segments based on the quantification of sequence- and disorder conservation (Varadi et al. 2015). The different categories of conservation can be associated with the differing roles of IDPs. For example, constrained disorder can be the result of the ability to undergo disorder-to-order transition that, in turn, can impose local structural constraints. Short linear motifs constitute a special case, where only the key amino acid positions are conserved. However, LMs can also show increased evolutionary plasticity and can (re)emerge relatively easily during evolution. PTM sites often show even less conservation. Taken together with the difficulty of aligning disordered regions, transfer of functional annotation is much more challenging compared to globular proteins.

Flexible disorder is common in the case of linkers and entropic chains. This issue was directly addressed in a study by Daughdrill and colleagues (Daughdrill et al. 2007). They analyzed the evolution and function of the disordered linker region connecting two globular domains in the 70 kDa subunit of replication protein A, RPA70 (Olson et al. 2005). Evolutionary rate studies showed large variability within the linker, with many sites evolving neutrally. Direct measures of backbone flexibility, such as residual dipolar coupling and the time of Brownian reorientation showed that the pattern of backbone flexibility is conserved despite large sequence variations. Several recent mutagenesis studies have pointed to an unconventional relationship between sequence and function in IDPs. In these studies, sequences of functional regions were scrambled, but function was found to be rather insensitive to randomization. The phenomenon is usually termed sequence independence (Ross et al. 2005; Tompa and Fuxreiter 2008), and has been demonstrated in the case of the transactivator domain of Gcn4p (Hope et al. 1988) and the chimeric transcription factor EWS fusion protein (Ng et al. 2007).

Prediction methods enable the large-scale analysis of intrinsic disorder in various proteomes. It was shown that disorder in general increases with increasing complexity of the organisms. On average, 2% of archaeal and 4% of bacterial and 33% of eukaryotic proteins were predicted to contain at least 30 residue long disordered

segment (Ward et al. 2004). Structural disorder was also more abundant in viruses than in prokaryotes (Tokuriki et al. 2009). There was, however, significant variation within kingdoms (Xue et al. 2010b; Pancsa and Tompa 2012). In addition, different kingdoms use conserved disordered for different functions. In prokaryotes, disordered regions are usually involved in complex formation, while eukaryotic and viral proteins take advantage of disordered regions in regulatory and signalling processes to form transient interactions.

## 6.9 Conclusions

In general, the prediction of function is more difficult than prediction of structure, because similar structures may carry out completely different functions. This is particularly true for IDPs, for which structure corresponds not simply to the lack of a well-defined 3D fold, but to an ensemble of interconverting conformational states of various transient, but function-related, short- and long-range structural elements. A range of bioinformatics predictors reliably predict the disordered state from amino acid sequence, and can also detect functional elements with reasonable accuracy. Attempts to predict the function of IDPs from sequence, however, lag far behind prediction of structure and probably millions of functional modules in IDPs await functional characterization. Given the functional importance of many IDPs, one may anticipate significant activity in this area in the near future.

**Acknowledgements** Z.D. acknowledges the support of the "Lendület" Grant from the Hungarian Academy of Sciences (LP2014-18) and OTKA grant (K108798). This work was supported by the Odysseus grant G.0029.12 from Research Foundation Flanders (FWO) to PT.

## References

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. Nat Genet 25(1):25–29 (The Gene Ontology Consortium). doi:10.1038/75556

Bell S, Klein C, Muller L, Hansen S, Buchner J (2002) p53 contains large unstructured regions in its native state. J Mol Biol 322(5):917–927

Bellay J, Han S, Michaut M, Kim T, Costanzo M, Andrews BJ, Boone C, Bader GD, Myers CL, Kim PM (2011) Bringing order to protein disorder through comparative genomics and genetic interactions. Genome Biol 12(2):R14. doi:10.1186/gb-2011-12-2-r14

Bhattacharyya RP, Remenyi A, Good MC, Bashor CJ, Falick AM, Lim WA (2006) The Ste5 scaffold allosterically modulates signaling output of the yeast mating pathway. Science 311 (5762):822–826. doi:10.1126/science.1120941

Blom N, Gammeltoft S, Brunak S (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. J Mol Biol 294(5):1351–1362. doi:10.1006/jmbi.1999.3310

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. Curr Opin Struct Biol 21(3):441–446. doi:10.1016/j.sbi.2011.02.005

Campen A, Williams RM, Brown CJ, Meng J, Uversky VN, Dunker AK (2008) TOP-IDP-scale: a new amino acid scale measuring propensity for intrinsic disorder. Protein Pept Lett 15(9): 956–963

Chen JW, Romero P, Uversky VN, Dunker AK (2006a) Conservation of intrinsic disorder in protein domains and families: I. A database of conserved predicted disordered regions. J Proteome Res 5(4):879–887. doi:10.1021/pr060048x

Chen JW, Romero P, Uversky VN, Dunker AK (2006b) Conservation of intrinsic disorder in protein domains and families: II. Functions of conserved disorder. J Proteome Res 5(4): 888–898. doi:10.1021/pr060049p

Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF (2013) From protein sequence to dynamics and disorder with DynaMine. Nat Commun 4:2741. doi:10.1038/ncomms3741

Cilia E, Pancsa R, Tompa P, Lenaerts T, Vranken WF (2014) The DynaMine webserver: predicting protein dynamics from sequence. Nucleic Acids Res 42(Web Server Issue):W264–270. doi:10.1093/nar/gku270

Cox CJ, Dutta K, Petri ET, Hwang WC, Lin Y, Pascal SM, Basavappa R (2002) The regions of securin and cyclin B proteins recognized by the ubiquitination machinery are natively unfolded. FEBS Lett 527(1–3):303–308

Das RK, Ruff KM, Pappu RV (2015) Relating sequence encoded information to form and function of intrinsically disordered proteins. Curr Opin Struct Biol 32:102–112. doi:10.1016/j.sbi.2015.03.008

Daughdrill GW, Narayanaswami P, Gilmore SH, Belczyk A, Brown CJ (2007) Dynamic behavior of an intrinsically unstructured linker domain is conserved in the face of negligible amino acid sequence conservation. J Mol Evol 65(3):277–288. doi:10.1007/s00239-007-9011-2

Davey NE, Shields DC, Edwards RJ (2006) SLiMDisc: short, linear motif discovery, correcting for common evolutionary descent. Nucleic Acids Res 34(12):3546–3554. doi:10.1093/nar/gkl486

Davey NE, Haslam NJ, Shields DC, Edwards RJ (2010) SLiMFinder: a web server to find novel, significantly over-represented, short protein motifs. Nucleic Acids Res 38(Web Server Issue): W534–539. doi:10.1093/nar/gkq440

Davey NE, Trave G, Gibson TJ (2011) How viruses hijack cell regulation. Trends Biochem Sci 36 (3):159–169. doi:10.1016/j.tibs.2010.10.002

Davey NE, Cowan JL, Shields DC, Gibson TJ, Coldwell MJ, Edwards RJ (2012a) SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. Nucleic Acids Res 40(21):10628–10641. doi:10.1093/nar/gks854

Davey NE, Van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, Budd A, Diella F, Dinkel H, Gibson TJ (2012b) Attributes of short linear motifs. Mol BioSyst 8(1):268–281. doi:10.1039/c1mb05231d

Demarest SJ, Martinez-Yamout M, Chung J, Chen H, Xu W, Dyson HJ, Evans RM, Wright PE (2002) Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. Nature 415(6871):549–553. doi:10.1038/415549a

Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13:6580–6603

Dinkel H, Van Roey K, Michael S, Davey NE, Weatheritt RJ, Born D, Speck T, Kruger D, Grebnev G, Kuban M, Strumillo M, Uyar B, Budd A, Altenberg B, Seiler M, Chemes LB, Glavina J, Sanchez IE, Diella F, Gibson TJ (2014) The eukaryotic linear motif resource ELM: 10 years and counting. Nucleic Acids Res 42(Database Issue):D259–266. doi:10.1093/nar/gkt1047

Disfani FM, Hsu WL, Mizianty MJ, Oldfield CJ, Xue B, Dunker AK, Uversky VN, Kurgan L (2012) MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. Bioinformatics 28(12):i75–i83. doi:10.1093/bioinformatics/bts209

Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005a) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434. doi:10.1093/bioinformatics/bti541

Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005b) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 347(4):827–839. doi:10.1016/j.jmb.2005.01.071

Dosztanyi Z, Chen J, Dunker AK, Simon I, Tompa P (2006) Disorder and sequence repeats in hub proteins and their implications for network evolution. J Proteome Res 5(11):2985–2995. doi:10.1021/pr060171o

Dosztanyi Z, Meszaros B, Simon I (2009) ANCHOR: web server for predicting protein binding regions in disordered proteins. Bioinformatics 25(20):2745–2746. doi:10.1093/bioinformatics/btp518

Dosztanyi Z, Meszaros B, Simon I (2010) Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform 11(2):225–243. doi:10.1093/bib/bbp061

Dunker AK, Obradovic Z, Romero P, Garner EC, Brown CJ (2000) Intrinsic protein disorder in complete genomes. Genome Inf 11:161–171 (Workshop on Genome Informatics)

Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, Oldfield CJ, Campen AM, Ratliff CM, Hipps KW, Ausio J, Nissen MS, Reeves R, Kang C, Kissinger CR, Bailey RW, Griswold MD, Chiu W, Garner EC, Obradovic Z (2001) Intrinsically disordered protein. J Mol Graph Model 19(1):26–59

Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. Biochemistry 41(21):6573–6582

Dyson HJ, Wright PE (2005) Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol 6(3):197–208. doi:10.1038/nrm1589

Edwards RJ, Davey NE, Shields DC (2007) SLiMFinder: a probabilistic method for identifying over-represented, convergently evolved, short linear motifs in proteins. PLoS ONE 2(10):e967. doi:10.1371/journal.pone.0000967

Elbaum M (2006) Materials science. Polymers in the pore. Science 314(5800):766–767. doi:10.1126/science.1135924

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. Nucleic Acids Res 42(Database Issue):D222–230. doi:10.1093/nar/gkt1223

Fisher CK, Stultz CM (2011) Constructing ensembles for intrinsically disordered proteins. Curr Opin Struct Biol 21(3):426–431. doi:10.1016/j.sbi.2011.04.001

Fontana A, Polverino de Laureto P, De Filippis V, Scaramella E, Zambonin M (1997) Probing the partly folded states of proteins by limited proteolysis. Fold Des 2(2):R17–R26. doi:10.1016/S1359-0278(97)00010-2

Fontes MR, Teh T, Kobe B (2000) Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. J Mol Biol 297(5):1183–1194. doi:10.1006/jmbi.2000.3642

Fowler DM, Koulov AV, Balch WE, Kelly JW (2007) Functional amyloid—from bacteria to humans. Trends Biochem Sci 32(5):217–224. doi:10.1016/j.tibs.2007.03.003

Fukuchi S, Amemiya T, Sakamoto S, Nobe Y, Hosoda K, Kado Y, Murakami SD, Koike R, Hiroaki H, Ota M (2014) IDEAL in 2014 illustrates interaction networks composed of intrinsically disordered proteins and their binding partners. Nucleic Acids Res 42(Database Issue):D320–325. doi:10.1093/nar/gkt1010

Fuxreiter M, Simon I, Friedrich P, Tompa P (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol 338(5):1015–1026. doi:10.1016/j.jmb.2004.03.017

Fuxreiter M, Tompa P, Simon I (2007) Local structural disorder imparts plasticity on linear motifs. Bioinformatics 23(8):950–956. doi:10.1093/bioinformatics/btm035

Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006) FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22(23):2948–2949. doi:10.1093/bioinformatics/btl504

Gill G, Ptashne M (1987) Mutants of GAL4 protein altered in an activation function. Cell 51 (1):121–126

Graham TA, Ferkey DM, Mao F, Kimelman D, Xu W (2001) Tcf4 can specifically recognize beta-catenin using alternative conformations. Nat Struct Biol 8(12):1048–1052. doi:10.1038/nsb718

Haarmann CS, Green D, Casarotto MG, Laver DR, Dulhunty AF (2003) The random-coil 'C' fragment of the dihydropyridine receptor II-III loop can activate or inhibit native skeletal ryanodine receptors. Biochem J 372(Pt 2):305–316. doi:10.1042/BJ20021763

Habchi J, Mamelli L, Darbon H, Longhi S (2010) Structural disorder within Henipavirus nucleoprotein and phosphoprotein: from predictions to experimental assessment. PLoS ONE 5 (7):e11684. doi:10.1371/journal.pone.0011684

Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME, Radivojac P, Uversky VN, Vidal M, Iakoucheva LM (2006) Intrinsic disorder is a common feature of hub proteins from four eukaryotic interactomes. PLoS Comput Biol 2(8):e100. doi:10.1371/journal.pcbi.0020100

He B, Wang K, Liu Y, Xue B, Uversky VN, Dunker AK (2009) Predicting intrinsic disorder in proteins: an overview. Cell Res 19(8):929–949. doi:10.1038/cr.2009.87

Hecker J, Yang JY, Cheng J (2008) Protein disorder prediction at multiple levels of sensitivity and specificity. BMC Genom 9(Suppl 1):S9. doi:10.1186/1471-2164-9-S1-S9

Hegyi H, Schad E, Tompa P (2007) Structural disorder promotes assembly of protein complexes. BMC Struct Biol 7:65. doi:10.1186/1472-6807-7-65

Holt C, Wahlgren NM, Drakenberg T (1996) Ability of a beta-casein phosphopeptide to modulate the precipitation of calcium phosphate by forming amorphous dicalcium phosphate nanoclusters. Biochem J 314(Pt 3):1035–1039

Hope IA, Mahadevan S, Struhl K (1988) Structural and functional characterization of the short acidic transcriptional activation region of yeast GCN4 protein. Nature 333(6174):635–640. doi:10.1038/333635a0

Hsu WL, Oldfield CJ, Xue B, Meng J, Huang F, Romero P, Uversky VN, Dunker AK (2013) Exploring the binding diversity of intrinsically disordered proteins involved in one-to-many binding. Protein Sci 22(3):258–273. doi:10.1002/pro.2207

Iakoucheva LM, Brown CJ, Lawson JD, Obradovic Z, Dunker AK (2002) Intrinsic disorder in cell-signaling and cancer-associated proteins. J Mol Biol 323(3):573–584

Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK (2004) The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res 32 (3):1037–1049. doi:10.1093/nar/gkh253

Ikura M, Ames JB (2006) Genetic polymorphism and protein conformational plasticity in the calmodulin superfamily: two ways to promote multifunctionality. Proc Natl Acad Sci U S A 103(5):1159–1164. doi:10.1073/pnas.0508640103

Jensen MR, Zweckstetter M, Huang JR, Blackledge M (2014) Exploring free-energy landscapes of intrinsically disordered proteins at atomic resolution using NMR spectroscopy. Chem Rev 114 (13):6632–6660. doi:10.1021/cr400688u

Jones DT, Ward JJ (2003) Prediction of disordered regions in proteins from position specific score matrices. Proteins 53(Suppl 6):573–578. doi:10.1002/prot.10528

Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 31(6):857–863. doi:10.1093/bioinformatics/btu744

Kalmar L, Homola D, Varga G, Tompa P (2012) Structural disorder in proteins brings order to crystal growth in biomineralization. Bone 51(3):528–534. doi:10.1016/j.bone.2012.05.009

Khan AN, Lewis PN (2005) Unstructured conformations are a substrate requirement for the Sir2 family of NAD-dependent protein deacetylases. J Biol Chem 280(43):36073–36078. doi:10.1074/jbc.M508247200

Kiss R, Bozoky Z, Kovacs D, Rona G, Friedrich P, Dvortsak P, Weisemann R, Tompa P, Perczel A (2008a) Calcium-induced tripartite binding of intrinsically disordered calpastatin to its cognate enzyme, calpain. FEBS Lett 582(15):2149–2154. doi:10.1016/j.febslet.2008.05.032

Kiss R, Kovacs D, Tompa P, Perczel A (2008b) Local structural preferences of calpastatin, the intrinsically unstructured protein inhibitor of calpain. Biochemistry 47(26):6936–6945. doi:10.1021/bi800201a

Kovacs D, Kalmar E, Torok Z, Tompa P (2008) Chaperone activity of ERD10 and ERD14, two disordered stress-related plant proteins. Plant Physiol 147(1):381–390. doi:10.1104/pp.108.118208

Kozlowski LP, Bujnicki JM (2012) MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins. BMC Bioinf 13:111. doi:10.1186/1471-2105-13-111

Kriwacki RW, Hengst L, Tennant L, Reed SI, Wright PE (1996) Structural studies of p21Waf1/Cip1/Sdi1 in the free and Cdk2-bound state: conformational disorder mediates binding diversity. Proc Natl Acad Sci U S A 93(21):11504–11509

Lacy ER, Filippov I, Lewis WS, Otieno S, Xiao L, Weiss S, Hengst L, Kriwacki RW (2004) p27 binds cyclin-CDK complexes through a sequential mechanism involving binding-induced protein folding. Nat Struct Mol Biol 11(4):358–364. doi:10.1038/nsmb746

Le Gall T, Romero PR, Cortese MS, Uversky VN, Dunker AK (2007) Intrinsic disorder in the protein data bank. J Biomol Struct Dyn 24(4):325–342. doi:10.1080/07391102.2007.10507123

Li X, Romero P, Rani M, Dunker AK, Obradovic Z (1999) Predicting protein disorder for N-, C-, and internal regions. Genome Inf 10:30–40 (Workshop on Genome Informatics)

Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31(13):3701–3708

Liu J, Rost B (2003) NORSp: predictions of long regions without regular secondary structure. Nucleic Acids Res 31(13):3833–3835

Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. J Mol Biol 322(1):53–64

Lobley A, Swindells MB, Orengo CA, Jones DT (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 3(8):e162. doi:10.1371/journal.pcbi.0030162

Lopez Garcia F, Zahn R, Riek R, Wuthrich K (2000) NMR structure of the bovine prion protein. Proc Natl Acad Sci U S A 97(15):8334–8339

Malhis N, Gsponer J (2015) Computational identification of MoRFs in protein sequences. Bioinformatics 31(11):1738–1744. doi:10.1093/bioinformatics/btv060

Manalan AS, Klee CB (1983) Activation of calcineurin by limited proteolysis. Proc Natl Acad Sci U S A 80(14):4291–4295

Mao AH, Lyle N, Pappu RV (2013) Describing sequence-ensemble relationships for intrinsically disordered proteins. Biochem J 449(2):307–318. doi:10.1042/BJ20121346

Mark WY, Liao JC, Lu Y, Ayed A, Laister R, Szymczyna B, Chakrabartty A, Arrowsmith CH (2005) Characterization of segments from the central region of BRCA1: an intrinsically disordered scaffold for multiple protein-protein and protein-DNA interactions? J Mol Biol 345(2):275–287. doi:10.1016/j.jmb.2004.10.045

Melamud E, Moult J (2003) Evaluation of disorder predictions in CASP5. Proteins 53(Suppl 6):561–565. doi:10.1002/prot.10533

Meszaros B, Tompa P, Simon I, Dosztanyi Z (2007) Molecular principles of the interactions of disordered proteins. J Mol Biol 372(2):549–561. doi:10.1016/j.jmb.2007.07.004

Meszaros B, Simon I, Dosztanyi Z (2009) Prediction of protein binding regions in disordered proteins. PLoS Comput Biol 5(5):e1000376. doi:10.1371/journal.pcbi.1000376

Meszaros B, Dosztanyi Z, Simon I (2012) Disordered binding regions and linear motifs—bridging the gap between two models of molecular recognition. PLoS ONE 7(10):e46829. doi:10.1371/journal.pone.0046829

Mi T, Merlin JC, Deverasetty S, Gryk MR, Bill TJ, Brooks AW, Lee LY, Rathnayake V, Ross CA, Sargeant DP, Strong CL, Watts P, Rajasekaran S, Schiller MR (2012) Minimotif Miner 3.0: database expansion and significantly improved reduction of false-positive predictions from consensus sequences. Nucleic Acids Res 40(Database Issue):D252–260. doi:10.1093/nar/gkr1189

Minezaki Y, Homma K, Kinjo AR, Nishikawa K (2006) Human transcription factors contain a high fraction of intrinsically disordered regions essential for transcriptional regulation. J Mol Biol 359(4):1137–1149. doi:10.1016/j.jmb.2006.04.016

Mittag T, Forman-Kay JD (2007) Atomic-level characterization of disordered protein ensembles. Curr Opin Struct Biol 17(1):3–14. doi:10.1016/j.sbi.2007.01.009

Mizianty MJ, Stach W, Chen K, Kedarisetti KD, Disfani FM, Kurgan L (2010) Improved sequence-based prediction of disordered regions with multilayer fusion of multiple information sources. Bioinformatics 26(18):i489–i496. doi:10.1093/bioinformatics/btq373

Monastyrskyy B, Fidelis K, Moult J, Tramontano A, Kryshtafovych A (2011) Evaluation of disorder predictions in CASP9. Proteins 79(Suppl 10):107–118. doi:10.1002/prot.23161

Monastyrskyy B, Kryshtafovych A, Moult J, Tramontano A, Fidelis K (2014) Assessment of protein disorder region predictions in CASP10. Proteins 82(Suppl 2):127–137. doi:10.1002/prot.24391

Mukhopadhyay R, Hoh JH (2001) AFM force measurements on microtubule-associated proteins: the projection domain exerts a long-range repulsive force. FEBS Lett 505(3):374–378

Neduva V, Russell RB (2006) DILIMOT: discovery of linear motifs in proteins. Nucleic Acids Res 34(Web Server Issue):W350–355. doi:10.1093/nar/gkl159

Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB (2005) Systematic discovery of new recognition peptides mediating protein interaction networks. PLoS Biol 3(12):e405. doi:10.1371/journal.pbio.0030405

Ng KP, Potikyan G, Savene RO, Denny CT, Uversky VN, Lee KA (2007) Multiple aromatic side chains within a disordered structure are critical for transcription and transforming activity of EWS family oncoproteins. Proc Natl Acad Sci U S A 104(2):479–484. doi:10.1073/pnas.0607007104

Nguyen Ba AN, Yeh BJ, van Dyk D, Davidson AR, Andrews BJ, Weiss EL, Moses AM (2012) Proteome-wide discovery of evolutionary conserved sequences in disordered regions. Sci Signal 5(215):rs1. doi:10.1126/scisignal.2002515

Oates ME, Romero P, Ishida T, Ghalwash M, Mizianty MJ, Xue B, Dosztanyi Z, Uversky VN, Obradovic Z, Kurgan L, Dunker AK, Gough J (2013) D(2)P(2): database of disordered protein predictions. Nucleic Acids Res 41(Database issue):D508–516. doi:10.1093/nar/gks1226

Obenauer JC, Cantley LC, Yaffe MB (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. Nucleic Acids Res 31(13):3635–3641

Olashaw N, Bagui TK, Pledger WJ (2004) Cell cycle control: a complex issue. Cell Cycle 3 (3):263–264

Oldfield CJ, Cheng Y, Cortese MS, Brown CJ, Uversky VN, Dunker AK (2005) Comparing and combining predictors of mostly disordered proteins. Biochemistry 44(6):1989–2000. doi:10.1021/bi047993o

Olson KE, Narayanaswami P, Vise PD, Lowry DF, Wold MS, Daughdrill GW (2005) Secondary structure and dynamics of an intrinsically unstructured linker domain. J Biomol Struct Dyn 23 (2):113–124. doi:10.1080/07391102.2005.10507052

Palopoli N, Lythgow KT, Edwards RJ (2015) QSLiMFinder: improved short linear motif prediction using specific query protein data. Bioinformatics 31(14):2284–2293. doi:10.1093/bioinformatics/btv155

Pancsa R, Tompa P (2012) Structural disorder in eukaryotes. PLoS ONE 7(4):e34687. doi:10.1371/journal.pone.0034687

Patil A, Nakamura H (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. FEBS Lett 580(8):2041–2045. doi:10.1016/j.febslet.2006.03.003

Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z (2006) Length-dependent prediction of protein intrinsic disorder. BMC Bioinf 7:208. doi:10.1186/1471-2105-7-208

Pentony MM, Jones DT (2010) Modularity of intrinsic disorder in the human proteome. Proteins 78(1):212–221. doi:10.1002/prot.22504

Pierce MM, Baxa U, Steven AC, Bax A, Wickner RB (2005) Is the prion domain of soluble Ure2p unstructured? Biochemistry 44(1):321–328. doi:10.1021/bi047964d

Pontius BW (1993) Close encounters: why unstructured, polymeric domains can increase rates of specific macromolecular association. Trends Biochem Sci 18(5):181–186

Potenza E, Di Domenico T, Walsh I, Tosatto SC (2015) MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins. Nucleic Acids Res 43(Database Issue):D315–320. doi:10.1093/nar/gku982

Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL (2005) FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21(16):3435–3438. doi:10.1093/bioinformatics/bti537

Prusiner SB (1998) Prions. Proc Natl Acad Sci U S A 95(23):13363–13383

Puntervoll P, Linding R, Gemund C, Chabanis-Davidson S, Mattingsdal M, Cameron S, Martin DM, Ausiello G, Brannetti B, Costantini A, Ferre F, Maselli V, Via A, Cesareni G, Diella F, Superti-Furga G, Wyrwicz L, Ramu C, McGuigan C, Gudavalli R, Letunic I, Bork P, Rychlewski L, Kuster B, Helmer-Citterich M, Hunter WN, Aasland R, Gibson TJ (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. Nucleic Acids Res 31(13):3625–3630

Radivojac P, Vucetic S, O'Connor TR, Uversky VN, Obradovic Z, Dunker AK (2006) Calmodulin signaling: analysis and prediction of a disorder-dependent molecular recognition. Proteins 63 (2):398–410. doi:10.1002/prot.20873

Romero P, Obradovic Z, Dunker AK (1999) Folding minimal sequences: the lower bound for sequence complexity of globular proteins. FEBS Lett 462(3):363–367

Romero P, Obradovic Z, Li X, Garner EC, Brown CJ, Dunker AK (2001) Sequence complexity of disordered protein. Proteins 42(1):38–48

Ross ED, Edskes HK, Terry MJ, Wickner RB (2005) Primary sequence independence for prion formation. Proc Natl Acad Sci U S A 102(36):12825–12830. doi:10.1073/pnas.0506136102

Russo AA, Jeffrey PD, Patten AK, Massague J, Pavletich NP (1996) Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. Nature 382 (6589):325–331. doi:10.1038/382325a0

Schlessinger A, Punta M, Rost B (2007) Natively unstructured regions in proteins identified from contact predictions. Bioinformatics 23(18):2376–2384. doi:10.1093/bioinformatics/btm349

Schlessinger A, Punta M, Yachdav G, Kajan L, Rost B (2009) Improved disorder prediction by combination of orthogonal approaches. PLoS ONE 4(2):e4433. doi:10.1371/journal.pone.0004433

Schwalbe M, Ozenne V, Bibow S, Jaremko M, Jaremko L, Gajda M, Jensen MR, Biernat J, Becker S, Mandelkow E, Zweckstetter M, Blackledge M (2014) Predictive atomic resolution descriptions of intrinsically disordered hTau40 and alpha-synuclein in solution from NMR and small angle scattering. Structure 22(2):238–249. doi:10.1016/j.str.2013.10.020

Schweers O, Schonbrunn-Hanebeck E, Marx A, Mandelkow E (1994) Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for beta-structure. J Biol Chem 269(39):24290–24297

Si K, Giustetto M, Etkin A, Hsu R, Janisiewicz AM, Miniaci MC, Kim JH, Zhu H, Kandel ER (2003a) A neuronal isoform of CPEB regulates local protein synthesis and stabilizes synapse-specific long-term facilitation in aplysia. Cell 115(7):893–904

Si K, Lindquist S, Kandel ER (2003b) A neuronal isoform of the Aplysia CPEB has prion-like properties. Cell 115(7):879–891

Sickmeier M, Hamilton JA, LeGall T, Vacic V, Cortese MS, Tantos A, Szabo B, Tompa P, Chen J, Uversky VN, Obradovic Z, Dunker AK (2007) DisProt: the database of disordered proteins. Nucleic Acids Res 35(Database issue):D786–793. doi:10.1093/nar/gkl893

Sigler PB (1988) Transcriptional activation. Acid blobs and negative noodles. Nature 333 (6170):210–212. doi:10.1038/333210a0

Su CT, Chen CY, Ou YY (2006) Protein disorder prediction by condensed PSSM considering propensity for order or disorder. BMC Bioinform 7:319. doi:10.1186/1471-2105-7-319

Tokuriki N, Oldfield CJ, Uversky VN, Berezovsky IN, Tawfik DS (2009) Do viral proteins possess unique biophysical features? Trends Biochem Sci 34(2):53–59. doi:10.1016/j.tibs.2008.10.009

Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27(10):527–533
Tompa P (2005) The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett 579(15):3346–3354. doi:10.1016/j.febslet.2005.03.072
Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37 (12):509–516. doi:10.1016/j.tibs.2012.08.004
Tompa P, Csermely P (2004) The role of structural disorder in the function of RNA and protein chaperones. Faseb J 18(11):1169–1175. doi:10.1096/fj.04-1584rev
Tompa P, Fuxreiter M (2008) Fuzzy complexes: polymorphism and structural disorder in protein-protein interactions. Trends Biochem Sci 33(1):2–8. doi:10.1016/j.tibs.2007.10.003
Tompa P, Szasz C, Buday L (2005) Structural disorder throws new light on moonlighting. Trends Biochem Sci 30(9):484–489. doi:10.1016/j.tibs.2005.07.008
Tompa P, Dosztanyi Z, Simon I (2006) Prevalent structural disorder in *E. coli* and *S. cerevisiae* proteomes. J Proteome Res 5(8):1996–2000. doi:10.1021/pr0600881
Tompa P, Fuxreiter M, Oldfield CJ, Simon I, Dunker AK, Uversky VN (2009) Close encounters of the third kind: disordered domains and the interactions of proteins. BioEssays 31(3):328–335. doi:10.1002/bies.200800151
Tompa P, Davey NE, Gibson TJ, Babu MM (2014) A million peptide motifs for the molecular biologist. Mol Cell 55(2):161–169. doi:10.1016/j.molcel.2014.05.032
Triezenberg SJ (1995) Structure and function of transcriptional activation domains. Curr Opin Genet Dev 5(2):190–196
Trombitas K, Greaser M, Labeit S, Jin JP, Kellermayer M, Helmes M, Granzier H (1998) Titin extensibility in situ: entropic elasticity of permanently folded and permanently unfolded molecular segments. J Cell Biol 140(4):853–859
Tucker MM, Robinson JB Jr, Stellwagen E (1981) The effect of proteolysis on the calmodulin activation of cyclic nucleotide phosphodiesterase. J Biol Chem 256(17):9051–9058
Tuite MF, Koloteva-Levin N (2004) Propagating prions in fungi and mammals. Mol Cell 14 (5):541–552. doi:10.1016/j.molcel.2004.05.012
Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11(4):739–756. doi:10.1110/ps.4210102
Uversky VN (2013) A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci 22(6):693–724. doi:10.1002/pro.2261
Uversky VN, Gillespie JR, Fink AL (2000) Why are "natively unfolded" proteins unstructured under physiologic conditions? Proteins 41(3):415–427
Uversky VN, Oldfield CJ, Dunker AK (2005) Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J Mol Recogn 18(5):343–384. doi:10.1002/jmr.747
Vacic V, Oldfield CJ, Mohan A, Radivojac P, Cortese MS, Uversky VN, Dunker AK (2007) Characterization of molecular recognition features, MoRFs, and their binding partners. J Proteome Res 6(6):2351–2366. doi:10.1021/pr0701411
van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, Fuxreiter M, Gough J, Gsponer J, Jones DT, Kim PM, Kriwacki RW, Oldfield CJ, Pappu RV, Tompa P, Uversky VN, Wright PE, Babu MM (2014) Classification of intrinsically disordered regions and proteins. Chem Rev 114(13):6589–6631. doi:10.1021/cr400525m
Van Roey K, Gibson TJ, Davey NE (2012) Motif switches: decision-making in cell regulation. Curr Opin Struct Biol 22(3):378–385. doi:10.1016/j.sbi.2012.03.004
Van Roey K, Dinkel H, Weatheritt RJ, Gibson TJ, Davey NE (2013) The switches. ELM resource: a compendium of conditional regulatory interaction interfaces. Sci Signal 6(269):rs7. doi:10.1126/scisignal.2003345
Van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, Gibson TJ, Davey NE (2014) Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. Chem Rev 114(13):6733–6778. doi:10.1021/cr400585q
Varadi M, Kosol S, Lebrun P, Valentini E, Blackledge M, Dunker AK, Felli IC, Forman-Kay JD, Kriwacki RW, Pierattelli R, Sussman J, Svergun DI, Uversky VN, Vendruscolo M, Wishart D, Wright PE, Tompa P (2014) pE-DB: a database of structural ensembles of intrinsically

disordered and of unfolded proteins. Nucleic Acids Res 42(Database Issue):D326–335. doi:10.1093/nar/gkt960

Varadi M, Guharoy M, Zsolyomi F, Tompa P (2015) DisCons: a novel tool to quantify and classify evolutionary conservation of intrinsic protein disorder. BMC Bioinf 16:153. doi:10.1186/s12859-015-0592-2

Via A, Gould CM, Gemund C, Gibson TJ, Helmer-Citterich M (2009) A structure filter for the Eukaryotic Linear Motif Resource. BMC Bioinf 10:351. doi:10.1186/1471-2105-10-351

Vucetic S, Brown CJ, Dunker AK, Obradovic Z (2003) Flavors of protein disorder. Proteins 52 (4):573–584. doi:10.1002/prot.10437

Waizenegger I, Gimenez-Abian JF, Wernic D, Peters JM (2002) Regulation of human separase by securin binding and autocleavage. Curr Biol 12(16):1368–1378

Wang L, Sauer UH (2008) OnD-CRF: predicting order and disorder in proteins using conditional random fields. Bioinformatics 24(11):1401–1402. doi:10.1093/bioinformatics/btn132

Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J Mol Biol 337(3):635–645. doi:10.1016/j.jmb.2004.02.002

Weinreb PH, Zhen W, Poon AW, Conway KA, Lansbury PT Jr (1996) NACP, a protein implicated in Alzheimer's disease and learning, is natively unfolded. Biochemistry 35 (43):13709–13715. doi:10.1021/bi961799n

Wootton JC (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. Comput Chem 18(3):269–285

Wootton JC, Federhen S (1996) Analysis of compositionally biased regions in sequence databases. Methods Enzymol 266:554–571

Wright PE, Dyson HJ (1999) Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293(2):321–331. doi:10.1006/jmbi.1999.3110

Xie Q, Arnold GE, Romero P, Obradovic Z, Garner E, Dunker AK (1998) The sequence attribute method for determining relationships between sequence and protein disorder. Genome Inf 9:193–200 (Workshop on Genome Informatics)

Xie H, Vucetic S, Iakoucheva LM, Oldfield CJ, Dunker AK, Uversky VN, Obradovic Z (2007) Functional anthology of intrinsic disorder. 1. Biological processes and functions of proteins with long disordered regions. J Proteome Res 6(5):1882–1898. doi:10.1021/pr060392u

Xue B, Dunbrack RL, Williams RW, Dunker AK (1804) Uversky VN (2010a) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim Biophys Acta 4:996–1010. doi:10.1016/j.bbapap.2010.01.011

Xue B, Williams RW, Oldfield CJ, Dunker AK, Uversky VN (2010) Archaic chaos: intrinsically disordered proteins in Archaea. BMC Syst Biol 4(Suppl 1):S1. doi:10.1186/1752-0509-4-S1-S1

Yang ZR, Thomson R, McNeil P, Esnouf RM (2005) RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21(16):3369–3376. doi:10.1093/bioinformatics/bti534

# Chapter 7
# Prediction of Protein Aggregation and Amyloid Formation

**Ricardo Graña-Montes, Jordi Pujols-Pujol, Carlota Gómez-Picanyol and Salvador Ventura**

**Abstract** Protein aggregation accounts for the onset of more than 40 human disorders, including neurodegenerative diseases like Alzheimer's and Parkinson's but also non-neuropathic pathologies like Diabetes type II or some types of cancers. In all these diseases, the toxic effect is associated with the self-assembly of proteins into insoluble amyloid fibrils displaying a common regular cross-β structure. Surprisingly, cells also exploit the amyloid fold for important physiological processes, from structure scaffolding to heritable information transmission. In addition, protein aggregation often occurs during the recombinant production and downstream processing of therapeutic proteins, becoming the main bottleneck in the marketing of these drugs. In this context, approaches aiming to predict the aggregation and amyloid formation propensities of proteins are receiving increasing interest, both because they can lead us to the development of novel therapeutic strategies and because they are providing us with a global understanding of the role of protein aggregation in physiological and pathological processes. Here we illustrate how our present understanding of the physico-chemical and structural basis of protein aggregation has crystalized in the development of algorithms able to forecast the aggregation properties of proteins both from their primary and tertiary structures. A detailed description of these computational approaches and their application is provided.

**Keywords** Amyloid · Protein aggregation · Amyloid-like Formation · Aggregation Prediction Methods · Sequence-based amyloid prediction · Structure-based amyloid prediction · Amyloid fibril · Aggregation prone regions (APRs) · Amyloidogenic stretch

R. Graña-Montes (✉) · J. Pujols-Pujol · C. Gómez-Picanyol · S. Ventura (✉)
Institut de Biotecnologia i Biomedicina and Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain
e-mail: r.granham@opmbx.org

S. Ventura
e-mail: Salvador.ventura@uab.cat



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_7

## 7.1 Introduction

For many years, the aggregation of proteins and polypeptides remained a neglected area of protein chemistry, as it was considered a process of a rather stochastic origin. It was not until the realization, starting in the early 1970s, that the insoluble deposits associated with different human diseases were primarily enriched in single specific polypeptides (Westermark 2005) that the analysis of the capability of these proteins to aggregate was addressed as a specific biophysical phenomenon. In this way, during the last three decades, the study of protein aggregation has evolved to become a vivid research topic whose implications span a variety of fields, including biochemistry, biomedicine, biotechnology, and nanotechnology. Three major areas may be currently identified where the analysis of protein aggregation reaches a greater impact—namely, in the study of a group of human pathologies known as conformational disorders, in the industrial manufacture of proteinaceous products, and in the development of novel bio-inspired nanomaterials. In first place, the formation of insoluble protein deposits in different tissues is linked to more than 40 human diseases—many of which are highly debilitating or even fatal—ranging from different classes of amyloidosis, neurodegenerative and prionic pathologies (like Alzheimer's, Parkinson's or Creutzfeldt-Jakob's diseases), to *Diabetes mellitus* type 2 or certain types of cancer (Selkoe 2003; Chiti and Dobson 2006; Invernizzi et al. 2012). Consequently, the analysis of the aggregation reactions of the proteins involved in this kind of diseases has attracted growing attention with the aim of developing methods to prevent or treat these devastating disorders. On the other hand, aggregation is the major bottleneck in the production of commercial proteins at an industrial level (Cromwell et al. 2006), an area which has gained a particular momentum in recent years because of the growing interest in the development of protein-based agents with therapeutic potential (Aggarwal 2009), and particularly of antibody-based therapies (Perchiacca and Tessier 2012; Lee et al. 2013). Finally, the characterization of the mechanical properties of amyloid-like structures (Knowles et al. 2007; Cherny and Gazit 2008; Knowles and Buehler 2011) has driven the development of a wide variety of applications for this type of assemblies as nanomaterials (Hauser et al. 2014; Shimanovich et al. 2014).

## 7.2 The Physico-chemical and Structural Basis of Protein Aggregation

The polypeptides known to accumulate in the insoluble deposits found in the majority of conformational disorders are characterized by their ability to display (either when purified from these deposits or after being synthesized and incubated) *in vitro* (a particular supramolecular structure named amyloid fibrils). This type of structure is distinguished by its compact and non-ramified, fibrous appearance under Transmission Electron Microscopy or Atomic Force Microscopy (Makin and

Serpell 2005). In the core of the fibrils, the polypeptide adopts a characteristic molecular architecture composed of two opposite β-sheets with their strands running perpendicular to the elongation axis of the fibrils. This arrangement was derived early from a common X-ray diffraction pattern found for amyloid fibrils formed by unrelated proteins (Sunde and Blake 1997), and it defines a supersecondary level of structure known as the cross-β conformation.

Despite the existence of a number of proteins, known as functional amyloids, that adopt amyloid structure in order to carry out their physiological function, (Gebbink et al. 2005; Fowler et al. 2007; Blanco et al. 2012), this is not the case for the proteins associated with different conformational disorders (Table 7.1), which experience an abnormal conformational conversion from their physiological native states to acquire the cross-β supersecondary structure. The latter proteins do not possess any overall sequential relationship, and they populate a wide diversity of native structures Fig 7.1 (Uversky and Fink 2004; Chiti and Dobson 2006; Invernizzi et al. 2012) ranging from intrinsically disordered proteins (e.g. the Amyloid β peptides - Aβ- and α-Synuclein—related to the Alzheimer's and Parkinson's diseases, respectively) to polypeptides able to adopt a stable tertiary structure, with either monomeric (e.g. Lysozyme and Prolactin—associated, with a systemic amyloidosis and with pituitary prolactinomas, respectively) or multimeric quaternary architectures (e.g. Transthyretin and Superoxide dismutase [Cu-Zn]—implicated in different amyloidoses and in amyotrophic lateral sclerosis, respectively) (see Fig.7.1). On the other hand, the conversion from the physiological conformation into amyloid-like structures is not limited to the discrete set of proteins associated to conformational diseases, but has been observed or induced for a large number of polypeptides, from different organisms belonging to all phyla, which do not possess any currently known relationship to disease (Rochet and Lansbury 2000; Stefani and Dobson 2003; Uversky and Fink 2004). Moreover, different types of proteinaceous aggregates with an apparently amorphous macroscopic appearance are known to share properties of amyloid-like fibrils (Carrió et al. 2005; de Groot et al. 2009), including the characteristic X-ray diffraction pattern of the cross-β conformation (Wang et al. 2008, 2010). In fact, the latter observation is closely related with the important issue of distinguishing between generic protein aggregation and specific formation of amyloid-like structures, two terms that are often used as if they were interchangeable. While it cannot be stated that protein aggregation always implies a gain in β-sheet conformation (as in the case of isoelectric or salt-induced protein precipitation), it is true, however, that such an enrichment is frequently associated with the formation of thermodynamically stable protein deposits, among which amyloid-like structure arises from an optimal quaternary arrangement of the cross-β conformation. In this context, it seems that the ability to attain cross-β structure constitutes a generic property of virtually every polypeptide—as was early postulated in terms of acquisition of amyloid-like structure (Dobson 2001, 2003)—simply because backbone-mediated intermolecular hydrogen-bonding constitutes the strongest contributor towards the adoption and stabilization of this conformation (Dobson 1999; Knowles et al. 2007; Cheon et al. 2007).

**Table 7.1** Proteins involved in the formation of amyloid or amyloid-like deposits in relationship with human disorders

| Precursor protein | UniProt accession | Disease | Conformation* |
|---|---|---|---|
| α-Synuclein | P37840 | Parkinson's disease<br>Dementia with Lewy bodies | Intrinsically disordered |
| β2-Microglobulin (wt) | P61769 | Hemodialysis-related amyloidosis | all-β |
| β2-Microglobulin (variants) | P61769 | Systemic amyloidosis | – |
| γ-Crystallin B | P07316 | Cataract | all-β |
| γ-Crystallin C | P07315 | Cataract | all-β |
| γ-Crystallin D | P07320 | Cataract | all-β |
| Amyloid β precursor protein (wt) | P05067 | Alzheimer's disease | Intrinsically disordered (residues 672-713) |
| Amyloid β precursor protein (variants) | P05067 | Alzheimer's disease<br>APP-related cerebral amyloid angiopathy | – |
| Androgen receptor (with polyQ expansion) | P10275 | X-linked 1 spinal and bulbar muscular atrophy | – |
| Apolipoprotein A-I (variants) | P02647 | Systemic amyloidosis | Coiled coil |
| Apolipoprotein A-II (variants) | P02652 | Systemic amyloidosis | Coiled coil |
| Apolipoprotein A-IV | P06727 | Systemic amyloidosis | Coiled coil |
| Ataxin-1 (with polyQ expansion) | P54253 | Spinocerebellar ataxia 1 | – |
| Ataxin-2 (with polyQ expansion) | Q99700 | Spinocerebellar ataxia 2 | – |
| Ataxin-3 (with polyQ expansion) | P54252 | Spinocerebellar ataxia 3 | – |
| Ataxin-7 (with polyQ expansion) | O15265 | Spinocerebellar ataxia 7 | – |
| Atrial natriuretic factor | P01160 | Isolated atrial amyloidosis | Intrinsically disordered |
| Atrophin-1 (with polyQ expansion) | P54259 | Dentatorubral-pallidoluysian atrophy | – |
| Calcitonin | P01258 | Medullary thyroid carcinoma | Intrinsically disordered |
| Corneodesmosin | Q15517 | Localized amyloidosis (cornified epithelia and hair follicles) | Intrinsically disordered (residues 60-171) |
| Cystatin-C (variants) | P01034 | Cystatin-C-related cerebral amyloid angiopathy | α+β |
| Fibrinogen α chain (variants) | P02671 | Familial visceral amyloidosis (systemic) | Coiled coil (residues 46-231)<br>α+β (residues 670-866) |

(continued)

**Table 7.1** (continued)

| Precursor protein | UniProt accession | Disease | Conformation* |
|---|---|---|---|
| Galectin-7 | P47929 | Localized cutaneous amyloidosis | all-β |
| Gelsolin (variants) | P06396 | Finnish type amyloidosis (systemic) | α+β |
| Huntingtin (with polyQ expansion) | P42858 | Huntington's disease | Intrinsically disordered |
| Immunoglobulin heavy chain | – | Systemic and localized amyloidosis | all-β |
| Immunoglobulin light chain | – | Systemic and localized amyloidosis | all-β |
| Insulin | P01308 | Injection-localized amyloidosis | Nearly all-α, disulphide-rich |
| Integral membrane protein 2B (ABri variant) | Q9Y287 | Familial British dementia | Intrinsically disordered (precursor polypeptide) |
| Integral membrane protein 2B (ADan variant) | Q9Y287 | Familial Danish dementia | Intrinsically disordered (precursor polypeptide) |
| Islet amyloid polypeptide | P10997 | Type II diabetes mellitus | Intrinsically disordered |
| Lactotransferrin | P02788 | Corneal amyloidosis associated with trichiasis | α/β |
| Leukocyte cell-derived chemotaxin-2 | O14960 | Systemic amyloidosis | – |
| Lysozyme C (variants) | P61626 | Familial visceral amyloidosis (systemic) | α+β |
| Major prion protein (wt) | P04156 | Creutzfeldt-Jakob's disease<br>Fatal familial insomnia | Intrinsically disordered (residues 23-121)<br>α+β (residues 122-230) |
| Major prion protein (variants) | P04156 | Creutzfeldt-Jakob's disease<br>Fatal familial insomnia<br>Gerstmann-Straussler's disease<br>Huntington's disease-like 1<br>Spongiform encephalopathy with neuropsychiatric features | – |
| Medin | Q08431 | Aortic medial amyloidosis | – |
| Microtubule-associated protein Tau | P10636 | Frontotemporal dementia<br>Pick's disease of brain | Intrinsically disordered |
| Odontogenic ameloblast-associated protein | A1E959 | Calcifying epithelial odontogenic tumors | – |

(continued)

**Table 7.1** (continued)

| Precursor protein | UniProt accession | Disease | Conformation* |
|---|---|---|---|
| Oncostatin-M-specific receptor subunit β | Q99650 | Primary localized cutaneous amyloidosis 1 | – |
| Prolactin | P01236 | Pituitary prolactinomas | all-α |
| Pulmonary surfactant-associated protein C | P11686 | Pulmonary alveolar proteinosis | – |
| Semenogelin-1 | P04279 | Localized amyloidosis (vesicular seminalis) | – |
| Serum amyloid A-1 protein | P0DJI8 | Systemic amyloidosis<br>Familial Mediterranean fever<br>Rheumatoid arthritis | All-α |
| Superoxide dismutase [Cu-Zn] (variants) | P00441 | Amyotrophic lateral sclerosis 1 | All-β |
| TATA-box-binding protein (with polyQ expansion) | P20226 | Spinocerebellar ataxia 17 | – |
| Transforming growth factor-β-induced protein ig-h3 (variants) | Q15582 | Avellino type corneal dystrophy<br>Groenouw type I corneal dystrophy<br>Type I lattice corneal dystrophy<br>Type IIIA lattice corneal dystrophy | α+β (residues 502–634) |
| Transthyretin (wt) | P02766 | Senile systemic amyloidosis | All-β |
| Transthyretin (variants) | P02766 | Familial amyloidotic polyneuropathy<br>Transthyretin-related amyloid cardiomyopathy<br>Carpal tunnel syndrome | – |

* Conformational properties, if known, of the precursor protein (or specified polypeptide regions) under native or close-to-native conditions.
Adapted from (Uversky and Fink 2004; Chiti and Dobson 2006; Invernizzi et al. 2012; Sipe et al. 2014)

It had long been speculated that the formation of amyloid-like structures is thermodynamically driven, since the adoption of an extended cross-β conformation provides a greater stability relative to a polypeptide's native state (although this could not be demonstrated until quite recently) because of the stabilization arising mainly from the massive enthalpic contribution of the main chain hydrogen-bonding network inherent to this fold (Baldwin et al. 2011). The generalization of the idea that different cross-β-enriched aggregated species constitute generic states which could virtually be populated by any polypeptide chain came to expand the picture of the conformational energy landscape accessible for a nascent polypeptide chain (Jahn and Radford 2005, 2008). In this way, after being synthesized, a given protein departing from a hypothetical collection of largely

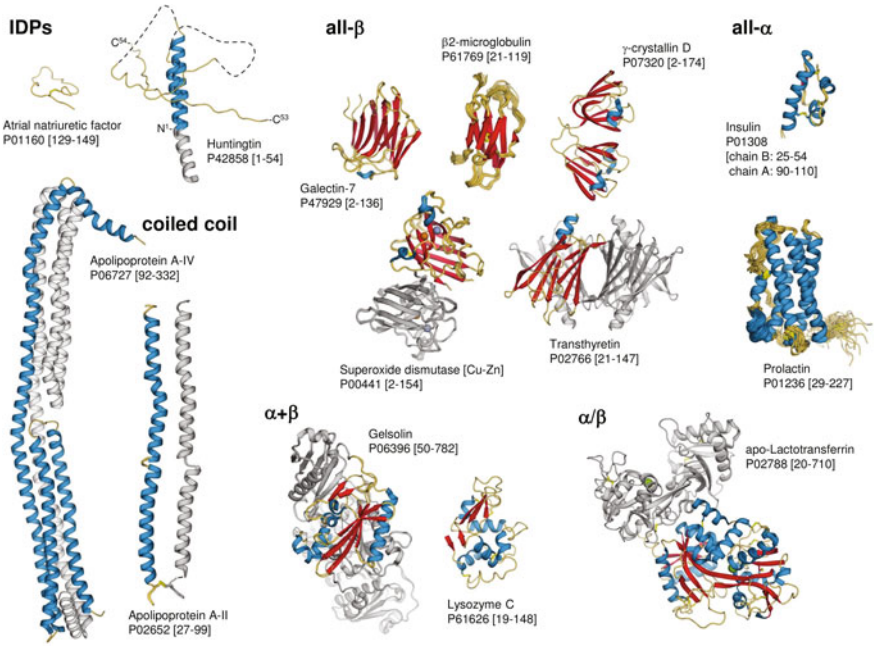



**Fig. 7.1** Native structures of several amyloidogenic proteins, accounting for the diversity of folds able to experience conformational transformation into amyloid. Cartoon representations show structures resolved experimentally, under close-to-native conditions, for a variety of amyloidogenic proteins associated to human conformational disorders (PDB codes: Atrial natriuretic factor—1YK0, Huntingtin—3IOR and 3IOT, Apolipoprotein A-IV—3S84, Apolipoprotein A-II—2OU1, Galectin-7—1BKZ, β2-Microglobulin—2XKS, γ-Crystallin D—1HK0, Superoxide dismutase [Cu-Zn]—1PU0, Transthyretin—1TTA, Insulin—1ZNJ, Prolactin—1RW5, Gelsolin—3FFN, Lysozyme C—1LZ1, apo-Lactotransferrin—1CB6); for Huntingtin, the structures shown correspond to fusions of its $N_{ter}$ poly-Q-rich region with Maltose-binding periplasmic protein (UniProt code P0AEX9). The UniProt accession code corresponding to the precursor polypeptide of each protein is specified, and the fragment of the precursor corresponding either to the mature protein or to the region involved in the amyloidogenic conversion is shown within brackets; unless the N- and C-termini are specifically noted, this numbering also corresponds to the polypeptide represented in the structure. Protein structures are grouped according to the structural class they belong to (shown in *bold*), which, in the case of globular proteins, corresponds to the SCOP fold classification scheme (Murzin et al. 1995); their diverse secondary structure content is highlighted: α-helical (*blue*), β-sheet (*red*), and loops (*yellow*) -fragments with unavailable structural data are depicted as dashed lines-. Different conformations arising from several NMR conformers resolved or from multiple identical chains in the crystallographic asymmetric unit are superimposed, when present, so as to emphasized the structural flexibility of native states. For proteins with a native homomeric quaternary or multidomain structure, only the essential fold is highlighted, as mentioned, and the rest of the structure is coloured grey. Disulfide bonds are represented as sticks (*bright yellow*) and complexed metal ions as spheres: copper (*brown*), zinc (*grey*), and chloride (*green*)

extended conformations—which can be defined as the unfolded state ensemble—may either remain populating such unstructured configurations to variable degrees, thus behaving as an Intrinsically Disordered Protein (IDP; Tompa 2012; Uversky 2013a; see also Chap. 6); it can collapse in order to adopt one or several structured or folded states (Anfinsen 1973; Dill et al. 2008; Bryan and Orban 2010)—which can be termed native states if attained under physiological conditions; or may, alternatively, misfold into different kinds of aggregated states with variable degrees of enrichment in cross-β conformation (Jahn and Radford 2008). Those folded and misfolded states can be significantly populated since they constitute deep local minima of the free energy landscape. Transitions between intrinsically disordered, folded, and misfolded states may take place upon binding to chemical cofactors or protein partners, environmental changes, or as the outcome of mutations (Gershenson et al. 2014; Flock et al. 2014). Even though the adoption of cross-β-enriched aggregated states would, in principle, be thermodynamically favoured, at least three factors can be identified that exert a kinetic control over their population—namely, the physico-chemical properties of the polypeptide chain, the competition between folding and aggregation, and the protein quality control machinery the cell has evolved. Protein folding and aggregation are regarded as competing processes because they are guided, at least at their initial stages, by similar principles (Kauzmann 1959; Cheon et al. 2007; Auer et al. 2008). However, while protein folding is directed by the establishment of specific intramolecular interactions (Lindorff-Larsen et al. 2005), aggregation is dominated by backbone-directed intermolecular contacts; in such a way that the efficient attainment of a folded native structure largely impairs the establishment of the aberrant intermolecular interactions that would lead to the formation of aggregated states (Monsellier and Chiti 2007). Additionally, theoretical studies on protein folding inspired by polymer statistical mechanics have shown that globular proteins have evolved to fold rapidly and cooperatively to their native states, because their energy landscapes are "minimally frustrated" (Wolynes 2008)—which implies that partially folded conformations tend to be short-lived. Furthermore, cells have evolved an intricate molecular machinery dedicated to preserving protein quality by either aiding polypeptides to attain their native state, preventing them from establishing non-functional intermolecular contacts (that could lead to the formation of aggregated species), or rescuing them from misfolded states (Hartl et al. 2011; Kim et al. 2013).

Regarding the physico-chemical properties of the polypeptide chain, which arise from its particular amino acidic composition, aside from encoding their native state ensemble (Anfinsen 1973), they also define the main ability of a protein to access different aggregated states—that is, the primary sequence determines the intrinsic aggregation propensity of a protein. This is supported by a large body of evidence based on the experimental analysis of amyloidogenic proteins and peptides (Hilbich et al. 1992; Esler et al. 1996; Wurth et al. 2002), as well as of model proteins able to form amyloid-like fibrils (Chiti et al. 2002b), which reveals how changes in the amino acid sequence may have a deep influence on their tendency to aggregate.

### 7.2.1 Intrinsic Determinants of Protein Aggregation

The sequential and mutagenic analysis of amyloidogenic proteins and peptides, particularly of those behaving as IDPs under native conditions (thus allowing the disentanglement between the forces promoting aggregation and those favouring folding into a 3-dimensional structure), has led to the identification of a series of properties of both single amino acids and amino acidic combinations which are relevant in determining the ability of a polypeptide to aggregate; consequently defining the intrinsic determinants of protein aggregation. Among them, hydrophobicity has been found to constitute a major force driving aggregation, as evidenced by the effect of substitutions of polar or charged residues by non-polar amino acids increasing the rate of aggregation, while the inverse changes tend to decrease the extent of aggregation or even have a disruptive effect (Hilbich et al. 1992; Esler et al. 1996; Wurth et al. 2002; Buell et al. 2009). Nonetheless, hydrophobicity alone has been judged insufficient to account for the impact of mutations on the propensity to aggregate (Chiti et al. 2003; Rousseau et al. 2006a). The tendency of amino acids to adopt a particular secondary structure is another important determinant of protein aggregation; consistent with the finding that the core of amyloid-like aggregates is enriched in cross-β conformation, both the enrichment in residues with a higher propensity to form β-sheet structure (Chiti et al. 2002a) and the pre-existence of β-strands in the native state (Pallarès et al. 2004) enhance the aggregation propensity of polypeptides. Consequently, amino acids with a low tendency to adopt β-sheet secondary structure such as Pro (which induces a bend in the polypeptide backbone), and Gly (due to the entropic cost associated to its fixation in secondary structure elements) tend to disfavour aggregation (Wood et al. 1995; Steward et al. 2002; Parrini et al. 2005). Furthermore, a variety of negative design strategies have been identified in all-β proteins in order to protect the peripheral strands flanking β-sheets (Richardson and Richardson 2002), which are at a higher risk of establishing non-functional intermolecular contacts for being free to establish hydrogen bonds with neighbouring molecules. The net charge of a polypeptide also influences the propensity to aggregate (Chiti et al. 2002a, 2003) since it defines the extent of repulsion between individual molecules, thus affecting the chances to establish the intermolecular contacts required for a protein to self-assemble and aggregate.

Although the above-mentioned factors emerge from the physico-chemical properties of individual amino acids, the linear combination of these properties along the primary sequence has a strong impact on the tendency of a polypeptide to aggregate. For example, the combinatorial design of polypeptide secondary structures has revealed that the alternation of hydrophobic and hydrophilic residues along the sequence facilitates the assembly into amyloid-like structures (West et al. 1999), likely because this pattern favours the formation of amphiphilic β-sheets. Quite interestingly, the statistical analysis of natural protein sequences revealed that this pattern is underrepresented, relative to other amino acid combinations, being

less frequent than it would be expected by chance (Broome and Hecht 2000). Similarly, continuous stretches with three or more hydrophobic residues are also underrepresented in natural protein sequences (Schwartz et al. 2001), which is consistent with hydrophobicity being a major force driving deleterious aggregation.

### 7.2.2 Extrinsic Determinants of Protein Aggregation

The intrinsic determinants accounting for protein aggregation are modulated by the specific environmental conditions, which impact kinetically, thermodynamically, and structurally the self-assembly process and the properties of the final aggregates, being an important source of polymorphism (Kodali and Wetzel 2007; Tycko 2014). The pH, the ionic strength, and the temperature of the system are the extrinsic determinants with greatest impact on the aggregation reaction of a protein (DuBay et al. 2004). First, pH influences the protonation state of residue side chains, thus modulating their physico-chemical properties, including the effective hydrophobicity, and the net charge of both individual amino acids and the whole protein molecule. Ionic strength acts at the protein net charge level since a higher ion concentration in solution favours the shielding of charged side chains, consequently reducing the repulsive effect between polypeptide molecules (Morel et al. 2010) and raising their probability to establish undesired intermolecular contacts. Finally, temperature has a strong influence on the conformational energy landscape of polypeptides, inducing changes in the relative free energy differences between local minima and in the kinetic barriers separating them, which might favour the preferential population of aggregated states. Indeed, temperature, as well as pH, may alter the network of interactions that sustain the native 3-dimensional structure, leading to the transient or permanent population of unstructured (either partially or globally) where aggregated states might result more easily accessible.

### 7.2.3 Specific Sequence Stretches Drive Aggregation

The specific physico-chemical properties of amino acids and their combination in linear patterns along the primary structure play a major role in the potential of a given polypeptide to aggregate and, thus, define the intrinsic determinants of protein aggregation. However, it has been shown that not all the protein sequence is equally important for the ability of a protein to aggregate—but, instead, there are short sequence fragments that promote and guide the formation of amyloid-like structures (Ventura et al. 2004; Ivanova et al. 2004). This principle defines the "amyloidogenic stretch" hypothesis, and such fragments are commonly referred to as "Hot Spots" or aggregation-prone regions (APRs). Consistent with the intrinsic determinants of protein aggregation, these segments are characterized by an

enrichment in hydrophobic amino acids—both aliphatic (Val, Leu, Ile) and aromatic (Phe, Trp, Tyr) (Rousseau et al. 2006b).

### 7.2.4 Structural Determinants of Amyloid-like Aggregation

As already discussed, the X-ray diffraction patterns of amyloids, amyloid-like fibrils, and different kinds of apparently amorphous aggregates share a cross-β super-secondary level of structure. However, the failure of these different kinds of protein aggregates to attain a sufficiently regular 3-dimensional assembly (even in the case of the apparently macroscopically ordered amyloid-like fibrils) hampered for a long time the description of amyloid structure at atomic detail. Fortunately, advances in solid state Nuclear Magnetic Resonance (ssNMR) (Petkova et al. 2002; Ritter et al. 2005) and in the microcrystallization of short amyloidogenic peptides (Makin et al. 2005; Nelson et al. 2005; Rodriguez et al. 2015) have elucidated the fine molecular architecture of the amyloid-like fibrils formed by different proteins and by peptides thereof. Most of the solved structures confirm a cross-β core composed of two opposite β-sheets running perpendicular to the axis of the fibril; although fibrils formed by certain amyloidogenic proteins adopt a β-helix structure instead, where three β-strands are arranged facing each other on every turn (Tycko 2011; Eisenberg and Jucker 2012; Tycko and Wickner 2013). The molecular complementarity required for the assembly of each pair of facing strands in the cross-β conformation is particularly highlighted by the crystallographic structures of amyloidogenic peptides (Sawaya et al. 2007), which reveal how the docking of facing strands defines a "steric zipper" formed by the inward-pointing side chains. At the same time, the structures of these peptides reflect the array of posible arrangements that can be adopted by β-strands to build the cross-β structure. The atomic detail provided by these experimental structures provide an outstanding framework to rationalize the role of the determinants of protein aggregation we have introduced before. In first place, it allows for an understanding of how the small sequence stretches defining APRs can guide and promote the formation of amyloid-like structure, since only a small portion of the polypeptide is strictly required in order to contribute a β-strand for the establishment of the cross-β core of the fibril, while the rest of the molecule may well remain exposed to the solvent or even attached as either a partially or completely structured fragment (Sambashivan et al. 2005). Next, the high degree of molecular complementarity required to build the cross-β conformation explains how, while the formation of amyloid-like structures is thermodynamically driven by the backbone-mediated hydrogen bond network, its assembly is limited by the requirement of amino acidic combinations able to provide appropriate physico-chemical properties and shape complementarity. In addition, residues whose side chains are responsible for the contacts between opposite strands sustaining the solvent-protected "steric zipper" tend to have an apolar character. Because the geometry of the β-conformation results in contiguous amino acids pointing out in opposite directions, this implies residues that do not

participate of the "steric zipper" would be located in the solvent-exposed face of the strand, explaining why a sequential pattern that alternates hydrophobic and hydrophilic amino acids is well accommodated by amyloid-like structures.

## 7.3 Prediction of Protein Aggregation from the Primary Sequence

We provide here a detailed description on how the elucidation of the physico-chemical, sequential, and structural determinants of protein aggregation into amyloid-like structures has been exploited to develop a variety of mathematical tools intended for the accurate prediction of the deposition propensities of polypeptides. Although these methods have also been employed for the analysis of the aggregation and propagation of prions and prion-like proteins, the singular features of this particular kind of amyloids have led to the development of specific tools for its prediction (Alberti et al. 2009; Toombs et al. 2012; Espinosa Angarica et al. 2013; Lancaster et al. 2014; Sabate et al. 2015; Zambrano et al. 2015a), whose underlying rationale lies out of the scope of this chapter.

The improved understanding of the determinants of protein aggregation described above, and the realization that they are mostly encoded in the primary sequence, has inspired the development of a variety of mathematical methods that aim to predict *in silico* the propensity of a given polypeptide chain to aggregate, requiring solely the knowledge of its primary structure. To date, more than 20 such computational tools have been made public (Table 7.2), each of them focusing on the analysis of a particular set of determinants of protein aggregation to perform its prediction. Depending on the nature of the determinants of protein aggregation evaluated and on the rationale of the approach employed in order to implement their predictions, the methods can be classified into three main families (Caflisch 2006; Belli et al. 2011). **Empirical or phenomenological** predictors are based on the experimental assessment of the different intrinsic determinants of protein aggregation. On the other hand, **structure-based** approaches rely on the analysis of the conformational compatibility of sequence stretches within the evaluated polypeptide against the structural determinants of amyloid-like structures. Most methods in this second class approximate such suitability by focusing on the assessment of the specific features β-strands or β-sheets adopt when they assemble into a cross-β supersecondary conformation. Finally, **consensus** methods depart from the premise that the analysis of a particular determinant, or a discrete set of determinants, is not sufficient for an accurate prediction of APRs. Therefore, these predictors attempt to identify these "Hot Spot" by defining a consensus prediction from the outcome of other methods, both phenomenological and structure-based.

Aside from the differences in the properties under evaluation (which define the class they are ascribed to), and in their mathematical implementation, the predictors may also vary in the type of output they provide—though, it commonly comprises the identification of APRs along the polypeptide sequence together with their

**Table 7.2** Methods for the prediction of protein aggregation relying on the analysis of the primary structure (linear predictors)

| | Method | Underlying principle | Level of development | URL | References |
|---|---|---|---|---|---|
| Phenomenological | Chiti et al. (2003) | Rationalization of the impact of mutations on the aggregation kinetics, based on hydrophobicity, secondary structure propensity, and net charge | Equation | | Chiti et al. (2003) |
| | Dubay et (al. 2004) | Refinement of the equation by Chiti et al. in order to predict aggregation rates by considering hydrophobicity, net charge, hydrophobic/hydrophilic patterns, pH, ionic strength, and polypeptide concentration | Equation | | DuBay et al. (2004) |
| | Pawar et al. (2005) | Adaptation of the expression by Dubay et al. so as to derive intrinsic aggregation propensity scales at different pH for the 20 naturally-occurring proteinogenic amino acids, on the basis of hydrophobicity, secondary structure propensity, hydrophobic/hydrophilic patterning, and net charge | Equation | | Pawar et al. (2005) |
| | Tartaglia et al. (2004) | Rationalization of the impact of mutations on the aggregation rate, according to β-sheet propensity, accessible surface area, π-stacking interactions, and dipolar moment of side chains | Equation | | Tartaglia et al.(2004) |
| | Tartaglia et al. (2005a) | Adaptation of the equation by Tartaglia et al. in order to predict β-aggregating stretches along the sequence, with discrimination of their preferred (either parallel or antiparallel) orientation | Equation | | Tartaglia et al. (2005a) |
| | Zyggregator | Development of the equation by Pawar et al. in order to implement the impact of gatekeeper residues along the sequence and the influence of structural protection against aggregation | Server | www-mvsoftware.ch.cam.ac.uk/ | Tartaglia and Vendruscolo (2008); Tartaglia et al. (2008) |
| | TANGO | Estimation of the population of different states, including β-aggegates, according to a partition function that considers amino acid physico-chemical properties and conformational preferences, as well as extrinsic physico-chemical parameters | Server | tango.crg.es/ | Fernandez-Escamilla et al. (2004) |
| | Idicula-Thomas & Balaji | Calculation of an amyloidogenic propensity score on the basis of tripeptide-based secondary structure propensity along the sequence, compositional bias towards order-promoting residues, and estimated protein half-life and thermostability | Equation | | Idicula-Thomas and Balaji (2005) |

(continued)

**Table 7.2** (continued)

| | Method | Underlying principle | Level of development | URL | References |
|---|---|---|---|---|---|
| | AGGRESCAN | Experimental determination of an *in vivo* aggregation propensity scale for the 20 naturally-occurring proteinogenic amino acids | Sever | bioinf.uab.es/aggrescan/ | Conchillo-Solé et al. (2007) |
| | SALSA | Calculation of an averaged tendency to adopt β-strand conformation, according to the Chou and Fasman secondary structure propensity scale | Equation | Available through AmylPred 2 | Zibaee et al. (2007) |
| | Pafig | Statistical selection of physico-chemical properties allowing to discriminate hexapeptides forming amyloid-like structure | Software | mobioinfor.cn/pafig/index.htm | Tian et al. (2009) |
| Structure-based | NetCSSP | Detection of hidden β-propensity through contact-dependent secondary structure prediction, employing artificial neural networks | Server | cssp2.sookmyung.ac.kr/ | Yoon and Welsh (2004, 2005); Yoon et al. (2007); Kim et al. (2009) |
| | SecStr | Consensus detection of coequal α and β conformation propensities by at least 3 of 6 different secondary structure predictors | Software | Available through AmylPred2 | Hamodrakas et al. (2007) |
| | FoldAmyloid | Determination, for the 20 naturally-occurring proteinogenic amino acids, of an "average packing density" and different H-bonding probability scales derived from protein structural data | Server | bioinfo.protres.ru/fold-amyloid/ | Galzitskaya et al.( 2006a); Garbuzynskiy et al. (2010) |
| | PASTA 2.0 | Calculation of β-pairing probability between polypeptide stretches, on the basis of interaction potentials statistically derived from amino acid pairs occurrences in experimentally-resolved β-sheets, either parallel or antiparallel | Server | protein.bio.unipd.it/pasta2/ | Trovato et al. (2006); Walsh et al. (2014) |
| | Saiki et al. | Calculation of a suitability score for sequence stretches to fit a predefined amyloid structural template, according to hydrophobic and H-bonding interactions between contiguous side chains in hydrogen-bonded β-strands (computed employing presupposed hydrophobic and H-bonding parameters for different groups of amino acids) | Equation | | Saiki et al. (2006) |
| | PIMA | Calculation of β-pairing interaction energy for polypeptide segments of variable length threaded onto an in-register (either parallel or antiparallel) β-sheet template, employing a physics-based energy potential | Equation | | Bui et al. (2008) |

(continued)

**Table 7.2** (continued)

| | Method | Underlying principle | Level of development | URL | References |
|---|---|---|---|---|---|
| | BETASCAN | Estimation of β-strand pairing propensity, according to probabilities of residue pairs to be H-bonded in amphiphilic β-sheets | Server | groups.csail.mit.edu/cb/betascan/ | Bryan et al. (2009) |
| | 3D Profile | Conformational modelling to structural templates derived from hexapeptides forming amyloid-like structure, employing a physics-based forcefield | Server | services.mbi.ucla.edu/zipperdb/ | Thompson et al.( 2006); Goldschmidt et al. (2010) |
| | Pre-Amyl | Conformational modelling to structural templates derived from the coordinates of the amyloid-like crystal formed by NNQQNY, employing statistically derived interaction potentials | Server | Available through AmylPred2 | Zhang et al. (2007) |
| | Amyloidogenic Pattern | Determination of a sequential pattern for amyloidogenicity, based on the intensive mutational analysis of the STVIIE peptide able to form amyloid-like structure | Amino acid pattern | Available through AmylPred2 | Lopez de la Paz and Serrano (2004) |
| | Waltz | Identification of amyloidogenic polypeptide regions based on a PSSM allowing to differentiate hexapeptides forming amyloid-like structure from non-forming ones; complemented with a parameter evaluating physico-chemical properties important for amyloid-like assembly, and another structural factor assessing conformational fitting to an amyloid-like template | Server | waltz.switchlab.org/ | Maurer-Stroh et al. (2010) |
| | FISH Amyloid | Discrimination between patterns of position-specific amino acid co-occurrence associated to amyloidogenic or non-amyloidogenic polypeptide stretches, employing a machine learning approach | Server | comprec.pwr.wroc.pl/fish/fish.php | Gasior and Kotulska (2014) |
| | GAP | Differential potentials for amino acid pairs in amyloid-like or β-amorphous hexapeptides, derived from position-specific pairing frequencies with discrimination of their relative orientation along the β-strand | Server | http://www.iitm.ac.in/bioinfo/GAP/ | Gromiha et al. (2012); Thangakani et al. (2013); Thangakani et al. (2014) |
| | AmyloidMutants | Evaluation of the population of different accessible states (restricted topologically to conform with known amyloid structural models) employing a partition function, according to statistically derived amino acid interaction potentials | Server | amyloid.csail.mit.edu/ | O'Donnell et al.( 2011) |

(continued)

**Table 7.2** (continued)

| | Method | Underlying principle | Level of development | URL | References |
|---|---|---|---|---|---|
| | STITCHER | Calculation of β-strand pairing probability (employing BETASCAN scores) and their most likely assembly into a topologically constrained model of natural amyloids following additional energetic rules, including Q/N- and π-stacking, van der Waals interactions, and inter-strand linker entropy | Server | stitcher.csail.mit.edu/ (broken) | Bryan et al. (2012) |
| Consensus | AmylPred2 | Consensus between the output from at least n/2 of n different aggregation predictors selected | Server | aias.biol.uoa.gr/AMYLPRED2/ | Frousios et al. (2009); Tsolis et al. (2013) |
| | MetAmyl | Statistical derivation of weighting parameters for a linear combination of outputs from other aggregation predictors | Server | metamyl.genouest.org/ | Emily et al. (2013) |

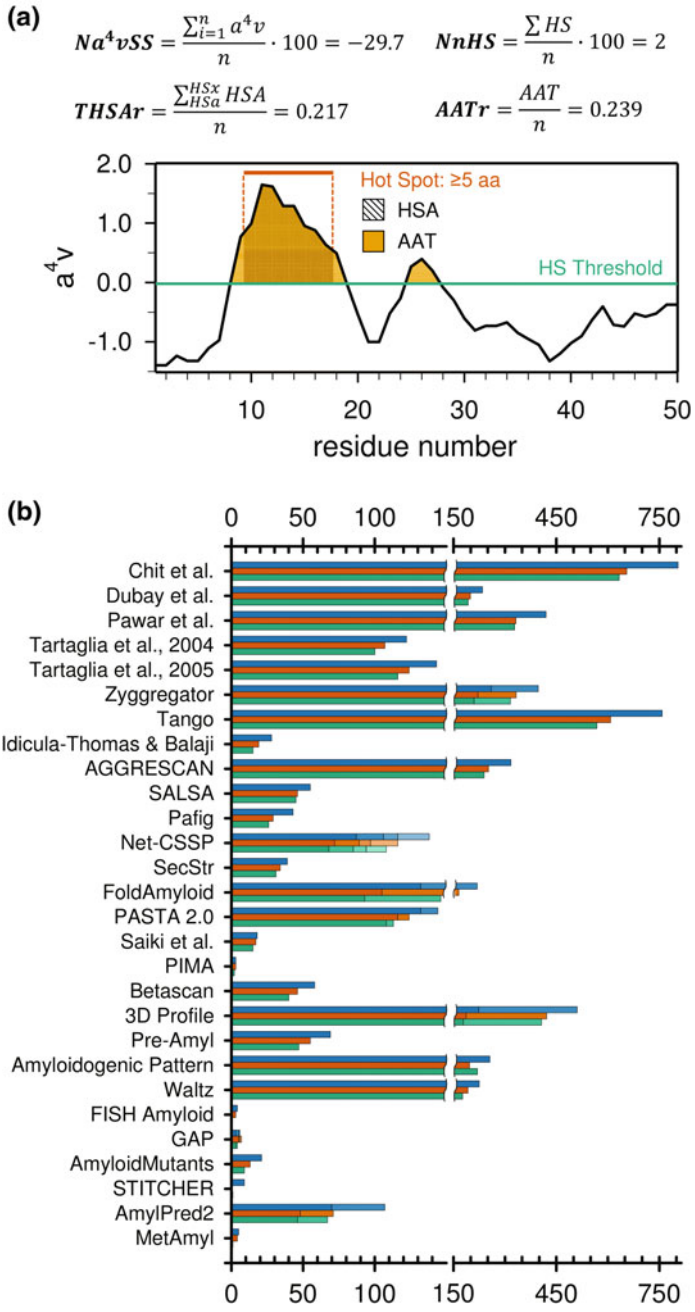
aggregative potential, as well as the relative or virtually absolute tendency to aggregate for the whole protein (Fig. 7.2a). Some approaches provide additional valuable estimates such as the nature of the pairing between β-strands (either parallel or antiparallel) in the β-sheets that could form the cross-β core of an amyloid-like fibril; or even attempt to forecast the quaternary assembly of the polypeptide chain in the amyloid-like structure.

In this section, we provide a brief description of the prediction methods that have been most widely exploited by the scientific community working in the field of protein aggregation (Fig. 7.2b), and of all those which have been employed to build up consensus predictors.

### *7.3.1 Phenomenological Approaches*

The first mathematical tool developed with the aim of predicting protein aggregation was an empirical equation derived from experimental data on the aggregation kinetics of different protein variants (Chiti et al. 2003). This equation allows the calculation for unstructured proteins or peptides of changes, upon mutation, in the rate of aggregation into amyloid-like structures—on the basis of the changes in hydrophobicity, in the propensity to convert from α-helical to β-sheet conformation, and in the net charge of the polypeptide caused by such mutation. A further refinement of this equation led to the development of an algorithm that allowed the calculation of fibril elongation rates, from fully or partially unfolded conformations, by considering seven variables—including intrinsic parameters such as hydrophobicity, net charge of the polypeptide, and the presence of alternating patterns of hydrophobic and hydrophilic residues, as well as extrinsic factors like the pH, ionic strength, and polypeptide concentration (DuBay et al. 2004). This algorithm was later adapted to estimate the intrinsic aggregation propensity of the 20 naturally-occurring amino acids. In this way, it was possible to calculate the aggregation propensity within a polypeptide sequence in a position-specific manner, by assigning to each residue the average intrinsic aggregation propensity of a sliding of amino acids centred on it (Pawar et al. 2005). This individual value of aggregation propensity is normalized relative to a reference value computed for random sequences with the same length than the analyzed sequence and with the amino acid frequencies found in the Swiss-Prot database. Therefore, this modification of the method by Dubay et al. allows the detection of APRs as those fragments of the sequence with consecutive residues possessing an intrinsic aggregation propensity above one standard deviation of the reference value.

Following a similar rationale, Tartaglia and co-workers developed a function to calculate the change in the aggregation rate upon mutation that, aside from the induced change in the propensity to adopt a β-sheet conformation and in charge, also takes into account changes in the accessible surface area, the number of aromatic residues (which influence the extent of π-stacking), and the dipolar moment of polar side chains (Tartaglia et al. 2004). An advantage of this method is the


(a)
$Na^4vSS = \frac{\sum_{i=1}^{n} a^4v}{n} \cdot 100 = -29.7$
$NnHS = \frac{\sum HS}{n} \cdot 100 = 2$
$THSAr = \frac{\sum_{HSa}^{HSx} HSA}{n} = 0.217$
$AATr = \frac{AAT}{n} = 0.239$
Hot Spot: ≥5 aa
HSA
AAT
HS Threshold
a4v
residue number
(b)
Chit et al.
Dubay et al.
Pawar et al.
Tartaglia et al., 2004
Tartaglia et al., 2005
Zyggregator
Tango
Idicula-Thomas & Balaji
AGGRESCAN
SALSA
Pafig
Net-CSSP
SecStr
FoldAmyloid
PASTA 2.0
Saiki et al.
PIMA
Betascan
3D Profile
Pre-Amyl
Amyloidogenic Pattern
Waltz
FISH Amyloid
GAP
AmyloidMutants
STITCHER
AmylPred2
MetAmyl

◂**Fig. 7.2** Typical output of a linear predictor of aggregation and spread within the field of the different methods developed for the prediction of protein aggregation. **a** The output of methods that rely on the analysis of the primary structure commonly comprises a profile along the sequence of a parameter scoring the tendency to aggregate, "Hot Spots" or APRs are usually identified as regions of the polypeptide above a minimal length that present values of such parameter overpassing a defined threshold. The prediction of the aggregative properties with AGGRESCAN (Conchillo-Solé et al. 2007) for a hypothetical protein sequence is shown as an example: the AGGRESCAN score for each position is computed as the average intrinsic aggregation propensity *in vivo* for a window of amino acids centred on the residue under scrutiny ($a^4v$), a "Hot Spot" is detected whenever a polypeptide stretch of 5 or more residues displays $a^4v$ values above the specified threshold. Several parameters can be derived from this profile that provide additional information on the propensity to aggregate, such as the normalized aggregation propensity score for the whole polypeptide ($Na^4vSS$) or the normalized number of "Hot Spots" found (NnHS), which allow a comparative analysis of the aggregative potential between individual proteins, sets of proteins or even entire proteomes. The normalized area above the threshold (AATr) and "Hot Spot" area (THSAr) advise, respectively, about the concentration of the aggregative potential along the sequence and the relevant strength of the APRs identified, these parameters can also be employed for comparisons between sets of proteins. **b** The impact and influence of the different predictors on the field of protein aggregation may be approximated through the citations statistics of their associated publications. The number of citations for the articles related to each method (as per Table 7.2) were retrieved from Google Scholar (*blue*), Scopus (*vermillion*) or Web of Science (*green*) between August 8th and 27th, 2015. Stacked bars represent different articles associated to the same predictor

absence of free parameters, thus allowing a broader generalization for the prediction of aggregation. This method was subsequently modified, as well, in order to allow for the prediction of absolute aggregation rates and the detection of APRs, introducing also the estimation of the preferred orientation (either parallel or antiparallel) of the β-aggregating segments (Tartaglia et al. 2005b).

Zyggregator implements the concept initially introduced by Chiti and co-workers to predict protein aggregation. It adds to the equation by Pawar et al. a parameter that accounts for the impact of gatekeeper residues against aggregation and (in contrast to the previous developments that only allowed to predict aggregation from fully or partially unfolded states) it also incorporates the influence of local structural stability (Tartaglia and Vendruscolo 2008; Tartaglia et al. 2008)—on the basis of the prediction of the flexibility and solvent accessibility of the polypeptide chain, as implemented in the CamP method (Tartaglia et al. 2007). Therefore, this algorithm allows to approximate, as well, the propensity of structured proteins to aggregate, although it cannot be strictly included among the algorithms which rely on tertiary structure analysis, which are described in Sect. 7.4.

The first method that allowed the evaluation of the tendency of a protein to aggregate from its sequence was the TANGO algorithm (Fernandez-Escamilla et al. 2004). This algorithm is based on a statistical mechanics concept where several states are defined—including the random coil and native conformations, as well as the α-helix, β-turn, and β-sheet aggregate states—which are characterized by specific physico-chemical properties, together with empirically and statistically derived conformational preferences. TANGO calculates the population of fragments of the sequence on each state according to a partition function whereby its

population is proportional to the energy of that fragment in such state. This energy is obtained taking also into account physico-chemical variables such as the pH, temperature, and ionic strength, or extrinsic factors like trifluoroethanol (TFE) concentration. TANGO predicts a segment of the sequence to possess a tendency to aggregate when its length is of at least 5 residues and it populates the β-sheet aggregate state with a probability higher than 5%. In this way, TANGO was also the first method to allow for the detection of APRs within protein sequences.

On the other hand, AGGRESCAN was the first predictor of protein aggregation that was specifically based on empirical data obtained *in vivo*. Prokaryotic cells have recently emerged as suitable model systems to study the mechanisms of amyloid formation (de Groot et al. 2009; Villar-Piqué and Ventura 2012). Although the formation of intracellular aggregates during recombinant protein expression was long considered to result from the nonspecific association of folding intermediates, leading to amorphous deposits, it has been shown that, indeed, the aggregates formed by different amyloidogenic proteins in bacteria display clear amyloid features, including cytotoxicity (Carrió et al. 2005; Dasari et al. 2011). By exploiting a strategy developed to determine protein aggregation in bacterial cells employing the green fluorescent protein GFP (Waldo and Standish 1999), a library of point mutants of the Aβ42 peptide fused to GFP was constructed—this allowed a scale of intrinsic aggregation propensity *in vivo* for the 20 naturally-occurring amino acids to be established (de Groot et al. 2006). AGGRESCAN predicts aggregation propensity from the primary sequence by computing for each amino acid the average aggregation propensity of a variable window, depending on the size of the protein, centred at this position (Conchillo-Solé et al. 2007). An APR or "Hot Spot" is identified whenever a stretch of 5 or more consecutive residues is detected within the polypeptide with computed aggregation propensities above the defined threshold, which corresponds to the average value of the aggregation propensity scale. The AGGRESCAN algorithm has been implemented as a web server and constitutes an extremely versatile tool—allowing the calculation of aggregation properties of either single polypeptides or large protein ensembles, and providing multiple parameters to establish comparisons between individual proteins or protein datasets, such as size-normalized absolute aggregation propensities and number of APRs per molecule, and indicators of the aggregative potency of the detected APRs.

A related phenomenological approach is represented by methods that rely on the assessment of previously established scales of physico-chemical properties of amino acids. The Simple ALgorithm for Sliding Averages (SALSA) assumes a strong correlation between the propensity to adopt β-strand conformation and the ability to form amyloid-like fibrillar structures (Zibaee et al. 2007). Therefore, it attempts to identify APRs as regions of the polypeptide sequence with a strong β-strand propensity. It does so by assigning to each residue the mean β-strand propensity of different averaging windows centred on it, according to the secondary structure propensities defined by Chou and Fasman (1974).

In turn, Pafig implemented a 2-round statistical scheme to select physico-chemical property scales that allow to significantly discriminate between two equally-sized populations in the Hexpepset dataset (Tian et al. 2009). One contains 1226 6-residue fragments derived from peptides or protein regions known to be involved in the formation amyloid-like fibrils; the other is composed of 876 6-residue stretches from peptides or polypeptide sequences which have been shown unable to form amyloid-like fibrils and 350 random hexapeptidic stretches outside the experimentally confirmed amyloidogenic determinants of Transthyretin, the Major prion protein, Apolipoprotein A-I, α-Synuclein and β2-Microglobulin. In a first selection round, a support vector machine was employed to select among an initial set of 531 physico-chemical property scales for the 20 naturally-occurring proteinogenic amino acids, as reported in the amino acid index database (Tomii and Kanehisa 1996; Kawashima et al. 2007), those allowing the separation of amyloidogenic from non-amyloidogenic hexapeptides with a certain level of accuracy. The final set of 41 property scales was subsequently defined employing a standard genetic algorithm, whose parameters were employed to compute the amyloid-like structural potential for the collection of 64 million possible combinations of 6-residue amino acid stretches. The certainty of classification provided by this machine learning strategy was assessed by defining a reliability index. Accordingly, an APR is detected by this method whenever a polypeptide sequence stretch matches an hexapeptide with high amyloidogenic propensity and a reliability index equal or above a fixed value.

### *7.3.2 Structure-Based Approaches*

The structure-based approaches rely on specific structural features associated with the formation of ordered amyloid-like aggregates. The first method of this kind was the NetCSSP algorithm, which exploits the concept of "chameleon sequences" or "conformational switches" as relevant conformational transition triggers in the formation of amyloid-like structure. This approach is based in the observation that the regular secondary structure adopted by a polypeptide is dependent on its tertiary contacts (Minor and Kim 1996), in such a way that certain sequences, that not adopt a β-conformation in the context of their structural environment, may encode a hidden β-propensity (Yoon and Welsh 2004). In this sense, Yoon and Welsh have developed the Contact-dependent Secondary Structure Prediction (CSSP) algorithm, employing artificial neural networks in order to detect sequence stretches with noticeable hidden β-propensity, which could act as potential "conformational switches" (Yoon and Welsh 2004; Yoon et al. 2007). A similar detection strategy is implemented by SecStr, a tool based on the consensus of six different methods for the prediction of secondary structure (Hamodrakas 1988). This algorithm cannot be strictly considered a predictor of aggregation propensity, since it was initially intended for the prediction of secondary structure, but its authors propose it can be useful in the detection of "conformational switches" (Hamodrakas et al. 2007).

Therefore, an APR is defined by SecStr whenever a stretch of the polypeptide sequence is simultaneously predicted to hold a similar propensity to adopt both α-helical and β-sheet conformations, following a consensus of at least 3 of the implemented methods.

FoldAmyloid exploits different concepts associated to the conformational properties of amyloid-like structures. This method exploits the "average packing density" approach, based originally on the consideration that the observed packing density of amino acids is associated with the conformational properties of the polypeptide chain (Galzitskaya et al. 2006a), in a way that amyloid-like structures are characterized by a notably high packing density. A mean packing density scale was defined for the 20 naturally-occurring amino acids by averaging their "observed packing density", defined as the number of contacts established with other residues. These contacts are computed by considering any neighbouring residue with at least one atom other than hydrogen within a 8Å radius of the amino acid under scrutiny, using a dataset of protein structures with less than 25% sequence identity and belonging to the four major structural classes (all-α, all-β, α/β and α+β) defined according to the SCOP scheme (Murzin et al. 1995). The method was extended to account for the relevance of hydrogen-bonding both as the main stabilizing force of the extended β-sheets forming the cross-β core of amyloid-like aggregates, and in the establishment of side chain stacking interactions by polypeptide sequences enriched in Gln and Asn (Michelitsch and Weissman 2000; Nelson et al. 2005). The same structural dataset used to derive the average packing density scale was also employed to calculate the statistics of the different types of hydrogen-bonding (backbone to backbone, between backbone and side chains or involving only side chains), these were subsequently translated into scales of probability for each individual amino acid to participate in the establishment of different classes of hydrogen bonds, either as donor or as acceptor (Garbuzynskiy et al. 2010). FoldAmyloid predicts APRs by exploiting these different scales to calculate a profile of averaged values along the sequence, employing a sliding window of 5-residues length by default. The method offers the possibility to select the scale to be considered for the calculation of the profile, either the average packing density or any of the distinct hydrogen-bonding scales, or hybrid scales resulting from different scales combinations. In the latter case, scale values are obtained by summing scale-specific normalized scores for each amino acid. An APR is detected according to FoldAmyloid when a stretch of consecutive residues presents scores in the averaged profile above a certain threshold. The cutoff values for each of the avaliable scales were statistically inferred as those allowing a better discrimination between peptides able to form amyloid-like structures and non-amyloidogenic ones within a dataset of peptides retrieved from the literature (Fernandez-Escamilla et al. 2004; Thompson et al. 2006).

In contrast to the previous methods, a majority of structure-based approaches are based on analysis of the specific features of β-sheet structure in the cross-β core of amyloid-like aggregates. One such method is the Prediction of Amyloid Structure Aggregation (PASTA) algorithm, which relies on the idea that β-conformation in amyloid structure corresponds to pairings of in-register β-sheets, either parallel or

antiparallel. PASTA calculates pairing energies for intermolecular β-sheets on the basis of individual pairing energies between each two amino acids facing each other in the β-sheet; such individual pairing energies have been statistically derived from the observed distribution of every possible pair of facing amino acids found in β-sheet conformation (either parallel or antiparallel), within a refined dataset of non-redundant crystal structures of globular proteins with diverse folds (Trovato et al. 2006). β-sheet energies are calculated by computing intermolecular pairings, both in a parallel or antiparallel fashion, between identical continuous stretches of variable length within the input sequence, while the rest of the polypeptide is considered disordered. The total pairing energies result from the sum of the individual pairing energies of each couple of contacting residues, after introducing a correction for the loss of entropy derived from the ordering of residues within these stretches. Those pairings of stretches with energies below a certain threshold are considered to present an increased likelihood to embody the cross-β core in amyloid-like structures and are identified as APRs. In contrast with other prediction tools, PASTA predicts, accordingly, the preferred pairing orientation—whether parallel or antiparallel—APRs would adopt in that cross-β core.

An analogous approach is employed by BETASCAN (Bryan et al. 2009), which aims to identify β-strand pairings with greater probability to build a β-sheet encompassing the cross-β core of an amyloid-like structure. The propensity of segments along the polypeptide chain to adopt a β-strand conformation, and the likelihood of pairing between couples of these streches are calculated according to probability scores for residue pairs to be H-bonded in amphiphilic β-sheets. These scores were derived from the analysis of selected non-redundant structures from the Protein Data Bank (excluding protein folds structurally similar to the known amyloid-like architectures, like β-helices, in order to avoid an unwanted bias), by defining those pairwise probabilities as a function of the residue side chains orientation in the β-sheet—either in the hydrophobic face or in the hydrophilic one.

Another type of structure-based methods are those which rely on the properties of the solved three-dimensional structures of small peptides forming amyloid-like structures (Nelson et al. 2005; Sawaya et al. 2007). This kind of short polypeptide segments are assumed to initially nucleate the amyloid-like aggregation reaction, becoming further embedded in the core that sustains the mature amyloid-like fibril (Ventura et al. 2004; Ivanova et al. 2004). The first within this class of approaches was the 3D profile method (Thompson et al. 2006), which was initially developed by defining an static ensemble of structural templates through the relative displacement, along the three orthogonal axes, of the opposing β-sheets defined by the 3-dimensional coordinates of the NNQQNY peptide amyloid-like crystal (Nelson et al. 2005). The conformational fitting to this amyloid-like structural ensemble is assessed by threading, into all the templates, every hexapeptidic stretch of the polypeptide sequence not containing a Pro or a Cys, and computing its energy employing the physics-based energy function implemented in the Rosetta Design program (Kuhlman and Baker 2000). The method was further developed to model all the backbone templates experimentally determined to be compatible with the cross-β conformation (Sawaya et al. 2007), not only that of the NNQQNY

amyloid-like structure, and was also improved at the computational efficiency level by introducing a fuzzy search algorithm which looks gradually for the structural template that provides the best fit for the threaded polypeptide segment sequence (in terms of spacing and relative translation of the two sheets composing the cross-β structure), instead of energetically evaluating the whole ensemble of structural templates. Moreover, the fitting to a given structural template is now computed not only on the basis of the sequence energy calculated with Rosetta Design, but also according to additional scores derived from (i) the assessment of shape complementarity (Lawrence and Colman 1993) between the residue side chains building the "steric zipper", and (ii) the analysis of solvent exclusion from the zipper through the calculation of the solvent accessible surface area (Lee and Richards 1971). This combined fitness score is computed again by threading onto the templates every hexapeptidic fragment of the sequence not containing Pro, and with their Cys, if present, substituted with Ser (to avoid issues associated with modelling disulphide bonding). A segment of the polypeptide chain is predicted to possess a high propensity towards the formation of amyloid-like structures when its analysis yields an energy score below a defined threshold, which has been set up on the basis of the energetic values computed for the structurally resolved hexapeptidic steric zippers.

An ensemble of static structural templates based on the atomic coordinates of the NNQQNY amyloid-like crystal, similar to that employed in the initial development of the 3D profile method, was also implemented in the Pre-Amyl prediction algorithm (Zhang et al. 2007). In this case, however, the conformational fitting of hexapeptidic sequence stretches threaded into each template is not evaluated employing a physics-based forcefield but, instead, according to statistically derived residue interaction potentials—derived from the number of contacts observed experimentally (within a certain distance radius) between every possible pair of residues, in crystallographic protein structures retrieved from the Protein Data Bank with sequence identity lower than 30% and resolution better than 2Å. This observed number of contacts is normalized relative to a theoretically-expected number of contacts within the same radius. A given hexapeptide stretch is considered amyloidogenic whenever the structural configuration it populates with lowest energy yields a computed value below a given energy threshold, which was established after an statistical analysis of the ability of the method to differentiate between hexapeptides assembling into amyloid-like fibrils and those that do not form them, that correspond to the AmylHex database (Thompson et al. 2006).

A different approximation based on the analysis of fibril-forming peptides concentrates on the effort to identify the position-dependent compositional determinants along the sequence either favouring or disfavouring the assembly into amyloid-like structures. A very first approach addressed this issue by performing an exhaustive mutational analysis on the ability to form amyloid-like fibrils by the STVIIE peptide, whose amyloidogenic properties has been previously tailored employing a computational method (López de la Paz et al. 2002). This study allowed the determination of a sequential pattern which defined the compositional requirements for amyloidogenicity (López de la Paz and Serrano 2004)—at least in the vicinity of the STVIIE sequence space. The sequences generated by this

analysis were later incorporated into the AmylHex database (Thompson et al. 2006), composed of two subsets, one corresponding to hexapeptides able to form amyloid-like fibrils and another comprising non-amyloidogenic peptides. However, since a large fraction of these datasets is constituted by sequences derived from STVIIE (most of them corresponding to point mutations of the original peptide), the database is likely reporting on a closely related sequence space; so that the general compositional features for the universe of hexapeptidic sequences able to populate amyloid-like structures cannot be inferred. In order to overcome this limitation, the AmylHex database was extended to increase its sequential diversity by further defining candidate amyloidogenic sequences. These were drawn either by (i) applying a sequential pattern derived from the original AmylHex database to proteins forming amyloid-like fibrils whose amyloidogenic determinants had not been previously determined, (ii) engineering double and triple substitutions in the hydrophobic core of the STVIIE peptide, or (iii) employing another sequential profile (derived from a preliminarily extended AmylHex database) to analyze human proteins *a priori* unrelated to amyloidosis, and retrieve sequences with the lowest similarity to the original hexapeptide. The ability of these newly retrived hexapeptides to assemble into amyloid-like structures of the newly retrieved hexapeptidic sequences was assesed experimentally in order to appropiately classify them. The resulting extended AmylHex database—comprising 116 peptides able to form amyloid-like fibrils (positive dataset) and 103 hexapeptides which do not aggregate into this class of structures (negative dataset)—was employed to build a position-specific scoring matrix (PSSM) intended to capture the general sequential determinants of the aggregation into amyloid-like structures. This PSSM is the main workhorse of the Waltz algorithm (Maurer-Stroh et al. 2010) which aims to specifically identify amyloidogenic regions within polypeptide sequences according to a compound scoring function. This function incorporates, first, a sequential parameter which measures the position-dependent compositional suitability to adopt amyloid-like structure along the sequence, according to the Waltz PSSM. Second, the function includes a parameter to weight a series of amino acids' physico-chemical properties relevant for position-specific amyloid-like assembly. These latter features were statistically selected, from a collection of databases comprising some 700 sets of normalized physico-chemical parameters for each of the 20 naturally-occuring amino acids (Tomii and Kanehisa 1996; Eisenhaber et al. 1998; Kawashima et al. 2007), following two stages. The scales were initially selected to correlate with amino acid frequencies of non-redundant subsets of the AmylHex database (independently for the positive and the negative datasets), while exhibiting the strongest deviation between the positive AmylHex dataset and the UniRef50 reference dataset (the latter expected to provide randomized values for the properties). Then, 19 properties scales were further selected using a genetic algorithm that searchs for a better discrimination between the positive and negative datasets (these being supplemented with random polypeptide sequences to avoid overprediction). Finally, the Waltz function also employs a third structural parameter that evaluates the conformational fitting of the sequence stretch under scrutiny to the structural template of the GNNQQNY fibril-forming peptide (Nelson

et al. 2005). This is done employing a position-specific pseudoenergy matrix built on the basis of the energy difference, upon threading onto the GNNQQNY structural template, between polyamino acids corresponding to every possible combination of the 20 naturally-occurring proteinogenic amino acids and the polyAla sequence; these energies are calculated using the FoldX energy function (Guerois et al. 2002), which is further described in the Sect. 7.4 of this chapter.

A different structural approach based on the knowledge of the structures of amyloid-like aggregates is implemented by AmyloidMutants (O'Donnell et al. 2011). Differently from the 3D profile method and Pre-Amyl, which evaluate the conformational fitting of the polypeptide sequence to predefined structural templates, this is a statistical mechanics-based predictor which attempts to explore the conformational landscape of amyloid-like states that can be populated by the polypeptide chain. In order to allow for a computationally feasible conformational search, the accessible states are topologically restricted to conform with the structural features determined experimentally for certain natural amyloid-like structures—in particular, those of the β-solenoid (Lührs et al. 2005; Wasmer et al. 2008) and the superpleated β-sheet (Kajava et al. 2004) models. This method aims to predict not solely the portions of the sequence able to form cross-β supersecondary elements but their possible quaternary arrangement as well. Interestingly, the topological constraints employed to define the accessible states allow the modelling of β-strand distortions observed experimentally in certain amyloid-like structures (Wasmer et al. 2008). However, in AmyloidMutants intrachain β-sheets are restricted to have a parallel arrangement, while it is known that cross-β spines with an antiparallel arrangement could exist in nature (Sawaya et al. 2007). The population of the accessible states is obtained, according to a Boltzmann distribution, depending on the polypeptide energy in each state, which is computed employing interaction potentials derived statistically from non-redundant structures of the Protein Data Bank. In the case of AmyloidMutants, those potentials are obtained as a function of features of the structural context which might be relevant for the population of amyloid-like states; such as amphipathicity, solvent accessibility, the proximity to β-strand or β-sheet edges, distortions of the β-conformation and stacking interactions of identical residues. This method is also capable of simultaneously assessing the impact of sequential variation on the population distribution of the accessible amyloid-like states by allowing the specification of discrete amino acid substitutions. The output provided by AmyloidMutants consists of the representative members of structural clusters corresponding to the populated states, as well as the conformation of the polypeptide sequence within them.

### *7.3.3 Consensus Methods*

Each of the methods described above exploits one or several of the specific determinants considered as relevant for the aggregation of polypeptides into

amyloid-like ordered structures. However, no unique tool has been developed yet that incorporates the ensemble of concepts implemented by the different approaches, despite their underlying principles might turn out as complementary for the prediction of polypeptides propensity to form β-sheet-enriched aggregates. Combining the outputs provided by different algorithms might increase the prediction accuracy—by improving the sensitivity towards the most relevant determinants (which may divergently promote the formation of different aggregated states), while minimizing, at the same time, the method-specific bias towards overprediction of certain types of aggregation. This rationale has been employed for the AMYLPRED server, which builds a consensus APRs prediction at a residue level by integrating the results provided by different previously published methods. The initial version (Frousios et al. 2009) incorporated 5 methods (the "average packing density" method which later inspired FoldAmyloid, SecStr, the amyloidogenic pattern defined by Lopez de la Paz and Serrano (2004), TANGO and Pre-Amyl) to construct the consensus prediction. A more recent release, AmylPred2 (Tsolis et al. 2013), can combine up to 11 different methods (the aforementioned ones complemented with AGGRESCAN, AmyloidMutants, SALSA, NetCSSP, Pafig and Waltz). Nonetheless, it is also true that for these different predictors, and particularly among those belonging to the same class, the predictions might be redundantly reporting on certain determinants of aggregation, thus biasing the identification of consensus APRs. To reduce this bias, AmylPred2 allows the selection of the specific methods, among the 11 available, to be employed to compute the consensus prediction. Therefore, based on expert knowledge the user could define a customized combination of methods offering a better complementarity; though, this goal requires a deep understanding of the rationale behind the methods.

MetAmyl employs a statistical approach in order to construct a consensus method aimed to achieve a better complementarity between the methods it incorporates while minimizing their redundancy (Emily et al. 2013). To this end, this consensus tool is formulated as a linear combination of different predictors, with a series of parameters weighting the outcome of each individual method. MetAmyl exploits the expanded AmylHex dataset (Thompson et al. 2006; Maurer-Stroh et al. 2010), described above, in order to fit the weights of the methods by maximizing the sensitivity and specificity of the ability to predict amyloid-like structure. In a first round of development, MetAmyl included the output of 11 predictors (TANGO, AGGRESCAN, SALSA, Pafig, PASTA, Waltz and the five different scores provided by FoldAmyloid) as individual variables to fit the weighting parameters of the linear algorithm. This first step allowed identifying the degree of overlapping in the information provided by the different predictors through the analysis of the level of correlation between methods. In this way, SALSA, Pafig, Waltz, and the "average packing density" scale from FoldAmyloid were found as those reflecting the most informative properties with a higher complementarity, and were selected in order to reduce the dimensionality of the MetAmyl algorithm. The weighting parameters were further refitted to the extended AmylHex dataset, employing only the latter predictors, in order to yield the final formulation of the algorithm.

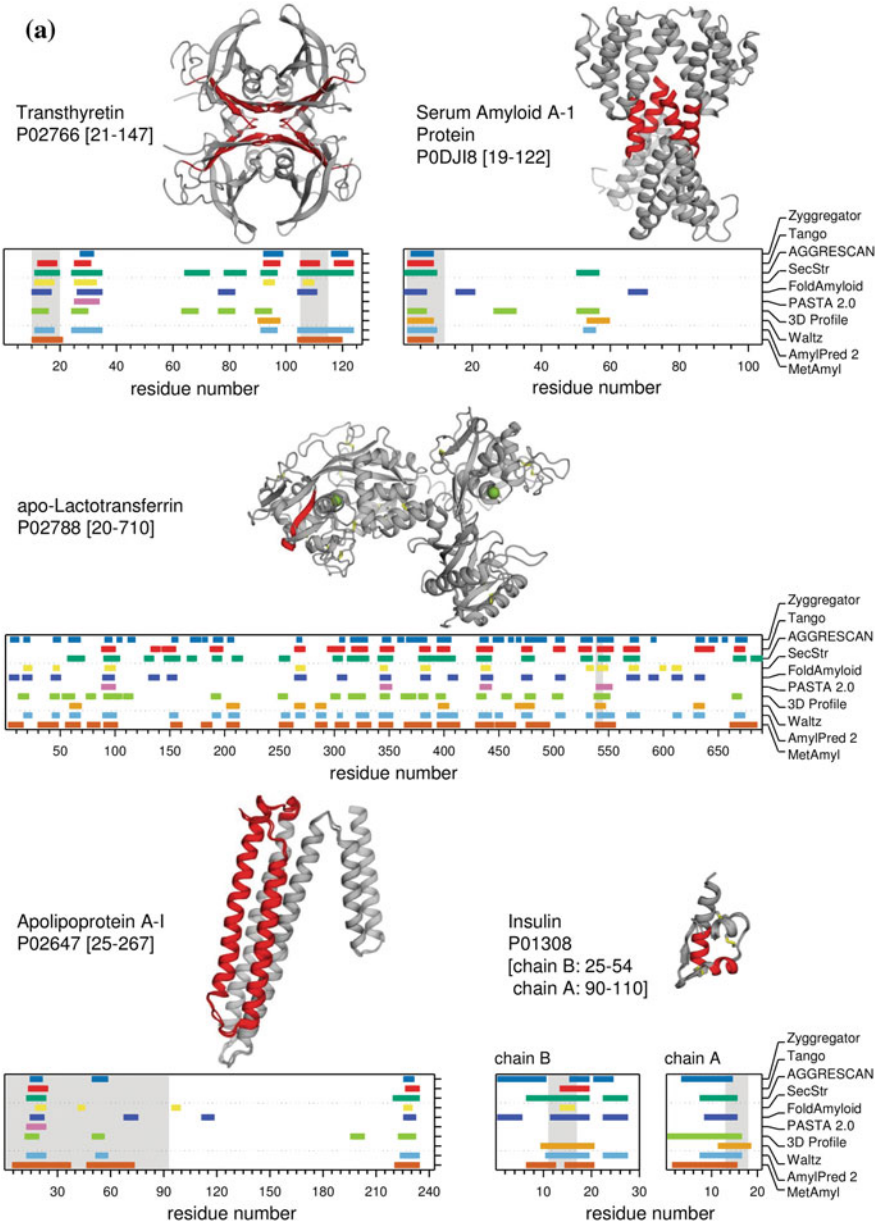
### 7.3.4 Applications of Sequence-Based Predictors

#### 7.3.4.1 Proteome-Wide Analyses

The computational tools described so far allow performing easy and straightforward analyses of the tendency to aggregate of individual proteins, as exemplified in Fig. 7.3, with the only requirement being knowledge of their primary structure. The outcome of such analyses can assist the user in the characterization of the aggregation process for a given protein of interest, and in the forecasting of the impact of sequential variations over this reaction. Additionally, some predictors have been built so as to allow high-throughput analyses of large sets of polypeptide sequences. Four of these methods—namely TANGO, AGGRESCAN, Zyggregator, and Waltz—have been intensively employed for the massive analysis of the aggregative properties of a variety of proteomes and protein datasets. The implementation of such large scale analyses has resulted in a deeper understanding of the determinants that influence the aggregation of polypeptides, it has concomitantly revealed signatures of a selective pressure acting on cellular proteomes along evolution in order to reduce their overall tendency to aggregate, and it has shown how this pressure has shaped protein sequences and structures (Monsellier and Chiti 2007; Reumers et al. 2009b; Castillo et al. 2011; Sanchez de Groot et al. 2012). These findings are further reviewed in this section.

An examination of the context where the APRs detected by these methods are placed, within the polypeptide chain, has revealed how these sequence stretches are frequently flanked either by charged residues like Arg, Lys, Asp, and Glu (whose function would be hampering the establishment of intermolecular interactions between APRs by providing repulsive charges) or by residues acting as β-sheet breakers, like Pro (Rousseau et al. 2006b). The presence of this type of residues (commonly referred to as "gatekeepers") at the flanks of APRs defines amino acidic patterns that are coincident with the substrate binding determinants of many chaperone classes. Accordingly, it has been suggested that the effect of the selective pressure against aggregation on protein sequences, leading to the emergence of "gatekeeper" residues, has concomitantly sculpted the binding specificity of those members of the protein quality control machinery that aid proteins to attain their native states or that block the establishment of non-functional intermolecular contacts.

High-throughput analyses of the aggregative properties of polypeptides agree that the presence of APRs represents an ubiquitous property of proteins, since, on average, the vast majority of polypeptides in a given proteome harbour at least one such predicted aggregation-prone stretch (Rousseau et al. 2006b; Conchillo-Solé et al. 2007; Reumers et al. 2009a). Nonetheless, the distribution of APRs is not uniform across the different classes of proteins embodying a certain proteome (Linding et al. 2004; Rousseau et al. 2006b; Conchillo-Solé et al. 2007). IDPs are particularly depleted in APR content, a feature intimately related to the compositional characteristics that differentiate these kind of polypeptides from globular


(a)
Transthyretin
P02766 [21-147]
Serum Amyloid A-1
Protein
P0DJI8 [19-122]
Zyggregator
Tango
AGGRESCAN
SecStr
FoldAmyloid
PASTA 2.0
3D Profile
Waltz
AmylPred 2
MetAmyl
residue number
apo-Lactotransferrin
P02788 [20-710]
Apolipoprotein A-I
P02647 [25-267]
Insulin
P01308
[chain B: 25-54
chain A: 90-110]
chain B
chain A

◂**Fig. 7.3** Detection, employing linear predictors, of APRs in human amyloidogenic proteins. APRs predicted by different methods that rely on the analysis of the primary structure are shown as *coloured bars* (missing bars might either imply no APR has been detected, the method failed to provide an outcome, or the protein under scrutiny constitutes a private precomputed entry not publicly available), *dotted lines* delimit groups of methods belonging to the same class (Phenomenological above, Structure-based in the middle and Consensus approaches in the lower section). In order to contrast the power and significance of the predictions, the amyloidogenic regions experimentally confirmed so far in the proteins analysed, as previously compiled by Hamodrakas and co-workers (Tsolis et al. 2013), are shown as *grey shadows* in the plot, and also mapped in *red* over the structures (when available). For each protein, the UniProt accession code of its precursor polypeptide is noted, and the region corresponding to the mature form, whose sequence was employed to run the predictions, is specified within brackets; structures, if represented, correspond to the mature polypeptide unless the N- and C-termini are explicitly indicated. The dynamic flexibility of native structures is highlighted by superimposing, when available, different NMR conformers solved or multiple unique chains in the crystallographic asymmetric unit. Disulphide bonds are depicted as *bright yellow* sticks and complexed chloride ions as *green* spheres. **a** and **b** Predictions for proteins with wild-type structures resolved experimentally under close-to-native conditions (PDB codes Transthyretin—1TTA, Serum amyloid A-1 protein—4IP8, apo-Lactotransferrin—1CB6, Apolipoprotein A-I—2A01, Insulin—1ZNJ, Prolactin—1RW5, Lysozyme C—1LZ1, Transforming growth factor-β-induced protein ig-h3—2LTB, Major prion protein—2LSB, β2-microglobulin—2XKS, Gelsolin—3FFN). For Transthyretin and Serum amyloid A-1 protein, the structure shown corresponds to their reported native quaternary assembly. **c** Predictions for proteins with experimentally resolved structures whose correspondence to the true native state is uncertain, either because they were resolved in the presence of detergents or large cofactors, or because they correspond to mutant sequences (PDB codes: β-amyloid protein 42—1Z0Q, Islet amyloid polypeptide—2L86, Cystatin-C—3NX0, Microtubule-associated protein Tau—2MZ7, α-synuclein—2KKW, Proapolipoprotein C-II—1O8T, Calcitonin—2JXZ). **d** Predictions for proteins without experimentally resolved structures available. For proteins in panels **c** and **d**, in order to estimate whether the APRs forecasted could be partially or totally exposed in their actual native states, the tendency along the polypeptide chain to adopt an ordered structure (Foldability) or to remain disordered was also predicted, according to the "average packing density" method (Galzitskaya et al. 2006a), employing FoldUnfold (Galzitskaya et al. 2006b); regions of the polypeptide chain with scores below a defined cut-off (shown as a *red line*) are predicted as disordered

proteins (Uversky 2002). More specifically, aggregation-prone residues are underrepresented in IDPs sequences, while an enrichment is observed in charged amino acids and β-breaker residues such as Pro (Tompa 2002), this provides them with a higher net charge and reduced hydrophobicity. Conversely, globular proteins present a significant number of detected APRs, but the analysis of their location in the native state of different sets of globular proteins with experimentally resolved structures reveals that these stretches tend to be buried (Linding et al. 2004; Buck et al. 2013), in many cases as part of the hydrophobic core, being therefore protected from the solvent and, thus, unable to establish deleterious intermolecular contacts under physiological conditions.

In spite of the ubiquitous presence of APRs in a large fraction of proteins from many, different proteomes, the predicted aggregation propensity is not homogenous across them but decreases with increasing complexity and longevity of the corresponding organism (Tartaglia et al. 2005b; Rousseau et al. 2006b). This observation

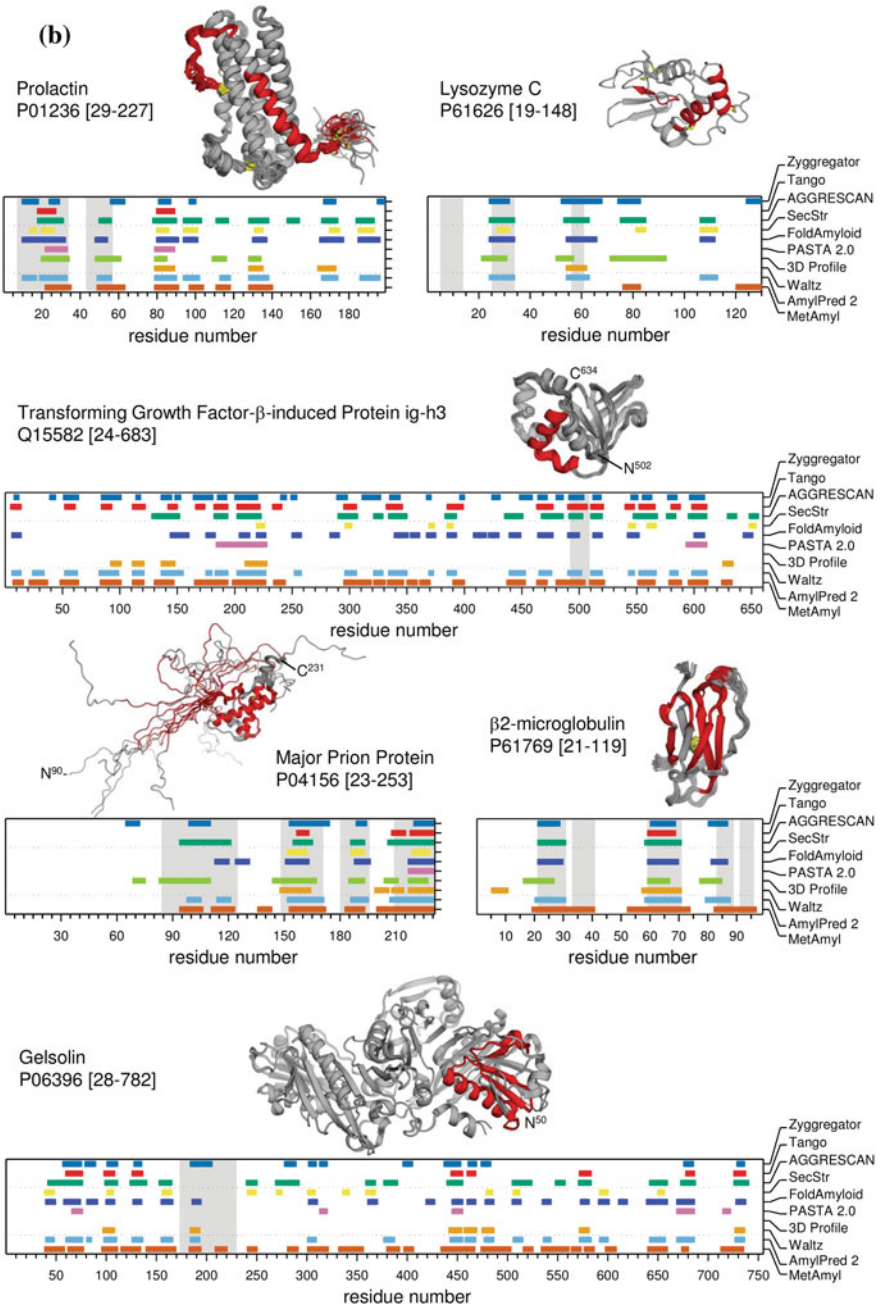



**Fig. 7.3** (continued)

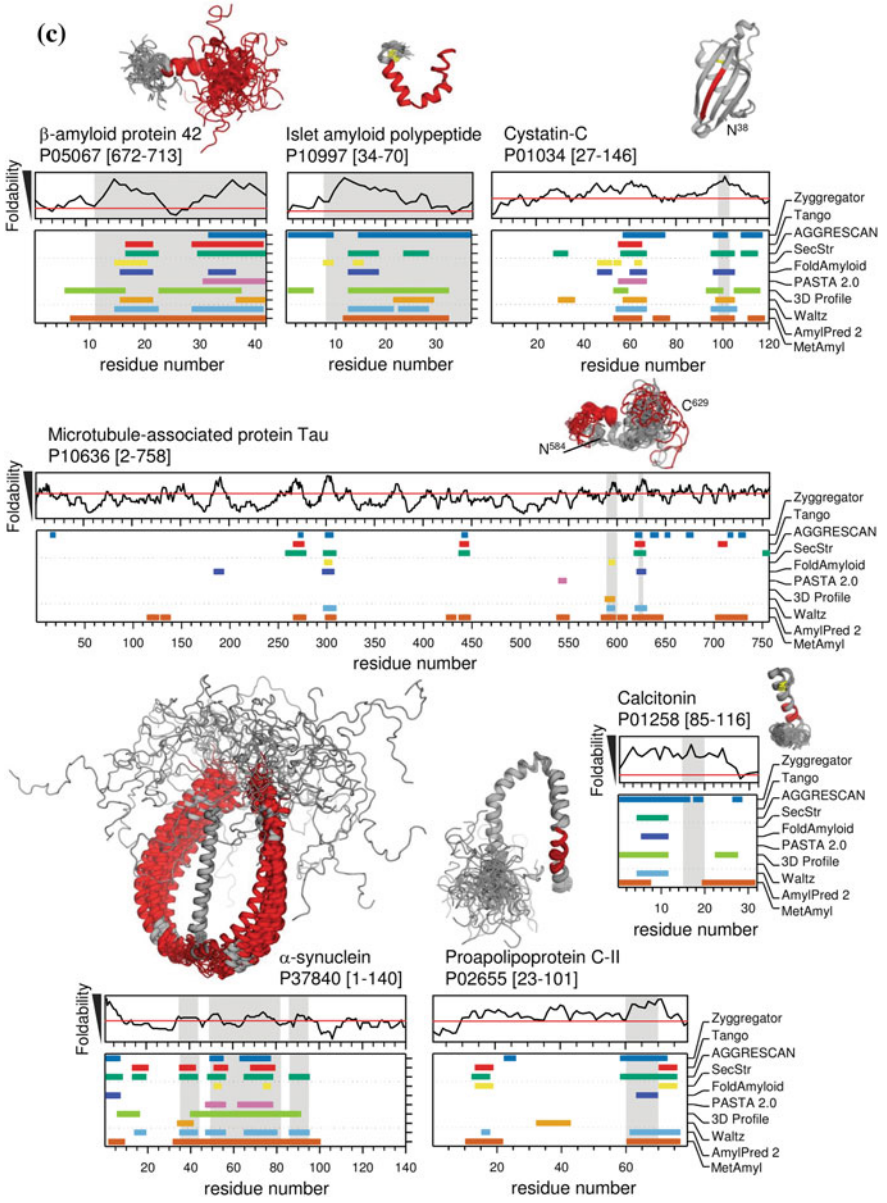



**Fig. 7.3** (continued)

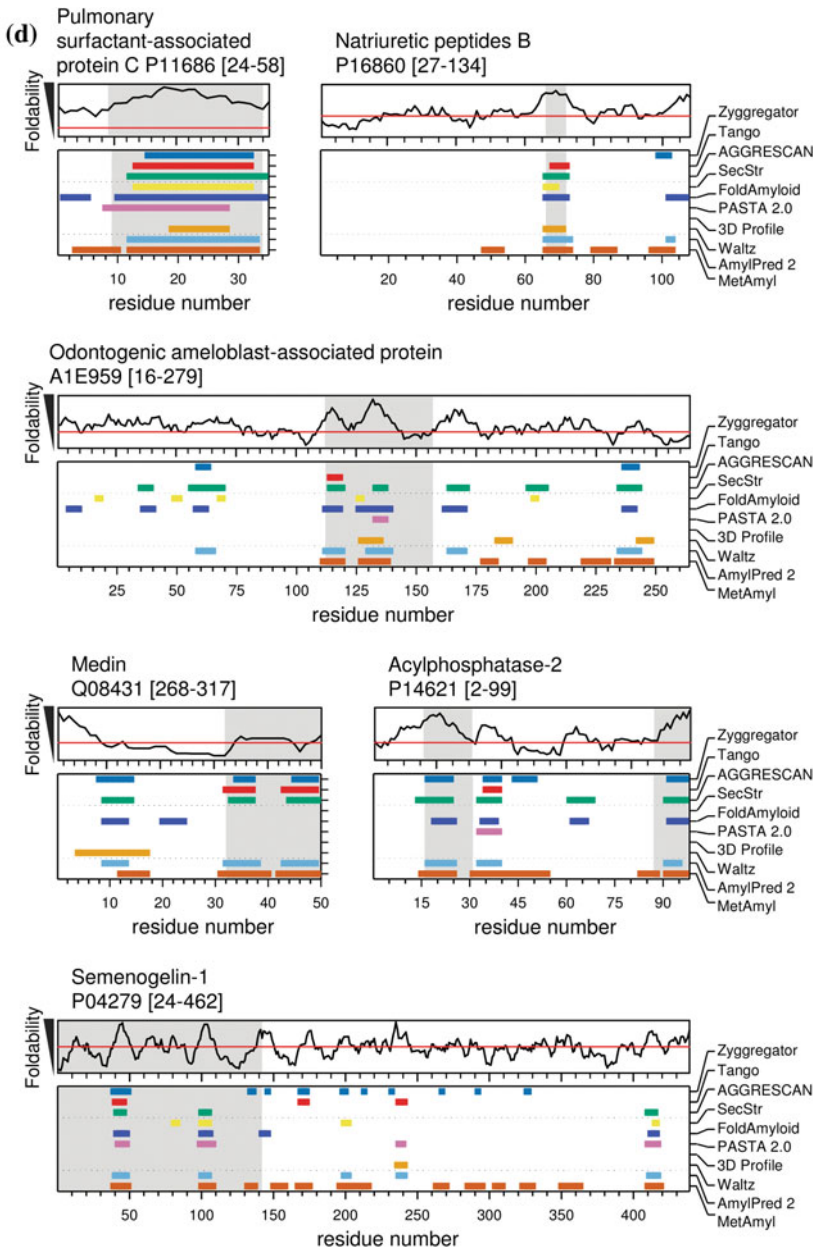



**Fig. 7.3** (continued)

has consistently been acknowledged employing unrelated aggregation predictors (Tartaglia et al. 2005b; Rousseau et al. 2006b; Graña-Montes et al. 2012b), and is taken as a robust evidence that evolutionary pressure acts on protein sequences in order to minimize their tendency to aggregate. Different detailed analyses performed with sequence-based methods have revealed apparent specific traits of this evolutionary selection against aggregation acting to shape protein sequences and structures, as well as their functional properties. For instance, the potency of the detected APRs was found to diminish with increasing polypeptide length (Monsellier et al. 2008)—this being consistent with theoretical estimates suggesting the rate of folding to the native state slows down as protein size increases (Ivankov et al. 2003), so that longer proteins are expected to populate partially unfolded states (where APRs may become substantially exposed) for a larger amount of time. On the other hand, protein aggregation is highly dependent on polypeptide concentration since it constitutes a high-order reaction, and, at the same time, the efficiency of the aggregation reaction is influenced by the degree of sequential identity between interacting molecules (Krebs et al. 2004; Wright et al. 2005). Consequently, proteins with a native homo-oligomeric architecture are at high a priori risk of aggregation because of the implicit high local concentration of identical polypeptide chains. However, proteins with this kind of quaternary structure generally present, indeed, a lower predicted aggregation propensity relative to other polypeptides of similar size (Chen and Dokholyan 2008). The same study also highlighted that proteins exerting essential functions present a lower predicted tendency to aggregate than non-essential polypeptides. These results suggest that a greater selective pressure to minimize their tendency to aggregate is experienced by proteins being at an inherently higher threat of aggregation or by those developing an essential functional role for the cell. Interestingly, the analysis of proteins with native oligomeric quaternary structures and protein-protein complexes reveals that the surfaces of interaction between polypeptides overlap spatially with detected APRs (Pechmann et al. 2009; Castillo and Ventura 2009). It has been proposed, accordingly, that the specific formation of stable quaternary structures and protein complexes could have evolved, aside from its straightforward functional implications, also as a protective strategy to avoid the establishment of non-functional unspecific intermolecular contacts, by exploiting the physical shielding of APRs. The significance of such a protective shielding strategy is further supported by the presence of disulphide bonds and attractive electrostatic interactions in the proximity of interfaces, enhancing both their stability and specificity (Pechmann et al. 2009). Several other mechanisms have been identified to have apparently been incorporated in cellular proteomes in order to confront the risk of aggregation. The incorporation of disulphide bonds in protein structures would be one such strategy, as derived from the observation that polypeptides born with disulphide bonds present a high predicted aggregation propensity compared to proteins devoid of these covalent links—such difference becomes substantially larger when the subset of extracellular proteins is considered alone (Mossuto et al. 2011; Graña-Montes et al. 2012a). These findings suggest the presence of disulphide bonds enables tolerance for a greater aggregative potential, particularly in

proteins that function in harsh environments (such as the extracellular space) where the maintenance of their native states might be compromised. That increased tolerance likely results from the disulphide cross-linking stabilizing effect on native conformations (Poland and Scheraga 1965; Lin et al. 1984; Grantcharova and Baker 2001; Graña-Montes et al. 2012a), but also because it constraints the kinetics of aggregation and leads to a reduced cytotoxicity of amyloid-like fibrils (Mossuto et al. 2011; Graña-Montes et al. 2012a). Likewise, the presence of intrinsically disordered regions—found frequently at the termini of globular proteins (Lobanov et al. 2010)—would represent yet another protective mechanism since the compositional bias of these regions provides them with a strikingly low predicted tendency to aggregate, compared to the ensemble of globular proteins (Graña-Montes et al. 2014). Although intrinsically disordered regions in a terminal position serve to develop a wide diversity of functions (Uversky 2013b), both theoretical and experimental models (Abeln and Frenkel 2008; De Simone et al. 2012; Santner et al. 2012) demonstrated that the structural flexibility of disordered regions adjacent to structured domains can exert a protective role against aggregation, by creating an excluded volume around the domains surface. This kind of dynamic behaviour, defining the entropic briste effect, thence prevents the establishment of spurious intermolecular. Remarkably, the anti-aggregative role of entropic bristles has been corroborated in the context of naturally-occurring protein domains (Graña-Montes et al. 2014), further supporting the evolutionary emergence of a widespread activity of these disordered termini against deleterious aggregation.

As mentioned before, aggregation kinetics are highly dependent on protein concentration, which suggests cellular protein levels should be tightly regulated in order to manage the risk of aggregation. Indeed, a correlation was observed between experimentally derived mRNA levels and predicted aggregation propensity in the human proteome (Tartaglia and Vendruscolo 2009), indicating that protein abundance is finely tuned at the expression level depending on the aggregation properties of polypeptides. The relationship between protein synthesis and the tendency to aggregate was later confirmed by the analysis of experimental data of gene expression, protein abundance, and protein solubility for polypeptides in the *Escherichia coli* proteome (Castillo et al. 2011). In this case, the best correlation with the aggregation properties of polypeptides is observed when real protein abundance is considered, thus indicating that are protein, and not mRNA, levels those under true evolutionary control. The detailed inspection of these data allows for the discrimination of polypeptide populations that appear to be under different selective pressure to avoid aggregation, because they exhibit contrasting tendencies to aggregate. Unsurprinsingly, the most abundant proteins experience a higher degree of anti-aggregative selection. After synthesis, polypeptide abundance is controlled by the degradation machinery—consequently, the analysis of human proteins has shown that long-lived polypeptides display an overall low aggregation propensity, whereas proteins with high turnover rates can tolerate an increased aggregation potential (De Baets et al. 2011), thus indicating that protein populations with a significant lifetime inside the cell are under stronger pressure to avoid the risk of aggregation. Along the same line, the control of protein concentration

becomes even more relevant when the extremely crowded nature of the cellular interior is taken into account (Zimmerman and Trach 1991; Ellis 2001), because macromolecular crowding is considered to increase dramatically the effective local protein concentration, as well as limiting biomacromolecule diffusibility. In fact, high molecular weight polymeric able to mimic the crowded intracellular environment (such as ficoll or polyethylene glycol) have been shown to accelerate the aggregation rates of polypeptides (van den Berg et al. 1999; Munishkina et al. 2004). Quite interestingly, the analysis of the aggregation propensity predicted either for bacterial (de Groot and Ventura 2010) or human (Tartaglia and Vendruscolo 2009) proteins, relative to their subcellular localization, reveals the predicted tendency to aggregate diminishes as the volume of the subcellular compartment decreases. This suggests that the impact of macromolecular crowding and reduced diffusibility on the risk of aggregation becomes more pronounced as the compartmental confinement accentuates and, accordingly, the pressure to decrease the intrinsic aggregation propensity appears stronger for proteins located in small cellular compartments.

Although the analyses of proteomes provide strong evidence for the existence of a selective pressure to avoid polypeptide aggregation, a persistent aggregative potential is still detected in a majority of protein sequences as most of them harbor at least one APR (Rousseau et al. 2006b; Conchillo-Solé et al. 2007; Reumers et al. 2009a). This suggests the purifying selection towards lower aggregation propensity in polypeptide sequences is limited to a certain extent, very likely because the maintenance of protein functionality requires sequential properties that overlap partially with the determinants of protein aggregation. In this context, the analysis of homogeneous ensembles of functionally or structurally related proteins has proved an extremely valuable strategy to identify specific functional constraints restricting the impact of selective pressure against aggregation, since this approach reduces significantly the sources of noise associated to the comparison of unrelated polypeptides. Following this rationale, independent analyses of a collection of proteins with enzymatic activity (Buck et al. 2013) and a set comprising the kinomes (ensembles of protein kinases) of different organisms (Graña-Montes et al. 2012b) have shown that, although catalytic amino acids do not tend to reside within aggregation-prone segments—since they usually present a polar or charged character—they are more frequently found in close proximity to APRs than would be expected by chance. This suggests the physico-chemical properties defining APRs concomitantly provide an appropriate environment for catalytic activity, which, in turn, restricts the action of purifying selection against these aggregation-prone stretches. Furthermore, the analysis of kinase domains also revealed that predicted amyloidogenic stretches usually map to regular secondary structure elements of the canonical protein kinase-like fold (Graña-Montes et al. 2012b), which is remarkably consistent with experimentally confirmed amyloidogenic stretches being found preferentially embedded within the regular secondary structure of the native states of different amyloidogenic proteins (Tzotzos and Doig 2010). Again, these findings suggest that the properties accounting for the maintenance of APRs are also relevant for the efficient attainment or preservation of the native structure, illustrating

further the unavoidable competition between folding and aggregation. Such a close interplay between the ability of a polypeptide to fold into a defined native (foldability) and its propensity to aggregate into non-functional conformations has also been highlighted by analysis of a representative collection of the ensemble of disulphide-rich domains (Fraga et al. 2014). The native states of this class of domains are essentially stabilized by the presence of disulphide bonds, and they populate long-lived, largely unstructured conformations along their slow folding reactions. The analysis of their characteristic features shows these kind of polypeptides escape the risk of aggregation through a compositional bias which translates into aggregative properties very similar to those of IDPs—with a reduced overall aggregation propensity, compared to the ensemble of globular proteins, as well as a very low number of APRs per sequence. Nonetheless, when APRs are still detected, these follow the same trend described before, by mapping more frequently to regular secondary structure elements. Here, the interplay between folding and aggregation is further emphasized by the observation that the differential oxidative folding efficiency of two structurally homologous disulphide-rich domains can be associated with the distinct amyloidogenicity of their regular secondary structure elements. Subsequently, this study arrived to the striking conclusion that the general traits of the oxidative folding pathways of disulphide-rich proteins can be successfully forecasted by the combined analysis of their predicted aggregation propensity and tendency to disorder. These findings indicate, once again, that the selective pressure to decrease the aggregation propensity of polypeptides is also restrained in globular proteins by the requirement to effectively achieve a defined three-dimensional conformation, provided the physico-chemical determinants of both efficient folding and aggregation propensity cannot be completely disentangled. The action of an excessive pressure against aggregation might compromise foldability, in such a way that additional stabilizing elements (such as disulphide bonds or external cofactors) might be required if a native globular conformation is to be maintained. The balance between the extent of anti-aggregative selection and the requirement for the efficient attainment of a three-dimensional structure would be defined by the combination that better preserves organismal fitness.

#### 7.3.4.2 Prediction of *in vivo* Protein Aggregation

Most of the algorithms described so far have been developed, and in some cases parameterized, based on properties of amyloid-like aggregates derived from the in vitro experimental characterization of aggregation reactions for a number of model proteins—being AGGRESCAN the sole exception, since it was developed from an experimentally derived scale of intrinsic aggregation propensity for the naturally-occurring proteinogenic amino acids.

Since the complex cellular environment strongly influences polypeptide deposition—indeed possess an intricate machinery to control this phenomenon—the question arises as to whether the previously described prediction tools are able to

predict protein aggregation *in vivo*. In order to address this question, Chiti and co-workers evaluated the ability of different publicly-available algorithms to predict the depositional properties of polypeptides inside the cell, by employing several datasets of proteins whose tendency to aggregate had been experimentally determined *in vivo* (Belli et al. 2011). In general terms, the predictors are substantially accurate in the forecasting of protein aggregation *in vivo* with phenomenological approaches performing globally better than structure-based methods. Such difference can be rationalized considering the constraints influencing the course of protein aggregation in the crowded cellular environment which certainly differ significantly from those in a test tube—here, the controlled environment and the absence of interference from other molecular components allow for more reproducible aggregation kinetics rendering highly ordered aggregated structures. Therefore phenomenological methods are, in principle, expected to capture the complexity of environmental conditions *in vivo* better than approaches, based exclusively on properties of the fine structure of late assembly products. Unsurprinsingly, AGGRESCAN (relying on its *in vivo* derived aggregation scale) is the algorithm yielding the best global performance across the different datasets analyzed. Interestingly, the good performance of AGGRESCAN, all considering the proteins in the testing datasets employed by Chiti and co-workers belong to different organisms, provides yet another piece of evidence for the suitability of *E. coli* as a model organism for the analysis of protein aggregation.

## 7.4 Prediction of Aggregation Propensity from the Tertiary Structure

The computational methods described so far make use only of the primary structure of polypeptides to perform predictions of their tendency to aggregate. In spite of the widely varying rationales behind each of these tools, they have been shown to forecast with significant overall accuracy, as mentioned, the actual change in the tendency to aggregate determined *in vivo* for a variety of datasets comprising mutational variants of different proteins (Belli et al. 2011).

The establishment of intermolecular contacts between APRs in polypeptides so as to build the cross-β spine of amyloid-like fibrils can be easily explained in the context of IDPs, because of the outspread nature of this kind of proteins. Conversely, since the APRs detected with linear predictors of aggregation map with high frequency to regions buried within the native state of globular proteins (Linding et al. 2004; Buck et al. 2013), the formation of amyloid-like structure by these polypeptides most likely requires significant conformational changes. Although different models have been proposed to rationalize the conversion of amyloidogenic proteins departing from globular native states into the cross-β conformation and its extension to form amyloid structures (Nelson and Eisenberg 2006), a better understanding of the mechanisms that trigger such conformational

conversion is still required. A couple of features appear of particular relevance for the amyloidogenic conversion, at least in its early stages, of proteins displaying a globular native state: (i) the modulation of the aggregative potential of specific APRs by their environment in the tertiary structure (that is, by residues or protein regions in close 3-dimensional vicinity), and (ii) variations in APRs' exposure to solvent due to local or global structural fluctuations, particularly regarding large scale variations that may arise from destabilizing mutations or harsh environmental conditions. The computational approaches that forecast the tendency to aggregate employing primary structure information only cannot deal with such conformational features. Consequently, the focus has recently moved towards the development of tools for the prediction of protein aggregation able to integrate the knowledge of the native tertiary and quaternary structures (Table 7.3).

A very first approach to predict the tendency to aggregate from the 3-dimensional structure of globular states was the Spatial Aggregation Propensity (SAP) method (Chennamsetty et al. 2009). This method was specifically aimed at the optimization of antibodies, which currently represent one of the most relevant groups of therapeutic agents. Since the production of these biomolecules is mainly

**Table 7.3** Characteristics of methods for the prediction of protein aggregation relying on the analysis of structural information (structural predictors)

| | | SAP | CamSol | Solubis | A3D |
|---|---|---|---|---|---|
| Structure-related features | Influence of the structural environment on the aggregation propensity/solubility, as a function of exposure | ✓ | ✓ | ✗ | ✓ |
| | Influence of the structural environment on the aggregation propensity/solubility, as a function of distance | ✗ | ✓ | ✗ | ✓ |
| | Modelling of structural dynamics | ✓ | ✗ | ✗ | ✓ |
| | Modelling of the impact of mutations on the structural stability | ✗ | ✗ | ✓ | ✓ |
| | Semi-automated mutational redesign for reduction of aggregation propensity/improvement of solubility | ✗ | ✓ | ✓ | ✗ |
| Level of development | | Equation | Server | YASARA plug-in | Server |
| URL | | – | www-mvsoftware.ch.cam.ac.uk | solubisyasara.switchlab.org | biocomp.chem.uw.edu.pl/A3D/ |

hampered by their typically low solubility, decreasing their tendency to aggregate seems an appropriate strategy to improve their industrial development. SAP exploits the great potential of the Molecular Dynamics (MD) computational tools (Dror et al. 2012; Papaleo 2015; see also Chap. 12) in order to simulate the structural fluctuations and molecular motions that underlie a wide variety of biological processes (Karplus and Kuriyan 2005; Dodson et al. 2008; Shaw et al. 2010; Lin et al. 2011; Dror et al. 2012), including the analysis of amyloid-like structures and of the different stages leading to their formation (Invernizzi et al. 2012). Concretely, SAP attempts to describe the dynamically averaged tendency to aggregate of a given protein structure or, in other words, the mean exposure of aggregation-prone patches at its surface—by exploring the polypeptide conformational space in the near-native ensemble through MD simulations in the tens of nanoseconds timescale. Although MD simulations performed in this timescale were found highly robust regardless of the force field employed (Rueda et al. 2007), this range of time appears to capture, essentially, fast relaxations corresponding to side chain motions (Henzler-Wildman and Kern 2007; van den Bedem and Fraser 2015); while backbone fluctuations, which would be consistent with transitions between local energy minima in the vicinity of the native conformation (native-like states), take place in the microsecond range (Shaw et al. 2010). It is the flux between native-like states, easily accessible because of thermal fluctuations, that can be expected (aside from mutations with a deep conformational impact) to influence more likely the aggregation propensity of polypeptides, rather than mere side chain motions at the protein surface. Therefore, the assessment of the tendency to aggregate of a globular protein would be better performed by means of simulations in the microsecond timescale. However, despite recent force field developments have dramatically increased both their reliability and consistency in order to perform MD simulations up to the millisecond time scale and beyond (Lindorff-Larsen et al. 2012), this kind of MD simulations remain highly costly in computational terms, even when dedicated computational resources are employed. Thus, their application in the prediction of aggregation propensity from tertiary and quaternary structure is not feasible yet, especially since the analysis of protein aggregation usually involves the evaluation of a large number of protein structures or sequential variants.

The impact of surface exposure and dynamic fluctuations on the tendency to aggregate is approximated in the SAP method by obtaining a score of aggregation proneness (Spatial Aggregation Propensity; SAP) per residue as the mean of the SAP values computed for each of the side chains' constituent atoms. The method relies on the expectation that the aggregation propensity of a particular side chain atom is modulated by that of any other neighbouring atom from side chains in spatial proximity. Consequently, SAP for a given atom is calculated by the summing the contribution (to its tendency to aggregate) of every side chain atom found within a sphere of 5Å radius centred on the atom of interest. Such contribution is obtained, for each atom within the sphere, as the product of the atom's solvent accessible area (SAA) with the hydrophobicity of the amino acid it belongs to. The atom's solvent exposure is computed relative to the SAA it would exhibit if its

corresponding amino acid (X) was fully exposed in a Ala-X-Ala tripeptide; and amino acid hydrophobicities are derived from a modified Black and Mould hydrophobicity scale (Black and Mould 1991), normalized to make Gly hydrophobicity equal to 0. The ultimate SAP score assigned to each atom results from averaging the values computed over all the structural conformations sampled along the MD simulation frames. In this way, the aggregation propensity per residue (evaluated in the context of its solvent exposure and considering the influence of structural fluctuations) is mapped on the protein surface, thus allowing a direct evaluation of exposed aggregation-prone patches and changes in APR structural shielding.

Although the development of SAP constituted a major breakthrough since it was the first tool that introduced the prediction of aggregation propensity from the 3-dimensional structure of globular proteins, it approximates the tendency to aggregate by considering fundamentally a hydrophobicity scale. However, it has long been known that hydrophobicity alone does not suffice for an accurate prediction of the potential to aggregate (Wurth et al. 2002; Chiti et al. 2003; Rousseau et al. 2006a). Still, SAP has inspired further developments of predictive algorithms which, based on the amyloidogenic stretch or "Hot Spot" hypothesis (Ventura et al. 2004; Ivanova et al. 2004), were initially intended to forecast aggregation departing solely from the knowledge of the primary structure. Nonetheless, the Zyggregator method already considered the impact on the aggregation propensity exerted by the structural shielding of APRs provided by native states—yet this was done employing the primary sequence, too, in order to predict the protection from hydrogen exchange (Tartaglia et al. 2007), which serves as an estimator of local structural stability along the sequence. The models implemented in Zyggregator have been recently redefined in order to develop the CamSol method, aiming to optimize the solubility of globular proteins by exploiting the knowledge of the tertiary structure, and with a particular focus on protein-based therapeutic agents (Sormanni et al. 2015). This algorithm relies on the assumption that, although solubility and aggregation propensity are related, they reflect distinct properties of the conformational energy landscape of the polypeptide chain—while solubility would constitute a thermodynamic property measuring the free energy difference between the native and aggregated states, the propensity to aggregate would represent a kinetic property describing the height of the free energy barrier between such states. Since CamSol has been designed to increase protein solubility through a rational and automated engineering of amino acid substitutions that preserve protein structure, the parameters in the original Zyggregator algorithm have been consequently modified to account for solubility (as a thermodynamic property) so as to reduce the bias towards the prediction of aggregation into amyloid-like structures. According to the authors, this is achieved by (i) substituting the hydrophobicity scale implemented in Zyggregator with another derived from the Wimley-White scale (Wimley and White 1996), (ii) modifying the preferences for $\alpha$ and $\beta$ conformation (which are now computed from PDB structures), and (iii) re-fitting the parameters that weight the properties taken into account for computing the intrinsic solubility (instead of the intrinsic aggregation propensity) of each

individual amino acid. Hence, the major differences between intrinsic aggregation propensity and intrinsic solubility are found for Pro and Gly residues, whose ability to disfavour the deposition into amyloid-like aggregates is well known (Wood et al. 1995; Steward et al. 2002; Parrini et al. 2005) because, as previously mentioned, they tend to disrupt β-conformation (Monsellier and Chiti 2007)—according to the CamSol method, however, these residues have a low impact on protein solubility.

In this way, CamSol is not strictly suitable for a direct assessment of protein aggregation. It is discussed in this section, however, because of the close relatedness between protein solubility and aggregation propensity, and, more importantly, because it represents one of the first tools to evaluate the influence exerted on the physico-chemical properties of a given amino acid by their neighbouring residues in the 3-dimensional protein structure. More specifically, CamSol first computes a solubility profile on the primary structure of the polypeptide (as done by Zyggregator) by assigning to each amino acid position the average of the intrinsic solubility score for a 7-residue window centred on it, and then adding both the impact of amino acid patterning along the sequence and of neighbouring charges. As discussed before, Zyggregator included the impact of patterning because certain amino acid arrangements (discussed previously amidst the determinants of protein aggregation) are known to facilitate aggregation into amyloid-like structures—its influence on protein solubility, though, is not as evident. The impact of side chain charges on solubility is more straightforward since, in aqueous environment, charged atoms can establish electrostatic interactions with water molecules, so solvation free energies of these residues are more favourable. At this point, CamSol introduces a first structural correction by modulating the effect of charges on the score assigned to a given amino acid in the solubility profile—as a function of both the distance (in the primary structure) between the charged residue and the centre of the window, and of the sign of the charges.

In CamSol, the solubility profile of the polypeptide chain is further corrected on the basis of the 3-dimensional coordinates, following a principle similar to that implemented in the SAP method, since the intrinsic solubility of neighbouring residues is considered to modulate the actual solubility of other amino acids along the structure. To this end, the contribution of any residue other than those defining the initial 7-residue window is estimated as the addition, for all these residues, of a solubility score, corrected by both weighting their relative exposure to solvent and their distance to the amino acid at the centre of the original window; such distance correction represents a novel feature relative to the approach introduced by SAP. The influence of exposure is defined by a sigmoid-like function, in such a way that residues with a relative solvent accessibility lower than 5% bear a weight of 0 (this means these residues are not considered as contributors to the solubility of their neighbours), the weight increases slowly up to a relative exposure around 20% and then linearly until reaching a maximum weight of 1 for amino acids with relative exposures equal to or above 50%. In the case of CamSol, the relative exposure is calculated employing the SAA of amino acid X in an extended Gly-X-Gly tripeptides as a reference. Meanwhile, the distance correction is provided by a function that decreases linearly with distance from 1 until reaching 0 at 8

Å or longer. Therefore, the structural context considered by CamSol to affect the solubility of a given residue is defined by a sphere of 8 Å radius centred on it, which, according to the authors, is equivalent to a projection of the 7-residue window over the 3-dimensional space. The aforementioned solubility score for any surrounding residue within the sphere that defines the structural context is, in turn, computed as the sum within a linear 7-residue window (centred on that neighbouring amino acid) of the product between residue intrinsic solubility and its relative exposure, which is then divided by the total relative solvent accessibility of the window—in this way, a solubility score is obtained somewhat averaged on the basis of exposure. This score for neighbouring residues is again corrected by the influence of amino acidic patterns, and by the "gatekeeping" effect of surrounding charges as well. The final structurally-corrected solubility value for each amino acid is obtained by summing the intrinsic score of that position in the linear solubility profile with the contribution of its structural context, and then multiplying this figure by the solvent exposure weight of the residue.

Once the structurally-corrected solubility profile is calculated, CamSol exploits it for a semi-automated rational redesign strategy aiming to identify amino acid substitutions or insertions able to increase the solubility of the target protein. This strategy first scans the structurally-corrected profile searching for poorly soluble sequence fragments. The identified stretches are then ranked in order to select the larger and less soluble ones as the better candidates for engineering an increase of their solubility. Next, the residues within the selected fragments are analyzed in order to identify the most suitable positions for the introduction of amino acid substitutions—defined as those being significantly exposed, but not essential either for protein function (this feature must be specified by the user) or for the maintenance of the structural stability (i.e. not being involved in relevant electrostatic interactions or hydrogen-bonding networks). Such requirements might imply, for instance, it is not most insoluble position that chosen for mutation however, since targeted amino acids are substituted by residues with a high intrinsic solubility, such as charged amino acids at neutral pH, the mutations introduced within the stretch might suffice (following the concept of solubility properties of a given residue being modulated by those of its neighbours) to significantly increase the solubility of the whole fragment. When functional or structural constraints impede the introduction of residue substitutions within the selected stretch, CamSol targets instead the flanks of the fragments for the insertion of amino acids which, by the mentioned proximity effect may enhance the solubility of the region. Once the target sites susceptible of substitution or insertion have been defined, the sequential variants incorporating changes at these positions are generated. In spite of the care taken by CamSol to delimit mutable positions, it additionally requires the user to set a maximum number of sites to be simultaneously engineered, thereby requiring expert knowledge to assist the protein redesign in avoiding the compromise of its stability by means of excessive mutation (this level of user intervention is circumvented in other methods by modelling the impact of sequential modifications on the structural stability). In order to assess the impact on solubility of the introduced substitutions CamSol computes the linear solubility profile, but not the

structurally-corrected profile since the program assumes that changes at selected sites would not perturb the polypeptide structure. A global solubility score is also provided for every variant, allowing an easier analysis of their mutational impact on solubility, which is obtained as a length-normalized summatory of the contribution to the solubility profile of each position with a score either above or below defined significance thresholds (of solubility or insolubility, respectively), while the intrinsic solubility of the remaining positions is not considered.

Additional methods specifically intended to forecast aggregation propensity in globular proteins that are based on well-established linear predictors have been recently released. One is Solubis, based on the previously described TANGO algorithm, which represented the first predictor purposely designed for the detection of APRs. Solubis does not represent a further development of the TANGO algorithm but is better described as a hybrid methodology that combines the prediction of the tendency to aggregate with the forecasting of the mutational impact on protein stability. In this way, Solubis exploits a different approach than SAP and CamSol, since it does not evaluate the modulation exerted on APR potential by the properties of their structural context, but instead relies on the observation that destabilization of the native state promotes the formation of different types of aggregates both in proteins associated to pathologies (Chan et al. 1996; Wall et al. 1999) and in model polypeptides (Chiti et al. 2000; Espargaró et al. 2008; Castillo et al. 2010). This supports the concept of the reduction in the conformational stability increasing the population of partially or even largely unfolded states where APRs might become substantially exposed, thence enabled to interact. In order to assess the mutational impact on protein stability the Solubis method integrates the FoldX algorithm, which has been extensively employed to model the impact of amino acid substitutions on protein structural stability for the analysis of a wide diversity of functional implications (Tokuriki et al. 2007; Kiel et al. 2008; Ashenberg et al. 2013; Fraga et al. 2014). FoldX implements an algorithm that weights the contribution of a variety of energetic terms which have been identified as the most relevant to account for protein stability (Guerois et al. 2002). Briefly, these include: Van der Waals interactions, the solvation energies for polar and non-polar amino acids, hydrogen bonds and electrostatic interactions, the entropic cost associated with the restriction of the configurational freedom of the backbone and side chain in the folded state, and the interaction between protein atoms and water molecules. The values for these energetic terms were obtained employing different approaches, such as the use of experimentally determined physico-chemical properties of amino acids (Nozaki and Tanford 1971; Levitt 1976; Radzicka and Wolfenden 1988; Roseman 1988) to approximate Van der Waals and solvation energies, the Coulomb and Debye-Hückel physical potentials employed to measure electrostatic interactions, statistical potentials based on secondary structure preferences of amino acids (Muñoz and Serrano 1994) to determine the entropic cost for backbone and side chain fixation, as well as experimental observations combined with theoretical estimates for the calculation of hydrogen-bonding and interactions with water. Some of these terms are modulated by employing scaling factors which take into account (i) the solvent accessibility of

amino acids, calculated according to the solvent contact method (Colonna-Cesari and Sander 1990), which computes residue exposure by considering the volumes of all the atoms surrounding a given atom, (ii) the steric clashes within a protein structure, and (iii) the effect of N-capping in α-helices. Among them, the most relevant scaling factor in the FoldX algorithm is solvent exposure, accounting for the observed lower impact of surface mutations on protein stability (Serrano et al. 1992; Matthews 1995) arising from a greater flexibility at this positions. The weights of the energy terms defining the ultimate FoldX energy function were fitted employing a large dataset of single point mutants from different proteins, whose associated stability changes had been experimentally determined. Once an input structure is provided, this function allows the prediction of stability changes upon mutation. To this end, FoldX attempts an optimization of the amino acids rotamer configurations in the mutant structure, and the energy difference is calculated by applying the same rotamer configurations to the wild-type (or other alternative reference) structure. In this way, FoldX is able to estimate mutational effects on stability by modelling the changes induced in side chain configurations—however, it cannot predict structural perturbations upon mutation since this method has not been designed to model backbone fluctuations

Solubis has been developed as a plug-in for the molecular graphics YASARA program (Krieger and Vriend 2014), which can be employed to perform a variety of analysis on protein structures, including modelling and simulation of their dynamics. The Solubis plug-in allows the calculation of the aggregation propensity for a polypeptide chain employing the TANGO algorithm and then plots the detected APRs on its 3-dimensional structure. Thus, although Solubis does not correct the tendency to aggregate by accounting for the physico-chemical properties of the structural environment, it still allows the user to evaluate the impact of structural shielding through a visual inspection of APRs location. Furthermore, this method also has an option to search for amino acid substitutions that reduce the tendency to aggregate, by mutating positions along the sequence to "gatekeeper" residues (Lys, Arg, Asp, Glu or Pro) (Rousseau et al. 2006b). However, since APRs are more frequently found buried within globular proteins and provided a high local β-sheet propensity is a sequential determinant that favours aggregation, there is an increased likelyhood that positions targeted for these substitutions might be located in the hydrophobic core or within β-strands. Hence, outmost caution must be taken when selecting mutable positions because "gatekeepers" are either amino acids charged at neutral pH (that may jeopardise protein structure by destabilizing the hydrophobic core) or Pro (whose tendency to disrupt β-conformation may compromise secondary structure elements). To circumvent the requirement of expert knowledge to determine the most suitable mutable sites, the Solubis mutagenic search workflow offers the option to perform a complementary analysis of the mutational impact on the structural stability, employing the FoldX algorithm. Following this scheme, Solubis assists the rational design of polypeptide variants bearing a lower aggregative potential, within a desired stability threshold expected to preserve the native conformation.

As a computational tool for the prediction of aggregation propensity from structural information, Solubis obviously requires a 3-dimensional structure as an input. An interesting alternative when experimentally determined protein structures are not available, is the generation of homology models (see Chap. 4). Both SAP and CamSol employed, to different extents, homology models in their case studies—although a specific validation between the performance of these methods with homology models and that with experimentally resolved structures would be necessary, it can be assumed a priori that employing homology models is an acceptable approximation when experimentally derived 3-dimensional coordinates are missing. In the case of Solubis, however, the authors explicitly discourage the use of homology models since this kind of structures are known to significantly affect accuracy of the FoldX algorithm in the estimation of stability changes upon mutation.

The method most recently developed as the upgrade of an existing algorithm employing the linear sequence only into a predictor of aggregation propensity exploiting structural information. Inspired by AGGRESCAN, which (as previously described) has emerged as a highly consistent predictor for the accurate forecasting of aggregation propensity *in vivo* (Belli et al. 2011), A3D exploits the same intrinsic aggregation propensity scale determined *in vivo* for each of the 20 naturally-occuring proteinogenic amino acids (de Groot et al. 2006). However, instead of employing an averaging window to calculate the aggregation propensity along the polypeptide sequence, this new development of the method corrects the intrinsic aggregation propensity of each amino acid in the 3-dimensional structure with the effect of its solvent exposure and the influence of its structural environment. In this sense, the algorithm evaluates, similarly to CamSol, both the impact of the structural shielding provided by residue burial, together with the influence of the properties of amino acids in the 3-dimensional vicinity; but, notably, A3D is specifically intended for the prediction of the tendency to aggregate. This method weights the role of amino acid exposure by correcting its intrinsic aggregation propensity with an exponential function (described below) of the relative surface accessibility. Here, the SAA is obtained as the contour defined by the centre of a 1.4 Å sphere (which approximates a water molecule) rolling over the Van der Waals surface of the protein (Lee and Richards 1971)—as implemented in the Naccess software (bioinf.manchester.ac.uk/naccess/)—employing the exposure of each amino acid X in extended Ala-X-Ala tripeptides as the reference solvent accessibility. The A3D score for each residue under inspection is then obtained by adding, to its exposure-corrected intrinsic aggregation propensity, the contribution of the structural context, which is estimated as the sum of the corrected aggregation propensity computed for every residue within a 10 Å radius from the Cα of the amino acid under evaluation. This aggregation proneness for each amino acid in the structural vicinity is calculated, in turn, by correcting its intrinsic aggregation propensity both with the exponential function of its solvent exposure and with another exponentially decreasing function of the distance between the neighbouring residue and the amino acid at the centre of the sphere. The exponential functions are defined so that relative solvent accessibility reaches a maximal weight of 1 when it

is equal to or above 55%, and a minimum of 0.1 at 10% exposure (with residues exhibiting a relative accessibility lower than 10% considered not to contribute to the structurally-corrected aggregation propensity). Analogously, neighbouring amino acids attain the highest weight of 1 at 1 Å or less from the centre of the sphere, while this weight decreases with distance until reaching 0.1 at 10 Å (amino acids outside the sphere are not considered to influence the aggregation propensity of the residue under scrutiny). Additionally, A3D offers the option to set a smaller sphere radius of 5 Å that, according to the authors, allows to assess the specific contribution of individual amino acids within APRs detected in the 3-dimensional structure.

Prior to the calculation of the structurally-corrected aggregation propensity, A3D performs an energy minimization of the input structure with the FoldX algorithm, intended to remove unfavourable energies arising from improper torsion angles, steric hindrance between amino acid side chains, and suboptimal rotamer configurations of residues in close vicinity. The A3D method also incorporates the FoldX energy function to allow the assessment of mutational effects on the stability of protein structures. Within the A3D workflow, the analysis of the mutational impact can be performed either before or after running the prediction of aggregation propensity on the input structure, depending on whether the aggregative properties of the wild-type (or alternative reference) structure are of interest.

The previously mentioned analyses may be performed on a static structural model, either provided directly as input or generated previously through FoldX (e.g. when only the assessment of a certain variant, bearing a specific substitution, is of interest). However, A3D also integrates the possibility to evaluate the impact of structural fluctuations on the aggregation propensity of the polypeptide. Averaging the properties of different conformers as in the SAP case is, perhaps, not be the best strategy for a realistic prediction because it could lead to underestimation of the potential of certain aggregation-prone conformer—i.e. an aggregation-prone state might trigger aggregation even when it is only transiently populated. In this sense, A3D produces a more valuable output by providing the model corresponding to the most aggregation-prone conformer. To this end, a "dynamic mode" of the method is available, which allows modelling of protein structural dynamics according to the CABS-flex protocol (Jamroz et al. 2013a). CABS-flex is a high-resolution coarse-grained molecular modelling approach that follows a Monte Carlo simulation scheme to sample backbone fluctuations. Through an extensive validation of its performance, CABS-FLEX has been shown to consistently reproduce the dynamic fluctuations of the near-native ensembles derived from all-atom MD simulations performed with a variety of force fields (Jamroz et al. 2013b). In "dynamic mode", A3D employs this robust computational tool in order to analyze an input or FoldX-mutated structure and then generate a collection of models describing the most representative fluctuations of the chain. Next, the structurally-corrected aggregation propensity is computed by A3D on these models, and the one with the highest A3D score is returned as the output approximating the most aggregation-prone state that is populated within the polypeptide's native-like conformational ensemble.

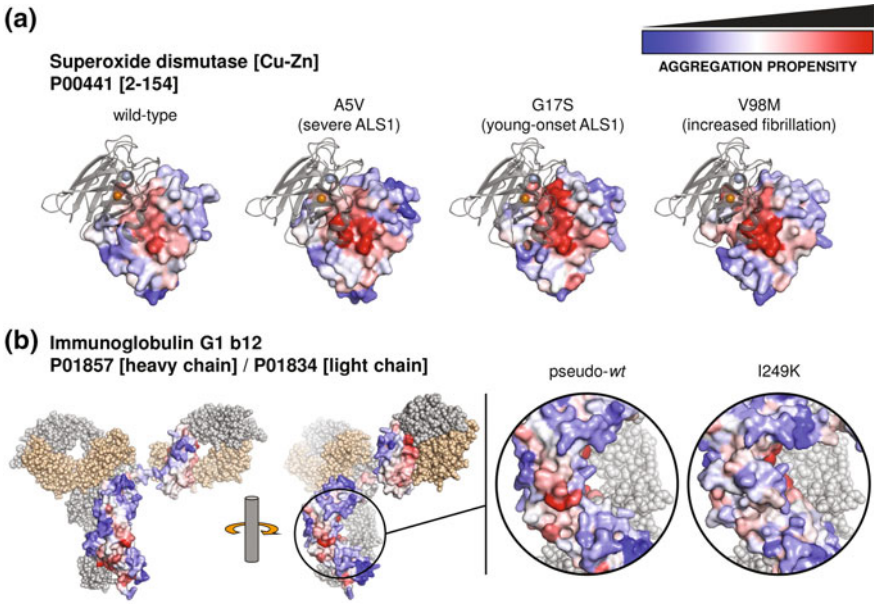



**Fig. 7.4** Potential applications of methods that integrate structural information for the prediction of protein aggregation are illustrated using A3D in "dynamic mode" (Zambrano et al. 2015b). **a** Structural methods are useful for the rationalization of the impact of mutations in the aggregation propensity of proteins that adopt a defined tertiary and/or quaternary structure: A3D reveals changes in the tendency to aggregate of Superoxide dismutase [Cu-Zn] (SOD1, PDB code: 1PU0) caused by the specified mutations, which are associated with either an increased rate of fibrillation or with severe forms of amyotrophic lateral sclerosis 1 (ALS1). These mutations raise the aggregation potential in the vicinity of the SOD1 dimerization interface; superimposing the modelled mutant monomeric unit (shown as surface representation) to the native quaternary structure (shown as *grey* cartoon), evidences how, upon disruption of the dimeric assembly (which might, indeed, be induced by the same mutations or result from a transient dissociation of the interaction surface), these amino acid substitutions increase the likelihood of aberrant intermolecular interactions. Metal ions complexed to SOD1 are depicted as spheres (copper in *brown* and zinc in *grey*) **b** Structural methods are also helpful to assist the redesign of protein-based drugs as exemplified by the analysis with A3D of a rationally-engineered substitution in the heavy chain constant region (represented as surface) of Immunoglobulin G1 b12 (IgG1 b12, PDB code: 1HZH), a potent neutralizing antibody against human immunodeficiency virus type 1 (HIV-1). This same structure was previously employed as a template for the structural modelling of antibodies analyzed with SAP (Chennamsetty et al. 2009). Here, the figure illustrates the potential of these methods for optimizing the formulation of proteinaceous therapeutic agents through the reduction of its tendency to aggregate by carefully engineering portions of the polypeptide without compromising function or stability. The light chains (*sand colour*) and heavy chains regions (*grey*) which have not been modelled with A3D are shown as spheres. UniProt codes are noted for SOD1 and for the light and heavy chains of IgG1 b12; the structure employed for A3D modelling of the latter presents two amino acidic substitutions relative to the canonical sequence deposited in UniProt

In this way, A3D encompasses in a single prediction tool the ensemble of features that are more relevant for the prediction of aggregation propensity in globular states (Fig. 7.4)—namely, the modulation of this propensity by the structural context, the assessment of the impact of mutations on both the tendency to aggregate and the stability of protein structure, and the evaluation of structural fluctuations. Regarding this latter point, the approach introduced in A3D for the modelling of protein conformational dynamics represents a significant advantage relative to the all-atomistic MD simulation performed by the SAP method, since the CABS-flex approach is able to equivalently reproduce the dynamic fluctuations in the near-native ensemble with a much higher computational efficiency. Furthermore, the dynamic modelling implemented in A3D might compensate for the previously mentioned issues arising in the assessment of the mutational impact on the structural stability with FoldX when the use of homology models cannot be avoided, though this potential remains to be explored. It is important to highlight that—because AGGRESCAN3D is ultimately based on an amino acid scale of intrinsic aggregation propensity determined *in vivo*—it is particularly suitable for the prediction of the aggregation propensities of therapeutic proteins, which are recombinantly produced employing heterologous expression systems.

## 7.5 Concluding Remarks

When we look back, the advances in the field of protein aggregation in the last fifteen years have been remarkable. This progress and the interest of the scientific community in the topic is clearly illustrated by more than 40,000 publications retrieved under the keyword "protein aggregation" in PubMed from the year 2000 onwards. A significant part of this progress owes to the ability of the computational methods described here to predict protein aggregation propensities with reasonable accuracy. These approaches have allowed the identification of the most dangerous regions of proteins linked to conformational diseases, the prediction of the impact of genetic mutations in these disorders, the analysis of the aggregation properties of entire proteomes, the design of therapeutic antibodies with highly improved solubility, and the invention of biomimetic materials—just to mention a few examples. Indeed, increasing evidence indicates that, apart from its involvement in disease, the amyloid fold is also exploited for evolutionarily selected biological functions, in diverse species, from bacteria to humans. The roles fulfilled by these so-called functional amyloids range from obligate macromolecular structures required for scaffolding and/or movement, to conditional amyloids (such as the yeast prions), whose aggregation can be triggered by environmental factors. Whether obligate or conditional, the natural selection of amyloid structure as a functional motif indicates that these properties are likely encoded in the sequence, and thus amenable for identification using aggregation prediction algorithms. Each of the discussed algorithms has its own pros and cons, and users should select the most suitable approach bearing in mind the specific problem they want to address—provided

these methods have been implemented using different theoretical or experimental knowledge and, therefore, they tend to capture different aspects accounting for the aggregation of natural or designed proteins. For this reason, the combination of existing methods in consensus algorithms has become a straightforward way of enhancing *in silico* predictive power. Nowadays, we are witnessing a new wave in which predictive strategies based on the analysis of protein sequences will be progressively substituted with tools able to deal with the much greater complexity of protein structures. The first 3D algorithms are just beginning to show their predictive power and are likely to become pivotal in providing novel mechanistic insights that, in turn, would allow development of even more precise computational tools in the very near future.

## References

Abeln S, Frenkel D (2008) Disordered flanks prevent peptide aggregation. PLoS Comput Biol 4: e1000241

Aggarwal S (2009) What's fueling the biotech engine–2008. Nat Biotechnol 27:987–993

Alberti S, Halfmann R, King O et al (2009) A systematic survey identifies prions and illuminates sequence features of prionogenic proteins. Cell 137:146–158

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Ashenberg O, Gong LI, Bloom JD (2013) Mutational effects on stability are largely conserved during protein evolution. Proc Natl Acad Sci USA 110:21071–21076

Auer S, Meersman F, Dobson CM, Vendruscolo M (2008) A generic mechanism of emergence of amyloid protofilaments from disordered oligomeric aggregates. PLoS Comput Biol 4: e1000222

Baldwin AJ, Knowles TPJ, Tartaglia GG et al (2011) Metastability of native proteins and the phenomenon of amyloid formation. J Am Chem Soc 133:14160–14163

Belli M, Ramazzotti M, Chiti F (2011) Prediction of amyloid aggregation in vivo. EMBO Rep 12:657–663

Black SD, Mould DR (1991) Development of hydrophobicity parameters to analyze proteins which bear post or cotranslational modifications. Anal Biochem 193:72–82

Blanco LP, Evans ML, Smith DR et al (2012) Diversity, biogenesis and function of microbial amyloids. Trends Microbiol 20:66–73

Broome BM, Hecht MH (2000) Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. J Mol Biol 296:961–968

Bryan AW, Menke M, Cowen LJ et al (2009) BETASCAN: probable beta-amyloids identified by pairwise probabilistic analysis. PLoS Comput Biol 5:e1000333

Bryan AW, O'Donnell CW, Menke M et al (2012) STITCHER: Dynamic assembly of likely amyloid and prion??-structures from secondary structure predictions. Proteins Struct Funct Bioinforma 80:410–420

Bryan PN, Orban J (2010) Proteins that switch folds. Curr Opin Struct Biol 20:482–488

Buck PM, Kumar S, Singh SK (2013) On the role of aggregation prone regions in protein evolution, stability, and enzymatic catalysis: insights from diverse analyses. PLoS Comput Biol 9:e1003291

Buell AK, Tartaglia GG, Birkett NR et al (2009) Position-dependent electrostatic protection against protein aggregation. Chem Bio Chem 10:1309–1312

Bui JM, Cavalli A, Gsponer Ö (2008) Identification of aggregation-prone elements by using interaction-energy matrices. Angew Chemie—Int Ed 47:7267–7269

Caflisch A (2006) Computational models for the prediction of polypeptide aggregation propensity. Curr Opin Chem Biol 10:437–444
Carrió M, González-Montalbán N, Vera A et al (2005) Amyloid-like properties of bacterial inclusion bodies. J Mol Biol 347:1025–1037
Castillo V, Espargaró A, Gordo V et al (2010) Deciphering the role of the thermodynamic and kinetic stabilities of SH3 domains on their aggregation inside bacteria. Proteomics 10:4172–4185
Castillo V, Graña-Montes R, Sabate R, Ventura S (2011) Prediction of the aggregation propensity of proteins from the primary sequence: aggregation properties of proteomes. Biotechnol J 6:674–685
Castillo V, Ventura S (2009) Amyloidogenic regions and interaction surfaces overlap in globular proteins related to conformational diseases. PLoS Comput Biol 5:e1000476
Chan W, Helms LR, Brooks I et al (1996) Mutational effects on inclusion body formation in the periplasmic expression of the immunoglobulin VL domain REI. Fold Des 1:77–89
Chen Y, Dokholyan NV (2008) Natural selection against protein aggregation on self-interacting and essential proteins in yeast, fly, and worm. Mol Biol Evol 25:1530–1533
Chennamsetty N, Voynov V, Kayser V et al (2009) Design of therapeutic proteins with enhanced stability. Proc Natl Acad Sci USA 106:11937–11942
Cheon M, Chang I, Mohanty S et al (2007) Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. PLoS Comput Biol 3:1727–1738
Cherny I, Gazit E (2008) Amyloids: Not only pathological agents but also ordered nanomaterials. Angew Chemie—Int Ed 47:4062–4069
Chiti F, Calamai M, Taddei N et al (2002a) Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. Proc Natl Acad Sci USA 99(Suppl 4):16419–16426
Chiti F, Dobson CM (2006) Protein misfolding, functional amyloid, and human disease. Annu Rev Biochem 75:333–366
Chiti F, Stefani M, Taddei N et al (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature 424:805–808
Chiti F, Taddei N, Baroni F et al (2002b) Kinetic partitioning of protein folding and aggregation. Nat Struct Biol 9:137–143
Chiti F, Taddei N, Bucciantini M et al (2000) Mutational analysis of the propensity for amyloid formation by a globular protein. EMBO J 19:1441–1449
Chou PY, Fasman GD (1974) Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins. Biochemistry 13:211–222
Colonna-Cesari F, Sander C (1990) Excluded volume approximation to protein-solvent interaction The solvent contact model. Biophys J 57:1103–1107
Conchillo-Solé O, de Groot NS, Avilés FX et al (2007) AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. BMC Bioinformatics 8:65
Cromwell MEM, Hilario E, Jacobson F (2006) Protein aggregation and bioprocessing. AAPS J 8: E572–E579
Dasari M, Espargaro A, Sabate R et al (2011) Bacterial inclusion bodies of Alzheimer's disease β-Amyloid peptides can be employed to study native-like aggregation intermediate states. Chem Bio Chem 12:407–423
De Baets G, Reumers J, Delgado Blanco J et al (2011) An evolutionary trade-off between protein turnover rate and protein aggregation favors a higher aggregation propensity in fast degrading proteins. PLoS Comput Biol 7:e1002090
de Groot NS, Aviles FX, Vendrell J, Ventura S (2006) Mutagenesis of the central hydrophobic cluster in Abeta42 Alzheimer's peptide. Side-chain properties correlate with aggregation propensities. FEBS J 273:658–668
de Groot NS, Sabate R, Ventura S (2009) Amyloids in bacterial inclusion bodies. Trends Biochem Sci 34:408–416

de Groot NS, Ventura S (2010) Protein aggregation profile of the bacterial cytosol. PLoS ONE 5: e9383

De Simone A, Kitchen C, Kwan AH et al (2012) Intrinsic disorder modulates protein self-assembly and aggregation. Proc Natl Acad Sci USA 109:6951–6956

Dill KA, Ozkan SB, Shell MS, Weikl TR (2008) The protein folding problem. Annu Rev Biophys 37:289–316

Dobson CM (2001) The structural basis of protein folding and its links with human disease. Philos Trans R Soc Lond B Biol Sci 356:133–145

Dobson CM (1999) Protein misfolding, evolution and disease. Trends Biochem Sci 24:329–332

Dobson CM (2003) Protein folding and misfolding. Nature 426:884–890

Dodson GG, Lane DP, Verma CS (2008) Molecular simulations of protein dynamics: new windows on mechanisms in biology. EMBO Rep 9:144–150

Dror RO, Dirks RM, Grossman JP et al (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu Rev Biophys 41:429–452

DuBay KF, Pawar AP, Chiti F et al (2004) Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. J Mol Biol 341:1317–1326

Eisenberg D, Jucker M (2012) The amyloid state of proteins in human diseases. Cell 148:1188–1203

Eisenhaber B, Bork P, Eisenhaber F (1998) Sequence properties of GPI-anchored proteins near the omega-site: constraints for the polypeptide binding site of the putative transamidase. Protein Eng 11:1155–1161

Ellis RJ (2001) Macromolecular crowding: obvious but underappreciated. Trends Biochem Sci 26:597–604

Emily M, Talvas A, Delamarche C (2013) MetAmyl: A METa-predictor for AMYLoid proteins. PLoS One

Esler WP, Stimson ER, Ghilardi JR et al (1996) Point substitution in the central hydrophobic cluster of a human?? amyloid congener disrupts peptide folding and abolishes plaque competence. Biochemistry 35:13914–13921

Espargaró A, Castillo V, de Groot NS, Ventura S (2008) The in vivo and in vitro aggregation properties of globular proteins correlate with their conformational stability: the SH3 case. J Mol Biol 378:1116–1131

Espinosa Angarica V, Ventura S, Sancho J (2013) Discovering putative prion sequences in complete proteomes using probabilistic representations of Q/N-rich domains. BMC Genom 14:316

Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol 22:1302–1306

Flock T, Weatheritt RJ, Latysheva NS, Babu MM (2014) Controlling entropy to tune the functions of intrinsically disordered regions. Curr Opin Struct Biol 26:62–72

Fowler DM, Koulov AV, Balch WE, Kelly JW (2007) Functional amyloid—from bacteria to humans. Trends Biochem Sci 32:217–224

Fraga H, Graña-Montes R, Illa R et al (2014) Association between foldability and aggregation propensity in small disulfide-rich proteins. Antioxid Redox Signal 21:368–383

Frousios KK, Iconomidou VA, Karletidi C-M, Hamodrakas SJ (2009) Amyloidogenic determinants are usually not buried. BMC Struct Biol 9:44

Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006a) Prediction of amyloidogenic and disordered regions in protein chains. PLoS Comput Biol 2:e177

Galzitskaya OV, Garbuzynskiy SO, Lobanov MY (2006b) FoldUnfold: web server for the prediction of disordered regions in protein chain. Bioinformatics 22:2948–2949

Garbuzynskiy SO, Lobanov MY, Galzitskaya OV (2010) FoldAmyloid: a method of prediction of amyloidogenic regions from protein sequence. Bioinformatics 26:326–332

Gasior P, Kotulska M (2014) FISH Amyloid—a new method for finding amyloidogenic segments in proteins based on site specific co-occurence of aminoacids. BMC Bioinformatics 15:54

Gebbink MFBG, Claessen D, Bouma B et al (2005) Amyloids–a functional coat for microorganisms. Nat Rev Microbiol 3:333–341
Gershenson A, Gierasch LM, Pastore A, Radford SE (2014) Energy landscapes of functional proteins are inherently risky. Nat Publ Gr 10:884–891
Goldschmidt L, Teng PK, Riek R, Eisenberg D (2010) Identifying the amylome, proteins capable of forming amyloid-like fibrils. Proc Natl Acad Sci USA 107:3487–3492
Grantcharova VP, Baker D (2001) Circularization changes the folding transition state of the src SH3 domain. J Mol Biol 306:555–563
Graña-Montes R, de Groot NS, Castillo V et al (2012a) Contribution of disulfide bonds to stability, folding, and amyloid fibril formation: the PI3-SH3 domain case. Antioxid Redox Signal 16:1–15
Graña-Montes R, Marinelli P, Reverter D, Ventura S (2014) N-terminal protein tails act as aggregation protective entropic bristles: the SUMO case. Biomacromolecules 15:1194–1203
Graña-Montes R, Sant'anna de Oliveira R, Ventura S (2012b) Protein aggregation profile of the human kinome. Front Physiol 3:438
Gromiha MM, Thangakani AM, Kumar S, Velmurugan D (2012) Sequence analysis and discrimination of amyloid and non-amyloid Peptides. pp 447–452
Guerois R, Nielsen JE, Serrano L (2002) Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. J Mol Biol 320:369–387
Hamodrakas SJ (1988) A protein secondary structure prediction scheme for the IBM PC and compatibles. Bioinformatics 4:473–477
Hamodrakas SJ, Liappa C, Iconomidou VA (2007) Consensus prediction of amyloidogenic determinants in amyloid fibril-forming proteins. Int J Biol Macromol 41:295–300
Hartl FU, Bracher A, Hayer-Hartl M (2011) Molecular chaperones in protein folding and proteostasis. Nature 475:324–332
Hauser CAE, Maurer-Stroh S, Martins IC (2014) Amyloid-based nanosensors and nanodevices. Chem Soc Rev 43:5326–5345
Henzler-Wildman K, Kern D (2007) Dynamic personalities of proteins. Nature 450:964–972
Hilbich C, Kisters-Woike B, Reed J et al (1992) Substitutions of hydrophobic amino acids reduce the amyloidogenicity of Alzheimer's disease beta A4 peptides. J Mol Biol 228:460–473
Idicula-Thomas S, Balaji PV (2005) Understanding the relationship between the primary structure of proteins and their amyloidogenic propensity: clues from inclusion body formation. Protein Eng Des Sel 18:175–180
Invernizzi G, Papaleo E, Sabate R, Ventura S (2012) Protein aggregation: mechanisms and functional consequences. Int J Biochem Cell Biol 44:1541–1554
Ivankov DN, Garbuzynskiy SO, Alm E et al (2003) Contact order revisited: influence of protein size on the folding rate. Protein Sci 12:2057–2062
Ivanova MI, Sawaya MR, Gingery M et al (2004) An amyloid-forming segment of beta2-microglobulin suggests a molecular model for the fibril. Proc Natl Acad Sci USA 101:10584–10589
Jahn TR, Radford SE (2005) The Yin and Yang of protein folding. FEBS J 272:5962–5970
Jahn TR, Radford SE (2008) Folding versus aggregation: polypeptide conformations on competing pathways. Arch Biochem Biophys 469:100–117
Jamroz M, Kolinski A, Kmiecik S (2013a) CABS-flex: server for fast simulation of protein structure fluctuations. Nucleic Acids Res 41:427–431
Jamroz M, Orozco M, Kolinski A, Kmiecik S (2013b) Consistent view of protein fluctuations from all-atom molecular dynamics and coarse-grained dynamics with knowledge-based force-field. J Chem Theory Comput 9:119–125
Kajava AV, Baxa U, Wickner RB, Steven AC (2004) A model for Ure2p prion filaments and other amyloids: the parallel superpleated beta-structure. Proc Natl Acad Sci USA 101:7885–7890
Karplus M, Kuriyan J (2005) Molecular dynamics and protein function. Proc Natl Acad Sci USA 102:6679–6685
Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14:1–63

Kawashima S, Pokarowski P, Pokarowska M et al (2007) AAindex: amino acid index database, progress report 2008. Nucleic Acids Res 36:D202–D205
Kiel C, Aydin D, Serrano L (2008) Association rate constants of ras-effector interactions are evolutionarily conserved. PLoS Comput Biol 4:e1000245
Kim C, Choi J, Lee SJ et al (2009) NetCSSP: web application for predicting chameleon sequences and amyloid fibril formation. Nucleic Acids Res 37:469–473
Kim YE, Hipp MS, Bracher A et al (2013) Molecular chaperone functions in protein folding and proteostasis. Annu Rev Biochem 82:323–355
Knowles TP, Fitzpatrick AW, Meehan S et al (2007) Role of intermolecular forces in defining material properties of protein nanofibrils. Science 318:1900–1903
Knowles TPJ, Buehler MJ (2011) Nanomechanics of functional and pathological amyloid materials. Nat Nanotechnol 6:469–479
Kodali R, Wetzel R (2007) Polymorphism in the intermediates and products of amyloid assembly. Curr Opin Struct Biol 17:48–57
Krebs MRH, Morozova-Roche LA, Daniel K et al (2004) Observation of sequence specificity in the seeding of protein amyloid fibrils. Protein Sci 13:1933–1938
Krieger E, Vriend G (2014) YASARA view—molecular graphics for all devices—from smartphones to workstations. Bioinformatics 30:1–2
Kuhlman B, Baker D (2000) Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA 97:10383–10388
Lancaster AK, Nutter-Upham A, Lindquist S, King OD (2014) PLAAC: a web and command-line application to identify proteins with Prion-Like Amino Acid Composition. Bioinformatics 30:2–3
Lawrence MC, Colman PM (1993) Shape complementarity at protein/protein interfaces. J Mol Biol 234:946–950
Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55:379–400
Lee CC, Perchiacca JM, Tessier PM (2013) Toward aggregation-resistant antibodies by design. Trends Biotechnol 31:612–620
Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107
Lin MM, Mohammed OF, Jas GS, Zewail AH (2011) Speed limit of protein folding evidenced in secondary structure dynamics. Proc Natl Acad Sci 108:16622–16627
Lin SH, Konishi Y, Denton ME, Scheraga HA (1984) Influence of an extrinsic crosslink on the folding pathway of ribonuclease A. Conformational and thermodynamic analysis of crosslinked (7-lysine, 41-lysine)-ribonuclease A. Biochemistry 23:5504–5512
Linding R, Schymkowitz J, Rousseau F et al (2004) A comparative study of the relationship between protein structure and β-aggregation in globular and intrinsically disordered proteins. J Mol Biol 342:345–353
Lindorff-Larsen K, Maragakis P, Piana S et al (2012) Systematic validation of protein force fields against experimental data. PLoS ONE 7:e32131
Lindorff-Larsen K, Røgen P, Paci E et al (2005) Protein folding and the organization of the protein topology universe. Trends Biochem Sci 30:13–19
Lobanov MY, Furletova EI, Bogatyreva NS et al (2010) Library of disordered patterns in 3D protein structures. PLoS Comput Biol 6:e1000958
López De La Paz M, Goldie K, Zurdo J et al (2002) De novo designed peptide-based amyloid fibrils. Proc Natl Acad Sci USA 99:16052–16057
Lopez de la Paz M, Serrano L (2004) Sequence determinants of amyloid fibril formation. Proc Natl Acad Sci 101:87–92
Lührs T, Ritter C, Adrian M et al (2005) 3D structure of Alzheimer's amyloid-beta (1-42) fibrils. Proc Natl Acad Sci USA 102:17342–17347
Makin OS, Atkins E, Sikorski P et al (2005) Molecular basis for amyloid fibril formation and stability. Proc Natl Acad Sci USA 102:315–320
Makin OS, Serpell LC (2005) Structures for amyloid fibrils. FEBS J 272:5950–5961

Matthews BW (1995) Studies on protein stability with T4 lysozyme. Adv Protein Chem 46:249–278
Maurer-Stroh S, Debulpaep M, Kuemmerer N et al (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. Nat Methods 7:237–242
Michelitsch MD, Weissman JS (2000) A census of glutamine/asparagine-rich regions: implications for their conserved function and the prediction of novel prions. Proc Natl Acad Sci 97:11910–11915
Minor DL, Kim PS (1996) Context-dependent secondary structure formation of a designed protein sequence. Nature 380:730–734
Monsellier E, Chiti F (2007) Prevention of amyloid-like aggregation as a driving force of protein evolution. EMBO Rep 8:737–742
Monsellier E, Ramazzotti M, Taddei N, Chiti F (2008) Aggregation propensity of the human proteome. PLoS Comput Biol 4:e1000199
Morel B, Varela L, Azuaga AI, Conejero-Lara F (2010) Environmental conditions affect the kinetics of nucleation of amyloid fibrils and determine their morphology. Biophys J 99:3801–3810
Mossuto MF, Bolognesi B, Guixer B et al (2011) Disulfide bonds reduce the toxicity of the amyloid fibrils formed by an extracellular protein. Angew Chem Int Ed Engl 50:7048–7051
Munishkina LA, Cooper EM, Uversky VN, Fink AL (2004) The effect of macromolecular crowding on protein aggregation and amyloid fibril formation. J Mol Recognit 17:456–464
Muñoz V, Serrano L (1994) Intrinsic secondary structure propensities of the amino acids, using statistical phi-psi matrices: comparison with experimental scales. Proteins 20:301–311
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540
Nelson R, Eisenberg D (2006) Structural models of amyloid-like fibrils. Adv Protein Chem 73:235–282
Nelson R, Sawaya MR, Balbirnie M et al (2005) Structure of the cross-beta spine of amyloid-like fibrils. Nature 435:773–778
Nozaki Y, Tanford C (1971) The solubility of amino in aqueous ethanol acids and two glycine dioxane solutions peptides. J Biol Chem 246:2211–2217
O'Donnell CW, Waldispühl J, Lis M et al (2011) A method for probing the mutational landscape of amyloid structure. Bioinformatics 27:34–42
Pallarès I, Vendrell J, Avilés FX, Ventura S (2004) Amyloid fibril formation by a partially structured intermediate state of alpha-chymotrypsin. J Mol Biol 342:321–331
Papaleo E (2015) Integrating atomistic molecular dynamics simulations, experiments, and network analysis to study protein dynamics: strength in unity. Front Mol Biosci 2:1–6
Parrini C, Taddei N, Ramazzotti M et al (2005) Glycine residues appear to be evolutionarily conserved for their ability to inhibit aggregation. Structure 13:1143–1151
Pawar AP, Dubay KF, Zurdo J et al (2005) Prediction of "aggregation-prone" and "aggregation-susceptible" regions in proteins associated with neurodegenerative diseases. J Mol Biol 350:379–392
Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M (2009) Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci USA 106:10159–10164
Perchiacca JM, Tessier PM (2012) Engineering Aggregation-Resistant Antibodies. Annu Rev Chem Biomol Eng 3:263–286
Petkova AT, Ishii Y, Balbach JJ et al (2002) A structural model for Alzheimer's beta -amyloid fibrils based on experimental constraints from solid state NMR. Proc Natl Acad Sci USA 99:16742–16747
Poland DC, Scheraga HA (1965) Statistical mechanics of noncovalent bonds in polyamino acids VIII covalent loops proteins. Biopolymers 3:379–399
Radzicka A, Wolfenden R (1988) Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution. Biochemistry 27:1664–1670

Reumers J, Maurer-Stroh S, Schymkowitz J, Rousseau F (2009a) Protein sequences encode safeguards against aggregation. Hum Mutat 30:431–437
Reumers J, Rousseau F, Schymkowitz J (2009b) Multiple evolutionary mechanisms reduce protein aggregation. Open Biol J 2:176–184
Richardson JS, Richardson DC (2002) Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 99:2754–2759
Ritter C, Maddelein M-L, Siemer AB et al (2005) Correlation of structural elements and infectivity of the HET-s prion. Nature 435:844–848
Rochet JC, Lansbury PT (2000) Amyloid fibrillogenesis: Themes and variations. Curr Opin Struct Biol 10:60–68
Rodriguez JA, Ivanova MI, Sawaya MR et al (2015) Structure of the toxic core of α-synuclein from invisible crystals. Nature 525(7570):486–490
Roseman MA (1988) Hydrophilicity of polar amino acid side-chains is markedly reduced by flanking peptide bonds. J Mol Biol 200:513–522
Rousseau F, Schymkowitz J, Serrano L (2006a) Protein aggregation and amyloidosis: confusion of the kinds? Curr Opin Struct Biol 16:118–126
Rousseau F, Serrano L, Schymkowitz JWH (2006b) How evolutionary pressure against protein aggregation shaped chaperone specificity. J Mol Biol 355:1037–1047
Rueda M, Ferrer-Costa C, Meyer T et al (2007) A consensus view of protein dynamics. Proc Natl Acad Sci USA 104:796–801
Sabate R, Rousseau F, Schymkowitz J, Ventura S (2015) What makes a protein sequence a prion? PLoS Comput Biol 11:e1004013
Saiki M, Konakahara T, Morii H (2006) Interaction-based evaluation of the propensity for amyloid formation with cross-?? structure. Biochem Biophys Res Commun 343:1262–1271
Sambashivan S, Liu Y, Sawaya MR et al (2005) Amyloid-like fibrils of ribonuclease a with three-dimensional domain-swapped and native-like structure. Nature 437:266–269
Sanchez de Groot N, Torrent M, Villar-Piqué A et al (2012) Evolutionary selection for protein aggregation. Biochem Soc Trans 40:1032–1037
Santner AA, Croy CH, Vasanwala FH et al (2012) Sweeping away protein aggregation with entropic bristles: intrinsically disordered protein fusions enhance soluble expression. Biochemistry 51(37):7250–7262
Sawaya MR, Sambashivan S, Nelson R et al (2007) Atomic structures of amyloid cross-beta spines reveal varied steric zippers. Nature 447:453–457
Schwartz R, Istrail S, King J (2001) Frequencies of amino acid strings in globular protein sequences indicate suppression of blocks of consecutive hydrophobic residues. Protein Sci 10:1023–1031
Selkoe DJ (2003) Folding proteins in fatal ways. Nature 426:900–904
Serrano L, Kellis JT, Cann P et al (1992) The folding of an enzyme. II. Substructure of barnase and the contribution of different interactions to protein stability. J Mol Biol 224:783–804
Shaw DE, Maragakis P, Lindorff-Larsen K et al (2010) Atomic-level characterization of the structural dynamics of proteins. Science 330:341–346
Shimanovich U, Efimov I, Mason TO et al (2014) Protein Microgels from Amyloid Fibril Networks. ACS Nano 9:43–51
Sipe JD, Benson MD, Buxbaum JN et al (2014) Nomenclature 2014: Amyloid fibril proteins and clinical classification of the amyloidosis. Amyloid 21:221–224
Sormanni P, Aprile FA, Vendruscolo M (2015) The CamSol method of rational design of protein mutants with enhanced solubility. J Mol Biol 427(2):478–490
Stefani M, Dobson CM (2003) Protein aggregation and aggregate toxicity: new insights into protein folding, misfolding diseases and biological evolution. J Mol Med (Berl) 81:678–699
Steward A, Adhya S, Clarke J (2002) Sequence conservation in Ig-like domains: the role of highly conserved proline residues in the fibronectin type III superfamily. J Mol Biol 318:935–940
Sunde M, Blake C (1997) The structure of amyloid fibrils by electron microscopy and X-ray diffraction. Adv Protein Chem 50:123–159

Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2004) The role of aromaticity, exposed surface, and dipole moment in determining protein aggregation rates. Protein Sci 13:1939–1941
Tartaglia GG, Cavalli A, Pellarin R, Caflisch A (2005a) Prediction of aggregation rate and aggregation-prone segments in polypeptide sequences. Protein Sci 14:2723–2734
Tartaglia GG, Cavalli A, Vendruscolo M (2007) Prediction of local structural stabilities of proteins from their amino acid sequences. Structure 15:139–143
Tartaglia GG, Pawar AP, Campioni S et al (2008) Prediction of aggregation-prone regions in structured proteins. J Mol Biol 380:425–436
Tartaglia GG, Pellarin R, Cavalli A, Caflisch A (2005b) Organism complexity anti-correlates with proteomic beta-aggregation propensity. Protein Sci 14:2735–2740
Tartaglia GG, Vendruscolo M (2008) The Zyggregator method for predicting protein aggregation propensities. Chem Soc Rev 37:1395–1401
Tartaglia GG, Vendruscolo M (2009) Correlation between mRNA expression levels and protein aggregation propensities in subcellular localisations. Mol BioSyst 5:1873–1876
Thangakani AM, Kumar S, Nagarajan R, Velmurugan D, Gromiha MM (2014) GAP: towards almost 100 percent prediction for β-strand-mediated aggregating peptides with distinct morphologies. Bioinformatics 30(14):1983–1990
Thangakani A, Kumar S, Velmurugan D, Gromiha M (2013) Distinct position-specific sequence features of hexa-peptides that form amyloid-fibrils: application to discriminate between amyloid fibril and amorphous β-aggregate forming peptide sequences. BMC Bioinformatics 14:S6
Thompson MJ, Sievers SA, Karanicolas J et al (2006) The 3D profile method for identifying fibril-forming segments of proteins. Proc Natl Acad Sci USA 103(11):4074–4078
Tian J, Wu N, Guo J, Fan Y (2009) Prediction of amyloid fibril-forming segments based on a support vector machine. BMC Bioinformatics 10:S45
Tokuriki N, Stricher F, Schymkowitz J et al (2007) The stability effects of protein mutations appear to be universally distributed. J Mol Biol 369:1318–1332
Tomii K, Kanehisa M (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng 9(1):27–36
Tompa P (2012) Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci 37:509–516
Tompa P (2002) Intrinsically unstructured proteins. Trends Biochem Sci 27:527–533
Toombs JA, Petri M, Paul KR, Kan GY, Ben-Hur A, Ross ED (2012) De novo design of synthetic prion domains. Proc Natl Acad Sci 109(17):6519–6524
Trovato A, Chiti F, Maritan A, Seno F (2006) Insight into the structure of amyloid fibrils from the analysis of globular proteins. PLoS Comput Biol 2:e170
Tsolis AC, Papandreou NC, Iconomidou VA, Hamodrakas SJ (2013) A consensus method for the prediction of 'aggregation-prone'peptides in globular proteins. PLoS ONE 8(1):e54175
Tycko R (2014) Physical and Structural Basis for Polymorphism in Amyloid Fibrils. Protein Sci 00:1–12
Tycko R (2011) Solid-state NMR studies of amyloid fibril structure. Annu Rev Phys Chem 62:279–299
Tycko R, Wickner RB (2013) Molecular structures of amyloid and prion fibrils: consensus versus controversy. Acc Chem Res 46:1487–1496
Tzotzos S, Doig AJ (2010) Amyloidogenic sequences in native protein structures. Protein Sci 19:327–348
Uversky VN (2013a) A decade and a half of protein intrinsic disorder: biology still waits for physics. Protein Sci 22:693–724
Uversky VN (2002) Natively unfolded proteins: a point where biology waits for physics. Protein Sci 11(4):739–756

Uversky VN (2013b) The most important thing is the tail: multitudinous functionalities of intrinsically disordered protein termini. FEBS Lett 587:1891–1901
Uversky VN, Fink AL (2004) Conformational constraints for amyloid fibrillation: the importance of being unfolded. Biochim Biophys Acta 1698:131–153
van den Bedem H, Fraser JS (2015) Integrative, dynamic structural biology at atomic resolution—it's about time. Nat Methods 12:307–318
van den Berg B, Ellis RJ, Dobson CM (1999) Effects of macromolecular crowding on protein folding and aggregation. EMBO J 18:6927–6933
Ventura S, Zurdo J, Narayanan S et al (2004) Short amino acid stretches can mediate amyloid formation in globular proteins: the Src homology 3 (SH3) case. Proc Natl Acad Sci USA 101:7258–7263
Villar-Piqué A, Ventura S (2012) Modeling amyloids in bacteria. Microb Cell Fact 11:166
Waldo GS, Standish BM, Berendzen J, Terwilliger TC (1999) Rapid protein-folding assay using green fluorescent protein. Nat Biotechnol 17(7):691–695
Wall J, Schell M, Murphy C et al (1999) Thermodynamic instability of human λ 6 Light chains: correlation with fibrillogenicity. Biochemistry 38:14101–14108
Walsh I, Seno F, Tosatto SCE, Trovato A (2014) PASTA 2.0: an improved server for protein aggregation prediction. Nucleic Acids Res 42:301–307
Wang L, Maji SK, Sawaya MR et al (2008) Bacterial inclusion bodies contain amyloid-like structure. PLoS Biol 6:e195
Wang L, Schubert D, Sawaya MR et al (2010) Multidimensional structure-activity relationship of a protein in its aggregated states. Angew Chem Int Ed Engl 49:3904–3908
Wasmer C, Lange A, Van Melckebeke H et al (2008) Amyloid fibrils of the HET-s(218–289) prion form a beta solenoid with a triangular hydrophobic core. Science 319:1523–1526
West MW, Wang W, Patterson J et al (1999) De novo amyloid proteins from designed combinatorial libraries. Proc Natl Acad Sci 96:11211–11216
Westermark P (2005) Amyloid Proteins. Wiley-VCH Verlag GmbH, Weinheim, Germany
Wimley WC, White SH (1996) Experimentally determined hydrophobicity scale for proteins at membrane interfaces. Nat Struct Biol 3:842–848
Wolynes PG (2008) The protein folding energy landscape: a primer. In: Muñoz V (ed) Protein folding, misfolding and aggregation. Royal Society of Chemistry, Cambridge, pp 49–69
Wood SJ, Wetzel R, Martin JD, Hurle MR (1995) Prolines and amyloidogenicity in fragments of the Alzheimer's peptide beta/A4. Biochemistry 34:724–730
Wright CF, Teichmann SA, Clarke J, Dobson CM (2005) The importance of sequence diversity in the aggregation and evolution of proteins. Nature 438(7069):878–881
Wurth C, Guimard NK, Hecht MH (2002) Mutations that reduce aggregation of the Alzheimer's Abeta42 peptide: an unbiased search for the sequence determinants of Abeta amyloidogenesis. J Mol Biol 319:1279–1290
Yoon S, Welsh WJ (2004) Detecting hidden sequence propensity for amyloid fibril formation. Protein Sci 13:2149–2160
Yoon S, Welsh WJ (2005) Rapid assessment of contact-dependent secondary structure propensity: relevance to amyloidogenic sequences. Proteins 60:110–117
Yoon S, Welsh WJ, Jung H, Do Yoo Y (2007) CSSP2: an improved method for predicting contact-dependent secondary structure propensity. Comput Biol Chem 31:373–377
Zambrano R, Conchillo-Sole O, Iglesias V et al (2015a) PrionW: a server to identify proteins containing glutamine/asparagine rich prion-like domains and their amyloid cores. Nucleic Acids Res 43(W1):W331–W337
Zambrano R, Jamroz M, Szczasiuk A et al (2015b) AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures. Nucleic Acids Res 8220211:1–8

Zhang Z, Chen H, Lai L (2007) Identification of amyloid fibril-forming segments based on structure and residue-based statistical potential. Bioinformatics 23:2218–2225

Zibaee S, Makin OS, Goedert M, Serpell LC (2007) A simple algorithm locates beta-strands in the amyloid fibril core of alpha-synuclein, Abeta, and tau using the amino acid sequence alone. Protein Sci 16:906–918

Zimmerman SB, Trach SO (1991) Estimation of macromolecule concentrations and excluded volume effects for the cytoplasm of Escherichia coli. J Mol Biol 222:599–620

# Chapter 8
# Prediction of Biomolecular Complexes

**Anna Vangone, Romina Oliva, Luigi Cavallo and Alexandre M.J.J. Bonvin**

**Abstract** Almost all processes in living organisms occur through specific interactions between biomolecules. Any dysfunction of those interactions can lead to pathological events. Understanding such interactions is therefore a crucial step in the investigation of biological systems and a starting point for drug design. In recent years, experimental studies have been devoted to unravel the principles of biomolecular interactions; however, due to experimental difficulties in solving the three-dimensional (3D) structure of biomolecular complexes, the number of available, high-resolution experimental 3D structures does not fulfill the current needs. Therefore, complementary computational approaches to model such interactions are necessary to assist experimentalists since a full understanding of how biomolecules interact (and consequently how they perform their function) only comes from 3D structures which provide crucial atomic details about binding and recognition processes. In this chapter we review approaches to predict biomolecular complexes, introducing the concept of molecular docking, a technique which uses a combination of geometric, steric and energetics considerations to predict the 3D structure of a biological complex starting from the individual structures of its constituent parts. We provide a mini-guide about docking concepts, its potential and challenges, along with post-docking analysis and a list of related software.

A. Vangone · A.M.J.J. Bonvin (✉)
Computational Structural Biology Group, Bijvoet Center for Biomolecular Research, Faculty of Science—Chemistry, Utrecht University, 3584 Utrecht, The Netherlands
e-mail: a.m.j.j.bonvin@uu.nl

R. Oliva
Department of Sciences and Technologies, University "Parthenope" of Naples, Centro Direzionale Isola C4, 80143 Naples, Italy

L. Cavallo
Kaust Catalysis Center, Physical Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_8

**Keywords** Protein-protein complexes · Protein-peptide complexes · Docking · Searching · Scoring · Data-driven docking · HADDOCK · CAPRI · Flexibility · Binding affinity · PRODIGY · CONSRANK

## 8.1 Introduction

Biomolecular complexes, such as protein-protein and protein-ligand ones, underlie almost all biological processes in the cell, such as DNA replication, transcription, translation, signaling pathways, immune system response, enzyme inhibition. To implement this wide diversity of (bio)chemical processes, proteins get in touch with other proteins, nucleic acids, sugars, lipids and various other molecules (Jones and Thornton 1996; Alberts 1998). The biological function of a protein is defined by its interactions in the cell. Inappropriate or altered (either inhibited and enhanced) interactions can lead to disease (Stites 1997; Sugiki et al. 2014). For these reasons, research aimed at understanding, disrupting or modulating protein-protein interactions (PPIs) is a crucial step in the investigation of almost all biological processes, ranging from enzyme catalysis and inhibition to signal transduction and gene expression. Accordingly, PPIs are currently receiving considerable attention as targets for rational drug design (González-Ruiz and Gohlke 2006; Metz et al. 2012; Nisius et al. 2012) and as therapeutic agents (Szymkowski 2005; Hwang and Park 2008; Zhou et al. 2013).

In recent years, experimental and theoretical work has been devoted to unravel the principles of protein-protein interactions (Phizicky and Fields 1995; Jones and Thornton 1996). The formation of biological complexes is driven by the free energy of the complex (mostly determined by physicochemical and geometrical interface properties) and the concentration of the protein components. The association of two proteins, in fact, relies on an encounter and possible rearrangement of the interacting surfaces, requiring co-localization in time and space. Generally proteins reside in crowded environments, with many potential binding partners with different surface properties; consequently, during evolution, the interaction surfaces are believed to have evolved to both optimize interaction efficacy and prevent undesired interactions (Ofran and Rost 2003).

In this scenario, it is a must to obtain 3D structural information in order to gain a complete understanding of both the biochemical nature of the process bringing the components together and to facilitate the design of compounds that might influence it. The structural characterization of a protein-protein interface includes in particular the identification of interatomic hydrogen bonds, salt bridges and hydrophobic interactions, the determination of the interaction surface area and possibly the identification of bridging water molecules (Northrup and Erickson 1992; Tsai et al. 1999). The combination of all this information defines the nature of the binding site and of the network of interactions, which makes it possible to pinpoint key residues and contacts for complex formation.

Obtaining 3D structures of biological complexes is therefore of supreme significance for the study of biomolecular interactions and all their possible pharmaceutical and medicinal applications. High-resolution atomic structures are obtained by X-ray crystallography and nuclear magnetic resonance (NMR), while methods like Small-Angle X-ray Scattering (SAXS) (Yang 2014; Chaudhuri 2015) or cryo-Electron Microscopy (cryo-EM) give low-resolution structural data, although the latter, thanks to recent developments in both detector technology and software, is now reaching near atomic resolution (Bai et al. 2015) with, for example, the recent <3 Å high-resolution structure of the ribosome-EF-Tu complex (Fischer et al. 2015). Experimental determination of biomolecules remains, however, difficult, time-consuming and costly (Chruszcz et al. 2010): with X-ray crystallography, dynamics and disorder can impede the crystallization, while (solution) NMR suffers from a size limitation when it comes to studying large macromolecular complexes; and both methods struggle with membrane-resident and membrane-associated complexes. For these reasons, there is relatively little structural information available about biomolecular complexes compared to proteins that exist as single chains or form permanent oligomers (Schreiber and Fersht 1996). As a result, the number of

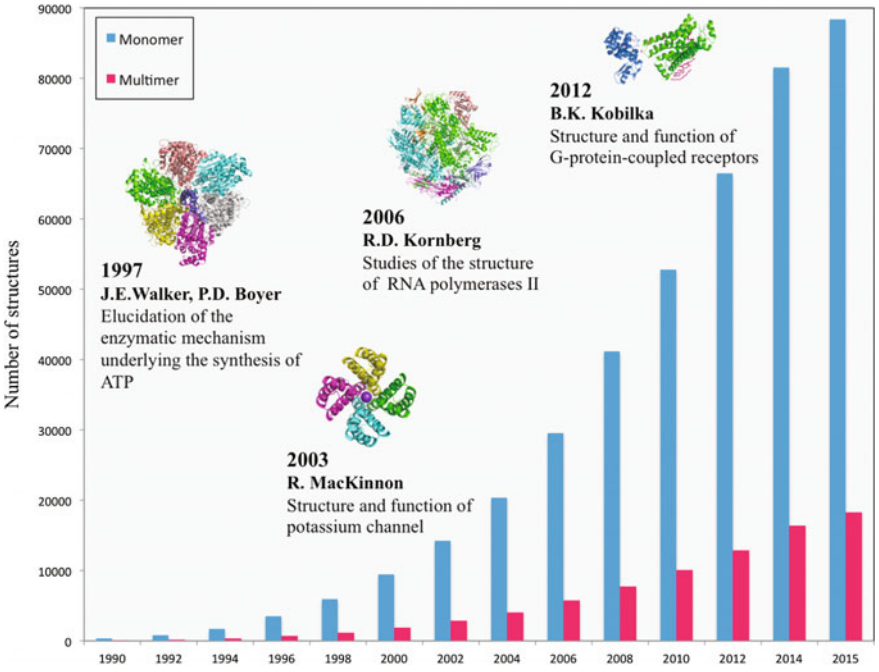



**Fig. 8.1** Yearly growth of protein structures number in Protein Data Bank (PDB) from 1990. The PDB was established in 1971, the total number of protein structures grew to 434 in 1990, reaching 106,650 structures on June 2015. The number of single protein structures is reported in *blue*, the number of multiple proteins systems is reported in *magenta*. Some of the Nobel Prized awarded for elucidation of structure (and function) of macromolecular systems are reported in the figure

solved complexes deposited in the Protein Data Bank (PDB) (Bernstein et al. 1977) (http://www.rcsb.org/) is still orders of magnitude smaller than that of individual proteins as shown in Fig. 8.1. Despite this disproportion, the growing number of available experimental structures for protein-protein complexes over the years has allowed statistical studies of the properties and physico-chemical forces that regulate protein-protein interactions (hydrophobicity, hydrogen bonding, electrostatic interactions, van der Waals interactions, and so on). These provide useful information in the development of computational strategies for structure prediction and characterization. Considering the experimental limitations discussed above, computational structural biology is now routinely considered an integral part of research.

Since the pioneering work of Janin and Wodak (Wodak and Janin 1978) who described, more than 30 years ago, the first automated algorithm to predict the 3D interaction between bovine pancreatic trypsin and its natural inhibitor, the docking field (with docking defined as the prediction of protein complexes structures starting from the structures of the free molecules) has advanced considerably (Schlick et al. 2011). The past decades have seen the emergence of a large variety of theoretical algorithms designed to predict the structures of protein-protein and protein-ligand complexes (Smith and Sternberg 2002; Bonvin 2006; Ritchie 2008; Vajda and Kozakov 2009; Moal et al. 2013a).

## 8.2 Docking

Molecular docking is a computational modeling technique that aims at predicting the 3D structure of a complex (bound form) given the structures of the individual molecules (unbound forms) (Fig. 8.2), hopefully revealing most of the relevant residue-residue contacts involved in the interaction (Smith and Sternberg 2002). It offers a tool for fundamental studies of biomolecular interactions and provides a structural basis for drug design. Docking approaches assume that the native complex is near the global minimum of the energy landscape. In fact, based on the thermodynamic hypothesis, at fixed temperature and pressure the Gibbs free energy of the macromolecule-solvent system reaches its global minimum at the native state of the macromolecule (Ruvinsky and Vakser 2008).

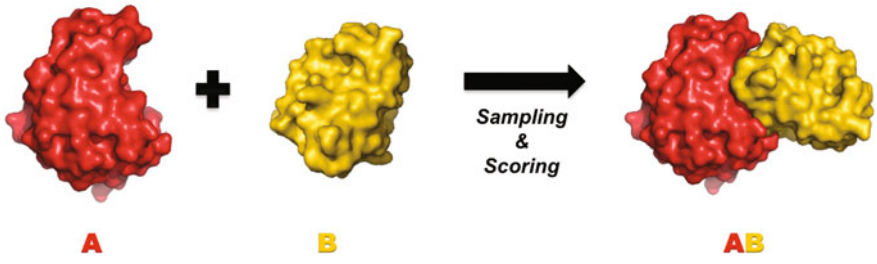



**Fig. 8.2** An illustration of protein-protein docking procedure starting from the unbound structures (*A* and *B*), into their final bound form (*AB*). (PDBcode: 1BRS (Buckle et al. 1994), chains *A* and *B*)

Progress in protein-protein docking performance has been monitored over the years with the community wide Critical Assessment of PRedicted Interactions (CAPRI) experiment (Janin et al. 2003; Lensink et al. 2007). Many rounds of blind predictions have highlighted the increasing accuracy of docking methods, in particular for some of them that consistently show good performance (Lensink and Wodak 2013; Lensink et al. 2016) (CAPRI results can be found at the url: http://www.ebi.ac.uk/msd-srv/capri/).

All current docking methods, despite their differences, start from the 3D structures of the unbound components (whether experimentally determined or computationally predicted) and incorporate two crucial steps (Halperin et al. 2002; Vajda and Kozakov 2009):

1. *Searching*, consisting in the generation of thousands of alternative poses to sample the conformational landscape;
2. *Scoring*, consisting in assessing the generated poses using a 'pseudo-energy' function in order to rank them and select the native-like solutions.

This separation into two stages is just one way of describing the docking approach, since sometimes there is no clear separation between these, or they may incorporate multiple different sub-steps. A fundamental point of any docking method is to be computationally efficient both in the search step and in its refinement and scoring scheme in order to be able to evaluate a huge number of candidate solutions and discriminate native-like binding modes from wrong ones in a reasonable computation time.

### *8.2.1 Step 1: Searching*

The search step involves an exhaustive sampling of the conformational space of one protein with respect to the other, resulting in a six-dimensional search (6D) in the case of rigid molecules. Almost all docking programs use a similar approach for the search step: one protein is fixed in space (usually the larger one, named receptor) and the second (named the ligand) is rotated and translated around the first. Various methods have been developed that can efficiently cover the entire conformational space (Vajda and Kozakov 2009) such as:

- *Fast Fourier transforms* (*FFT*)*-based docking*. Despite the huge size of the conformational space to be sampled, the search can be efficiently performed through several FFT calculations, as originally introduced by Katchalski-Katzir et al. (1992). FFT-based methods represent the proteins on a Cartesian grid, with some degree of inter-protein penetration between the ligand and the receptor allowed to account for small conformational changes of mainly side-chains. The shape complementarity is measured using Fourier correlation. Additional terms can be encoded into measure for example electrostatic and hydrophobic matching. Adding such terms in the scoring typically requires multiple FFT

evaluation per pose. Widely used nowadays (Comeau et al. 2007; Pierce et al. 2011; Jiménez-García et al. 2013), such methods efficiently perform an exhaustive rigid-body search.

- *Geometric hashing docking*. First developed in the area of computer vision and implemented in docking by Wolfson and colleagues (Fischer et al. 1992; Mashiach et al. 2010b), this approach allows efficient searching by dividing the biomolecular surface into patches and matching them across the interacting molecules.
- *Spherical harmonics-based docking*. Pioneered by Ritchie and co-workers (Ritchie and Kemp 2000; Macindoe et al. 2010), this uses spherical polar Fourier correlations to accelerate the search, describing the protein shapes as a combination of spherical harmonic functions and calculating the relative orientations via scalar products of rotated and translated coefficient vectors.

Those methods can evaluate very large numbers of interaction poses in a relatively short time amount, making efficient use of computational resources (CPU cores), but other algorithms, although less computationally efficient, can reach high performance as well. HADDOCK (Dominguez et al. 2003) for example uses a gradient-based search method in Cartesian space (rigid-body energy minimization), targeting specific patches on the molecular surface deemed favorable by the energy function used. ATTRACT (Zacharias 2005) pioneered normal-mode analysis into the searching phase and SwarmDock (Moal and Bates 2010) incorporated it into a Particle Swarm Optimization meta-heuristic to perform docking while optimizing conformation, position and orientation simultaneously.

Table 8.1 reports a list of the top-performing docking approaches in CAPRI. For a more complete compilation of the existing docking programs see the latest CAPRI assessment, for recent reviews on the topic see (Moreira et al. 2010; Rodrigues and Bonvin 2014).

### 8.2.2 Step 2: Scoring

While the goal of sampling is to generate a set of poses, ideally with the highest number of correct conformations (although non-exhaustive sampling might not allow to do that), the goal of scoring is to single out the near-native ones within the pool of models generated. Due to the high complexity of the energetics governing the interaction, scoring is a critical step in docking: for such a reason, a separate challenge to test scoring methods has been added to CAPRI (Lensink et al. 2007).

In an ideal scoring process, one or more descriptors of the docking poses allows to derive a score, which nicely correlates with the model quality (in terms of similarity to the true solution), thus unambiguously distinguishing correct solutions from incorrect ones (Fig. 8.3a–c).

However, current scoring functions are far from reaching perfection, although the CAPRI experiment shows that they are constantly improving (Lensink and

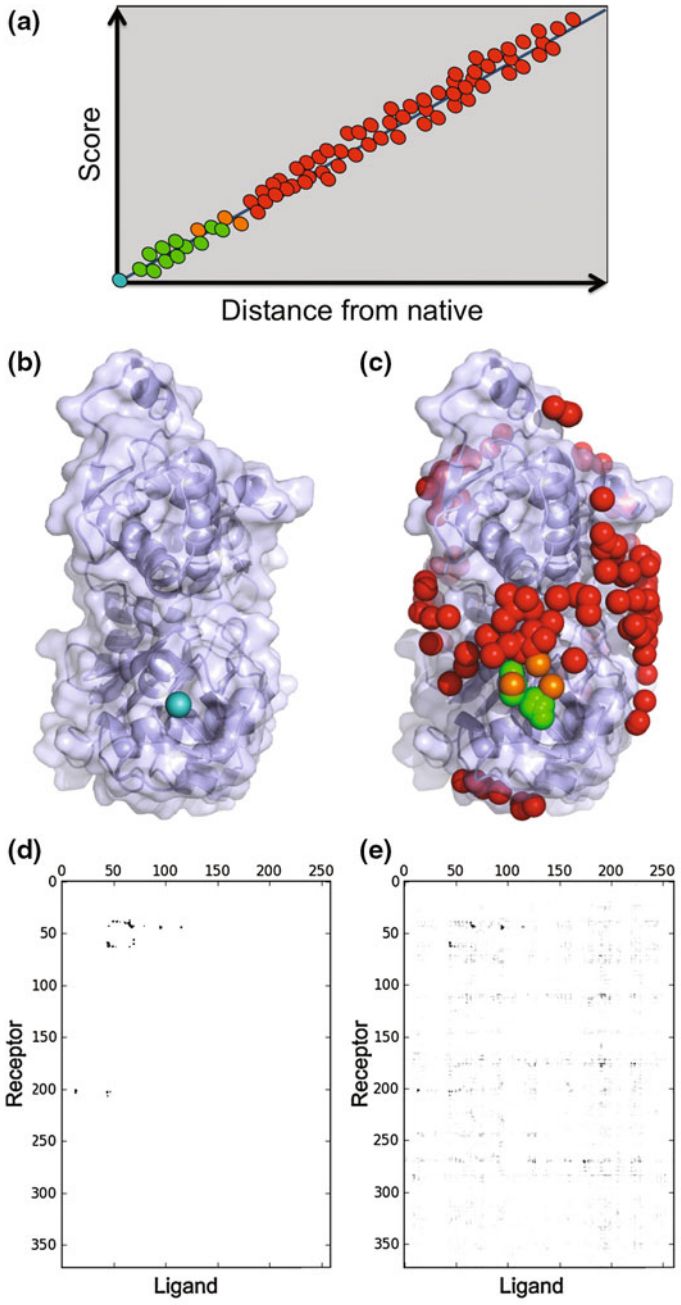
**Table 8.1** List of protein-protein docking algorithms

| Program name | Searching protocol details | Web-server |
|---|---|---|
| ATTRACT (Zacharias 2005) | Energy minimization in translational and rotational space using NMA to allow conformational changes upon binding | None |
| ClusPro (Comeau et al. 2004b) | Rigid-body search via FFT | http://cluspro.bu.edu |
| GRAMM-X (Tovchigrechko and Vakser 2006) | Grid-based FFT rigid-body docking | http://vakser.compbio.ku.edu/resources/gramm/grammx/ |
| HADDOCK (de Vries et al. 2010) | Rigid-body energy minimization followed by semi-flexible refinement in torsion angle space | http://haddocking.org |
| HEX server (Macindoe et al. 2010) | Spherical harmonics, polar FFT | http://hexserver.loria.fr |
| PatchDock (Schneidman-Duhovny et al. 2005) | Geometric hashing | http://bioinfo3d.cs.tau.ac.il/PatchDock |
| pyDock (Cheng et al. 2007) | Rigid-body search via FFT | http://life.bsc.es/servlet/pydock/home |
| RosettaDock (Lyskov and Gray 2008) | Low-resolution, rigid-body MC search | http://antibody.graylab.jhu.edu/docking |
| SwarmDock (Moal and Bates 2010) | Local docking and particle swarm optimization of position and orientation, NMA | http://bmm.cancerresearchuk.org/~SwarmDock/ |
| ZDOCK (Chen et al. 2003) | FFT-based rigid-body serach | http://zdock.umassmed.edu |

*FFT* fast Fourier transform, *MC* Monte Carlo, *NMA* Normal Mode Analysis

Wodak 2010, 2013). Traditionally, scoring functions for protein-protein docking poses rely on two approaches, both of them widely tested in CAPRI blind tests where they were shown to perform competitively. The first approach uses a linear combination of energy terms, which can be physics-based and/or empirical, such as van der Waals, electrostatics and desolvation energies, buried surface area and terms accounting for shape complementarity (Gray et al. 2003; Cheng et al. 2007; de Vries et al. 2007; Venkatraman et al. 2009; Gong et al. 2010). Weights used in the linear combination are usually optimized to distinguish native-like solutions from non native-like ones.

The second traditional approach is statistics-based or "knowledge-based", as it uses properties derived from experimental structures of protein-protein complexes. Such properties are usually embodied in atom-atom or residue-residue potentials, derived from the statistical occurrences observed in the analyzed database of complexes by means of an inverse Boltzmann equation (the higher the population, the lower the energy) (Moont et al. 1999; Jiang et al. 2002; Lu et al. 2003; Huang and Zou 2008; Kowalsman and Eisenstein 2009; Khashan et al. 2012).


(a)
Score
Distance from native
(b)
(c)
(d)
Receptor
Ligand
0
50
100
150
200
250
300
350
(e)
Receptor
Ligand

◂**Fig. 8.3** **a** Scheme of an ideal scoring process: the score strongly correlates with the distance of the model from the native structure (same color scheme of **b** and **c**). **b**, **d** Actin-DNase I complex [PDB ID: 1ATN (Kabsch et al. 1990)]: surface representation of the receptor (actin, *light blue*) with sphere representation of the center of mass of the ligand (DNase I, teal) interface (**b**) and intermolecular contact map generated by COCOMAPS server (Vangone et al. 2011) (**d**). **c**, **e** An ensemble of 185 predicted docking poses for 1ATN: surface representation of the receptor (*light blue*) with sphere representation of the centre of mass of the model ligand interface (*green*: correct; *red*: incorrect; *orange*: intermediate **c** and 'consensus map' (**e**)

This approach (Viswanath et al. 2013), like the energy-based one, can also take advantage of a training process on extended sets of docking poses, to distinguish correct from incorrect solutions.

The above approaches are, however, not mutually exclusive and in several scoring functions they are indeed combined into a hybrid approach (Pierce and Weng 2007; Andrusier et al. 2007; Vreven et al. 2011). Some of these methods also take advantage of machine learning algorithms in the training process to derive best coefficients to combine the different scoring terms (Champ and Camacho 2007; Fink et al. 2011).

It is important to mention that, as now generally accepted, a native structure is not an isolated event in the global energy landscape and thus native-like models are expected to form "funnels", i.e. clusters of similar low energy solutions. The clustering is often done based on RMSD comparisons between models, but can also efficiently be performed based on the fraction of common contact as introduced by Rodrigues et al. (2012). On these bases, some scoring methods try to characterize funnel-like energy structures on the global energy landscape (Kozakov et al. 2008; London and Schueler-Furman 2008; Moal and Bates 2010; Torchala et al. 2013), also using the concept of transient complex (Qin and Zhou 2013), while others, after scoring, perform a clustering of models in an ensemble of low-energy conformations and select the top ones based on the cluster population (Comeau et al. 2004a). The above approaches implicitly use the concept of consensus, i.e. similarity within an ensemble of docking models. More recently, a "pure" consensus method, CONSRANK, based on the frequency of inter-residue contacts in an ensemble of docking solutions, has been proposed for the ranking of docking solutions. Blind testing in CAPRI Round 30 showed it to perform competitively with classical energy- and knowledge-based approaches.

Other approaches to the scoring include methods using evolutionary information (Tress et al. 2005; Andreani et al. 2013; Xue et al. 2014) and methods using experimental information on the complex, when available (de Vries et al. 2007; Gajda et al. 2010; Moreira et al. 2015). For recent reviews on the topic see (Moal et al. 2013a, b).

### 8.2.3 Data-Driven Docking

Although important progresses in the searching and scoring procedures have been achieved, one of the most useful approaches to improve the quality of the docking simulations is the use of biological information about the interaction regions of the complex when available. As clearly inferable from the latest CAPRI assessment reports (Lensink and Wodak 2013), information (experimentally or computationally derived) on regions and residues involved in the interaction is one of the key points for the improvement of a docking simulation. Many docking programs offer the possibility to integrate data, for example as a scoring bias or as a filter to select solutions at the end, to exclude from the search regions not involved in the interaction or to drive the docking towards the areas known to be involved.

HADDOCK, one of the top performing docking program in the last CAPRI rounds (Lensink and Wodak 2013; Lensink et al. 2016), is the pioneer of data (or information)-driven docking and, in contrast to other docking methods that usually incorporate data at some stage of the protocol, HADDOCK is the only program that uses such data throughout the entire protocol (see Sect. 8.3.1). In HADDOCK the data (experimental and/or predicted) are incorporated into the calculation as an additional restraint energy term, as distance [i.e. mutagenesis, nuclear Overhauser effect, chemical cross-links, electron paramagnetic resonance distances, or even co-evolution-derived distances (Hopf et al. 2014)], orientation [e.g. NMR residual dipolar coupling (van Dijk et al. 2005), pseudo-contact shifts (Schmitz and Bonvin 2011)] or relaxation anisotropy (van Dijk et al. 2006) restraints (Schmitz et al. 2012) or even recently shape information [e.g. cryo-EM data (van Zundert et al. 2015)]. HADDOCK implements the concept of highly ambiguous distance restraints to incorporate information which define patches of interacting residues but no specific pairwise interactions between them (like in the case of NMR chemical shift perturbations).

Most traditionally successful methods in CAPRI also offer the possibility to integrate data into the protocol: FFT-based approaches [ClusPro (Comeau et al. 2004b), GRAMM-X (Tovchigrechko and Vakser 2006), pyDock (Cheng et al. 2007), ZDOCK (Chen et al. 2003) and HEX (Macindoe et al. 2010)] use data to bias the score toward models that satisfy it, or as a filter at the end. Thus, SwarmDock (Moal and Bates 2010) uses the data to pre-orientate the molecules such as the identified or predicted interfaces face each other while PatchDock (Schneidman-Duhovny et al. 2005) allows the definition of interacting or non-interacting regions, and also the setting of distance constraints. The RosettaDock (Lyskov and Gray 2008) program includes data as distance-filters to bias the Monte Carlo search whereas the most recent version of ATTRACT now also supports ambiguous distance restraints and allows docking using Cryo-EM density maps (de Vries and Zacharias 2012). Finally, ZDOCK (Chen et al. 2003) includes specific knowledge-based scoring functions in the protocol.

The quality of models coming out of data-driven docking approaches will depend on the quality of the data used. The most common experiments that give

information about interface residues involved in the binding are mutagenesis, NMR chemical shift perturbation and cross-saturation and hydrogen/deuterium exchange, while techniques such as nuclear Overhauser effect in NMR and cross-link experiments in mass spectromety provide distance information. This experimental information can be complemented or even replaced by bioinformatics predictions. These are mostly based on the study of sequence/structure conservation of key residues, co-evolution principles allowing to derive residue pairs in predicted proximity, propensity of residues to be surface-exposed, or the combination of such information as consensus and partner-specific methods (Neuvirth et al. 2004; de Vries et al. 2006; Porollo and Meller 2006; Negi et al. 2007; Qin and Zhou 2007; Ashkenazy et al. 2010; Ahmad and Mizuguchi 2011; de Vries and Bonvin 2011; Zhang et al. 2011; Zellner et al. 2012; Xue et al. 2014). However, the predictions have to be analyzed critically and combined with experimental information when available.

## 8.3 The Challenges of Docking: Flexibility and Binding Affinity

### *8.3.1 Changes upon Binding: The Flexible Docking Challenge*

Although docking programs have improved their performance over the years according to CAPRI, predicting the structure of biomolecular complexes remains a difficult problem with, at the moment, two major challenges: the identification of correct solutions within a pool of models (scoring) and the treatment of proteins with substantial conformational change upon binding (flexibility).

Proteins are not rigid, and during the association process they usually undergo conformational changes that include both backbone and side-chains movements (Betts and Sternberg 1999). As a result, the conformation of the proteins within the complex/bound form might be different from the one they have in the free form. Therefore, incorporating flexibility in docking algorithms is necessary to predict the native associations and reach high accuracy of the solutions. In the cases where structural changes occurring upon binding are minimal, the difference between bound and free forms can be neglected so the rigid body docking is sufficient. A major problem here is that, in general, one can not know a priori if conformational changes will take place or not, nor their extent. Properly dealing with flexibility in docking is therefore one of the main challenges in the field (Smith et al. 2005a; Bonvin 2006; Lensink et al. 2007).

A major problem of incorporating flexibility in docking, compared to performing rigid-body docking only, is the considerable increase in the number of degrees of freedom and, consequently, in the search space. This also often goes together with a higher rate of false-positive solutions, since all might be refined to some local

energy minimum, which thus complicates the identification of correct solutions (Andrusier et al. 2008).

Flexibility can be introduced at several levels:

- *Implicitly*. Implicit flexibility can be incorporated by soft-docking, by smoothing the protein surface or allowing some degrees of interpenetration or overlap of atoms (Palma et al. 2000; Heifetz and Eisenstein 2003) [although one of the drawbacks of such an approach is that severe steric clashes can be introduced (Smith et al. 2005b)], or with cross-docking by performing rigid-body docking of ensembles of conformations, taken for example from NMR structures or MD simulations or any other conformational sampling method (de Groot et al. 1997). Depending on the implementation this can lead to a significant increase in computing time. It has, on the other hand, the advantage that rather large conformational changes can be pre-sampled in that way.
- *Explicitly*. In the past few years, flexibility has been explicitly introduced into the docking process by allowing side-chains and/or backbone to move. The docking programs allowing side-chain flexibility (Fernández-Recio et al. 2003; Zacharias 2005; de Vries et al. 2007; Lyskov and Gray 2008; de Vries et al. 2010) use different approaches, like Monte Carlo (MC) optimization of the ligand (ICM-DISCO) (Fernández-Recio et al. 2003), sampling the known populated rotamers of the side-chains followed by energy minimization steps (ATTRACT) (Zacharias 2005), using MD simulated annealing for refinement of both receptor and ligand side-chains (HADDOCK) (de Vries et al. 2010), or repacking and optimization of side-chains in a MC search (RosettaDock) (Lyskov and Gray 2008).

In contrast with side-chains flexibility, which is easier to model, backbone flexibility is currently one of the main challenges in docking.

In addition to conformational changes upon binding, some programs have been developed to tackle the challenge of large domain motions, such as the flexible multi-domain docking approach proposed by Karaca and Bonvin (Karaca and Bonvin 2011) that can describe large domain motion-type conformational changes. The proper treatment of flexibility in protein-protein docking and also for peptide docking (see Sect. 8.4) remains an active area of research. In small-molecule docking (like protein-ligand docking), in which flexibility plays a major role, the problem is more tractable, but no less challenging (Brooijmans and Kuntz 2003; Erickson et al. 2004).

### 8.3.2 *The 'Perfect' Scoring Function and the Binding Affinity Problem*

Scoring approaches typically attempt to fish the most likely model of a complex from a set of poses but are not designed to predict how strongly the proteins bind,

i.e. their free energy of binding $\Delta G_{binding}$, or whether they bind at all [as showed by cross-docking simulations (Sacquin-Mora et al. 2008; Wass et al. 2011a, b, Martin and Lavery 2012)]. That is because scoring (ranking) and binding affinity prediction ($\Delta G$) are two different things. The $\Delta G_{binding,}$ or Gibbs free energy of the complex can be determined by measuring the dissociation constant as:

$$\Delta G = RT \ln K_d$$

where R is the gas constant, T is the temperature and $K_d$ is the dissociation constant. It reflects the natural inclination of molecules entities to associate and is a key thermodynamic quantity for understanding recognition and association phenomena, and possible dysfunctions thereof.

Accurately predicting binding free energies with a general scoring function, while a very ambitious goal, would revolutionize the efficiency of docking methods. Different methods aimed at predicting binding affinity in protein complexes have been proposed throughout the years, taking into account different structural and energetic features of the complex and varying greatly in terms of accuracy and computational cost. Based on the initial observation of Chothia and Janin (1975) in the 1970s and described by Horton and Lewis (1992) in 1992, the buried surface area (BSA), i.e. the surface that is buried upon complex formation, has been the first descriptor to be related to the binding affinity. Since then, many methods have been proposed. Exact methods such as free energy perturbation and thermodynamics integration can be very accurate, but due to their computational costs their application is extremely limited (mostly to low throughput studies and mainly for small drug binding or mutations). Methods based on empirical functions (empirical, force field-based potentials, statistical potentials, scoring functions used in docking) are much faster (Jiang et al. 2002; Ma et al. 2002; Zhang et al. 2005; Audie and Scarlata 2007; Su et al. 2009; Bai et al. 2011; Qin et al. 2011; Moal and Bates 2012; Tian et al. 2012; Luo et al. 2014; Kastritis et al. 2014). However, even if some have been very successful on small training sets (Horton and Lewis 1992; Audie and Scarlata 2007), most published models still fail to systematically predict the binding affinity (Kastritis and Bonvin 2010, 2013a, b) for large datasets or discriminate between binders from non-binders (Sacquin-Mora et al. 2008; Fleishman et al. 2011). Usually, factors such as conformational changes occurring upon binding, allosteric regulation, solvent and co-factor effects, which may contribute to the binding strength, are neglected, which entails their main weaknesses. Using a large binding affinity benchmark consisting of 144 complexes (Kastritis et al. 2011) [updated version of the benchmark now available in (Vreven et al., 2015)], Kastritis et al. (2014) demonstrated that non-interacting surface properties like percentages of charged and polar residues do also contribute to binding affinity. These rather surprising finding were corroborated in a recent study by Cazal et al. in which this contribution from the non-interaction surface was reproduced (Marillet et al. 2015). New advances were made by Vangone and Bonvin (Vangone and Bonvin 2015; Xue et al. 2016) who recently showed that the network of contacts made at the

interface in a protein-protein complex is a better structural descriptor of the binding affinity then the BSA.

## 8.4 Protein-Peptide Docking

In eukaryotes more than 40% of the interactions are estimated to be mediated by peptides, for example in signal transduction, protein degradation, transcription regulation and immune response (Petsalaki and Russell 2008). Due to their involvement in many biological pathways, peptide interactions are implicated in many diseases and in cancer (Petsalaki and Russell 2008; Naider and Anglister 2009), making them of high interest in the development of new therapeutics and for drug design (Vaara 2009). Indeed, alongside small-molecule inhibitors, peptides are large enough to competitively inhibit protein-protein interactions and can mimic protein binding domains. However, the experimental structure determination of protein-peptide recognition remains a challenging task also in this case, mainly due of two factors: peptides are highly flexible and they usually show transient interactions with the substrate. From a structural point of view, peptides are short chains ranging from 5 to 30 amino acids, often lacking a well-defined fold in their free form. They might not necessarily be independent molecules, but can appear as disordered regions of proteins (for example at termini), and they can show multiplicity in their interaction, as for example in the case of the tumor suppressor p53 (Russell and Gibson 2008).

Complementary computational prediction methods like docking are therefore urgently required to model those systems, as also reflected by the recent addition of protein-peptides cases in CAPRI. Peptides' peculiar characteristics represent, however, a unique challenge for computational predictions. Conventional protein-protein docking struggles with the high flexibility of peptides while ligand-docking protocols have only been successfully applied to short peptide, due to the significant higher number of peptide rotatable bonds than in drug-like small molecules (Hetényi and van der Spoel 2002; Sousa et al. 2006; Rubinstein and Niv 2009; London et al. 2013). Over the last years a number of new algorithms or ad hoc adaptations have been developed with the aim of modelling protein-peptide complexes (Petsalaki et al. 2009; Antes 2010; Raveh et al. 2010; Ben-Shimon and Eisenstein 2010; Raveh et al. 2011; Donsky and Wolfson 2011; Dagliyan et al. 2011; Trellet et al. 2013; Verschueren et al. 2013; Lavi et al. 2013; Ben-Shimon and Niv 2015; Kurcinski et al. 2015). Similarly to protein-protein docking, there are two main steps: (i) identification of the binding site on the protein surface (which might include the use of experimental or bioinformatics data when available; see also Chap. 10) and (ii) docking and refinement of the peptide into the binding site.

Several high-resolution approaches have been successfully applied to unbound protein-peptide datasets. FlexPepDock (Raveh et al. 2010, 2011), the first generic algorithm released to model protein-peptide complexes, uses fragment-based sampling for the generations of different peptide backbone conformations, and then

allows full flexibility of the peptide and to the protein side chains within a defined docking site. HADDOCK (Trellet et al. 2013) overcomes the problem of the indetermination of the peptide free structure by using as input an ensemble of three different conformation of the peptide: alpha-helix, polyproline-II and extended. Taken together, these conformations cover about 80% of the observed protein-peptide structures in the PDB (Diella et al. 2008). This is followed by flexible refinement step in which more flexibility is given to the peptide. This protocol mimics the conformational selection mechanism/induced fit recognition mechanism (Weikl and Deuster 2009; Hammes et al. 2009; Csermely et al. 2010; Changeux and Edelstein 2011). Lately Blaszczyk and co-workers implemented CABS-dock, an ab initio protein-peptide modelling approach (Kurcinski et al. 2015) that performs the search for the binding site and docking (giving flexibility) simultaneously using a coarse grained representation of the system. Additional ab initio algorithms or tools aimed to predict candidate sites of interaction on the protein surface (Trabuco et al. 2012) have been implemented lately to overcome the lack of information on the peptide binding site (Ben-Shimon and Niv 2015). which is, in addition to the high flexibility of peptides, the main challenge to overcome in protein-peptide docking.

Despite the recent progresses, this is a field that still is its infancy with further development and extensive evaluation required, for example in CAPRI challenges. For further information please check (London et al. 2013; Trellet et al. 2015).

## 8.5 Post-docking: Interface Prediction from Docking Results and Use of Docking-Derived Contacts for Clustering and Ranking

It is now over ten years since Fernandez-Recio et al. (2004) proposed to predict residues at the protein-protein interface from results of docking simulations (Fig. 8.4a). They analyzed the rigid-body docking energy landscape in several training sets, in search of protein recognition areas, showing that the energy profile for the ensemble of found docked poses can be used to determine accurately interaction sites on protein surfaces. In particular, they defined a normalized interface propensity (NIP) parameter, which represents the tendency of a given residue to be located at the interface, based on the buried surface area in docking poses from rigid docking simulations. Based on the NIP definition, more recently Fernadez-Recio and Grosdidier derived a method for hot-spot residues prediction, achieving up to 80% positive predictive value (Grosdidier and Fernández-Recio 2008).

In 2010, based on their experience as assessors in the CAPRI experiment, Lensink and Wodak confirmed the potential of docking techniques for the prediction of protein interfaces (Lensink et al. 2014). By analyzing docking models submitted in CAPRI by 76 participants for 46 interfaces in 20 targets, they found

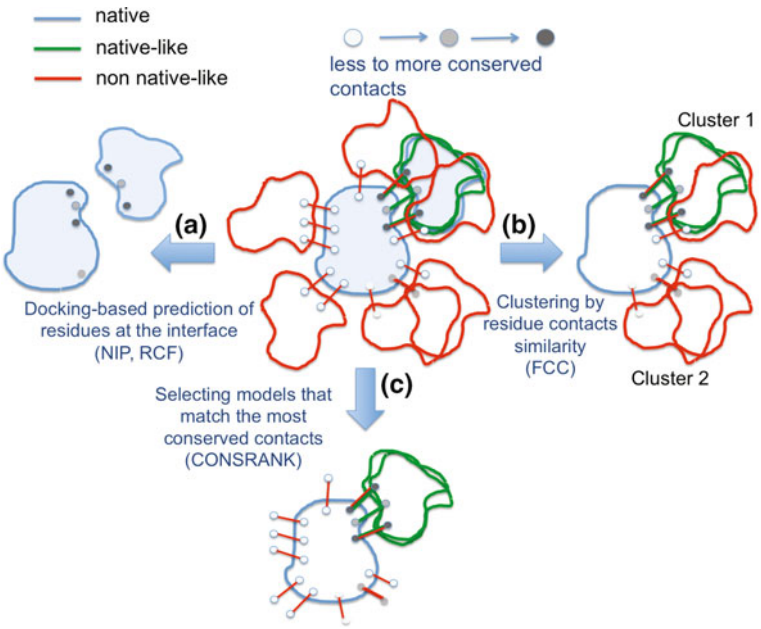



**Fig. 8.4** Scheme of the use of docking results for: **a** predicting residues at the interface, and for: **b** models clustering and **c** ranking. Figure adapted from Oliva et al. (2013)

that the best performing groups were able to predict residues at the interface with precision and sensitivity levels around 60% for the majority of the analyzed cases, thus reaching a performance competitive with the most successful non-docking based methods in the field. The main finding of this analysis was thus that, apparently, models ranked highly by docking procedures are more enriched in correct interfaces than in correct complexes. In fact, prediction of correct interfaces is also contributed by incorrect (according to the CAPRI assessment) models, which were found to feature one quarter of correct interfaces (with precision and sensitivity above 50%), contributing to 70% of the overall correct interface predictions.

de Vries and Bsonvin also showed that, after improving the performance of docking predictions with HADDOCK by a consensus monomer-based interface prediction, the interface prediction itself could be further improved by post-prediction based on top-scored docking results (de Vries and Bonvin 2011). Following these findings, Weng and colleagues recently developed RCF (residue contact frequency), another method to predict interface residues from models generated by docking algorithms (Hwang et al. 2014) (Fig. 8.4a). They used RCF to predict the binding interfaces of proteins that bind to multiple partners, finding that it correctly predicts interface residues unique for the respective binding partners. They also showed that the combination of RCF with monomer-based interface prediction methods, through a support vector machine, improved performance compared to both separated approaches. RCF was also used by the Weng's group to

analyze their docking results in the CAPRI rounds 20–26, where selection of final models for submission was in fact guided by RCF (Vreven et al. 2013).

Besides the identification of residues likely involved in the interface from results of docking simulations, specific inter-residue contacts observed in docking poses have been recently used to guide their clustering, analysis and ranking. As the native structure of a complex is not expected to be an isolated position in the energy landscape, docking experiments often incorporate one clustering step in their protocols, which is classically based on time-consuming (live memory, RAM) and size-dependent RMSD measures (Janin 2010). In this context, Bonvin and colleagues proposed the use of the fraction of common contacts (FCC) within models as a similarity description to base their clustering on (Rodrigues et al. 2012) (Fig. 8.4b). They showed that FCC is an efficient measure of the structure similarity for protein complexes, greatly reducing the computation time while generating clusters of similar quality with the state-of-the art RMSD-based methods. Further, it is particularly suited for flexible docking approaches, multicomponent assemblies and heterogeneous systems like protein-DNA complexes.

Oliva and colleagues proposed instead to analyse an ensemble of protein-protein docking models, by deriving a consensus based on the conservation within them of the inter-residue contacts (Vangone et al. 2012). Such a consensus can also be visualized as a "consensus contact map", i.e. an intermolecular contact map where the conservation of contacts is reported on a gray scale (see an example in Fig. 8.3e, compared to the intermolecular contact map of the corresponding crystal structure, Fig. 8.3d). Analysis of prediction sets of docking models for seven CAPRI targets showed that a significant fraction of native contacts was included within the contacts with highest conservation rate, even in the cases where only a small percentage of solutions were correct. This suggests that incorrect models can contribute to the correct prediction not only of residues, but also of specific inter-residue contacts at the complexes interface. A natural extension of this approach was the development of CONSRANK (CONSensus-RANKing) (Oliva et al. 2013; Vangone et al. 2013), a consensus method for the scoring of docking models, which ranks models based on their ability to match the most conserved contacts in the ensemble they belong to (Fig. 8.4c).

### *8.5.1 Web Tools for the Post-docking Processing*

As discussed previously (Sect. 8.2.2), a scoring/filtering step is normally included in a docking procedure. However, to date no program can provide a single docking solution with a high enough confidence to be correct. Docking programs instead generally provide the user with an ensemble of models, corresponding to a subset (usually refined) of the solutions they generated in the conformational sampling step, which possibly contain native-like models. These models have thus to be analyzed to attempt to single out the correct ones. Some tools have been specifically devoted to the post-docking processing, i.e. the analysis, scoring and ranking of

**Table 8.2** List of available web servers for the post-docking processing

| Server name | Algorithm | Analyses | URL |
|---|---|---|---|
| CCharPPI (Moal et al. 2015) | Energy/knowledge-based | 109 parameters including FireDock, PyDock, RosettaDock, SIPPER & ZRANK scores | http://life.bsc.es/pid/ccharppi/ |
| CONSRANK (Chermak et al. 2014) | Consensus-based | Contacts analysis and visualization; re-scoring | https://www.molnac.unisa.it/BioTools/consrank/ |
| DOCKRANK (Xue et al. 2014) | Evolution-based | Prediction of the interface; re-scoring | http://einstein.cs.iastate.edu/DockRank/ |
| FastContact (Champ and Camacho 2007) | Energy/knowledge-based | Energy minimization; prediction of residue contact free energies; re-scoring | http://structure.pitt.edu/servers/fastcontact/ |
| FiberDock (Mashiach et al. 2010a, 2008) | Energy/knowledge-based | Flexible refinement; re-scoring | http://bioinfo3d.cs.tau.ac.il/FiberDock/ http://bioinfo3d.cs.tau.ac.il/FireDock/ |
| FILTREST3D (Gajda et al. 2010) | User-defined restraints from experimental data | Re-scoring | http://filtrest3d.genesilico.pl/filtrest3d/ |
| FunHunt (London and Schueler-Furman 2008) | Energy-based | Characterization of local energy landscape | http://funhunt.furmanlab.cs.huji.ac.il/ |
| PROCOS (Fink et al. 2011) | Energy/knowledge-based | Re-scoring | http://compdiag.uni-regensburg.de/procos/ |

models representing the output of docking programs. Several of these post-processing tools are publicly available as web servers and are listed in Table 8.2, together with the corresponding URLs. The scoring approaches they mainly rely on, reflecting the ones described above (Sect. 8.2.2), are also reported in Table 8.2.

## 8.6 Concluding Remarks

In view of the growing interest in protein-protein interactions for pharmaceutical and medical applications, and the persistent disproportion between experimental structures available for single proteins and multiple protein systems, the relevance of molecular docking as the method of choice for modelling the structure of protein-protein complexes is set to increase.

In the last 15 years, the CAPRI blind assessment has shown that docking techniques can be successfully applied to a variety of cases, with biological information on the interface, when available, further improving results, by driving the search of allowed configurations and helping in filtering out incorrect solutions. At the same time, the development of web servers characterized by a user-friendly interface, for performing both docking predictions and post-docking analyses, is in fact making the use of this technique accessible also to a non-specialized audience.

That notwithstanding, to further extend its confident applicability to critical cases, protein-protein docking needs to face a number of challenges in the near future. First of all, the flexibility of the two interacting proteins has to be more confidently coped with, possibly by exploring novel approaches to the sampling of the conformational space. In this regard, it is remarkable that, in the latest CAPRI rounds, scorer groups have been shown to achieve overall a better prediction performance than predictor groups. In other words, the same groups typically recognized more correct solutions from ensembles of models obtained by a variety of techniques, rather than from their own generated models ensemble. This suggests that the bottleneck in a docking procedure still resides in an efficient sampling of the conformational space and that application of different docking strategies to the target system could help overcoming the issue—a kind of consensus docking strategy using various approaches. Other challenges that need to be addressed include a reliable identification of native-like models, with possibly an estimation of the binding affinity of the complex. In addition, when one of the interacting partners is a peptide, docking protocols have to deal with further challenges, such as the high flexibility and the undefined folding of peptides.

Finally, the prediction of the 3D structure of a biomolecular complex, which is fundamental for understanding biological processes, can also help in advancement of related fields. Indeed, it is becoming increasingly clear that results of docking simulations can also be used as an intermediate step for other applications, such as the interface prediction itself, which can be very valuable for experimentalists to guide their work (e.g. to target mutagenesis to interesting regions on the surface of a protein). Further, three-dimensional structural information can also be useful to identify pair on interacting proteins/peptide motifs with the final goal to predict the full network of protein-protein interactions governing the cells (Zhang et al. 2012; Chen et al. 2015).

**Acknowledgements** AV was supported by Marie Skłodowska-Curie Individual Fellowship H2020 MSCA-IF-2015 [BAP-659025]. RO was supported by Regione Campania [LR5-AF2008].

## References


Ahmad S, Mizuguchi K (2011) Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS ONE 6:e29104. doi:10.1371/journal.pone.0029104

Alberts B (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell 92:291–294. doi:10.1016/s0092-8674(00)80922-8

Andreani J, Faure G, Guerois R (2013) InterEvScore: a novel coarse-grained interface scoring function using a multi-body statistical potential coupled to evolution. Bioinformatics 29:1742–1749. doi:10.1093/bioinformatics/btt260

Andrusier N, Mashiach E, Nussinov R, Wolfson HJ (2008) Principles of flexible protein-protein docking. Proteins 73:271–289. doi:10.1002/prot.22170

Andrusier N, Nussinov R, Wolfson HJ (2007) FireDock: fast interaction refinement in molecular docking. Proteins 69:139–159. doi:10.1002/prot.21495

Antes I (2010) DynaDock: a new molecular dynamics-based algorithm for protein-peptide docking including receptor flexibility. Proteins 78:1084–1104. doi:10.1002/prot.22629

Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38:W529–W533. doi:10.1093/nar/gkq399

Audie J, Scarlata S (2007) A novel empirical free energy function that explains and predicts protein-protein binding affinities. Biophys Chem 129:198–211. doi:10.1016/j.bpc.2007.05.021

Bai H, Yang K, Yu D, Zhang C, Chen F, Lai L (2011) Predicting kinetic constants of protein-protein interactions based on structural properties. Proteins 79:720–734. doi:10.1002/prot.22904

Bai X-C, McMullan G, Scheres SHW (2015) How cryo-EM is revolutionizing structural biology. Trends Biochem Sci 40:49–57. doi:10.1016/j.tibs.2014.10.005

Ben-Shimon A, Eisenstein M (2010) Computational mapping of anchoring spots on protein surfaces. J Mol Biol 402:259–277. doi:10.1016/j.jmb.2010.07.021

Ben-Shimon A, Niv MY (2015) AnchorDock: blind and flexible anchor-driven peptide docking. Structure 23:929–940. doi:10.1016/j.str.2015.03.010

Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542. doi:10.1016/s0022-2836(77)80200-3

Betts MJ, Sternberg MJ (1999) An analysis of conformational changes on protein-protein association: implications for predictive docking. Protein Eng 12:271–283. doi:10.1093/protein/12.4.271

Bonvin AMJJ (2006) Flexible protein-protein docking. Curr Opin Struct Biol 16:194–200. doi:10.1016/j.sbi.2006.02.002

Brooijmans N, Kuntz ID (2003) Molecular recognition and docking algorithms. Annu Rev Biophys Biomol Struct 32:335–373. doi:10.1146/annurev.biophys.32.110601.142532

Buckle AM, Schreiber G, Fersht AR (1994) Protein-protein recognition: crystal structural analysis of a barnase-barstar complex at 2.0-A resolution. Biochemistry 33:8878–8889. doi:10.1021/bi00196a004

Champ PC, Camacho CJ (2007) FastContact: a free energy scoring tool for protein-protein complex structures. Nucleic Acids Res 35:W556–W560. doi:10.1093/nar/gkm326

Changeux J-P, Edelstein S (2011) Conformational selection or induced fit? 50 years of debate resolved. F1000 Biol Rep 3:19. doi:10.3410/B3-19

Chaudhuri BN (2015) Emerging applications of small angle solution scattering in structural biology. Protein Sci 24:267–276. doi:10.1002/pro.2624

Chen R, Li L, Weng Z (2003) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87. doi:10.1002/prot.10389

Chen TS, Petrey D, Garzon JI, Honig B (2015) Predicting peptide-mediated interactions on a genome-wide scale. PLoS Comput Biol 11:e1004248. doi:10.1371/journal.pcbi.1004248

Cheng TM-K, Blundell TL, Fernández-Recio J (2007) pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. Proteins 68:503–515. doi:10.1002/prot.21419

Chermak E, Petta A, Serra L, Vangone A, Scarano V, Cavallo L, Oliva R (2014) CONSRANK: a server for the analysis, comparison and ranking of docking models based on inter-residue contacts. Bioinformatics 31:1481–1483. doi:10.1093/bioinformatics/btu837

Chothia C, Janin J (1975) Principles of protein–protein recognition. Nature 256:705–708. doi:10.1038/256705a0

Chruszcz M, Domagalski M, Osinski T, Wlodawer A, Minor W (2010) Unmet challenges of structural genomics. Curr Opin Struct Biol 20:587–597. doi:10.1016/j.sbi.2010.08.001

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004a) ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 20:45–50. doi:10.1093/bioinformatics/btg371

Comeau SR, Gatchell DW, Vajda S, Camacho CJ (2004b) ClusPro: a fully automated algorithm for protein-protein docking. Nucleic Acids Res 32:W96–W99. doi:10.1093/nar/gkh354

Comeau SR, Kozakov D, Brenke R, Shen Y, Beglov D, Vajda S (2007) ClusPro: performance in CAPRI rounds 6–11 and the new server. Proteins 69:781–785. doi:10.1002/prot.21795

Csermely P, Palotai R, Nussinov R (2010) Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. Trends Biochem Sci 35:539–546. doi:10.1016/j.tibs.2010.04.009

Dagliyan O, Proctor EA, D'Auria KM, Ding F, Dokholyan NV (2011) Structural and dynamic determinants of protein-peptide recognition. Structure 19:1837–1845. doi:10.1016/j.str.2011.09.014

de Groot BL, van Aalten DM, Scheek RM, Amadei A, Vriend G, Berendsen HJ (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29:240–251. doi:10.1002/(SICI)1097-0134(199710)29:2<240:AID-PROT11>3.0.CO;2-O

de Vries SJ, Bonvin AMJJ (2011) CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS ONE 6:e17695. doi:10.1371/journal.pone.0017695

de Vries SJ, van Dijk ADJ, Bonvin AMJJ (2006) WHISCY: what information does surface conservation yield? Application to data-driven docking. Proteins 63:479–489. doi:10.1002/prot.20842

de Vries SJ, van Dijk ADJ, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AMJJ (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733. doi:10.1002/prot.21723

de Vries SJ, van Dijk M, Bonvin AMJJ (2010) The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 5:883–897. doi:10.1038/nprot.2010.32

de Vries SJ, Zacharias M (2012) ATTRACT-EM: a new method for the computational assembly of large molecular machines using cryo-EM maps. PLoS ONE 7:e49733. doi:10.1371/journal.pone.0049733

Diella F, Haslam N, Chica C, Budd A, Michael S, Brown NP, Trave G, Gibson TJ (2008) Understanding eukaryotic linear motifs and their role in cell signaling and regulation. Front Biosci 13:6580–6603. doi:10.2741/3175

Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737. doi:10.1021/ja026939x

Donsky E, Wolfson HJ (2011) PepCrawler: a fast RRT-based algorithm for high-resolution refinement and binding affinity estimation of peptide inhibitors. Bioinformatics 27:2836–2842. doi:10.1093/bioinformatics/btr498

Erickson JA, Jalaie M, Robertson DH, Lewis RA, Vieth M (2004) Lessons in molecular recognition: the effects of ligand and protein flexibility on molecular docking accuracy. J Med Chem 47:45–55. doi:10.1021/jm030209y

Fernández-Recio J, Totrov M, Abagyan R (2003) ICM-DISCO docking by global energy optimization with fully flexible side-chains. Proteins 52:113–117. doi:10.1002/prot.10383

Fernández-Recio J, Totrov M, Abagyan R (2004) Identification of protein-protein interaction sites from docking energy landscapes. J Mol Biol 335:843–865. doi:10.1016/j.jmb.2003.10.069

Fink F, Hochrein J, Wolowski V, Merkl R, Gronwald W (2011) PROCOS: computational analysis of protein-protein complexes. J Comput Chem 32:2575–2586. doi:10.1002/jcc.21837

Fischer D, Bachar O, Nussinov R, Wolfson H (1992) An efficient automated computer vision based technique for detection of three dimensional structural motifs in proteins. J Biomol Struct Dyn 9:769–789. doi:10.1080/07391102.1992.10507955

Fischer N, Neumann P, Konevega AL, Bock LV, Ficner R, Rodnina MV, Stark H (2015) Structure of the *E. coli* ribosome-EF-Tu complex at <3 Å resolution by Cs-corrected cryo-EM. Nature 520:567–570. doi:10.1038/nature14275

Fleishman SJ, Whitehead TA, Strauch E-M, Corn JE, Qin S, Zhou H-X, Mitchell JC, Demerdash ONA, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko J-S, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovalı ŞK, Ozbek P, Tal NB, Haliloglu T, Hwang H, Vreven T, Pierce BG, Weng Z, Pérez-Cano L, Pons C, Fernández-Recio J, Jiang F, Yang F, Gong X, Cao L, Xu X, Liu B, Wang P, Li C, Wang C, Robert CH, Guharoy M, Liu S, Huang Y, Li L, Guo D, Chen Y, Xiao Y, London N, Itzhaki Z, Schueler-Furman O, Inbar Y, Potapov V, Cohen M, Schreiber G, Tsuchiya Y, Kanamori E, Standley DM, Nakamura H, Kinoshita K, Driggers CM, Hall RG, Morgan JL, Hsu VL, Zhan J, Yang Y, Zhou Y, Kastritis PL, Bonvin AMJJ, Zhang W, Camacho CJ, Kilambi KP, Sircar A, Gray JJ, Ohue M, Uchikoga N, Matsuzaki Y, Ishida T, Akiyama Y, Khashan R, Bush S, Fouches D, Tropsha A, Esquivel-Rodríguez J, Kihara D, Stranges PB, Jacak R, Kuhlman B, Huang S-Y, Zou X, Wodak SJ, Janin J, Baker D (2011) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol 414:289–302. doi:10.1016/j.jmb.2011.09.031

Gajda MJ, Tuszynska I, Kaczor M, Bakulina AY, Bujnicki JM (2010) FILTREST3D: discrimination of structural models using restraints from experimental data. Bioinformatics 26:2986–2987. doi:10.1093/bioinformatics/btq582

Gong X, Wang P, Yang F, Chang S, Liu B, He H, Cao L, Xu X, Li C, Chen W, Wang C (2010) Protein-protein docking with binding site patch prediction and network-based terms enhanced combinatorial scoring. Proteins 78:3150–3155. doi:10.1002/prot.22831

González-Ruiz D, Gohlke H (2006) Targeting protein-protein interactions with small molecules: challenges and perspectives for computational binding epitope detection and ligand finding. Curr Med Chem 13:2607–2625. doi:10.2174/092986706778201530

Gray JJ, Moughon S, Wang C, Schueler-Furman O, Kuhlman B, Rohl CA, Baker D (2003) Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J Mol Biol 331:281–299. doi:10.1016/s0022-2836(03)00670-3

Grosdidier S, Fernández-Recio J (2008) Identification of hot-spot residues in protein-protein interactions by computational docking. BMC Bioinformatics 9:447. doi:10.1186/1471-2105-9-447

Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: an overview of search algorithms and a guide to scoring functions. Proteins 47:409–443. doi:10.1002/prot.10115

Hammes GG, Chang Y-C, Oas TG (2009) Conformational selection or induced fit: a flux description of reaction mechanism. Proc Natl Acad Sci 106:13737–13741. doi:10.1073/pnas.0907195106

Heifetz A, Eisenstein M (2003) Effect of local shape modifications of molecular surfaces on rigid-body protein-protein docking. Protein Eng 16:179–185. doi:10.1093/proeng/gzg021

Hetényi C, van der Spoel D (2002) Efficient docking of peptides to proteins without prior knowledge of the binding site. Protein Sci 11:1729–1737. doi:10.1110/ps.0202302

Hopf TA, Schärfe CPI, Rodrigues JPGLM, Green AG, Kohlbacher O, Sander C, Bonvin AMJJ, Marks DS (2014) Sequence co-evolution gives 3D contacts and structures of protein complexes. Elife. doi:10.7554/eLife.03430

Horton N, Lewis M (1992) Calculation of the free energy of association for protein complexes. Protein Sci 1:169–181. doi:10.1002/pro.5560010117

Huang S-Y, Zou X (2008) An iterative knowledge-based scoring function for protein-protein recognition. Proteins 72:557–579. doi:10.1002/prot.21949

Hwang H, Vreven T, Weng Z (2014) Binding interface prediction by combining protein-protein docking results. Proteins 82:57–66. doi:10.1002/prot.24354

Hwang I, Park S (2008) Computational design of protein therapeutics. Drug Discov Today Technol 5:e43–e48. doi:10.1016/j.ddtec.2008.11.004

Janin J (2010) Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol BioSyst 6:2351–2362. doi:10.1039/c005060c

Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ, Critical Assessment of PRedicted Interactions (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52:2–9. doi:10.1002/prot.10381

Jiang L, Gao Y, Mao F, Liu Z, Lai L (2002) Potential of mean force for protein-protein interaction studies. Proteins 46:190–196. doi:10.1002/prot.10031

Jiménez-García B, Pons C, Fernández-Recio J (2013) pyDockWEB: a web server for rigid-body protein-protein docking using electrostatics and desolvation scoring. Bioinformatics 29:1698–1699. doi:10.1093/bioinformatics/btt262

Jones S, Thornton JM (1996) Principles of protein-protein interactions. Proc Natl Acad Sci USA 93:13–20. doi:10.1073/pnas.93.1.13

Kabsch W, Mannherz HG, Suck D, Pai EF, Holmes KC (1990) Atomic structure of the actin: DNase I complex. Nature 347:37–44. doi:10.1038/347037a0

Karaca E, Bonvin AMJJ (2011) A multidomain flexible docking approach to deal with large conformational changes in the modeling of biomolecular complexes. Structure 19:555–565. doi:10.1016/j.str.2011.01.014

Kastritis PL, Bonvin AMJJ (2013a) Molecular origins of binding affinity: seeking the Archimedean point. Curr Opin Struct Biol 23:868–877. doi:10.1016/j.sbi.2013.07.001

Kastritis PL, Bonvin AMJJ (2013b) On the binding affinity of macromolecular interactions: daring to ask why proteins interact. J R Soc Interface 10:20120835. doi:10.1098/rsif.2012.0835

Kastritis PL, Bonvin AMJJ (2010) Are scoring functions in protein-protein docking ready to predict interactomes? Clues from a novel binding affinity benchmark. J Proteome Res 9:2216–2225. doi:10.1021/pr9009854

Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AMJJ, Janin J (2011) A structure-based benchmark for protein-protein binding affinity. Protein Sci 20:482–491. doi:10.1002/pro.580

Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ (2014) Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. J Mol Biol 426:2632–2652. doi:10.1016/j.jmb.2014.04.017

Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA (1992) Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci USA 89:2195–2199. doi:10.1073/pnas.89.6.2195

Khashan R, Zheng W, Tropsha A (2012) Scoring protein interaction decoys using exposed residues (SPIDER): a novel multibody interaction scoring function based on frequent geometric patterns of interfacial residues. Proteins 80:2207–2217. doi:10.1002/prot.24110

Kowalsman N, Eisenstein M (2009) Combining interface core and whole interface descriptors in postscan processing of protein-protein docking models. Proteins 77:297–318. doi:10.1002/prot.22436

Kozakov D, Schueler-Furman O, Vajda S (2008) Discrimination of near-native structures in protein-protein docking by testing the stability of local minima. Proteins 72:993–1004. doi:10.1002/prot.21997

Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S (2015) CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. Nucleic Acids Res 43:W419–W424. doi:10.1093/nar/gkv456

Lavi A, Ngan CH, Movshovitz-Attias D, Bohnuud T, Yueh C, Beglov D, Schueler-Furman O, Kozakov D (2013) Detection of peptide-binding sites on protein surfaces: the first step toward the modeling and targeting of peptide-mediated interactions. Proteins 81:2096–2105. doi:10.1002/prot.24422

Lensink MF, Méndez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd Edition. Proteins 69:704–718. doi:10.1002/prot.21804

Lensink MF, Moal IH, Bates PA, Kastritis PL, Melquiond ASJ, Karaca E, Schmitz C, van Dijk M, Bonvin AMJJ, Eisenstein M, Jiménez-García B, Grosdidier S, Solernou A, Pérez-Cano L, Pallara C, Fernández-Recio J, Xu J, Muthu P, Praneeth Kilambi K, Gray JJ, Grudinin S, Derevyanko G, Mitchell JC, Wieting J, Kanamori E, Tsuchiya Y, Murakami Y, Sarmiento J, Standley DM, Shirota M, Kinoshita K, Nakamura H, Chavent M, Ritchie DW, Park H, Ko J, Lee H, Seok C, Shen Y, Kozakov D, Vajda S, Kundrotas PJ, Vakser IA, Pierce BG, Hwang H, Vreven T, Weng Z, Buch I, Farkash E, Wolfson HJ, Zacharias M, Qin S, Zhou H-X, Huang S-Y, Zou X, Wojdyla JA, Kleanthous C, Wodak SJ (2014) Blind prediction of interfacial water positions in CAPRI. Proteins 82:620–632. doi:10.1002/prot.24439

Lensink MF, Velankar S, Kryshtafovych A, Huang SY, Schneidman-Duhovny D, Sali A, Segura J, Fernandez-Fuentes N, Viswanath S, Elber R, Grudinin S, Popov P, Neveu E, Lee H, Baek M, Park S, Heo L, Rie Lee G, Seok C, Qin S, Zhou HX, Ritchie DW, Maigret B, Devignes MD, Ghoorah A, Torchala M, Chaleil RA, Bates PA, Ben-Zeev E, Eisenstein M, Negi SS, Weng Z, Vreven T, Pierce BG, Borrman TM, Yu J, Ochsenbein F, Guerois R, Vangone A, Rodrigues JP, van Zundert G, Nellen M, Xue L, Karaca E, Melquiond AS, Visscher K, Kastritis PL, Bonvin AM, Xu X, Qiu L, Yan C, Li J, Ma Z, Cheng J, Zou X, Shen Y, Peterson LX, Kim HR, Roy A, Han X, Esquivel-Rodriguez J, Kihara D, Yu X, Bruce NJ, Fuller JC, Wade RC, Anishchenko I, Kundrotas PJ, Vakser IA, Imai K, Yamada K, Oda T, Nakamura T, Tomii K, Pallara C, Romero-Durana M, Jimenez-Garcia B, Moal IH, Fernandez-Recio J, Joung JY, Kim JY, Joo K, Lee J, Kozakov D, Vajda S, Mottarella S, Hall DR, Beglov D, Mamonov A, Xia B, Bohnuud T, Del Carpio CA, Ichiishi E, Marze N, Kuroda D, Roy Burman SS, Gray JJ, Chermak E, Cavallo L, Oliva R (2016) Prediction of homo- and hetero-protein complexes by protein docking and template-based modeling: a CASP-CAPRI experiment. Proteins. doi:10.1002/prot.25007

Lensink MF, Wodak SJ (2013) Docking, scoring, and affinity prediction in CAPRI. Proteins 81:2082–2095. doi:10.1002/prot.24428

Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084. doi:10.1002/prot.22818

London N, Raveh B, Schueler-Furman O (2013) Peptide docking and structure-based characterization of peptide binding: from knowledge to know-how. Curr Opin Struct Biol 23:894–902. doi:10.1016/j.sbi.2013.07.006

London N, Schueler-Furman O (2008) FunHunt: model selection based on energy landscape characteristics. Biochem Soc Trans 36:1418–1421. doi:10.1042/BST0361418

Lu H, Lu L, Skolnick J (2003) Development of unified statistical potentials describing protein-protein interactions. Biophys J 84:1895–1901. doi:10.1016/S0006-3495(03)74997-2

Luo J, Guo Y, Zhong Y, Ma D, Li W, Li M (2014) A functional feature analysis on diverse protein-protein interactions: application for the prediction of binding affinity. J Comput Aided Mol Des 28:619–629. doi:10.1007/s10822-014-9746-y

Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36:W233–W238. doi:10.1093/nar/gkn216

Ma XH, Wang CX, Li CH, Chen WZ (2002) A fast empirical approach to binding free energy calculations based on protein interface information. Protein Eng 15:677–681. doi:10.1093/protein/15.8.677

Macindoe G, Mavridis L, Venkatraman V, Devignes M-D, Ritchie DW (2010) HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 38:W445–W449. doi:10.1093/nar/gkq311

Marillet S, Boudinot P, Cazals F (2015) High resolution crystal structures leverage protein binding affinity predictions

Martin J, Lavery R (2012) Arbitrary protein-protein docking targets biologically relevant interfaces. BMC Biophys 5:7. doi:10.1186/2046-1682-5-7

Mashiach E, Nussinov R, Wolfson HJ (2010a) FiberDock: a web server for flexible induced-fit backbone refinement in molecular docking. Nucleic Acids Res 38:W457–W461. doi:10.1093/nar/gkq373

Mashiach E, Schneidman-Duhovny D, Andrusier N, Nussinov R, Wolfson HJ (2008) FireDock: a web server for fast interaction refinement in molecular docking. Nucleic Acids Res 36:W229–W232. doi:10.1093/nar/gkn186

Mashiach E, Schneidman-Duhovny D, Peri A, Shavit Y, Nussinov R, Wolfson HJ (2010b) An integrated suite of fast docking algorithms. Proteins 78:3197–3204. doi:10.1002/prot.22790

Metz A, Ciglia E, Gohlke H (2012) Modulating protein-protein interactions: from structural determinants of binding to druggability prediction to application. Curr Pharm Des 18:4630–4647. doi:10.2174/138161212802651553

Moal IH, Bates PA (2010) SwarmDock and the use of normal modes in protein-protein docking. Int J Mol Sci 11:3623–3648. doi:10.3390/ijms11103623

Moal IH, Bates PA (2012) Kinetic rate constant prediction supports the conformational selection mechanism of protein binding. PLoS Comput Biol 8:e1002351. doi:10.1371/journal.pcbi.1002351

Moal IH, Jiménez-García B, Fernández-Recio J (2015) CCharPPI web server: computational characterization of protein-protein interactions from structure. Bioinformatics 31:123–125. doi:10.1093/bioinformatics/btu594

Moal IH, Moretti R, Baker D, Fernández-Recio J (2013a) Scoring functions for protein–protein interactions. Curr Opin Struct Biol 23:862–867. doi:10.1016/j.sbi.2013.06.017

Moal IH, Torchala M, Bates PA, Fernández-Recio J (2013b) The scoring of poses in protein-protein docking: current capabilities and future directions. BMC Bioinformatics 14:286. doi:10.1186/1471-2105-14-286

Moont G, Gabb HA, Sternberg MJ (1999) Use of pair potentials across protein interfaces in screening predicted docked complexes. Proteins 35:364–373. doi:10.1002/(SICI)1097-0134(19990515)35:3<364:AID-PROT11>3.0.CO;2-4

Moreira IS, Fernandes PA, Ramos MJ (2010) Protein-protein docking dealing with the unknown. J Comput Chem 31:317–342. doi:10.1002/jcc.21276

Moreira IS, Martins JM, Coimbra JTS, Ramos MJ, Fernandes PA (2015) A new scoring function for protein-protein docking that identifies native structures with unprecedented accuracy. Phys Chem Chem Phys 17:2378–2387. doi:10.1039/c4cp04688a

Naider F, Anglister J (2009) Peptides in the treatment of AIDS. Curr Opin Struct Biol 19:473–482. doi:10.1016/j.sbi.2009.07.003

Negi SS, Schein CH, Oezguen N, Power TD, Braun W (2007) InterProSurf: a web server for predicting interacting sites on protein surfaces. Bioinformatics 23:3397–3399. doi:10.1093/bioinformatics/btm474

Neuvirth H, Raz R, Schreiber G (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. J Mol Biol 338:181–199. doi:10.1016/j.jmb.2004.02.040

Nisius B, Sha F, Gohlke H (2012) Structure-based computational analysis of protein binding sites for function and druggability prediction. J Biotechnol 159:123–134. doi:10.1016/j.jbiotec.2011.12.005

Northrup SH, Erickson HP (1992) Kinetics of protein-protein association explained by Brownian dynamics computer simulation. Proc Natl Acad Sci USA 89:3338–3342. doi:10.1073/pnas.89.8.3338

Ofran Y, Rost B (2003) Analysing six types of protein-protein interfaces. J Mol Biol 325:377–387. doi:10.1016/s0022-2836(02)01223-8

Oliva R, Vangone A, Cavallo L (2013) Ranking multiple docking solutions based on the conservation of inter-residue contacts. Proteins 81:1571–1584. doi:10.1002/prot.24314

Palma PN, Krippahl L, Wampler JE, Moura JJ (2000) BiGGER: a new (soft) docking algorithm for predicting protein interactions. Proteins 39:372–384. doi:10.1002/(SICI)1097-0134(20000601)39:4<372:AID-PROT100>3.0.CO;2-Q

Petsalaki E, Russell RB (2008) Peptide-mediated interactions in biological systems: new discoveries and applications. Curr Opin Biotechnol 19:344–350. doi:10.1016/j.copbio.2008.06.004

Petsalaki E, Stark A, García-Urdiales E, Russell RB (2009) Accurate prediction of peptide binding sites on protein surfaces. PLoS Comput Biol 5:e1000335. doi:10.1371/journal.pcbi.1000335

Phizicky EM, Fields S (1995) Protein-protein interactions: methods for detection and analysis. Microbiol Rev 59:94–123

Pierce B, Weng Z (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67:1078–1086. doi:10.1002/prot.21373

Pierce BG, Hourai Y, Weng Z (2011) Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS ONE 6:e24657. doi:10.1371/journal.pone.0024657

Porollo A, Meller J (2006) Prediction-based fingerprints of protein-protein interactions. Proteins 66:630–645. doi:10.1002/prot.21248

Qin S, Pang X, Zhou H-X (2011) Automated prediction of protein association rate constants. Structure 19:1744–1751. doi:10.1016/j.str.2011.10.015

Qin S, Zhou H-X (2013) Using the concept of transient complex for affinity predictions in CAPRI rounds 20-27 and beyond. Proteins 81:2229–2236. doi:10.1002/prot.24366

Qin S, Zhou H-X (2007) meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 23:3386–3387. doi:10.1093/bioinformatics/btm434

Raveh B, London N, Schueler-Furman O (2010) Sub-angstrom modeling of complexes between flexible peptides and globular proteins. Proteins 78:2029–2040. doi:10.1002/prot.22716

Raveh B, London N, Zimmerman L, Schueler-Furman O (2011) Rosetta FlexPepDock ab-initio: simultaneous folding, docking and refinement of peptides onto their receptors. PLoS ONE 6:e18934. doi:10.1371/journal.pone.0018934

Ritchie DW (2008) Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 9:1–15. doi:10.2174/138920308783565741

Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39:178–194. doi:10.1002/(SICI)1097-0134(20000501)39:2<178:AID-PROT8>3.0.CO;2-6

Rodrigues JPGLM, Bonvin AMJJ (2014) Integrative computational modeling of protein interactions. FEBS J 281:1988–2003. doi:10.1111/febs.12771

Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis P, Karaca E, Melquiond ASJ, Bonvin AMJJ (2012) Clustering biomolecular complexes by residue contacts similarity. Proteins 80:1810–1817. doi:10.1002/prot.24078

Rubinstein M, Niv MY (2009) Peptidic modulators of protein-protein interactions: progress and challenges in computational design. Biopolymers 91:505–513. doi:10.1002/bip.21164

Russell RB, Gibson TJ (2008) A careful disorderliness in the proteome: sites for interaction and targets for future therapies. FEBS Lett 582:1271–1275. doi:10.1016/j.febslet.2008.02.027

Ruvinsky AM, Vakser IA (2008) Interaction cutoff effect on ruggedness of protein-protein energy landscape. Proteins 70:1498–1505. doi:10.1002/prot.21644

Sacquin-Mora S, Carbone A, Lavery R (2008) Identification of protein interaction partners and protein-protein interaction sites. J Mol Biol 382:1276–1289. doi:10.1016/j.jmb.2008.08.002

Schlick T, Collepardo-Guevara R, Halvorsen LA, Jung S, Xiao X (2011) Biomolecularmodeling and simulation: a field coming of age. Q Rev Biophys 44:191–228. doi:10.1017/S0033583510000284

Schmitz C, Bonvin AMJJ (2011) Protein-protein HADDocking using exclusively pseudocontact shifts. J Biomol NMR 50:263–266. doi:10.1007/s10858-011-9514-4

Schmitz C, Melquiond ASJ, de Vries SJ, Karaca E, van Dijk M, Kastritis PL, Bonvin AMJJ (2012) Protein–protein docking with HADDOCK. Towards Mech Syst Biol 520–535. doi:10.1002/9783527644506.ch32

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) PatchDock and SymmDock: servers for rigid and symmetric docking. Nucleic Acids Res 33:W363–W367. doi:10.1093/nar/gki481

Schreiber G, Fersht AR (1996) Rapid, electrostatically assisted association of proteins. Nat Struct Biol 3:427–431. doi:10.1038/nsb0596-427

Smith GR, Fitzjohn PW, Page CS, Bates PA (2005a) Incorporation of flexibility into rigid-body docking: applications in rounds 3–5 of CAPRI. Proteins 60:263–268. doi:10.1002/prot.20568
Smith GR, Sternberg MJE (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12:28–35. doi:10.1016/s0959-440x(02)00285-3
Smith GR, Sternberg MJE, Bates PA (2005b) The relationship between the flexibility of proteins and their conformational states on forming protein-protein complexes with an application to protein-protein docking. J Mol Biol 347:1077–1101. doi:10.1016/j.jmb.2005.01.058
Sousa SF, Fernandes PA, Ramos MJ (2006) Protein-ligand docking: current status and future challenges. Proteins 65:15–26. doi:10.1002/prot.21082
Stites WE (1997) Proteinminus signProtein interactions: interface structure, binding thermodynamics, and mutational analysis. Chem Rev 97:1233–1250. doi:10.1021/cr960387h
Su Y, Zhou A, Xia X, Li W, Sun Z (2009) Quantitative prediction of protein-protein binding affinity with a potential of mean force considering volume correction. Protein Sci 18:2550–2558. doi:10.1002/pro.257
Sugiki T, Fujiwara T, Kojima C (2014) Latest approaches for efficient protein production in drug discovery. Expert Opin Drug Discov 9:1189–1204. doi:10.1517/17460441.2014.941801
Szymkowski DE (2005) Creating the next generation of protein therapeutics through rational drug design. Curr Opin Drug Discov Devel 8:590–600
Tian F, Lv Y, Yang L (2012) Structure-based prediction of protein-protein binding affinity with consideration of allosteric effect. Amino Acids 43:531–543. doi:10.1007/s00726-011-1101-1
Torchala M, Moal IH, Chaleil RAG, Agius R, Bates PA (2013) A Markov-chain model description of binding funnels to enhance the ranking of docked solutions. Proteins 81:2143–2149. doi:10.1002/prot.24369
Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310–W314. doi:10.1093/nar/gkl206
Trabuco LG, Lise S, Petsalaki E, Russell RB (2012) PepSite: prediction of peptide-binding sites from protein surfaces. Nucleic Acids Res 40:W423–W427. doi:10.1093/nar/gks398
Trellet M, Melquiond ASJ, Bonvin AMJJ (2013) A unified conformational selection and induced fit approach to protein-peptide docking. PLoS ONE 8:e58769. doi:10.1371/journal.pone.0058769
Trellet M, Melquiond ASJ, Bonvin AMJJ (2015) Information-driven modeling of protein-peptide complexes. Methods Mol Biol 1268:221–239. doi:10.1007/978-1-4939-2285-7_10
Tress M, de Juan D, Graña O, Gómez MJ, Gómez-Puertas P, González JM, López G, Valencia A (2005) Scoring docking models with evolutionary information. Proteins 60:275–280. doi:10.1002/prot.20570
Tsai CJ, Kumar S, Ma B, Nussinov R (1999) Folding funnels, binding funnels, and protein function. Protein Sci 8:1181–1190. doi:10.1110/ps.8.6.1181
Vaara M (2009) New approaches in peptide antibiotics. Curr Opin Pharmacol 9:571–576. doi:10.1016/j.coph.2009.08.002
Vajda S, Kozakov D (2009) Convergence and combination of methods in protein-protein docking. Curr Opin Struct Biol 19:164–170. doi:10.1016/j.sbi.2009.02.008
van Dijk ADJ, Fushman D, Bonvin AMJJ (2005) Various strategies of using residual dipolar couplings in NMR-driven protein docking: application to Lys48-linked di-ubiquitin and validation against 15N-relaxation data. Proteins 60:367–381. doi:10.1002/prot.20476
van Dijk ADJ, Kaptein R, Boelens R, Bonvin AMJJ (2006) Combining NMR relaxation with chemical shift perturbation data to drive protein-protein docking. J Biomol NMR 34:237–244. doi:10.1007/s10858-006-0024-8
van Zundert GCP, Melquiond ASJ, Bonvin AMJJ (2015) Integrative modeling of biomolecular complexes: HADDOCKing with cryo-electron microscopy data. Structure 23:949–960. doi:10.1016/j.str.2015.03.014
Vangone A, Bonvin AM (2015) Contacts-based prediction of binding affinity in protein-protein complexes. Elife 4:e07454. doi:10.7554/eLife.07454
Vangone A, Cavallo L, Oliva R (2013) Using a consensus approach based on the conservation of inter-residue contacts to rank CAPRI models. Proteins 81:2210–2220. doi:10.1002/prot.24423

Vangone A, Oliva R, Cavallo L (2012) CONS-COCOMAPS: a novel tool to measure and visualize the conservation of inter-residue contacts in multiple docking solutions. BMC Bioinformatics 13(Suppl 4):S19. doi:10.1186/1471-2105-13-S4-S19

Vangone A, Spinelli R, Scarano V, Cavallo L, Oliva R (2011) COCOMAPS: a web application to analyze and visualize contacts at the interface of biomolecular complexes. Bioinformatics 27:2915–2916. doi:10.1093/bioinformatics/btr484

Venkatraman V, Yang YD, Sael L, Kihara D (2009) Protein-protein docking using region-based 3D Zernike descriptors. BMC Bioinformatics 10:407. doi:10.1186/1471-2105-10-407

Verschueren E, Vanhee P, Rousseau F, Schymkowitz J, Serrano L (2013) Protein-peptide complex prediction through fragment interaction patterns. Structure 21:789–797. doi:10.1016/j.str.2013.02.023

Viswanath S, Ravikant DVS, Elber R (2013) Improving ranking of models for protein complexes with side chain modeling and atomic potentials. Proteins 81:592–606. doi:10.1002/prot.24214

Vreven T, Hwang H, Weng Z (2011) Integrating atom-based and residue-based scoring functions for protein-protein docking. Protein Sci 20:1576–1586. doi:10.1002/pro.687

Vreven T, Pierce BG, Hwang H, Weng Z (2013) Performance of ZDOCK in CAPRI rounds 20-26. Proteins 81:2175–2182. doi:10.1002/prot.24432

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, Chaleil R, Jimenez-Garcia B, Bates PA, Fernandez-Recio J, Bonvin AM, Weng Z (2015) Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. J Mol Biol 427:3031–3041. doi:10.1016/j.jmb.2015.07.016

Wass MN, David A, Sternberg MJE (2011a) Challenges for the prediction of macromolecular interactions. Curr Opin Struct Biol 21:382–390. doi:10.1016/j.sbi.2011.03.013

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A (2011b) valencia2011. Mol Syst Biol 7:1–8. doi:10.1038/msb.2011.3

Weikl TR, von Deuster C (2009) Selected-fit versus induced-fit protein binding: kinetic differences and mutational analysis. Proteins 75:104–110. doi:10.1002/prot.22223

Wodak SJ, Janin J (1978) Computer analysis of protein-protein interaction. J Mol Biol 124:323–342. doi:10.1016/0022-2836(78)90302-9

Xue LC, Rodrigues JP, Kastritis PL, Bonvin AM, Vangone A (2016) PRODIGY: a web server for predicting the binding affinity of protein-protein complexes. Bioinformatics. doi:10.1093/bioinformatics/btw514

Xue LC, Jordan RA, El-Manzalawy Y, Dobbs D, Honavar V (2014) DockRank: ranking docked conformations using partner-specific sequence homology-based protein interface prediction. Proteins 82:250–267. doi:10.1002/prot.24370

Yang S (2014) Methods for SAXS-based structure determination of biomolecular complexes. Adv Mater Weinheim 26:7902–7910. doi:10.1002/adma.201304475

Zacharias M (2005) ATTRACT: protein-protein docking in CAPRI using a reduced protein model. Proteins 60:252–256. doi:10.1002/prot.20566

Zellner H, Staudigel M, Trenner T, Bittkowski M, Wolowski V, Icking C, Merkl R (2012) PresCont: predicting protein-protein interfaces utilizing four residue properties. Proteins 80:154–168. doi:10.1002/prot.23172

Zhang C, Liu S, Zhu Q, Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 48:2325–2335. doi:10.1021/jm049314d

Zhang QC, Deng L, Fisher M, Guan J, Honig B, Petrey D (2011) PredUs: a web server for predicting protein interfaces using structural neighbors. Nucleic Acids Res 39:W283–W287. doi:10.1093/nar/gkr311

Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490:556–560. doi:10.1038/nature11503

Zhou P, Wang C, Ren Y, Yang C, Tian F (2013) Computational peptidology: a new and promising approach to therapeutic peptide design. Curr Med Chem 20:1985–1996. doi:10.2174/09298673113201500005

# Part II
# From Structures to Functions

# Chapter 9
# Function Diversity Within Folds and Superfamilies

**Benoit H. Dessailly, Natalie L. Dawson, Sayoni Das and Christine A. Orengo**

**Abstract** The structural genomics initiatives significantly increased the numbers of three-dimensional structures available for proteins of unknown function. However, the extent to which structural information helps understanding function is still a matter of debate. Here, the value of detecting structural relationships at different levels (typically, fold and superfamily) for transferring functional annotations between proteins is reviewed. First, function diversity of proteins sharing the same fold is investigated, and it is shown that although the identification of a fold can in some cases provide clues on functional properties, the diversity of functions within a fold can be such that this information is very limited for some particularly diverse folds (e.g. super-folds). Next, since structural data can help detecting homology in the absence of sequence similarity, function diversity between proteins from the same superfamily (homologous proteins) is analysed. The evolutionary causes and the mechanisms that have generated the observed functional diversity between related proteins are discussed, and helpful tools for the correlated analysis of structure, function and evolution are reviewed.

**Keywords** Protein structure · Protein function · Function annotation · Function prediction · Protein function diversity · Protein evolution · Protein function evolution · Functional sites · Protein folds · Superfolds · Protein superfamilies · Structural genomics initiatives · Homology

B.H. Dessailly · N.L. Dawson · S. Das · C.A. Orengo (✉)
Department of Structural and Molecular Biology, University College London, London WC1E 6BT, UK
e-mail: c.orengo@ucl.ac.uk



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_9

## 9.1 Defining Function

Before discussing how the detection of fold or superfamily relationships can help determining the function of a protein, it is necessary to define clearly the meaning of the term *function* in this chapter and, in particular, to delineate the aspects of function that can be inferred best using structural information.

Function is a relatively vague concept that covers many different aspects of the activity of a protein. Furthermore, the aspects covered by that single word vary with the different fields of protein science. For example, a physiologist may describe the function of a protein in terms of its impact on the global phenotype (e.g. "inducer of cell death"), whereas a biochemist would generally define the function of the protein he studies on the basis of its molecular interactions or catalytic activity (e.g. "Receptor-interacting serine/threonine-protein kinase"). Because of these different usages of the word, it is very difficult to provide a universal and widely accepted definition of function.

However, it is not essential to come up with such a definition. The Gene Ontology (GO) consortium have proposed a framework with which they have been able to define or, most importantly, categorise the functions of proteins in a widely accepted way (Ashburner et al. 2000). In GO, three different aspects of function are considered and defined separately. According to GO, the cellular component describes the biological structures to which the protein belongs (e.g. nucleus or ribosome); the biological process corresponds to the processes or pathways in which the protein is involved (e.g. metabolism, signal transduction or cell differentiation); the molecular function of a protein is the ensemble of activities it can undertake (e.g. binding, catalysis or transport).

Three-dimensional structures of proteins mostly shed light on catalytic mechanisms and potential interactions with other molecules, both aspects which are covered by the *molecular function* category. Consequently, it is essentially molecular function that is considered when dealing with structure-function relationship as is the case in this chapter.

Several databases and annotation systems are available for the description of the molecular function of proteins, and are very helpful for studying structure-function relationships, notably on an automated basis. Probably the oldest system for describing the molecular function of proteins is the Enzyme Commission numbering scheme (EC) in which enzymatic reactions are hierarchically classified using a four-digit system, where each level describes increasingly detailed aspects of the reaction, from the general type of catalytic activity (oxidoreductase, hydrolase, etc.) to the specific molecule that acts as substrate of the reaction (Nomenclature Committee of the IUBMB 1992). In order to address long-standing limitations of the EC classification, two new databases have recently been set up to classify enzymes and their reactions: EzCatDB ("Enzyme catalytic-mechanism Database") (Nagano 2005) and MACiE ("Mechanism, Annotation and Classification in Enzymes") (Holliday et al. 2011). Both of these databases focus on the description and classification of enzymatic reaction mechanisms rather than the reactions themselves, since it has been argued

that a reaction-based classification like the EC system is not necessarily appropriate as a classification of the corresponding enzymes (O'Boyle et al. 2007). Complementary to these databases, the Catalytic Site Atlas provides detailed information on the specific amino acid residues that directly participate in catalytic mechanisms, for enzymes of known structure (Porter et al. 2004; Furnham et al. 2014). Several databases provide further description of all protein residues involved in binding biologically important molecules such as substrates and cofactors (Dessailly et al. 2008; Lopez et al. 2011). Other widely-used annotation systems for protein function include KEGG, which was initially aimed at describing metabolic pathways and biological reaction networks, and has now extended into a more widely-scoped classification system of biological functions (Kanehisa et al. 2014); FUNCAT (the Functional Catalogue), which classifies protein functions into a unique hierarchical tree (Ruepp et al. 2004); Reactome, which focuses on characterising human biological metabolic pathways (Croft et al. 2014); and MetaCyc, a database of primary and secondary metabolic pathways from all kingdoms of life (Caspi et al. 2014). KEGG and FUNCAT have traditionally been more focused on the biological processes in which proteins are involved rather than their molecular activities, but both of these databases can nevertheless provide very useful clues regarding what is referred to as *molecular function* in GO.

## 9.2 From Fold to Function

### *9.2.1 Definition of a Fold*

#### 9.2.1.1 General Understanding

The fold adopted by a protein is generally understood as the global arrangement of its main elements of secondary structures, both in terms of their relative orientations and of their topological connections. A major difficulty directly arises from this general statement since there are no objective rules to decide which are the *main elements of secondary structure* to be considered for defining the fold (Grishin 2001).

One objective of this chapter is to describe how knowledge of relationships between proteins, such as sharing the same fold, helps in transferring functional annotations from well-characterised proteins to proteins of unknown function. As will be discussed further in Sect. 9.3, the main assumption made in the process of transferring annotations between proteins is that evolutionarily related (i.e. homologous) proteins generally tend to share functional properties. But proteins adopting the same fold are not necessarily homologous. It has been argued that proteins can attain a given fold independently by convergent evolution, because only a limited number of folds are physically acceptable (Russell et al. 1997). For example, it is not clear whether all superfamilies of proteins that adopt the TIM-like $(\beta/\alpha)_8$ barrel fold are evolutionarily related, as definitive evidence in that sense has not been found (Nagano et al. 2002).

#### 9.2.1.2 Practical Definitions

Several databases have been set-up to classify protein structures into a comprehensive framework of structural relationships (Table 9.1). The practical definition of a fold used in the most-widely used of these databases is given below. As will emerge from the following definitions, the concept of fold is generally applied to domains rather than full-length proteins, but the definition of a domain can vary between databases.

**CATH**—The CATH database is a hierarchical classification of protein domain structures (Orengo et al. 1997; Sillitoe et al. 2015). The highest level of classification assigns protein domains to 3 different *classes* based on their global content in secondary structures. Within CATH classes, protein domains are classified into different *architectures* that describe the orientation of secondary structures without considering their connectivity. Domains in a given architecture are further sub-classified into different *topologies*, depending on how secondary structures are connected to one another. It is this *topology* level that fits most closely to the general notion of a fold described above. In practise, assignment of domains to the topologies in CATH is performed automatically with the structural alignment program SSAP (Orengo and Taylor 1996) and empirically derived cut-offs.

**SCOP**—Like CATH, the Structural Classification of Proteins (SCOP) is a hierarchical classification of protein domain structures (Murzin et al. 1995; Pethica et al. 2012), but the levels of classification differ between the two databases. As in CATH, the highest level of classification in SCOP is the structural *class*, but SCOP defines four different classes whereas CATH defines three. The next level of classification is the *fold*; two protein domains are assigned to the same fold if they share the same major elements of secondary structure arranged in a similar orientation, and with the same topological connections. This definition corresponds well to the definition of the topology level in CATH but, in practise, assignments of individual domains can differ between the two databases because of the degree of subjectivity in each definition (i.e. which secondary structure elements are to be considered *major*), and of the protocols used to assign the domains (automated in CATH, mostly manual in SCOP).

**SCOP2**—With the large increase in structural data deposited in the PDB, more remote evolutionary relationships have been detected. These have in turn have revealed complex relationships between domain structures in some fold groups and homologous superfamilies, which have led to the production of a SCOP2 prototype. This database still organises protein domains using structural and evolutionary relationships but instead uses them to form a network rather than a hierarchy (Andreeva et al. 2014, 2015).

**FSSP**—A purely objective definition of folds has been offered by FSSP (families of structurally similar proteins) (Holm and Sander 1996a). In FSSP, pair-wise structural alignments were performed for a set of representative and non-redundant PDB structures using the structural alignment program DALI (Holm and Sander 1993). Hierarchical clustering was applied using the scores obtained from these pair-wise structural alignments thus generating a so-called fold tree, from which *fold families* were automatically defined by cutting the tree at different levels of similarity.

**ECOD**—The Evolutionary Classification of Protein Domains, or ECOD (Cheng et al. 2014), is a new resource by the Grishin group that currently consists of proteins with experimentally determined structures. Similar to CATH and SCOP, protein domains are hierarchically classified using evolutionary relationship information, however this database focuses on finding remote homologues and it is also updated every week.

There are 5 levels to the ECOD hierarchy. The top level is the 'Architecture', which is comparable to the A-level in CATH as it represents domains grouped according to the secondary structure composition and their arrangement in 3D space. The second, 'X-group', level does not have an equivalent in either CATH or SCOP. It represents groups of domains that are thought to be homologues but do not yet have enough supporting evidence apart from structural similarity. The H-group, is comparable to the homologous superfamily level in CATH and SCOP and domains are classified to these groups using sequence, structure, and function information. The next level down is the T-group, which is made up of groups having similar topological connections, such as in CATH's T-level. Finally, there are the family F-groups. These groups of domains have significant sequence similarity and are largely made up of mapped Pfam families (Finn et al. 2014) and HHsearch-based clusters (Söding 2005).

**SCOPe**—SCOP and SCOP2 are not up-to-date with the latest version of the PDB. To overcome this issue, an extended version of SCOP, SCOPe, has been introduced by the Chandonia group (Fox et al. 2014). SCOPe uses a combination of automatic and manual curation methods to classify more recent PDB structures and also corrects some errors in SCOP. The ASTRAL database is also incorporated and updated. A sequence-based approach is used to classify recently deposited PDB protein chains into SCOPe using the structural classification hierarchy in SCOP. New protein chains are searched against SCOP using BLAST to look for previously classified domains that are significantly similar in sequence and the aligned match must also have high coverage to the query sequence (Fox et al. 2014).

#### 9.2.1.3 Paradigm Shift

Structure is generally better conserved than sequence in evolution, and many proteins display common structural characteristics. As more and more three-dimensional protein structures were being solved in the mid-nineties, structural classification systems became necessary in order to make some sense out of the increasing amount of data. This lead to the development of the above-mentioned hierarchical classifications of protein structures. The realisation that global structural motifs, such as the $(\beta/\alpha)_8$ barrels or the 4-helix bundles, were observed in proteins that were unrelated in sequence, lead to the notion of fold that we have just described. Until recently, folds have been understood as recurrent global structural motifs that incidentally act as practical divisions of the protein structure space. Implicit in that view is the idea that fold space is discrete, in the sense that (a) each protein has a unique fold, which it will share with other related proteins, and which

will distinguish it from most other unrelated proteins (though accounting for the existence of analogous folds, see Sect. 9.2.2.1); and (b) that each fold is structurally different and constitutes an isolated and non-overlapping structural group from the others (Kolodny et al. 2006).

But as more and more structural data becomes available, notably via structural genomics initiatives, the perception of the fold is changing in favour of a view of fold space that is continuous rather than discrete (Harrison et al. 2002). It is now becoming widely recognised that homologous proteins can actually adopt different folds (Grishin 2001; Kolodny et al. 2006), and that some proteins can adopt multiple, changeable folding motifs depending on time and conditions (Andreeva and Murzin 2006). This has consequences on the usability of the fold for function prediction; the main argument for using fold similarities when inferring function is that proteins sharing the same fold may often display remote homologies that would not be detectable otherwise, and that homologous proteins should in turn tend to perform related functions (Moult and Melamud 2000). It necessarily follows from the finding that the relationship between fold and homology is not clear, that the relationship between fold and function is likely to be fuzzy as well. However, recent results obtained using the ensemble of currently available structural data in CATH suggest that the majority of folds are structurally coherent and significantly distinct from other folds (Cuff et al. 2009); and indeed, as will be shown presently, fold similarities can provide some clues on function similarities between proteins (Martin et al. 1998).

## *9.2.2 Prediction of Function Using Fold Relationships*

This section focuses on functional properties that can be inferred using features that do not imply homology, i.e. functional properties that tend to arise by convergent evolution; issues regarding functional inference based on homology relationships are addressed in Sect. 9.3 of this chapter.

In general, the determination of a protein structure and its fold will allow a researcher to run a plethora of structure-based function prediction methods that would not be available if the structure was not known. Some of these methods rely on the principle that knowing the structure allows one to detect global homologies that are not apparent at the sequence level (Lee et al. 2007). But other approaches are only making use of purely structural properties that are expected to be relevant for a protein to perform its molecular function, with no evolutionary consideration. Many of these methods are covered by several other chapters in this book (see Chaps. 10, 11, 13 and 14). Here, only situations that directly relate to knowledge of the *fold* are discussed.

### 9.2.2.1 Folds with a Single Function

A newly solved protein structure can be used to search for fold similarities with previously known structures, via structure comparison programs that generally

assess the significance of detected structural similarities using specific scoring schemes. Several of these programs are publicly available (Table 9.1) and have been recently benchmarked using a large dataset of known structure similarities built from CATH (Kolodny et al. 2005; Redfern et al. 2007). Such programs include DALI (Holm and Sander 1996b), FATCAT (Ye and Godzik 2003), SSM (Krissinel and Henrick 2004), CE (Shindyalov and Bourne 1998) and CATHEDRAL (Redfern et al. 2007). If the new structure is from a protein of unknown function, the next step if fold similarity has been detected is to evaluate whether functional annotations can be transferred from structurally similar proteins.

Some folds are adopted only by homologous proteins whereas other folds may have arisen partly by convergent evolution. These folds are coined homologous and analogous folds, respectively (Moult and Melamud 2000). Similarly, some folds appear homogeneous in terms of functions whereas others are adopted by proteins with widely divergent functions. For example, about $\sim 10\%$ of the folds in the current version of CATH (v4.0) have 100 or more functions associated with them. It is generally assumed that homologous folds are more functionally homogeneous than analogous folds (Moult and Melamud 2000). Obviously, if a fold is associated to a unique function *X*, the recognition of that fold in a protein of unknown function would directly allow to annotate that protein with function *X*. But in practise, the situation is more complex because a functionally diverse fold can misleadingly appear to be related to only one function due to sampling bias.

In any case, there are documented cases where fold identification has helped predicting the function of a protein (Moult and Melamud 2000). For example, the three-dimensional structure of the ycaC gene product from *Escherichia Coli* revealed a fold similar to that adopted by a family of amidohydrolases, and further investigation indicated that this protein had a similar catalytic apparatus as other proteins sharing that fold (Colovos et al. 1998; Moult and Melamud 2000). Increasing numbers of successful examples of function prediction via fold identification are being documented in the context of structural genomics that globally aim at solving large numbers of protein structures (Adams et al. 2007). In most cases, however, successful function prediction does not result from fold identification only, but rather from a combination of fold relationship with other evidence such as sequence motif recognition or functional site similarities.

#### 9.2.2.2 Supersites

Generally, three-dimensional structures are very helpful for identifying protein functional sites, i.e., the subsets of residues that are crucial for the molecular function of the protein. Functional sites mostly consist of binding sites (sets of protein residues that interact with ligands) (Dessailly et al. 2008) or catalytic sites (sets of residues that directly participate in an enzymatic reaction) (Porter et al. 2004).

One reason why structures are useful for detecting functional site(s) is that the latter tend to occupy well-conserved topological locations in the structure.

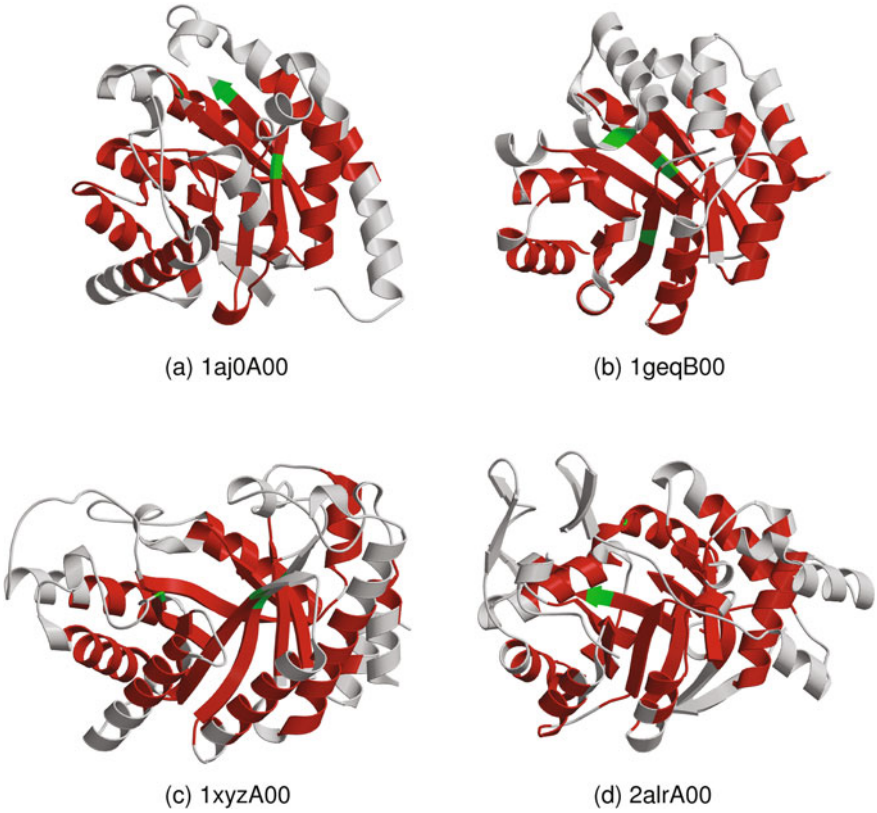



**Fig. 9.1** Supersites in the $(\beta/\alpha)_8$ TIM-like barrel fold. Cartoon illustrations of 4 proteins adopting the $(\beta/\alpha)_8$ barrel fold, which have been classified in different CATH (and SCOP) superfamilies: (**a**) *E. coli* Dihydropteroate Synthase (CATH domain ID: 1aj0A00), (**b**) *P. furiosus* Tryptophan Synthase alpha-subunit (CATH domain ID: 1geqB00), (**c**) *C. thermocellum* Endo-1,4-beta-xylanase Z (CATH domain ID: 1xyzA00), and (**d**) *H. sapiens* Aldehyde Reductase (CATH domain ID: 2alrA00). The four structures have been superposed using CORA (Orengo 1999). They are shown into a similar orientation, and common elements between the four structures are coloured in red. The positions of the catalytic residues in these 4 proteins (as defined in the Catalytic Site Atlas) are coloured green. In spite of major structural differences and the absence of evidence for homology between these proteins, the catalytic sites always locate around the C-terminal end of the core β-strands. Figures of three-dimensional structures were drawn using Molscript (Kraulis 1991) and rendered using Raster3D (Merritt and Bacon 1997)

Furthermore, even when no definitive evidence supports homology between proteins that share a given fold, functional sites still tend to locate in similar regions of the three-dimensional structures. Such functional sites are called supersites and have been shown to occur in a large number of *analogous folds* (or *superfolds*, see Sect. 9.2.3.1), that is folds shared by non-homologous proteins (Russell et al. 1998). Figure 9.1 describes a very well-known example of supersite: the catalytic site of proteins adopting the $(\beta/\alpha)_8$ barrel fold, in which the catalytic residues

invariably occur at the C-terminal ends of the β-strands in the central parallel β-sheet, although the particular β-strands to which they belong may vary (Nagano et al. 2002).

#### 9.2.2.3 Superfolds

Folds that are adopted by proteins from many different superfamilies, and that generally display remarkable functional diversity, have been called "superfolds" (Orengo et al. 1994). Striking examples of such superfolds comprising proteins with many different functions include the TIM-like $(\beta/\alpha)_8$ barrel fold which is adopted by proteins from more than 29 diverse superfamilies (Nagano et al. 2002); and the Rossmann-fold, which is adopted by proteins from 130 CATH superfamilies (CATH v4.0), several of which are functionally diverse. Even though they represent a very small fraction of known folds, these superfolds seem to account for a disproportionate fraction of proteins in known genomes (Lee et al. 2005). Superfolds also cause a major problem for function prediction using fold recognition since proteins sharing such a fold do not necessarily share the same function. The existence of such folds and their considerable coverage of the protein world has prompted caution regarding the usefulness of detecting fold relationships for function prediction.

## 9.3 Function Diversity Between Homologous Proteins

In general, detecting homology (superfamily relationship) is much more helpful for function prediction than structural similarity alone (fold relationship). In this section, the relation between function diversity and structural homology is examined and it is shown that even when homology is identified, many obstacles remain when attempting to transfer functional annotations from one protein to another.

### *9.3.1 Definitions*

Before explaining how function diverges within superfamilies, it is necessary to define clearly what a *superfamily* is, and how it is used in practise. The term *family*, which is used throughout this section, is also introduced.

#### 9.3.1.1 General Understanding

A *superfamily* is an ensemble of proteins that are thought to be evolutionarily related. Superfamily relationships can be determined by sequence similarities,

which are detected using either traditional sequence alignment methods or more sensitive HMM searches (Reid et al. 2007). In the absence of sequence similarity, remote homologies can also be uncovered from structure and/or function similarities. But contrary to the situation with sequence similarity, there is no widely accepted approach to assess whether a structural or functional similarity is statistically significant. Because of that, the cut-offs used to define superfamily relationships can be arbitrary and somewhat subjective. Today, several databases such as CATH and SCOP have come up with standard and widely-accepted definitions of what superfamilies are (see Sect. 9.3.1.2). But in all of these, some degree of subjectivity in the assignment of proteins to superfamilies remain, as hinted by the facts that they still rely on manual validation for this specific process, and that incompatible assignments are made in the different databases for a number of domains (Greene et al. 2007; Andreeva et al. 2007). It is worth noting that both CATH and SCOP now pre-classify new protein structures using automated protocols, but final assignment to superfamilies still ultimately involves manual processing.

The concept of a *family* is vaguer. Nowadays, a family is commonly understood as a sub-classification of homologous proteins according to some criteria. For example, a sequence family at a particular level of sequence similarity groups together all proteins that share at least that level of sequence similarity; a functional family groups together homologues that have the same function; an orthologous family groups together orthologues; etc. Depending on the focus of the databases, the definition of a *family* will vary.

#### 9.3.1.2 Practical Definitions

Only databases that consider structural data are described here.

**CATH and Gene3D**—In the CATH classification, domains in a given *topology* (see Sect. 9.2.1.2) are further classified in the same Homologous superfamily (*H-level*) if they are believed to have a common ancestor. Two domains are considered homologous if they satisfy at least two of the following criteria: (a) structural similarity, assessed using empirically-derived cut-offs; (b) sequence similarity, assessed using standard sequence comparison methods and HMM sequence searches; and (c) functional similarity, identified using manual analysis. Gene3D expands this classification to proteins of unknown structure, by scanning sequences against a library of CATH profile-HMM's, thus matching parts of these sequences to CATH homologous superfamilies (Yeats et al. 2008; Lees et al. 2014).

CATH superfamilies are further divided into functional families or FunFams that groups together sequence homologues that share the same function or sub-function within a superfamily (Sillitoe et al. 2015; Das et al. 2015). This is done by hierarchical agglomerative clustering of all sequence homologues of each CATH superfamily using the GeMMA algorithm (Lee et al. 2010) to generate a clustering tree, followed by an optimal partitioning of the tree using the FunFHMMer algorithm (Das et al. 2015) which exploits sequence patterns. GeMMA first clusters the

sequence homologues at 90% sequence identity into S90 clusters using CD-HIT (Fu et al. 2012), and builds multiple sequence alignments for each cluster using MAFFT (Katoh and Standley 2013). It then exploits COMPASS (Sadreyev and Grishin 2003) to compare sequence profiles derived from the sequence alignments of pairs of clusters present at each iteration of the clustering. After each iteration, the cluster profiles matching above a threshold are merged and alignments are generated for the new clusters. These iterations continue till a single cluster remains generating a bottom-up hierarchical clustering tree built from the leaf nodes to the root. FunFHMMer identifies highly conserved positions and specificity-determining positions in sequence alignments to distinguish between families that perform different functions and ensure functional coherence in the identified families. Residues in multiple sequence alignments that are highly conserved among all the sequence relatives are generally important for stability, folding or carrying out a common function of the domain whereas specificity-determining positions i.e. residues that are differentially conserved in groups of sequences sharing a function or sub-function in a multiple sequence alignment are generally implicated in functional divergence (Rausell et al. 2010). The functional purity of the CATH FunFams has been demonstrated by their utility to transfer functional annotations for query sequences. This was validated by the international function prediction experiment, CAFA (Critical Assessment of Function Annotation experiment) (Radivojac et al. 2013) where FunFHMMer was ranked among the top 5 function prediction methods. A comprehensive summary of the relationships between FunFams in a superfamily can be visualised using CATH superfamily networks (Fig. 9.2) where FunFams are represented by nodes and the edge distances correspond to the sequence similarity between the FunFams. CATH superfamilies have also been divided into coarser sequence families that are defined at different cut-offs of sequence identity. A cut-off of 35% sequence identity is used to define non-redundant groups of proteins (s35 families).

**SCOP and Superfamily**—For SCOP superfamilies, homologies are determined by sequence similarity or by manual comparison of structural and functional features (Andreeva et al. 2007). This manual assignment provides the community with a curated expert classification of domain structures, but suffers from the concomitant drawback that any manual process is inevitably prone to subjective decisions. Domains are classified into the same SCOP family if they are "clearly evolutionarily related". In practise, this definition generally means that protein domains are grouped into the same family if they share pair-wise residue identities of more than 30%. However, some domains are classified into the same SCOP families in the absence of high sequence identities if similar structures and functions provide definitive evidence of common ancestry. This has the advantage of allowing for some flexibility in the assignment of homology relationships but also gives more room for subjectivity in the process. The Superfamily database expands SCOP to proteins of unknown structure by annotating sequences with SCOP descriptions at the family and superfamily level (Wilson et al. 2007; Oates et al. 2015). As with Gene3D, Superfamily uses SCOP-based HMM profiles to assign matches in sequences.

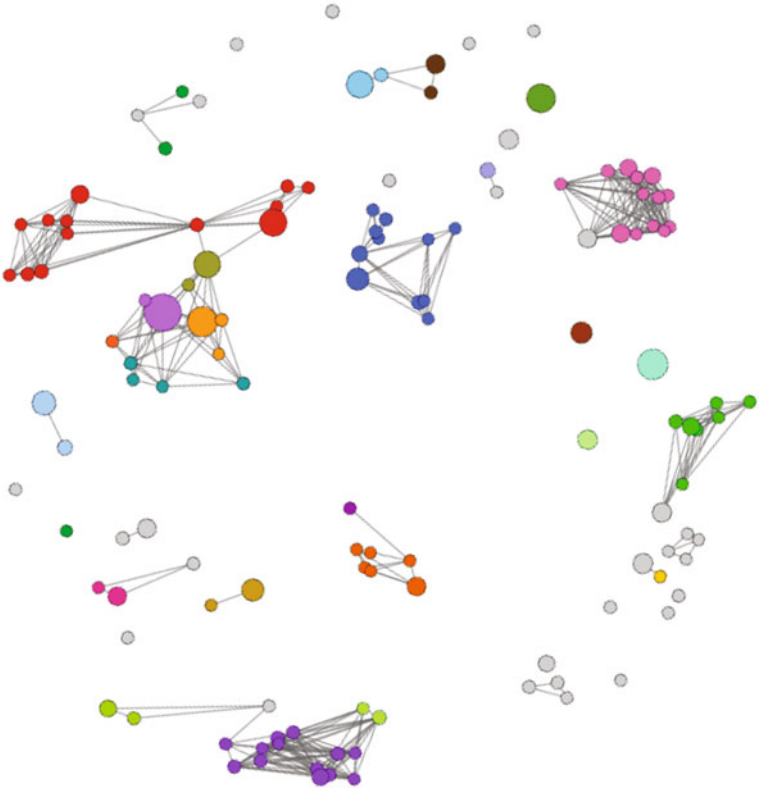


**Fig. 9.2** Visualization of functional family (FunFam) relationships in a CATH superfamily (CATH code: 3.40.50.620) using networks. Each node in the network corresponds to a FunFam, where the size of the node reflects the relative size of the FunFam. The edges in the network represent the sequence similarities between the FunFam profile HMMs using Profile Comparer (PRC; Madera 2008). The nodes are coloured according to their enzyme commission (EC) numbers, where grey nodes indicate FunFams without any EC annotation, including non-enzymes. The nodes are linked by edges if the similarity of their profile HMMs is within a threshold PRC score of 50

**SFLD**—The Structure-Function Linkage Database (SFLD) has been developed more recently with the specific aim of studying the structure-function relationships amongst homologous enzymes. It currently covers a relatively small set of superfamilies, as compared with CATH and SCOP, but provides a detailed description of the evolution of function within these superfamilies. The SFLD imposes that enzymes within superfamilies should not only be homologous but must share a mechanistic attribute in the catalytic reaction using conserved structural elements (Akiva et al. 2014). SFLD families consist of enzymes that perform the same overall reaction in a given superfamily.

**FunTree**—The FunTree resource (Furnham et al. 2012a, b) uses sequence, structure, phylogenetics, chemical and reaction mechanism data to functionally

annotate and analyse CATH enzyme domain superfamilies. Through this resource, information is provided to help users understand how particular enzymes have evolved new functions. Enzyme domain superfamilies are selected for analysis by using CATH to first find protein domains with structural data and then secondly, the MACiE database (Holliday et al. 2011) is used to determine whether the domain is part of an enzyme. As functional divergence can occur due to mutations within a single domain or from a change in the multi-domain architecture, FunTree generates two types of data clusters for analysis: one based only on the single superfamily domain, and the second type uses all of the domains in the protein sequences. Sequence data from UniProtKB/SwissProt (The UniProt Consortium 2014) and CATH-Gene3D is used to build phylogenetic trees, which are combined with functional information from the MACiE and Catalytic Site Atlas (Furnham et al. 2014) databases and displayed online.

### *9.3.2 Evolution of Protein Superfamilies*

Ultimately, the criterion to group proteins together in superfamilies is that the genes encoding them descend from a common ancestor gene. The processes by which an ancestor gene gives rise to two (or more) copies of itself are commonly referred to under the term *duplication*.

By definition, a duplication event gives rise to homologous genes. But further distinctions can be made. Genes that descend from a common ancestor gene via duplication within a given genome and in the absence of an accompanying speciation process are known as *paralogues*. Genes that descend from a common ancestor gene via duplication of the genome itself during speciation are known as *orthologues*. It is generally assumed that orthologous genes tend to preserve the function of the ancestor gene, due to a strong selective pressure to ensure that the ancestral function is still performed in both descendant species (Tatusov et al. 1997). Based on this assumption, some authors even define orthologues as homologues that have the same function in different species. Several databases have been set up to define orthologous genes from different sets of organisms (Dolinski and Botstein 2007). On the contrary, the presence of multiple copies of a given gene within a genome, i.e. paralogues, could arguably often result in one of the copies being under strong selective pressure to maintain the original function thus allowing more freedom for divergence for the other copies. The process by which one copy of a duplicated gene conserves the function of the ancestor gene, whereas the other copies evolve alternative functions is known as *neofunctionalisation*. In the absence of selective pressure on these additional copies, a frequent outcome of evolutionary divergence is the loss of some of them into *pseudo-genes*, which are gene relics no longer expressed (Harrison and Gerstein 2002). This evolutionary process is called *nonfunctionalisation*. *Subfunctionalisation* is a third evolutionary process which refers to cases where multiple functions of an ancestral gene are divided between

the paralogues. In any case, paralogues are often considered to be more functionally diverse than orthologues because of their larger freedom to diverge.

In whatever order they occurred, the subsequent events of duplication into orthologues or paralogues that took place during biological history have resulted in the current protein superfamilies. Not all superfamilies seem to have been equally successful in this process, as some of them are known to account for disproportionately large numbers of genes in fully sequenced genomes (Marsden et al. 2006; Chothia and Gough 2009). To date, reasons for evolutionary success disparity of the different superfamilies are not clear, and arguments relating to structural and functional properties, or evolutionary dynamics have been proposed (Goldstein 2008). It can be expected that older superfamilies, having had more time to diverge and explore different functions, should generally be more extended in present time. For example, the HUP superfamily (CATH code 3.40.50.620), which on account of phylogenetic considerations is believed to trace back to the RNA world, displays a very wide array of seemingly unrelated functions (Aravind et al. 2002); whereas several recent superfamilies that are observed exclusively in eukaryotic species are often restricted to very specific sets of functions. However, the age doesn't seem to be the main factor explaining the varying sizes of superfamilies. In a recent analysis of evolutionary dynamics of gene families that contain genes with essential functions (termed *E-families*) and gene families that do not contain such genes (termed *N-families*), it was proposed that paralogues in E-families are more likely to evolve new functions than those in N-families thus suggesting that the function of ancestral genes in a family is a key determinant of its evolutionary success (Shakhnovich and Koonin 2006). As will be shown in the next section, other arguments to explain the variable success of protein superfamilies may derive from the mechanisms that have been proposed to explain function evolution.

### *9.3.3 Function Divergence During Protein Evolution*

The traditional approach for annotating a protein of unknown function is to look for homologies between that protein and other well-characterised proteins, and to transfer the functional annotations from the latter to the former, assuming that proteins that descend from a common ancestor should share some degree of common functionality (Whisstock and Lesk 2003). But it is now a well-established fact that this approach is error-prone and that its incautious application results in unmanageable propagation of erroneous annotations in databases (Devos and Valencia 2001).

The major source of errors in this process is that the assumption following which homologous proteins have similar functions is inaccurate (Devos and Valencia 2000). There are now numerous documented cases of related proteins with very different functions, including the long-known example of hen egg-white lysozyme and mammalian α-lactalbumin that share more than 35% sequence identity and have very similar structures. Yet, it is reasonable to assume that the larger the

evolutionary distance between two homologous proteins, the lower the probability of these proteins sharing the same function. Several studies have attempted to determine sequence identity cut-offs that would safely guarantee conservation of function between pairs of homologues, but results are somewhat discordant and the issue is still under debate (Todd et al. 2001; Rost 2002; Tian and Skolnick 2003; Sangar et al. 2007; Addou et al. 2009). One likely explanation for the difficulty to derive universal sequence identity cut-offs for reliable transfer of function annotations between homologues is the above-mentioned fact that different superfamilies have very different patterns of sequence and function divergence. Accordingly, many recent studies focus on the analysis of sequence-structure-function relationships in specific superfamilies or subsets thereof, and may reveal highly valuable insights as to how the variations in sequence and structure correlate with variations in function.

In the following section, we describe function variation within superfamilies in more detail, with particular emphasis on the mechanisms thought to bring about this variation.

#### 9.3.3.1 Function Diversity at the Superfamily Level

The sequences of proteins classified in the same superfamily have sometimes diverged beyond levels that can be detected by standard sequence alignment methods. Even though three-dimensional structures are generally accepted to be far more conserved than sequences during evolution, major differences can still be observed between the structures of remote homologues. Such structural differences can arise from insertions/deletions (*indels*) of large elements of secondary structures or even several of these. A recent study of indels amongst homologous structures showed that it is not uncommon for successive insertions of secondary structures to occur in the same location of the fold of a protein during evolution, thus giving rise to so-called *nested indels* (Jiang and Blouin 2007). Another analysis of insertions within CATH superfamilies showed that not only do inserted secondary structures tend to co-locate in the fold but that the resulting embellishments often occur close to functionally important regions such as enzyme catalytic sites or protein-protein interfaces (Reeves et al. 2006); this observation indicates a correlation between structural and functional changes.

Insertions of new elements of secondary structure near the active site will most likely change the function, but more subtle changes such as residue substitutions of important catalytic residues will also result in functional differences. Recent analysis found cases in 16 enzyme domain superfamilies in CATH where the catalytic residues changed across functional families, even though their members performed the same enzyme chemistry (Furnham et al. 2015). The “Aldolase Class I” CATH superfamily (CATH ID 3.20.20.70) for example has four functional families that perform the same aldehyde lyase enzyme chemistry, however they each use different catalytic machineries (see Fig. 9.3). Changes in domain context can also result in drastic changes in the role of proteins, so that even if some aspect of

molecular function is conserved, it can hardly be said of the proteins that their function is the same (Todd et al. 2001; Hegyi and Gerstein 2001). This is the case of the PBP-like domains of eukaryotic and prokaryotic glutamate receptors, which bind the same ligand in a similar topological location, but widely differ in their function at the cellular level (see Fig. 9.4).

The long-term evolutionary processes via which function can diverge between homologues are numerous and difficult to summarise. Nevertheless, in a recent attempt to understand and categorise such processes, Bashton and Chothia have described and illustrated a subset of these to understand how the function of homologous domains can change depending on whether they are found in the context of single-domain proteins or combined with other domains in multi-domain proteins (Bashton and Chothia 2007). Examples of the processes identified include cases where the domain function is modified by its combination with other domains

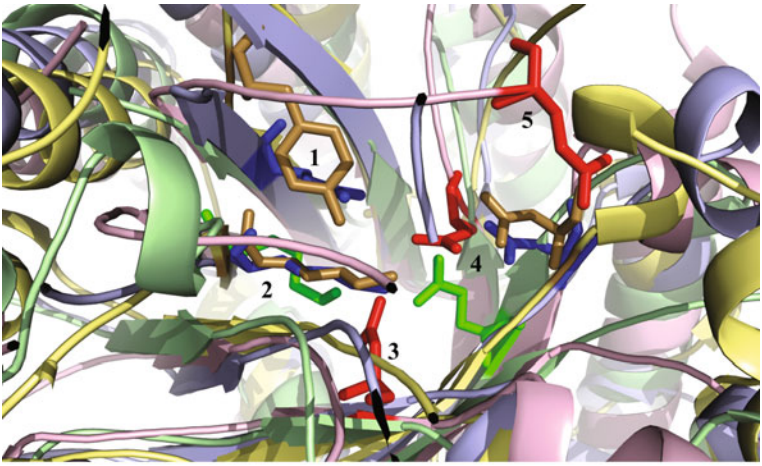



**Fig. 9.3** Comparison of the structural positions and functional properties of catalytic residues in four domains with the same enzyme chemistry but different catalytic machineries. A superposition of CATH domains from the Aldolase Class I superfamily (CATH ID 3.20.20.70) that catalyse aldehyde lyase activity (EC 4.1.2.-): 1aldA00 (*light blue*), 1ok4A00 (*light yellow*), 1fq0A00 (*light green*), and 1b57A00 (*light pink*). The catalytic residues (*dark blue*, *brown*, *green*, and *red*) from these four functional family representative domains (1aldA00, 1ok4A00, 1fq0A00, and 1b57A00, respectively) cluster into five spatial sites and one can assign a common functional property to each cluster

that modify its substrate specificity, or cases where the fusion of domains results in multi-functional proteins in which each domain is responsible for a particular function.

The above-mentioned occurrence of structural changes in the vicinity of functional regions points at the resulting functional diversity that is to be expected between superfamily members. And indeed, results from several studies indicate that remote homologues within superfamilies often perform very different functions (Todd et al. 2001). Most of these studies are focused on the evolution of function within particular superfamilies that generally show exceptional functional diversification, and prominent examples of which include haloacid dehalogenases (Burroughs et al. 2006), short-chain dehydrogenases/reductases (Favia et al. 2008), enolases (Gerlt and Babbitt 2001), HUP domains (Aravind et al. 2002) or "Two dinucleotide binding domains" flavoproteins (tDBDF's) (Ojha et al. 2007). The study of these different groups of proteins has revealed a large variety of processes by which function diverges between relatives, and these processes will now be considered separately with examples.

***Mechanistically Diverse Superfamilies***

A subset of much studied superfamilies constitute the core of the data in the Structure-Function Linkage Database (Akiva et al. 2014) and in spite of their functional diversity, and respecting the criteria of inclusion in SFLD (see Sect. 9.3.1.2), all members of these superfamilies share a common mechanistic attribute in the diverse reactions they catalyse.

The SFLD is in fact specifically aimed at describing these mechanistically diverse enzyme superfamilies and provides a classification of evolutionarily related enzymes notably based on similarities in their functional mechanisms. For example, the SFLD superfamily of haloacid dehalogenases groups together enzymes that can process a vast variety of substrates, but always act via the formation of a covalent enzyme-substrate intermediate through a conserved aspartate (Glasner et al. 2006), that in turn facilitates cleavage of C–Cl, P–C or P–O bonds. The haloacid dehalogenase superfamily contains 1285 unique sequences classified in 20 different families, each of which catalyses a unique reaction (e.g. histidinol phosphatases—EC number 3.1.3.15; or trehalose phosphatases—EC number 3.1.3.12). Some families are grouped together into sub-groups that constitute a convenient intermediate level whose definition varies between superfamilies.

Currently, the SFLD only covers 12 superfamilies. But the conservation of parts of the reaction chemistry within superfamilies appears very common, being observed in 22 out of the 31 enzyme superfamilies that were studied by Todd et al. (2001). In contrast, substrate specificity was *not* conserved in 20 of these superfamilies (see below).

The occurrence of a common mechanistic step in mechanistically diverse superfamilies suggests that enzymes in these superfamilies have maintained aspects of their catalytic mechanism in the course of their evolutionary diversification. Such situations hint at an evolutionary scenario in which enzymes evolve new functions, via duplication and recruitment, by maintaining partial reaction mechanisms (rather

than partial substrate specificity, see below), thus resulting in the mechanistically diverse superfamilies observed nowadays (Gerlt and Babbitt 2001).

A recent large-scale analysis of 379 enzyme domain superfamilies in CATH also confirmed conservation of reaction mechanism chemistry within some superfamilies (Furnham et al. 2015). The study also examined how function had diverged within these enzyme superfamilies by quantifying changes in the reaction mechanisms with EC-BLAST (Rahman et al. 2014) and used the scores as a proxy for a change in enzyme chemistry. To examine whether a change in enzyme chemistry between functional families was typically associated with a change in catalytic machinery, the similarity in reaction mechanism was compared with the similarity in catalytic residues. No clear correlation was found and examples of all combinations were found. Extreme outliers were discussed, for example: two domains from different functional families within a superfamily that performed very different reaction mechanisms using the same catalytic machinery (e.g. the catalytic domains from L-lactase dehydrogenase in yeast and glycolate oxidase in spinach); and on the other hand, domains that performed the same reaction mechanism using very different catalytic machinery (e.g. the domains with aldehyde lyase activity in Fig. 9.3) (Furnham et al. 2015).

***Specificity Diverse Superfamilies***

An alternative scenario for the divergent evolution of enzymatic functions within superfamilies is one in which an ancestral enzyme with broad specificity duplicates and the descendant copies specialise in binding more specific substrates. In such a scenario, substrate specificity is the dominant factor for function evolution in the superfamily. In their extensive analysis of enzyme superfamilies, Todd et al. showed that in most cases, reaction mechanisms were more conserved than substrate specificities between homologous enzymes. Out of 28 superfamilies that were involved in substrate binding, 10 displayed no conservation of the substrate whatsoever, and another 10 had very varied substrates with only a small common chemical moiety such as a peptide bond (Todd et al. 2001).

The expectation that substrate specificity might be conserved between homologous enzymes in a superfamily derives from Horowitz's proposal on the backward evolution of metabolic pathways (Horowitz 1945). This hypothesis suggests that when the substrate of an enzyme becomes depleted, an organism possessing a new enzyme that is able to produce that substrate from a precursor compound which is available will have a selective advantage over others, and the new enzyme will be fixed by evolution thus giving rise to an initial 2-step metabolic pathway. A similar evolutionary process can then take place for the other steps of the extant pathway. According to this scenario, pathway evolution goes backward as compared with the direction of the metabolic flow (Rison and Thornton 2002). Because the original enzyme has the ability to bind a substrate molecule that is the same as the product of the new enzyme, it has been suggested that this common property may be used as a basis for the evolution of the latter enzyme. Following this idea, all enzymes within a metabolic pathway would be homologous, and the enzyme catalysing the final step of the pathway would be the most ancient. In addition, the evolution of

these enzymes would have been driven by their substrate selectivity, and this would result in a tendency of extant superfamilies to display commonalities in substrate specificity. Possible examples of backward evolution have been collected, including that of the tryptophan biosynthesis pathway in which several enzymes that catalyse sequential steps are clearly homologous (Gerlt and Babbitt 2001; Todd et al. 2001).

However, results from several studies suggest that this hypothetical process has actually played a marginal role in the evolution of metabolism, which instead, would have resulted mostly from a chemistry-driven recruitment of enzymes between pathways (Rison and Thornton 2002). Indeed, superfamilies in which the substrate selectivity is conserved seem rare in comparison with those cases where the catalytic mechanism is conserved. Interestingly, the TIM-barrel phosphoenolpyruvate-binding enzymes superfamily, which was the only superfamily with absolutely conserved substrate specificity in the analysis of Todd et al. (2001), proved to be amongst the superfamilies with most diverse cognate ligands in a more recent study (Bashton et al. 2006), suggesting that the data used in the previous analysis may have been misleading due to its scarcity.

PROCOGNATE (Bashton et al. 2008) is a very useful tool for the analysis of ligand diversity bound by the different enzymes within a superfamily. The PROCOGNATE database maps enzymes to their cognate ligands, i.e. the ligands that the enzymes bind in vivo. Indeed, ligand data from PDB structures poses a problem: frequently, non-specific ligands bind to the enzymes in their active site thus mimicking the real ligand that binds in vivo (Dessailly et al. 2008). These contaminants make it difficult to automatically study ligand diversity in proteins of known structures as it is not obvious how to distinguish them from the biological ligands. PROCOGNATE is organised around the superfamilies (CATH, SCOP or PFAM) to which the enzymes belong. It is therefore useful for determining ligand diversity for any given superfamily of interest. For example, searching PROCOGNATE for the mechanistically diverse haloacid dehalogenase superfamily (CATH code 3.40.50.1000) returns a list of 57 cognate PDB ligands and 17 cognate KEGG compounds that bind to enzymes in that superfamily. The ancient and diverse HUP-domain superfamily (CATH code 3.40.50.620) is associated with 92 PDB ligands and 29 KEGG ligands in PROCOGNATE. These 29 KEGG ligands are shown in Fig. 9.5 and illustrate the diversity of molecules that can be bound by evolutionarily related proteins.

***Functional Changes Due To Changes in the Environmental Context***

Functional changes between duplicated copies of a protein can also arise not so much from changes within the protein itself, but rather from changes in the environmental conditions in which the different copies are active. For example, the recruitment of a protein in new locations of an organism may theoretically result in its encounter with small molecules that were not present in the original environment of the ancestor protein, and the recruited protein may display unexpected ability to bind these newly available ligands. Likewise, the molecular function of a protein may change if other proteins in its environment undergo mutations which result in

the possibility for new interactions or, on the contrary, in some protein-protein interactions becoming no longer possible.

A known example of functional changes between homologous enzymes that is related to changes in the environment is described in the literature for the "Two dinucleotide binding domains" flavoproteins, where diversification of function across the superfamily has resulted from the conscription of different protein partners acting as electron acceptors, via a conserved mode of protein-protein interactions (Ojha et al. 2007).

***Enzyme—non-enzyme***

A source of functional diversity in superfamilies that is not often discussed in the literature is that arising from the loss/gain of catalytic capability between homologues. Indeed, the analysis of non-enzymatic proteins is not as straightforward as that of enzymes, for which several annotation systems and analysis tools are now well-established (e.g. EC, KEGG and CSA; see Sect. 9.1). Non-enzymatic proteins are nevertheless frequently found in so-called enzymatic families. The processes by which a protein loses catalytic capabilities are fairly straightforward as the mere loss of a single crucial catalytic residue by substitution will generally lead to a loss of the enzymatic activity (Todd et al. 2002). The superfamily of HUP domains (CATH code 3.40.50.620) consists mostly of enzymes, but contains a few isolated examples of proteins with no known catalytic activity. For example, subunits of electron transferring flavoproteins constitute a separate functional family and display significant sequence, structure, and function alterations from other members of the superfamily (Aravind et al. 2002). An example at another level within that superfamily is that of the cryptochrome DASH, a non-enzyme that shows striking similarities with evolutionarily related DNA repair photolyases in terms of DNA binding and redox-dependent function, but also major differences notably in the active site (Brudler et al. 2003). There are also examples of superfamilies that are largely dominated by non-enzymes, such as the Periplasmic-Binding-Protein like domains (CATH code 3.40.190.10) in which many distinct functional families are identified on the basis of the molecules to which they bind, or of their role in the context of the cell, e.g. transporters or surface receptors.

***Extreme examples of functionally diverse superfamilies***

From the above discussion on mechanistically diverse and specificity diverse superfamilies, it appears that most superfamilies maintain some degree of functional commonality between members in spite of their divergence. This is to be expected since superfamilies consist of evolutionarily related proteins by definition, and the rules of parsimony make it reasonable to assume that homologous proteins may retain at least some aspect of their function in the course of evolution. However, examples of superfamilies also exist in which such commonalities have not been uncovered yet. In the previously mentioned analysis of large and diverse superfamilies by Todd et al., one superfamily—the Hexapeptide Repeat Proteins—displayed neither commonalities in catalytic mechanism nor in substrate selectivity (Todd et al. 2001). Another example of superfamily for which any functional similarity fails to emerge between members is that of the HUP-domains. Figure 9.6

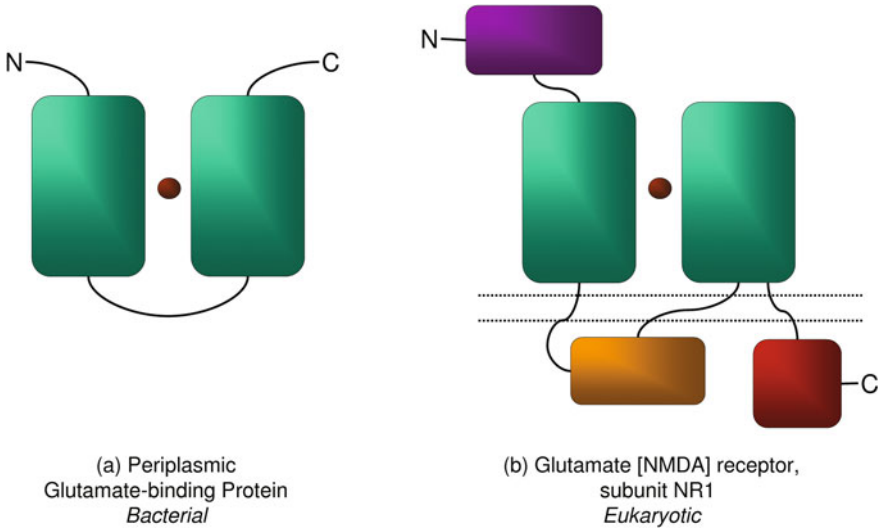



**Fig. 9.4** Multi-domain architectures of (**a**) Periplasmic Glutamate-binding protein from Gram-negative bacteria and (**b**) subunit NR2 of Glutamate [NMDA] receptor from Rat. Individual domains are represented as rectangles. N- and C-termini are represented with capital letters "N" and "C", respectively. The ligand L-glutamate is represented as a *brown sphere*. The cellular membrane in (**b**) is displayed as a *double dotted line*. The domains between which L-glutamate binds are coloured *green*. These domains are homologous to one another, both within and between the 2 proteins (CATH superfamily 3.40.190.10). These 2 proteins have very different functions, as suggested by their very different multi-domain architectures: (**a**) bacterial periplasmic glutamate-binding protein consists only of the 2 domains involved in binding glutamate and freely transports the latter across the periplasm (Takahashi et al. 2004). (**b**) Glutamate [NMDA] receptor (subunit NR2) is part of a transmembrane channel that plays a major role in excitatory neurotransmission; it consists of 5 globular domains and its binding to L-glutamate participates in opening the channel for cation influx (Furukawa et al. 2005). Even though the pair of green domains in these 2 proteins are homologous and share the ability to bind L-glutamate in a similar location of their structure, they undoubtedly have very different functions

summarises the functional diversity in that superfamily, together with representative structures for the main functional groups. Yet, due to the difficulty to apprehend function, it may well be that even within these extremely diverse superfamilies, functional commonalities that are not apparent at this stage will come to light as more data is collected and studied.

#### 9.3.3.2 Function Diversity Between Close Homologues

The above sections described the amount of functional diversity that is to be expected within protein superfamilies, with particular emphasis on remote homologues. But functional diversity is also observed between closer homologues, and

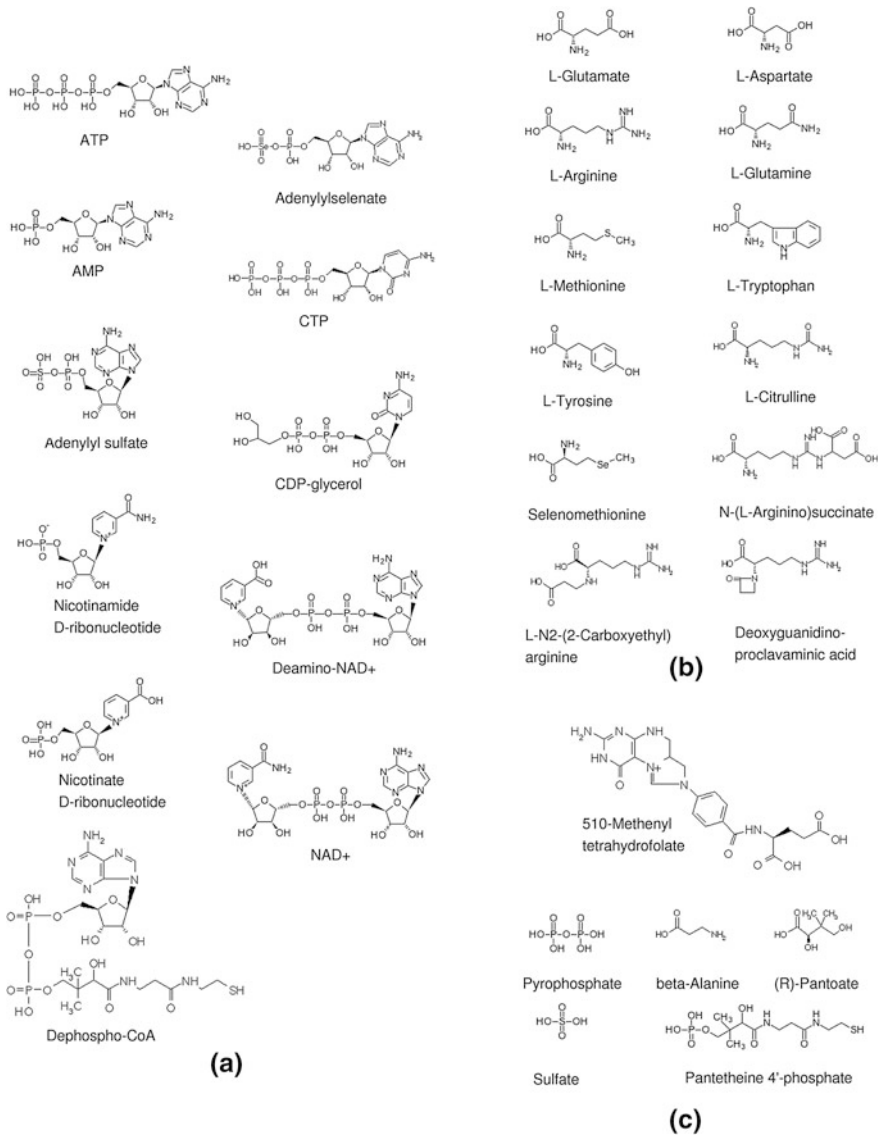



**Fig. 9.5** KEGG cognate ligands identified in PROCOGNATE as binding HUP domains (CATH superfamily 3.40.50.620). Three major categories of ligands are distinguished for clarity: (**a**) adenine-containing ligands and derivatives thereof, (**b**) amino-acids and derivatives thereof, and (**c**) diverse ligands that cannot be classified in either of the above two categories. Many more molecules (92) are found to bind HUP domains in the PDB but are not shown here. This figure shows that evolutionarily related domains are able to bind to a diverse range of molecules

sometimes even between exactly identical proteins seen in diverse contexts. For example, relatives can have multiple catalytic activities not necessarily of equal efficiency, as in promiscuous enzymes (Khersonsky and Tawfik 2010); or moonlighting functions whereby proteins perform completely different functions to their native activity sometimes involving different sites (Jeffery 1999). Promiscuity can be often be the starting point for the evolution of a new function and under natural selection, these enzymes can give rise to specialist enzymes by a variety of different mechanisms including domain insertions (Pandya et al. 2014), rearrangements in the catalytic metal ions (Baier and Tokuriki 2014) or binding of alternative cofactors (Baier et al. 2015) (Fig. 9.7).

Well-known examples of moonlighting proteins are eye lens crystallins, which are identical in sequence to liver enolase and lactate dehydrogenase (Piatigorsky

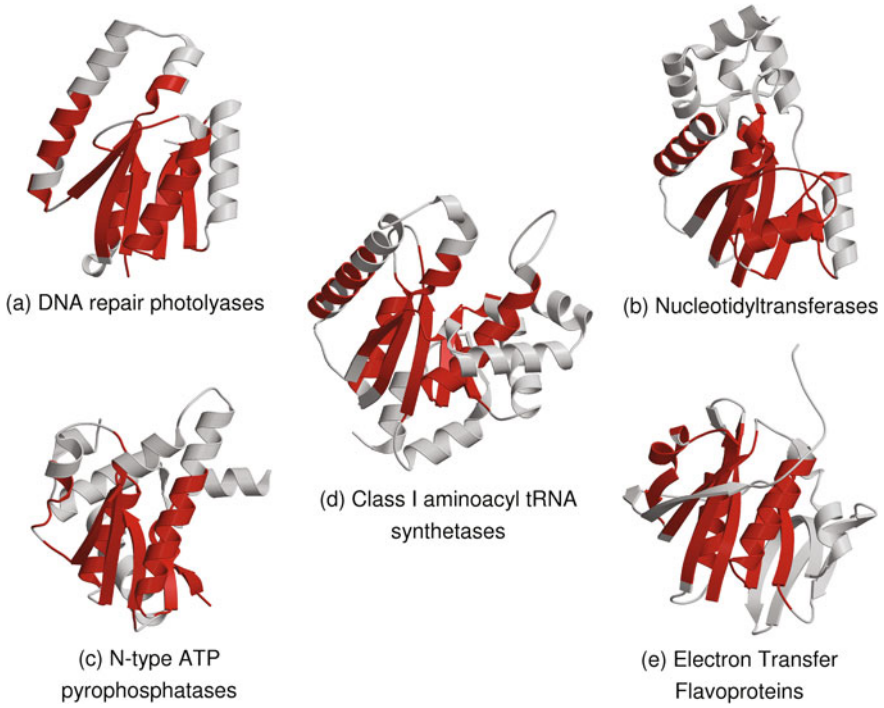



**Fig. 9.6** Diversity of structures and functions in the HUP-domains superfamily (CATH code 3.40.50.620). HUP-domains adopt a Rossmann-like fold and have been shown to be very ancient (Aravind et al. 2002). Together, they form a very large superfamily with many different functions. In this figure, representative structures of the major functional groups in this superfamily are displayed in cartoons. These structures were multiply aligned with CORA (Orengo 1999) and the multiple alignment was used to derive the common core of the domain. Residues that constitute the core are coloured *red* in each structure. The CATH domains that were used as representatives of each functional groups are: (**a**) 1dnpA01 for DNA repair photolyases, (**b**) 1ej2A00 for nucleotidyltransferases, (**c**) 1gpmA02 for N-type ATP pyrophosphatases, (**d**) 1n3lA01 for class I aminoacyl tRNA synthetases and (**e**) 1o97D01 for electron transfer flavoproteins

**Table 9.1** URLs and short descriptions of databases and tools of interest mentioned in the text

| Name | URL | Description |
|---|---|---|
| CATH | http://www.cathdb.info | Structural classification of proteins |
| SCOP | http://scop.mrc-lmb.cam.ac.uk/scop/ | Structural classification of proteins |
| SCOP2 | http://scop2.mrc-lmb.cam.ac.uk | Structural classification of proteins |
| SCOPe | http://scop.berkeley.edu/about/ver=2.05 | Structural classification of proteins |
| ECOD | http://prodata.swmed.edu/ecod/ | Structural classification of proteins |
| SFLD | http://sfld.rbvi.ucsf.edu/ | Functional classification of enzyme superfamilies |
| FunTree | http://www.funtree.info/FunTree/ | Exploring the evolution of protein function with sequence, structure and phylogenetics |
| PROCOGNATE | http://www.ebi.ac.uk/thornton-srv/databases/procognate/index.html | Mapping of domains to their cognate ligands |
| Gene Ontology | http://www.geneontology.org | Controlled vocabulary of protein functions |
| EC | http://www.chem.qmul.ac.uk/iubmb/enzyme/ | Classification of enzymatic reactions |
| EzCatDB | http://ezcatdb.cbrc.jp/EzCatDB/ | Database of enzyme catalytic mechanisms |
| MACiE | http://www.ebi.ac.uk/thornton-srv/databases/MACiE/ | Database of enzyme reaction mechanisms |
| KEGG | http://www.genome.jp/kegg/ | Integrated representation of genes, gene products and pathways |
| FUNCAT | http://mips.helmholtz-muenchen.de/funcatDB/ | Annotation scheme of protein functions |
| DALI | http://ekhidna.biocenter.helsinki.fi/dali_server | Structure alignment |
| FATCAT | http://fatcat.burnham.org/ | Flexible structure alignment |

et al. 1994; Whisstock and Lesk 2003). The multiple roles of moonlighting proteins are not restricted to certain organisms or protein families, nor do they have a common mechanism through which they switch between different functions. However, the functional diversity of moonlighting proteins is not caused by gene fusion, splice variance, varying post-transcriptional modifications, or multiple proteolytic fragments. Experimentally identified moonlighting proteins have been shown to switch functions as a consequence of changes in cellular locations within and outside the cell, expression in different cell types, oligomerisation states, ligand binding locations, binding partners and complex formation (Jeffery 1999, 2004). Moreover, orthologous proteins in different organisms do not necessarily share

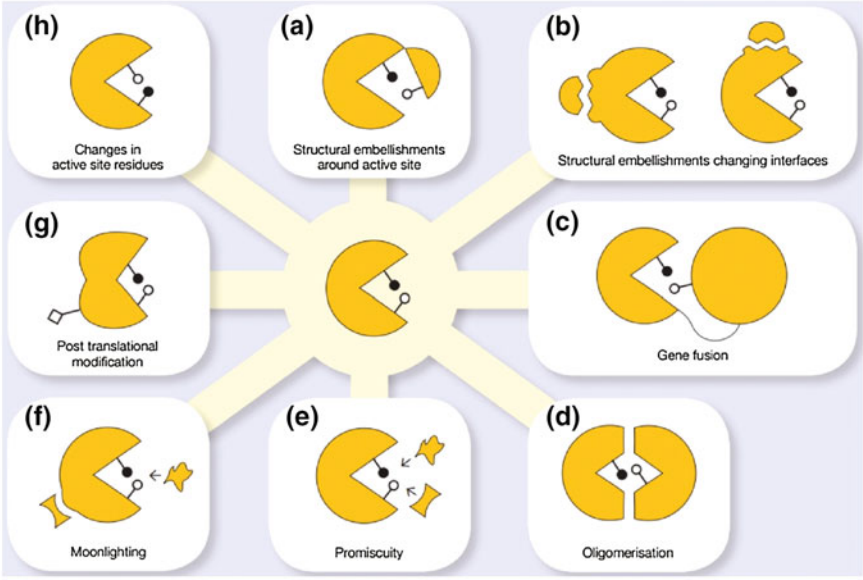



**Fig. 9.7** The various mechanisms, one or a combination of which can give rise due to functional diversity of proteins, are: (**a**) structural embellishments around active site, (**b**) structural embellishments changing interfaces, (**c**) gene fusion, (**d**) oligomerisation, (**e**) promiscuity, (**f**) moonlighting, (**g**) post-translational modification and (**h**) changes in active site residue. Note that for the mechanism panels (**a**), (**c**) and (**d**), one of the enzyme active site residue is contributed by its domain partner

moonlighting functions. Currently, there exist two manually-curated databases of moonlighting proteins, MultitaskProtDB (Hernández et al. 2014) and MoonProt (Mani et al. 2014), each of which lists more than 280 moonlighting proteins known in the literature. However, the rapid increase in the number of identified moonlighting proteins suggest that the phenomenon may be common in all kingdoms of life.

Furthermore, increasing evidence indicate that enzymes carry in them the potential for functional changes, in that they are generally able to catalyse promiscuous reactions in addition to the main, generally highly specific, reaction they are responsible for (Khersonsky et al. 2006). These extreme cases of function diversity between proteins displaying no or very low differences in sequence and structure are mentioned here in order to convey further the notion that the relationship between sequence, structure and function diversity is definitely a highly complex one, and that simple and reliable rules to predict function from sequence and structure are difficult to derive.

## 9.4 Conclusion

In this chapter, the relationship between function and structural similarity is explored. It is first shown that proteins sharing the same fold do not necessarily share the same function, but that knowledge of the structure and fold is often helpful for function annotation. The definition of a fold is discussed, with particular emphasis on the recent conceptual shift towards a continuous rather than discrete view of fold space.

Proteins sharing the same fold are not necessarily homologous. On the contrary, superfamilies are defined as groups of evolutionarily related proteins. But even within superfamilies, proteins are likely to perform different functions. Diverse processes to explain the evolution of superfamilies, and of protein function within them have been considered in the literature, and these processes are commented upon here. It is shown that even though evolutionarily related proteins do not necessarily share the same function, common elements of functionality are generally likely to remain between them. For example, mechanistically diverse superfamilies consist of enzymes that share a common mechanistic attribute in the enzymatic reactions they catalyse.

The relationship between protein function, structure and homology is complex, and perfect prediction of one of these attributes from any of the others is still not yet possible without errors. Nevertheless, identification of fold similarities or structural homologies between proteins is clearly helpful in function prediction, and the increase in structure, sequence and function data from the various—*omics* initiatives promises to greatly improve our understanding of the relationships between these attributes.

## Bibliography

Adams MA, Suits MDL, Zheng J, Jia Z (2007) Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. Proteomics 7:2920–2932. doi:10.1002/pmic.200700099

Addou S, Rentzsch R, Lee D, Orengo CA (2009) Domain-based and family-specific sequence identity thresholds increase the levels of reliable protein function transfer. J Mol Biol 387:416–430. doi:10.1016/j.jmb.2008.12.045

Akiva E, Brown S, Almonacid DE et al (2014) The structure-function linkage database. Nucleic Acids Res 42:D521–D530. doi:10.1093/nar/gkt1130

Andreeva A, Murzin AG (2006) Evolution of protein fold in the presence of functional constraints. Curr Opin Struct Biol 16:399–408. doi:10.1016/j.sbi.2006.04.003

Andreeva A, Howorth D, Chandonia JM et al (2007) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36:D419–D425. doi:10.1093/nar/gkm993

Andreeva A, Howorth D, Chothia C et al (2014) SCOP2 prototype: a new approach to protein structure mining. Nucleic Acids Res 42:D310–D314. doi:10.1093/nar/gkt1242

Andreeva A, Howorth D, Chothia C et al (2015) Investigating protein structure and evolution with SCOP2. Curr Protoc Bioinform 49:1.26.1–1.26.21. doi:10.1002/0471250953.bi0126s49

Aravind L, Anantharaman V, Koonin EV (2002) Monophyly of class I aminoacyl tRNA synthetase, USPA, ETFP, photolyase, and PP-ATPase nucleotide-binding domains: implications for protein evolution in the RNA. Proteins 48:1–14. doi:10.1002/prot.10064

Ashburner M, Ball CAA, Blake JAA et al (2000) Gene ontology: tool for the unification of biology. Nat Genet 25:25–29. doi:10.1038/75556

Baier F, Tokuriki N (2014) Connectivity between catalytic landscapes of the Metallo-β-Lactamase superfamily. J Mol Biol 426:2442–2456. doi:10.1016/j.jmb.2014.04.013

Baier F, Chen J, Solomonson M et al (2015) Distinct metal isoforms underlie promiscuous activity profiles of metalloenzymes

Bashton M, Chothia C (2007) The generation of new protein functions by the combination of domains. Structure 15:85–99. doi:10.1016/j.str.2006.11.009

Bashton M, Nobeli I, Thornton JM (2006) Cognate ligand domain mapping for enzymes. J Mol Biol 364:836–852. doi:10.1016/j.jmb.2006.09.041

Bashton M, Nobeli I, Thornton JM (2008) PROCOGNATE: a cognate ligand domain mapping for enzymes. Nucleic Acids Res 36:D618–D622. doi:10.1093/nar/gkm611

Brudler R, Hitomi K, Daiyasu H et al (2003) Identification of a new cryptochrome class. Structure, function, and evolution. Mol Cell 11:59–67

Burroughs AM, Allen KN, Dunaway-Mariano D, Aravind L (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. J Mol Biol 361:1003–1034. doi:10.1016/j.jmb.2006.06.049

Caspi R, Altman T, Billington R et al (2014) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 42:D459–D471. doi:10.1093/nar/gkt1103

Cheng H, Schaeffer RD, Liao Y et al (2014) ECOD: an evolutionary classification of protein domains. PLoS Comput Biol 10:e1003926. doi:10.1371/journal.pcbi.1003926

Chothia C, Gough J (2009) Genomic and structural aspects of protein evolution. Biochem J 419:15–28. doi:10.1042/BJ20090122

Colovos C, Cascio D, Yeates TO (1998) The 1.8 A crystal structure of the ycaC gene product from Escherichia coli reveals an octameric hydrolase of unknown specificity. Structure 6:1329–1337

Croft D, Mundo AFF, Haw R et al (2014) The Reactome pathway knowledgebase. Nucleic Acids Res 42:D472–D477. doi:10.1093/nar/gkt1102

Cuff A, Redfern OC, Greene L et al (2009) The CATH hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. Structure 17:1051–1062. doi:10.1016/j.str.2009.06.015

Das S, Lee D, Sillitoe I et al (2015) Functional classification of CATH superfamilies: a domain-based approach for protein function annotation. Bioinformatics btv398:1–8. doi:10.1093/bioinformatics/btv398

Dessailly BH, Lensink MF, Orengo CA, Wodak SJ (2008) LigASite—a database of biologically relevant binding sites in proteins with known apo-structures. Nucleic Acids Res. doi:10.1093/nar/gkm839

Devos D, Valencia A (2000) Practical limits of function prediction. Proteins Struct Funct Genet 107:98–107

Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17:429–431

Dolinski K, Botstein D (2007) Orthology and functional conservation in eukaryotes. Annu Rev Genet 41:465–507. doi:10.1146/annurev.genet.40.110405.090439

Favia AD, Nobeli I, Glaser F, Thornton JM (2008) Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. J Mol Biol 375:855–874. doi:10.1016/j.jmb.2007.10.065

Finn RD, Bateman A, Clements J et al (2014) Pfam: the protein families database. Nucleic Acids Res 42:D222–D230. doi:10.1093/nar/gkt1223

Fox NK, Brenner SE, Chandonia J-MM (2014) SCOPe: structural classification of proteins–extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 42:D304–D309. doi:10.1093/nar/gkt1240

Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. doi:10.1093/bioinformatics/bts565

Furnham N, Sillitoe I, Holliday GL et al (2012a) FunTree: a resource for exploring the functional evolution of structurally defined enzyme superfamilies. Nucleic Acids Res 40:D776–D782. doi:10.1093/nar/gkr852

Furnham N, Sillitoe I, Holliday GL et al (2012b) Exploring the evolution of novel enzyme functions within structurally defined protein superfamilies. PLoS Comput Biol 8:e1002403 +. doi:10.1371/journal.pcbi.1002403

Furnham N, Holliday GL, de Beer TAP et al (2014) The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42:D485–D489. doi:10.1093/nar/gkt1243

Furnham N, Dawson NL, Rahman SA et al (2015) Large-scale analysis exploring evolution of catalytic machineries and mechanisms in enzyme superfamilies. J Mol Biol. doi:10.1016/j.jmb.2015.11.010

Furukawa H, Singh SK, Mancusso R, Gouaux E (2005) Subunit arrangement and function in NMDA receptors. Nature 438:185–192

Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 70:209–246. doi:10.1146/annurev.biochem.70.1.209

Glasner M, Gerlt J, Babbitt P (2006) Evolution of enzyme superfamilies. Curr Opin Chem Biol 10:492–497. doi:10.1016/j.cbpa.2006.08.012

Goldstein RA (2008) The structure of protein evolution and the evolution of protein structure. Curr Opin Struct Biol 18:170–177. doi:10.1016/j.sbi.2008.01.006

Greene LH, Lewis TE, Addou S et al (2007) The CATH domain structure database: new protocols and classification levels give a more comprehensive resource for exploring evolution. Nucleic Acids Res 35:D291–D297. doi:10.1093/nar/gkl959

Grishin NV (2001) Fold change in evolution of protein structures. J Struct Biol 134:167–185

Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. J Mol Biol 318:1155–1174

Harrison A, Pearl F, Mott R et al (2002) Quantifying the similarities within fold space. J Mol Biol. doi:10.1016/S0022-2836(02)00992-0

Hegyi H, Gerstein M (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. Genome Res 11:1632–1640. doi:10.1101/gr.183801

Hernández S, Ferragut G, Amela I et al (2014) MultitaskProtDB: a database of multitasking proteins. Nucleic Acids Res 42:D517–D520. doi:10.1093/nar/gkt1153

Holliday GL, Andreini C, Fischer JD et al (2011) MACiE: exploring the diversity of biochemical reactions. Nucleic Acids Res 40:gkr799–D789. doi:10.1093/nar/gkr799

Holm L, Sander C (1993) Protein structure comparison by alignment of distance matrices. J Mol Biol 233:123–138. doi:10.1006/jmbi.1993.1489

Holm L, Sander C (1996a) The FSSP database: fold classification based on structure-structure alignment of proteins. Nucleic Acids Res 24:206–209

Holm L, Sander C (1996b) Mapping the protein universe. Science 273:595–603

Horowitz NH (1945) On the evolution of biochemical syntheses. Proc Natl Acad Sci USA 31:153–157

Jeffery CJ (1999) Moonlighting proteins. Tr Bioch Sci 24:8–11

Jeffery CJ (2004) Moonlighting proteins: complications and implications for proteomics research. Drug Discov Today TARGETS 3:71–78. doi:10.1016/S1741-8372(04)02405-3

Jiang H, Blouin C (2007) Insertions and the emergence of novel protein structure: a structure-based phylogenetic study of insertions. BMC Bioinform 8:444. doi:10.1186/1471-2105-8-444

Kanehisa M, Goto S, Sato Y et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42:D199–D205. doi:10.1093/nar/gkt1076

Katoh K, Standley DM (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. doi:10.1093/molbev/mst010

Khersonsky O, Tawfik DS (2010) Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 79:471–505
Khersonsky O, Roodveldt C, Tawfik D (2006) Enzyme promiscuity: evolutionary and mechanistic aspects. Curr Opin Chem Biol 10:498–508. doi:10.1016/j.cbpa.2006.08.011
Kolodny R, Koehl P, Levitt M (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. J Mol Biol 346:1173–1188. doi:10.1016/j.jmb.2004.12.032
Kolodny R, Petrey D, Honig B (2006) Protein structure comparison: implications for the nature of "fold space", and structure and function prediction. Curr Opin Struct Biol 16:393–398. doi:10.1016/j.sbi.2006.04.007
Kraulis PJ (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Crystallogr 24:946–950
Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60:2256–2268. doi:10.1107/S0907444904026460
Lee D, Grant A, Marsden RL, Orengo C (2005) Identification and distribution of protein families in 120 completed genomes using Gene3D. Proteins Struct Funct Bioinforma. doi:10.1002/prot.20409
Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8:995–1005. doi:10.1038/nrm2281
Lee DA, Rentzsch R, Orengo C (2010) GeMMA: functional subfamily classification within superfamilies of predicted protein structural domains. Nucleic Acids Res 38:720–737. doi:10.1093/nar/gkp1049
Lees JG, Lee D, Studer RA et al (2014) Gene3D: multi-domain annotations for protein sequence and comparative genome analysis. Nucleic Acids Res 42:D240–D245. doi:10.1093/nar/gkt1205
Lopez G, Maietta P, Rodriguez JM et al (2011) Firestar–advances in the prediction of functionally important residues. Nucleic Acids Res 39:W235–W241. doi:10.1093/nar/gkr437
Madera M (2008) Profile comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics 24:2630–2631
Mani M, Chen C, Amblee V et al (2014) MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res gku954
Marsden RL, Ranea JAG, Sillero A et al (2006) Exploiting protein structure data to explore the evolution of protein function and biological complexity. Philos Trans R Soc B Biol Sci. doi:10.1098/rstb.2005.1801
Martin AC, Orengo CA, Hutchinson EG et al (1998) Protein folds and functions. Structure 6:875–884
Merritt EA, Bacon DJ (1997) [26] Raster3D: photorealistic molecular graphics. Methods Enzymol 277:505–524
Moult J, Melamud E (2000) From fold to function. Curr Opin Struct Biol 10:384–389
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247:536–540. doi:10.1016/S0022-2836(05)80134-2
Nagano N (2005) EzCatDB: the enzyme catalytic-mechanism database. Nucleic Acids Res 33:D407–D412. doi:10.1093/nar/gki080
Nagano N, Orengo CA, Thornton JM (2002) One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 321:741–765
Nomenclature Committee of the IUBMB (1992) Enzyme nomenclature: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology. Academic Press, San Diego, California
O'Boyle NM, Holliday GL, Almonacid DE, Mitchell JBO (2007) Using reaction mechanism to measure enzyme similarity. J Mol Biol 368:1484–1499. doi:10.1016/j.jmb.2007.02.065

Oates ME, Stahlhacke J, Vavoulis DV et al (2015) The SUPERFAMILY 1.75 database in 2014: a doubling of data. Nucleic Acids Res 43:D227–D233. doi:10.1093/nar/gku1041

Ojha S, Meng EC, Babbitt PC (2007) Evolution of function in the "two dinucleotide binding domains" flavoproteins. PLoS Comput Biol 3:e121 +. doi:10.1371/journal.pcbi.0030121

Orengo CA, Taylor WR (1996) SSAP: sequential structure alignment program for protein structure comparison. In: Russell FD (ed) Methods in enzymology. Academic Press, Cambridge

Orengo CA, Jones DT, Thornton JM (1994) Protein domain superfolds and superfamilies

Orengo CA (1999) CORA—topological fingerprints for protein structural families. Protein Sci 8:699–715

Orengo CA, Michie AD, Jones S et al (1997) CATH—a hierarchic classification of protein domain structures. Structure 5:1093–1108

Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN (2014) Enzyme promiscuity: engine of evolutionary innovation. J Biol Chem 289:30229–30236. doi:10.1074/jbc.R114.572990

Pethica RB, Levitt M, Gough J (2012) Evolutionarily consistent families in SCOP: sequence, structure and function. BMC Struct Biol 12:27. doi:10.1186/1472-6807-12-27

Piatigorsky J, Kantorow M, Gopal-Srivastava R, Tomarev SI (1994) Recruitment of enzymes and stress proteins as lens crystallins. EXS 71:241–250

Porter CT, Bartlett GJ, Thornton JM (2004) The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32:D129–D133. doi:10.1093/nar/gkh028

Radivojac P, Clark WT, Oron TR et al (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10:221–227. doi:10.1038/nmeth.2340

Rahman SA, Cuesta SM, Furnham N et al (2014) EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods 11:171–174. doi:10.1038/nmeth.2803

Rausell A, Juan D, Pazos F, Valencia A (2010) Protein interactions and ligand binding: from protein subfamilies to functional specificity. Proc Natl Acad Sci 107:1995–2000. doi:10.1073/pnas.0908044107

Redfern OC, Harrison A, Dallman T et al (2007) CATHEDRAL: a fast and effective algorithm to predict folds and domain boundaries from multidomain protein structures. PLoS Comput Biol 3:e232 +. doi:10.1371/journal.pcbi.0030232

Reeves G, Dallman T, Redfern O et al (2006) Structural diversity of domain superfamilies in the CATH database. J Mol Biol 360:725–741. doi:10.1016/j.jmb.2006.05.035

Reid AJ, Yeats C, Orengo CA (2007) Methods of remote homology detection can be combined to increase coverage by 10% in the midnight zone. Bioinformatics 23:2353–2360. doi:10.1093/bioinformatics/btm355

Rison SCG, Thornton JM (2002) Pathway evolution, structurally speaking. Curr Opin Struct Biol 12:374–382. doi:10.1016/s0959-440x(02)00331-7

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318:595–608

Ruepp A, Zollner A, Maier D et al (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. Nucleic Acids Res 32:5539–5545. doi:10.1093/nar/gkh894

Russell RB, Saqi MA, Sayle RA et al (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. J Mol Biol 269:423–439. doi:10.1006/jmbi.1997.1019

Russell RB, Sasieni PD, Sternberg MJ (1998) Supersites within superfolds. Binding site similarity in the absence of homology. J Mol Biol 282:903–918. doi:10.1006/jmbi.1998.2043

Sadreyev R, Grishin N (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. J Mol Biol 326:317–336

Sangar V, Blankenberg DJ, Altman N, Lesk AM (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC Bioinform 8:294. doi:10.1186/1471-2105-8-294

Shakhnovich BE, Koonin EV (2006) Origins and impact of constraints in evolution of gene families. Genome Res 16:1529–1536. doi:10.1101/gr.5346206

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11:739–747. doi:10.1093/protein/11.9.739

Sillitoe I, Lewis TE, Cuff A et al (2015) CATH: comprehensive structural and functional annotations for genome sequences. Nucleic Acids Res 43:D376–D381. doi:10.1093/nar/gku947

Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951–960. doi:10.1093/bioinformatics/bti125

Takahashi H, Inagaki E, Kuroishi C, Tahirov TH (2004) Structure of the Thermus thermophilus putative periplasmic glutamate/glutamine-binding protein. Acta Crystallogr Sect D Biol Crystallogr 60:1846–1854

Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278:631–637

The UniProt Consortium (2014) UniProt: a hub for protein information. Nucleic Acids Res 43:D204–D212. doi:10.1093/nar/gku989

Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333:863–882

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307:1113–1143. doi:10.1006/jmbi.2001.4513

Todd AE, Orengo CA, Thornton JM (2002) Sequence and structural differences between enzyme and nonenzyme homologs. Structure 10:1435–1451

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36:307–340

Wilson D, Madera M, Vogel C et al (2007) The SUPERFAMILY database in 2007: families and functions. Nucleic Acids Res 35:D308–D313. doi:10.1093/nar/gkl910

Ye Y, Godzik A (2003) Flexible structure alignment by chaining aligned fragment pairs allowing twists. Bioinformatics 19:ii246–ii255. doi:10.1093/bioinformatics/btg1086

Yeats C, Lees J, Reid A et al (2008) Gene3D: comprehensive structural and functional annotation of genomes. Nucleic Acids Res. doi:10.1093/nar/gkm1019

# Chapter 10
# Function Prediction Using Patches, Pockets and Other Surface Properties

**Daniel J. Rigden**

**Abstract** With few exceptions protein functions depend sensitively upon their interactions with other biomolecules. Thus, the surface of a protein is of particular interest for function annotation: definition of the protein surface in experimental or modelled protein structure enables the application of a wide range of structural bioinformatic tools for function prediction. The development of such tools has been significantly accelerated in recent years as a response to the flux of information from Structural Genomics programs which, at least in part, have deliberately targeted mysterious protein families of unknown function about which conventional homology-based protein function annotation can say little or nothing (Bateman et al. in Acta Crystallographica Section F: Structural Biology and Crystallization Communications 66:1148–1152, 2010). As this chapter will illustrate, the underlying principles behind the resulting toolset vary impressively but, ultimately, most are based upon discovering putative interaction sites through detecting ways in which they differ somehow from protein surface in general. These differences may be physicochemical, electrostatic or steric in nature, or be of evolutionary origin. Predictions can be strengthened by observing concordant results from orthogonal methods. Indeed, many programs now improve performance by combining multiple factors in their calculations. Some methods find functional sites in general, others provide direct evidence supporting specific biochemical functions. This chapter will not attempt a comprehensive historical overview of the area, rather aiming to guide the user to the current state of the art while acknowledging key methodology papers. Methods that are readily available will be favoured, particularly those implemented at servers and those for which plug-ins for popular molecular visualisation tools exist.

**Keywords** Surface patches · Pockets · Cavities · Hydrophobicity · Electrostatics · Sequence conservation · Channels and Tunnels · Binding sites · Catalytic sites

D.J. Rigden (✉)
Institute of Integrative Biology, University of Liverpool,
Crown St., Liverpool L69 7ZB, UK
e-mail: drigden@liverpool.ac.uk



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_10

## 10.1 Definitions of Protein Surfaces

Most of the methods in this chapter depend on defining and describing a protein surface. It is therefore appropriate to briefly introduce the commonly encountered definitions of a protein surface (Fig. 10.1). The simplest is known as the van der Waals surface and is straightforwardly defined as the outermost surface of a set of overlapping atomic spheres, one for each atom in the protein, each having the corresponding van der Waals radius of the atom in question (Fig. 10.1a). It is most often seen as the space-filling representation of a molecule in visualisation software. It is not often used as a representation of the protein surface since much of the empty space between atoms is inaccessible to solvent atoms.

Two related definitions of protein surface are used more frequently. In each the surface is defined with reference to surrounding solvent by rolling a solvent molecule, generally modelled as a sphere with radius 1.4 Å, over the protein's van der Waal's structure. The first, the Molecular surface, also known as the Connolly surface (Connolly 1983), is defined as the surface traced by contact points between the protein and the rolling solvent. It reveals the surface that is available for interaction with solvent or other molecules. The surface is composed of a contact surface and a reentrant surface where the solvent molecule is in touch with one or multiple protein atoms, respectively. By doing so it defines a solvent-excluded volume that sums the van der Waal's volume and interstitial volumes, including those on the interior side of reentrant surfaces (Fig. 10.1b).

The most common definition of the protein surface is the solvent accessible surface (SAS; Lee and Richards 1971). As with the Molecular surface, a solvent molecule rolls over the van der Waal's surface but this time the surface is describe by the centre of the solvent molecule yielding a larger surface (Fig. 10.1c).

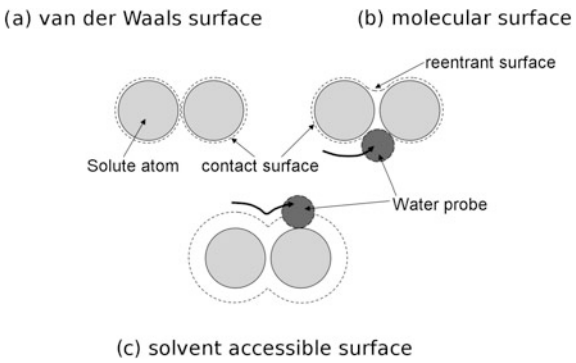



**Fig. 10.1** Illustrations of three commonly used molecular surfaces (*dotted lines*): **a** the van der Waals surface, **b** the molecular (Connolly) surface and **c** the solvent accessible surface. In (**a**) and (**b**) the *shaded atoms* are shown at slightly less than their van der Waals radius in order to reveal the surface better. Adapted from a figure originally published in Burgoyne and Jackson (2009); published with kind permission of © Springer Science+Business Media B.V. All Rights Reserved

## 10.2 Surface Patches

Regions whose characteristics depart from those generally expected of protein surfaces make a good starting point for prediction of protein functional sites. The majority of a protein surface is likely to lack specific function, serving for example only to help solubilise a globular protein or maintain favourable lipid interactions in the case of transmembrane stretches of a membrane-resident protein. This major portion need only have the appropriate general physicochemical properties. On the other hand, a functional site may need to buck these trends possessing, for example, the pronounced hydrophobic characteristics required for binding of a similarly hydrophobic ligand, even in the context of the generally hydrophilic surface of a globular protein and despite a consequent destabilisation of the protein fold. Furthermore, the astonishing range of molecules bound by proteins exerting their biological roles means that binding sites for different natural ligands can vary dramatically: thus bioinformatic analysis of protein surfaces in diverse respects is important for a full understanding of protein function. In the case of enzymes, the atypical characteristics of the catalytic site can be seen as a tradeoff between stability and activity (Beadle and Shoichet 2002): indeed spotting residues whose replacement is predicted to electrostatically stabilise structure can be used to predict catalytic sites (Elcock 2001) (see later). Furthermore, an important binding site will be constrained by evolution in terms of shape and potential interactions such that mutations may not be tolerated.

### *10.2.1 Hydrophobic Patches*

It has long been known that the surfaces of proteins found in an aqueous environment are largely composed of hydrophilic amino-acids. Indeed, the entropic benefits of burying hydrophobic amino-acids in the core of the protein structure drive protein folding. Nevertheless, functional surfaces of a soluble protein may have to be at least partially hydrophobic in order to provide a complementary site for a given ligand. The program QUILT (Lijnzaad et al. 1996), not implemented as a server but available to download (Table 10.1), seeks the largest patch on the SAS composed of carbon or sulphur atoms. A randomisation is used to estimate the significance of patches found. In initial tests the largest patches it found coincided with known binding regions of proteins with hydrophobic ligands such as lipase and LIV-binding protein: for other proteins the hydrophobic nature of the largest patch was generally less clearly correlated to known function. Interestingly, however, a dimer interface was picked out for triose phosphate isomerase: it is now well-understood that protein-protein interfaces tend to be more hydrophobic than protein surface in general and this observation contributes to many predictive methods (see Sect. 10.5). Among subsequent applications, QUILT was used to describe the surface of a model of an emulsifying protein from sunflower seed.

**Table 10.1** Selected webservers and other methods for function annotation by analysis of protein surface properties. URLs indicate servers except those italicised which are for software downloads

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| *Patches* | | | | |
| Hydrophobicity | QUILT | Measures hydrophobic patches as carbon and sulphur contributions to the solvent accessible surface | *http://bioinformatics.holstegelab.nl/publications/lijnzaad/quilt* | Lijnzaad et al. (1996) |
| Electrostatics | DELPHI | Finite difference Poisson-Boltzmann | http://compbio.clemson.edu/delphi_webserver | Rocchia et al. (2002) |
| | APBS | Adaptive Poisson-Boltzmann | http://nbcr-222.ucsd.edu/pdb2pqr_1.8 | Baker et al. (2001) |
| | eF-site/eF-surf | Poisson-Boltzmann | http://ef-site.hgc.jp/eF-site | Kinoshita and Nakamura (2004) |
| | Patch Finder Plus | Uses APBS | http://pfp.technion.ac.il | Shazman et al. (2007) |
| | webPIPSA | Uses e.g. APBS | http://pipsa.eml.org | Richter et al. (2008) |
| | Bluues | Uses approximate generalized born model | http://protein.bio.unipd.it/bluues | Walsh et al. (2012) |
| Conservation | ConSurf; ConSurf-DB | Phylogenetic tree-based inference of conserved residues and options for visualisation | http://consurf.tau.ac.il; http://consurfdb.tau.ac.il | Ashkenazy et al. (2010, 2016), Goldenberg et al. (2009) |
| | FuncPatch | Phylogenetic tree-based inference of conserved residues, combined with a Bayesian treatment of correlated substitution rates at spatially nearby positions | http://info.mcmaster.ca/yifei/FuncPatch | Huang and Golding (2015) |
| | EVtrace | Phylogenetic tree-based inference of conserved residues and options for visualisation | http://mammoth.bcm.tmc.edu | Morgan et al. (2006), Ward et al. (2009), Wilkins et al. (2012) |

(continued)

**Table 10.1** (continued)

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| | INTREPID | Phylogenetic tree-based inference of conserved residues | http://phylogenomics.berkeley.edu/intrepid | Sankararaman et al. (2009) |
| | SDPsite | Finds clusters of conserved residues and putative specificity-determining residues | http://bioinf.fbb.msu.ru/SDPsite | Kalinina et al. (2009) |
| Surface statistics | STP | Surface atom triplet propensities | http://opus.bch.ed.ac.uk/stp | Mehio et al. (2010) |
| | LISE | Surface atom triangle propensities and conservation | http://lise.ibms.sinica.edu.tw | Xie and Hwang (2012), Xie et al. (2013) |
| Multiple properties | HotPatch | Hydrophobicity, electrostatic properties, surface roughness, concavity and combinations thereof. | http://hotpatch.mbi.ucla.edu | Pettit et al. (2007) |
| *Pockets* | | | | |
| Geometric description | PASS | The surface is covered in probes. After removal of those in convex regions, clusters define pockets | *http://www.ccl.net/cca/software/UNIX/pass/overview.html* | Brady and Stouten (2000) |
| | GHECOM | Pockets are defined as accessible to a small probe placed on the protein surface, but not to a large probe | http://strcomp.protein.osaka-u.ac.jp/ghecom | Kawabata (2010) |
| | CASTp | Based on Delaunay triangulation | http://sts.bioe.uic.edu/castp | Dundas et al. (2006) |
| | Fpocket | Uses alpha spheres to identify pockets. Can track pockets during MD trajectories | http://bioserv.rpbs.univ-paris-diderot.fr/services/fpocket | Schmidtke et al. (2010) |
| | KVFinder | Two-probe grid-based method allowing user-defined splitting of pockets | *http://lnbio.cnpem.br/bioinformatics/main/software* | Oliveira et al. (2014) |
| Tunnels and channels | MOLEonline 2.0 | Discovers and describes tunnels leading to buried cavities and transmembrane channels | http://mole.upol.cz | Berka et al. (2012) |
| | MolAxis | Discovers and describes tunnels leading to buried cavities and transmembrane channels | http://bioinfo3d.cs.tau.ac.il/MolAxis | Yaffe et al. (2008) |

(continued)

**Table 10.1** (continued)

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| | PoreWalker | Discovers and describes transmembrane channels | http://www.ebi.ac.uk/thornton-srv/software/PoreWalker | Pellegrini-Calace et al. (2009) |
| | CAVER | Discovers tunnels and channels | *http://www.caver.cz* | Chovancova et al. (2012) |
| Distinguishing functional pockets | LIGSITE$^{CSC}$ | Cavity detection using the Connolly surface combined with sequence conservation | http://gopubmed2.biotec.tu-dresden.de/cgi-bin/index.php | Huang and Schroeder (2006) |
| | ConCavity | Cavities scored by sequence conservation | http://compbio.cs.princeton.edu/concavity | Capra et al. (2009) |
| | PDBinder | Pockets are analysed for conservation and by binding propensities for residue triplets in the protein of interest | *http://cbm.bio.uniroma2.it/pdbinder/usage.html* | Bianchi et al. (2013) |
| | VASP-E | Calculation and comparison of volumes representing electrostatic potential. | *http://www.cse.lehigh.edu/~chen/software.htm* | Chen (2014) |
| | Depth | Measures cavity depth, optionally including conservation information | http://mspc.bii.a-star.edu.sg/tankp | Tan et al. (2013) |
| | SuMo | Defines and matches 3D arrangements of chemical groups | http://sumo-pbil.ibcp.fr | Jambon et al. (2003) |
| | ProBis | Matches graph representations of surface features | http://probis.cmm.ki.si | Konc and Janezic (2010) |
| | SMAP | Aligns profiles that represent binding sites in a sequence-order independent fashion | http://nbcr-222.ucsd.edu/opal2/services/SMAPDBSearch | Ren et al. (2010) |
| | Patch-Surfer | Matches pockets represented as multiple local patches capturing geometry, hydrophobicity and electrostatic potentials | http://kiharalab.org/patchsurfer2.0 | Sael and Kihara (2012) |
| | IsoMIF Finder | Matches six molecular interaction fields | http://bcb.med.usherbrooke.ca/imfi | Chartier et al. (2016) |
| | metaPocket 2.0 | Meta server, largely based on pocket detection by geometric criteria | http://projects.biotec.tu-dresden.de/metapocket | Zhang et al. (2011) |

(continued)

**Table 10.1** (continued)

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| | eMatchSite | Matches to sequence order-independent local binding site alignments. Particularly designed for protein models | http://brylinski.cct.lsu.edu/ematchsite | Brylinski (2014) |
| | GalaxySite | Compounds from superimposable structures are docked and ranked as candidate ligands for the protein of interest | http://galaxy.seoklab.org/cgi-bin/submit.cgi?type=SITE | Heo et al. (2014) |
| | ProBIS-ligands | Predicts ligand binding poses based on binding sites matched by PROBIS | http://probis.cmm.ki.si/ligands | Konc and Janezic (2014) |
| | FINDSITE-LHM | Docks and clusters ligands matching predicted binding sites | http://cssb.biology.gatech.edu/findsitelhm | Brylinski and Skolnick (2009) |
| | COFACTOR | Docks and clusters ligands matching predicted binding sites | http://zhanglab.ccmb.med.umich.edu/COFACTOR | Roy et al. (2012) |
| | COACH | Consensus prediction applicable to both pockets and patches | http://zhanglab.ccmb.med.umich.edu/COACH/ | Yang et al. (2013) |
| Catalytic residues | MEPI | Catalytic site residue propensities, micro-environment and geometry. A further score includes conservation | http://protein.cau.edu.cn/mepi | Han et al. (2012) |
| | DISCERN | Multiple factors, including structure- and sequence-derived characteristics | http://phylogenomics.berkeley.edu/intrepid | Sankararaman et al. (2010) |
| | POOL | THEMATICS, pockets and (optionally) conservation | http://www.pool.neu.edu/wPOOL/index2.jsp | Somarowthu et al. (2011) |
| | EXIA2 | Relative side chain orientation, backbone flexibility and (optionally) conservation | http://203.64.84.196 | Chien and Huang (2012), Lu et al. (2014) |
| Protein-protein interfaces | cons-PPISP | Uses sequence profiles and solvent accessibility for residues and their neighbours | http://pipe.scs.fsu.edu/ppisp.html | Zhou and Shan (2001) |
| | meta-PPISP | Metaserver using cons-PPISP, PINUP and Promate predictions | http://pipe.scs.fsu.edu/meta-ppisp.html | Qin and Zhou (2007) |

(continued)

**Table 10.1** (continued)

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| | CPORT | Consensus prediction from six different methods | http://haddock.science.uu.nl/services/CPORT | de Vries and Bonvin (2011) |
| | PRISE | Atomic composition, residue type and solvent exposure of a central surface residue plus its neighbours | http://prise.cs.iastate.edu/index.py | Jordan et al. (2012) |
| | VORFFIP | Considers a variety of structural, energetic, evolutionary and crystallographic parameters | http://www.bioinsilico.org/VORFFIP | Segura et al. (2011) |
| Nucleic acids | iDBPs | Conservation, electrostatics and secondary structure | http://idbps.tau.ac.il | Nimrod et al. (2009) |
| | DISPLAR | Protein-DNA interface propensities, conservation and solvent accessibility | http://pipe.scs.fsu.edu/displar.html | Tjong and Zhou (2007) |
| | DP-dock | Docking compatibility with B-DNA | http://cssb.biology.gatech.edu/skolnick/webservice/DP-dock/index.html | Gao and Skolnick (2009) |
| | DR_bind | Geometry, electrostatics and conservation | http://dnasite.limlab.ibms.sinica.edu.tw | Chen et al. (2012b) |
| | BindUP | Electrostatics and other structural features | http://bindup.technion.ac.il/ | Paz et al. (2016) |
| | RBscore | Electrostatic potential, solvent accessibility and sequence conservation followed by neighbouring network analysis | http://ahsoka.u-strasbg.fr/rbscore/ | Miao and Westhof (2015) |
| | KYG | Residue propensities at RNA-binding sites, single and doublet, plus conservation | http://cib.cf.ocha.ac.jp/KYG | Kim et al. (2006) |
| | aaRNA | Geometry, conservation and residue composition | http://sysimm.ifrec.osaka-u.ac.jp/aarna | Li et al. (2014b) |
| | NABind | Electrostatics and triplet interface propensity | http://lilab.ecust.edu.cn/NABind | Sun et al. (2016) |

(continued)

**Table 10.1** (continued)

| Resource classification | Resource name | Method description | URL | References |
|---|---|---|---|---|
| Methods for other classes of ligands | ISMBLab | Probability density distributions of interacting atoms | http://ismblab.genomics.sinica.edu.tw/index.php | Tsai et al. (2012) |
| | SiteHound | Clusters predictions from interaction energy maps | http://scbx.mssm.edu/sitehound/sitehound-web/Input.html | Hernandez et al. (2009) |
| Druggability | DoGSiteScorer | Predicts druggability of pockets and subpockets using both local and global features | http://dogsite.zbh.uni-hamburg.de | Volkamer et al. (2012) |
| | FTMAP | Docks, scores and clusters small molecular probes | http://ftmap.bu.edu | Ngan et al. (2012) |

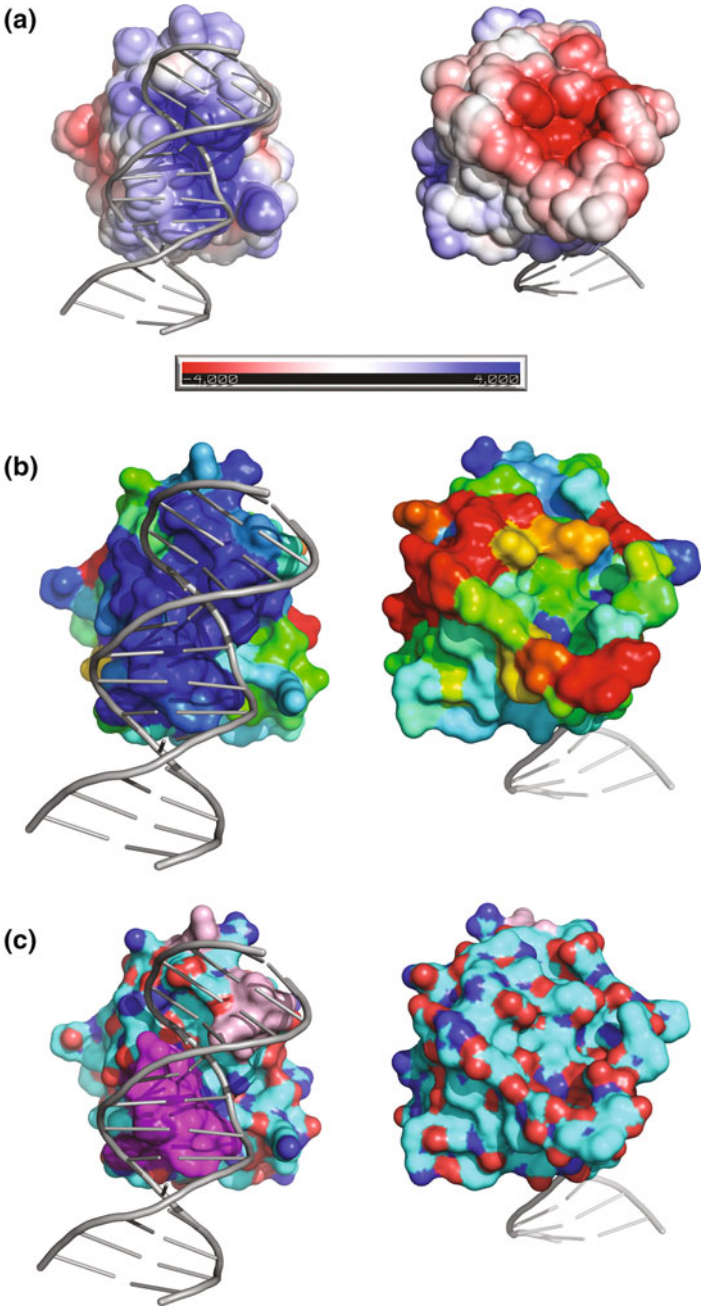
In that case the presence of a large hydrophobic patch was putatively linked to the emulsifying activity (Pandya et al. 2000).

Interesting recent work has quantified the importance of hydrophobic surface patches for enzyme adsorption onto lignin films. The existence of such patches, rather than overall hydrophobic surface area, was found to be predictive of the lignin interaction (and hence inhibition) of enzymes involved in the breakdown of plant cell wall polysaccharides (Sammond et al. 2014). Such findings should enable enzyme cocktails with better performance to be devised or designed.

### *10.2.2 Electrostatics*

Analysis of the electrostatic field of a protein can help explain and predict its function. This is most commonly done by mapping local electrostatic potential to a representation of the protein surface, but overall properties such as the electric dipole and quadrupole moments can also be calculated and have predictive value. The best known application is for DNA or RNA binding proteins: since these nucleic acids have a pronounced negative charge on their sugar-phosphate backbones proteins that bind to them often do so at positively-charged surfaces e.g. (Ohlendorf and Matthew 1985); see Fig. 10.2a. Specialised methods for prediction of patches binding nucleic acids are considered later (Sect. 10.6). Similarly, enzymes with charged substrates may use long-range electrostatic interactions to assist in substrate binding e.g. (Warwicker 1986).

The best known packages for calculation of protein electrostatic properties are APBS, Adaptive Poisson-Boltzmann Solver (Baker et al. 2001), and DelPhi (Rocchia et al. 2002). Both of these interface to the Chimera molecular visualisation software (Pettersen et al. 2004), while access to APBS for PyMOL users is facilitated by a plug-in (http://www.pymolwiki.org/index.php/Apbsplugin). Each is also available via a server (Table 10.1) with visualisation after calculation. Additional, browser-based visualisation of electrostatic properties is offered by a variety of servers, including eF-site/eF-surf for deposited PDB structures or uploaded files respectively (Kinoshita and Nakamura 2004), Patch Finder Plus (Shazman et al. 2007), webPIPSA (Richter et al. 2008) and Bluues (Walsh et al. 2012). Notably, the latter two allow useful comparisons to be made. Bluues allows for modelling of a point mutation in a known structure and analysis of the consequent changes in electrostatic potential and other factors such as pKa values of nearby residues. WebPIPSA focuses on potentials alone, but allows easy comparison of multiple structure with sophisticated methods for analysing the results. This comparative approach can be used to rationalise and predict enzyme kinetic parameters (Gabdoulline et al. 2007).


(a)
-4.000
4.000
(b)
(c)

◂**Fig. 10.2** Predictions of binding sites for the DNA-binding protein *Magnaporthe oryzae* PCG2 (PDB code 4ux5). In each component, the DNA duplex is shown as *grey cartoon* and the protein as its solvent-accessible surface (**a**) or molecular surface (**b**, **c**). Each component contains a component orientated to face the DNA-binding site on the *left* and, for comparison, a view of the opposite face of the molecule on the *right*. **a** Electrostatic analysis using APBS (Baker et al. 2001) and visualised using the PyMOL plugin. The protein surface is coloured *blue* (positive charge) or *red* (negative charge). The unit of the scale is $k_BT/e_c$ where $k_B$ is the Boltzmann constant, $T$ is the temperature, and $e_c$ is the charge of the electron. **b** Conservation analysis from the ConSurf server (Ashkenazy et al. 2010; Goldenberg et al. 2009). The surface is coloured using a spectrum from *blue*, most conserved, to *red*, least conserved. **c** Combined results from two structure-based methods to predicted DNA-binding residues. *Magenta* shows predictions from both DISPLAR (Tjong and Zhou 2007) and DR_BIND (Chen et al. 2012b), *light purple* residues predicted by DISPLAR alone and *deep purple* residues only predicted by DR_BIND. The advantage of applying multiple methods is immediately evident since each makes unique correct predictions

### *10.2.3 Sequence Conservation*

Sequence conservation is one of the most powerful and general factors for inferring the importance of residues in a protein: positions at which mutations have been disallowed or restricted to similar amino-acids over long evolutionary time periods can be generally interpreted as being important for protein folding or function. Localization of such positions at the protein surface suggests a potential functional role in inter-molecular interactions, rather than the structural role that would be inferred for a conserved, buried position. Of particular interest may be residues conserved only among orthologous subsets within a large superfamily, not across the whole superfamily, often called specificity-determining residues (SDRs) (reviewed in Chagoyen et al. 2016).

Perhaps the best known protein conservation tool is ConSurf (Ashkenazy et al. 2010, 2016), also available with precalculated results for PDB entries as ConSurf-DB (Goldenberg et al. 2009). The method involves a pipeline of a sequence database search (that can be initiated by a structure of interest), sequence alignment, phylogenetic tree inference and tree based estimation of conservation from site-specific evolutionary rates. Example results are shown in Figs. 10.2b and 10.3b. The server allows choices to be made at all steps including, crucially, selection of database search results to include in the calculation. This selection allows tailoring of the search to either encompass all members of a superfamily or to restrict the calculation to members of a defined sub-group within a superfamily thereby highlighting SDRs. A sophisticated but rapid treatment of sequence conservation is provided by the FuncPatch server (Huang and Golding 2015). It can exploit the reasonable assumption that nearby residues in a functional patch will be subject to similar substitution rates. The user provides an alignment which again, as the authors point out, can alternatively contain a wide selection of sequences or focus on a sub-group of orthologues.

The approach, of finding positions whose variation correlates with phylogenetic tree structure—potential SDRs, has been well-explored. The well-known Evolutionary Trace (ET) method (Lichtarge et al. 1996), available at a server

(Table 10.1), is fundamentally a sequence-based method, but the results can be visualised in the context of a representative known structure (Nemoto et al. 2013). The ET Viewer server (Morgan et al. 2006) allows access to precalculated results for structures in the PDB online or by a PyMOL session download and associated plugin (http://mammoth.bcm.tmc.edu/pyetv). Extensions to the ET method propose 3D motifs composed of predicted functional residues that can be compared to other structures and databases by the methods discussed in Chap. 11 (Ward et al. 2009). Similar methods, available at the SDPsite server, automatically detect spatial clusters of both broadly conserved positions and sub-tree conserved putative SDRs using a known protein structure thereby improving prediction of functional sites (Kalinina et al. 2009).

The INTREPID algorithm (Sankararaman and Sjolander 2008), also available as a server (Sankararaman et al. 2009), is broadly comparable to the ET method, but applies a different statistical model to analysis of the tree, enabling the effective use of more divergent sequence alignments. Accordingly, it proved to predict functional residues better than the ConSurf and ET algorithms of the time. INTREPID data are widely used as a component of multi-factorial methods for catalytic site prediction.

Interesting recent work describes how amino-acid positions conserved within a group of orthologues can be distinguished from those conserved in a broader family of proteins (Lee et al. 2015). It involves one filter based on statistical comparison of putative within-orthologue conserved positions to residues found more broadly, discarding those commonly found elsewhere, and another that exploits structural knowledge by selecting solvent-exposed positions. This results in a set of specificity-conserved sites, in one case picking out heparin binding sites found only in antithrombins, but not more broadly in the larger serpin family (Lee et al. 2015).

A less common means of assessing the importance of protein residues is to measure the strength of purifying selection pressure on them at the DNA level as the ratio of non-synonymous to synonymous substitutions (Nei and Gojobori 1986; Suzuki 2004). Although much more labour-intensive, this approach has some advantages, particularly an ability to work effectively with smaller numbers of sequences. One interesting application mapped surfaces of membrane proteins, revealing stronger selection pressure on interfaces for some bound lipids e.g. cholesterol than on lipid-facing surfaces in general (Adamian et al. 2011).

### *10.2.4 Surface Atom Triplet Propensities*

Two comparable methods exploit a purely statistical description of the protein surface in terms of protein triplets or triangles, composed of 13 different atom types. The STP method (Mehio et al. 2010) calculates points at which a probe (of 1.4 Å radius, simulating a water molecule) simultaneously touches three protein atoms. Purely empirically, it is observed that triplets composed of certain combinations of atom types are more commonly found at ligand binding sites than others. Triplet propensities to be found at binding sites drive a rapid calculation the result of which

is visualised as a colour-coded surface. Interestingly, some preferences for certain triplets of particular ligand atom types could be discerned. The results were comparable to or better than site finding through pocket finding by geometric or energetic criteria (see below). These statistics are independent of geometric or evolutionary information: thus they require only a single experimentally-determined protein structure, and are equally applicable to surface patches, shallow peptide-binding sites or pockets (Mehio et al. 2010).

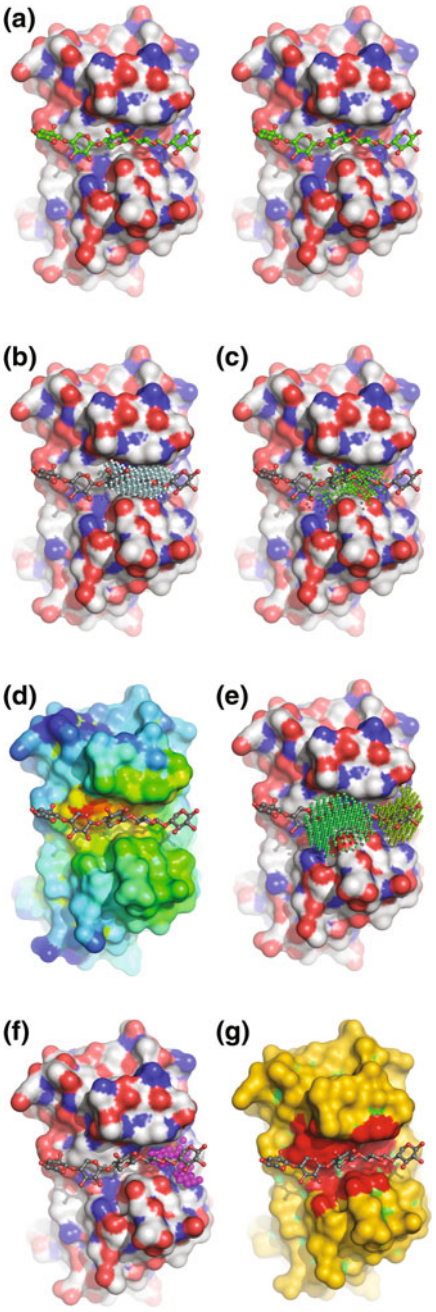
The later LISE method (Xie and Hwang 2012; Xie et al. 2013) finds triangles of certain maximum dimensions that would be capable of contacting two ligand atoms. As with STP, propensities for each triangle to be located at binding sites are calculated but, in contrast, conservation information from PSI-BLAST is employed and makes an additional small contribution to success. Again different to STP's surface mapping, the results are grid points above the protein surface predicted to be likely sites of occupancy by ligands. Use of these two methods is illustrated in Fig. 10.3d, e.

### *10.2.5 Multiple Properties*

Recognition of the many distinct special properties that functional patches may have, in comparison to protein surface in general, led to the development of HotPatch which finds unusual patches based on diverse criteria (Pettit et al. 2007). These include hydrophobicity, electrostatic properties, surface roughness and concavity. Different properties proved to be effective for detecting different kinds of functional patches, benchmarked against different classes of enzymes and proteins that binding different classes of ligands, nucleic acids, lipids, carbohydrates etc. For most challenges, neural networks capturing multiple properties out-performed single characteristics in pinpointing functional patches.

## 10.3 Pockets

It has long been recognised that protein pockets or cavities are well-suited to small molecule binding, the invagination of the protein surface providing increased opportunities for intermolecular interactions for affinity or specificity of binding. For enzymes, such an arrangement can also allow the necessarily precise orientation of substrate at the catalytic site, as well as advantageously lowering the local dielectric constant (Fersht 1985). The power of a purely geometric analysis of the protein surface for identifying sites of interest was amply illustrated by a survey of cavities in single-chain enzyme structures (Laskowski et al. 1996) showing that in a remarkable 83% of cases the catalytic site was located in the largest cavity. This correlation was particularly strong where the cavity was unusually large compared to the other cavities found in the same protein.


(a)
(b)
(c)
(d)
(e)
(f)
(g)

◂**Fig. 10.3** Predictions of binding sites for a Family 15 Carbohydrate Binding Module from *Cellvibrio japonicus* (PDB code 1gny). **a** Stereo view of xylopentaose (*balls* and *sticks*) bound to the protein surface (*coloured* by atom type). The carbohydrate is bound partly within a surface crevice, interactions with the protein including both hydrogen bonds and hydrophobic interactions with solvent-exposed aromatic residues. In this case only a small number of homologous sequences are available so that evolutionary conservation does not help locate the binding site. The GHECOM (**b**; Kawabata 2010) pocket detection picks out the small binding crevice, while the metaPocket results (**c**; Zhang et al. 2011) extend the prediction to cover a slightly more open region of the binding site. In each case, small spheres indicate predicted pocket regions. The surface statistics methods STP (**d**; Mehio et al. 2010; surface coloured from *red* (high) to *blue* (low) propensity) and LISE (**e**; Xie and Hwang 2012; Xie et al. 2013; two sites with volumes represented by differently *coloured* small spheres) both provide very useful complementary predictions, each covering parts of the binding site not detected by pocket predictions. Detection of carbohydrate binding sites by using a hydroxyl group probe also works well: second cluster from SiteHound (**f**; Hernandez et al. 2009; site represented by *small magenta spheres*) and the top patch from the ISMBLab server (**g**; Tsai et al. 2012; predicted binding residues coloured *red*) both correspond to the experimentally observed site

### *10.3.1 Geometric Descriptions of Pockets*

The geometric definition of surface pockets has a long history, but new algorithms continue to emerge, seeking greater efficiency and offering useful capabilities e.g. for splitting pockets into distinct sub-pockets. The oldest and simplest pocket finding methods placed the protein structure in a grid and then sought grid points that were defined in some way as being within the protein. In Pocket (Levitt and Banaszak 1992), for example, cavities were defined as empty regions enclosed on both sides (along some dimension) by protein. LIGSITE (Hendlich et al. 1997) introduced sampling along diagonals and a finer grid size to smooth the surface of the pockets obtained and to reduce the orientation-dependence of the results. This method was later developed into the popular LIGSITE$^{CSC}$ (Huang and Schroeder 2006; see Sect. 10.3.3 and Table 10.1). A recent grid-based method, KVFinder, adopts a two probe approach, defining pockets as volumes accessible to a small probe—given a radius of 1.4 Å to emulate a water molecule—but not to a variably sized larger probe (Oliveira et al. 2014). Available via a PyMOL plug-in, the method notably allows users the ability to segment pockets into distinct functional sub-pockets and measure their volume.

Grid-based methods have the disadvantage of producing slightly different volume measurements according to grid spacing and the orientation of the structure with respect to a frame of reference. Grid-free methods can be divided into those that use probes to define the protein surface and those based on representing the surface using Voronoi diagrams. SURFNET (Laskowski 1995) defines cavities by placing spheres between pairs of protein atoms and progressively reducing the sphere diameter until protein atom overlaps are eliminated. Spheres for which the radius drops below 1 Å are discarded, the retained remainder, with radii between 1 and 4 Å, defining the surface cavities. SURFNET is available through the integrated ProFunc server (see Chap. 13) and contributes to metaPocket 2.0

(Zhang et al. 2011), and a version of the software can be accessed from within Chimera. The PASS method (Brady and Stouten 2000) iteratively covers the protein surface with probes, eliminating between each cycle probes that are less buried. This burial is defined using the number of protein atoms nearby, a number that will be lower at convex surface regions than in surface cavities. At the end of the process, clusters of probes remain within protein pockets. It is available for download and accessible through the metaPocket 2.0 server. Similarly to KVFinder, the PHECOM method (Kawabata and Go 2007) employs two probes. The protein surface is covered with a small probe after which locations that overlap with positions that could be taken up by a large probe are eliminated. In this way only buried small probe locations, inaccessible to a large probe, are spared. Interestingly, different radii of the larger probe were optimal for detection of binding sites for different ligand classes. A grid-based extension of the method GHECOM (Kawabata 2010) is available as a server (Table 10.1). Its use is illustrated in Fig. 10.3b.

Approaches based on Voronoi diagrams include CASTp (Liang et al. 1998). This works with a mathematically equivalent Delaunay triangulation representation of the protein surface, identifying pockets as collections of empty triangles by the discrete flow method. Conveniently, the method produces a clear description of the boundary between bulk solvent and pocket and, lacking a grid basis, is rotationally invariant. In addition to a server (Dundas et al. 2006) (Table 10.1), the method has a useful PyMOL plug-in (http://sts.bioe.uic.edu/castp/pymol.php) while precalculated server results for structures deposited in the PDB can be read directly into Chimera. A recent method Fpocket (Le Guilloux et al. 2009) works with alpha spheres, spheres that contact four atoms but which contain no atom. Such alpha spheres of very small radius could be located within the protein, large spheres only accommodated exterior to the protein. In an interesting intermediate range of radii these spheres can occupy surface cavities and so their clustering can be used to identify pockets. A final ranking follows the elimination of small and hydrophilic pockets. The method relates to Voronoi diagrams since Voronoi vertices correspond to the alpha sphere centres. An Fpocket server extends the methodology using the MDpocket package to allow tracking of pocket volumes during Molecular Dynamics trajectories (Schmidtke et al. 2010). Cavities can be automatically identified or the user may focus on a region of particular interest.

### *10.3.2 Channels and Tunnels*

A number of servers are specialised in the related tasks of finding the tunnels that link enzyme catalytic sites to bulk solvent or the transmolecular channels that allow transit through membrane transporters or pores. These analyses can facilitate an in-depth understanding of protein function, but can also be predictive since the size and characteristics of membrane channels, for example, will naturally define their permeability to different ligands.

The MOLEonline 2.0 (Berka et al. 2012), MolAxis (Yaffe et al. 2008) and BetaCavityWeb (Kim et al. 2015) servers can each accept a single structure and then define tunnels leading to buried protein cavities, either automatically identified or specified by the user. The residues lining the tunnels are specified and an estimate made of the bottleneck radius i.e. the radius of the tunnel at its most restricted point. The corresponding MOLE 2.0 version is also available as a standalone program and as a PyMOL plugin (http://webchem.ncbr.muni.cz/Platform/App/Mole). Application of MOLE 2.0 has provided interesting insights into the frequency and composition of channels, and how channel characteristics can be correlated to protein function (Pravda et al. 2014; Sehnal et al. 2013). PoreWalker (Pellegrini-Calace et al. 2009) is a tool specifically designed for identifying channels through transmembrane proteins. It provides profiles of channel radius along the pore axis and measurements of a potassium channel supported the proposed existence of a selectivity filter allowing passage only to dehydrated $K^+$ ions. The transient and dynamic nature of these tunnels and channels means that analytical tools will ideally accept multiple conformations from, for example, Molecular Dynamics simulations. This is the case for the CAVER software, available as a freestanding application CAVER Analyst (Kozlikova et al. 2014) and via a PyMOL plug-in (http://www.caver.cz/). Impressively, analysis of the dynamic behaviour of bottleneck radii in tunnels determined for a haloalkane dehalogenase structure was in very good agreement with kinetic data previously obtained for proteins mutated at bottleneck positions (Chovancova et al. 2012).

### *10.3.3 Distinguishing Functional Pockets*

Once discovered geometrically, the key question then becomes how to further distinguish between pockets that are likely to bind ligands and those that have arisen by chance. This is likely to be a particular issue for smaller pockets since atypically large pockets are very likely to be functional (Laskowski et al. 1996). Identifying which (sub-)pockets are capable of binding small ligands has two distinct applications, one in predicting or understanding the binding of natural ligands, the second in prediction of druggability. Druggability refers to the ability of proteins, and especially the functionally important regions of those proteins, to bind to small molecules with drug-like properties. These properties are somewhat distinct from those of naturally occurring compounds (Feher and Schmidt 2003) and so some specialised predictors have been developed as discussed in Sect. 10.7.

Some of the methods used to predict functional pockets among candidate surface invaginations are the general ones previously encountered above. Thus, mapping of evolutionary conservation on to the protein surface will highlight functionally important pockets through their conservation. This was implemented, for example, by the ConCavity method (Capra et al. 2009) and by the LIGSITE$^{CSC}$ algorithm (Huang and Schroeder 2006) each of which was demonstrated to perform better than a purely structural approach. webPDBinder (Bianchi et al. 2013) adds a third,

novel property to geometry and conservation in making its predictions of ligand-binding pockets. This represents the frequency with which similarities between residue triplets in the protein structure of interest, each residue being represented by three main chain and three side chain points in three dimensions, are found in databases of binding and non-binding pockets. Triplets over-represented in binding versus non-binding pockets are inferred as making the pocket in which they are found in the structure of interest more likely to be ligand binding.

The method Q-Sitefinder (Laurie and Jackson 2005) implemented an energy-based, rather than purely geometric, discovery of pockets. The protein of interest was placed in a grid and van de Waals interaction potential energy calculated for all intersections that do not overlap with the protein. Clustering of the positions with the most favourable interaction energy produces a ranking of pockets in terms of probability of ligand binding. Comparable probe-based interaction methods can be used, with probes of diverse chemical types, to help predict specific binding sites for certain kinds of ligands e.g. carbohydrates or phosphorylated compounds (see Sect. 10.6).

Another factor, not yet mentioned, that distinguishes ligand binding pockets is their depth: most such cavities will possess residues that are both solvent accessible and deep i.e. determined to be distant from bulk solvent. This approach, optionally including additional sequence conservation information, is available at the Depth server (Tan et al. 2013) (Table 10.1). Similarly, the electrostatic properties of identified pockets have been given special treatment by the VASP-E method (Chen 2014). This defines 3D volumes representing electrostatic potentials and can compare volumes between homologous sequences, rationalising observed substrate specificities and allowing prediction of key specificity-determining residues. A further property that distinguishes functional properties is desolvation. The dPredGB method (Schneider and Zacharias 2012), for example, has been shown to improve detection of binding pockets by adding desolvation to purely geometric pocket finding criteria. Finally, it is worth mentioning the MetaPocket server (Zhang et al. 2011) which integrates results from several distinct algorithms (see also Fig. 10.3c). These largely employ distinct algorithms for geometry-based discovery of pockets, but some methods additionally factor in conservation and physicochemical properties.

## *10.3.4 Predicting Ligands for Pockets*

### 10.3.4.1 Pocket Matching

Some methods go further than predicting ligand binding and attempt to predict *which* ligands are likely to bind to a given pocket or sub-pocket. This can involve an attempt to spot broad similarities between new pockets of interest and others already characterized as binding a particular ligand, as reviewed in (Jalencas and Mestres 2013). These programs are comparable to the 3D motif methods discussed

in Chap. 11. A recent review has examined small molecule binding pockets on a large scale, considering how a single pocket can recognise diverse ligands or, conversely, how a single ligand can bind to similar or quite different pockets on different proteins (Gao and Skolnick 2013). Among pocket matching methods with currently available servers, SuMo (Jambon et al. 2003) describes and matches 3D arrangements of chemical groups in a fashion that is independent of both amino-acid identity and residue order. ProBIS represents functional groups of protein surfaces as graphs and produces results that identify the superimposable sub-graphs between graphs for query and database proteins (Konc and Janezic 2010). A recent extension, ProBIS-ligands (Konc and Janezic 2014), uses matching binding sites to predict the binding mode of a ligand for a protein of interest based on the pose observed in another (see also below). SMAP works by aligning profiles that represent binding sites in a sequence-order independent fashion (Xie et al. 2009). Importantly, recent developments demonstrate improvements in matching known binding sites in the PDB to homology models (Brylinski 2014), using sequence-order independent binding site alignments, with tolerance of modelling-induced distortions in the sites. A particularly sophisticated treatment of pocket characteristics is implemented by Patch-Surfer (Sael and Kihara 2012). 3D Zernike descriptors are used to efficiently capture characteristics of circular patches within putative ligand-binding pockets enabling rapid database searching. These characteristics include geometric properties (surface shape and visibility i.e. concavity or convexity), hydrophobicity and electrostatic potential. Benchmarking confirmed that two important advantages result from this local, patch-based characterisation: identification of pocket similarity even if the two related structures are in different conformations or if the sites are flexible; and ligand prediction based on analogous pockets in the absence of sequence or structure similarity. A recent method, IsoMIF (Chartier and Najmanovich 2015) (also available as a server Chartier et al. 2016) calculates in-pocket molecular interaction fields (MIFs) using six probes—hydrophobic, aromatic, H-bond donor, H-bond acceptor, positive charge and negative charge—and assesses similarities to user-provided pockets or larger databases such as a set of drug-target complexes. Similar MIFs are indicative of potentially similar native binding properties. Finally, although these methods are based on clear physicochemical similarities between evolutionarily unrelated binding sites for similar ligands e.g. mononucleotides (Kinoshita et al. 1999), other work has highlighted the sometimes surprisingly different environments the same compound encounters in different binding sites (Kahraman et al. 2010).

#### 10.3.4.2 Docking for Function Prediction

Just as pockets may be probed with atoms or small groups (e.g. with Q-Sitefinder, see above) they may be targeted for docking of whole molecules. A detailed discussion of small molecule docking algorithms is not appropriate here, but their use to annotate protein function is worth mentioning. This has been established in a series of papers that have docked libraries of molecules to proteins, considering as

putative natural ligands those which are predicted to interact most favourably (Hermann et al. 2006). An early prominent example concerned a member of the amidohydrolase superfamily, of previously unknown function, from *Thermotoga maritima*. Given this superfamily membership, around 4000 potential substrates were identified from the KEGG database (Kanehisa 2002) and modelled in their high-energy tetrahedral transition states. The best-scoring molecules upon docking frequently contained an adenosine substructure leading to the eventual biochemical characterisation of the enzyme in question as a 5-methylthioadenosine/*S*-adenosylhomocysteine deaminase (Hermann et al. 2007). Importantly, other papers demonstrate the value of homology models of target structure for function annotation by docking, in one case to assign substrate specificity among dipeptide epimerases (Lukk et al. 2012). An interesting recent variation of this theme employed docking of multiple potential intermediates in order to better understand specific reaction routes taken by triterpenoid synthases (Tian et al. 2014). In these studies, well-known docking programs such as DOCK (Lang et al. 2009) and Glide (Friesner et al. 2004) have been used but they can also be supplemented with more CPU-intensive quantum mechanical/molecular mechanical (QM/MM) calculations to provide structure-based predictions of enzyme specificity e.g. (Tian et al. 2013).

A number of comparable methods discover potential ligand binding sites through analysis of structures of proteins homologous to that of interest. In GalaxySite (Heo et al. 2014), compounds found bound to PDB structures of homologous proteins are treated as candidate ligands for the new structure of interest and their poses considered as starting points for refinement. Compounds that refine to favourable binding modes are considered as potential true ligands of the compound of interest. More advanced methods like FINDSITE (Brylinski and Skolnick 2009) and COFACTOR (Roy et al. 2012) find additional candidate binding sites in the protein of interest, and thereby additional candidate ligands through, for example, 3D motif matching (Chap. 11). After refinement, candidate ligands are clustered, scored and used for function prediction. As reflected in the name sometimes applied to these tools, Ligand Homology Modelling, performance on modelled rather than experimental structures has been a particular focus, and even poorer models have useful predictive power (Skolnick et al. 2013).

## 10.4 Prediction of Catalytic Residues

As mentioned above, a geometric analysis of the protein surface is surprisingly effective at picking out pockets containing catalytic sites. Within catalytic sites, which would also be expected to be strongly conserved of course, a variety of methods aim to predict the identity of key catalytic residues (Zhang et al. 2009). Among them are methods to spot previously seen arrangements of catalytic residues through the use of motifs (see Chap. 11).

Other methods can exploit the statistical over-representation of certain amino-acids among catalytic residues (Bartlett et al. 2002). Amino-acids with acidic

or basic side chains are the most common with His, having a pKa value near neutrality, being the most over-represented. Small or hydrophobic amino-acids, lacking suitable chemistry in their side chains, rarely participate in catalytic sites. Other individual characteristics that can be used to help predict catalytic residues are relatively low solvent accessibility compared to non-catalytic polar residues (Bartlett et al. 2002), relatively high numbers of contacts with other residues (del Sol et al. 2006), relative centrality (closeness to the protein's centre; Ben-Shimon and Eisenstein 2005) and relatively high rigidity (Sacquin-Mora et al. 2007; Yuan et al. 2003). Striking recent success has also been achieved by a measure of relative side chain orientation, catalytic residues tending to point towards the centre of the catalytic site (Chien and Huang 2012).

Two interesting methods have detected catalytic residues through in silico calculation of biophysical properties. The first detects catalytic residues through their often being charged residues (see above) positioned in an electrostatically destabilising environment (Elcock 2001). Thus, residues with the most positive calculated electrostatic free energies are often found to be catalytic or otherwise functionally important. The second method, theoretical microscopic titration curves (THEMATICS), exploits the observation that catalytic residues possessing ionisable groups are often positioned in environments that perturb their protonation behaviour. pKa values are often found to deviate from those typically observed for the amino-acid in question and perturbed residues can persist in a given partially protonated state over an unusual span of pH ranges (Ondrechen et al. 2001).

Current methods for predicting catalytic residues typically exploit several distinct characteristics, exploiting statistical or Artificial Intelligence means to produce improved consensus predictions. For example, the MEPI server (Han et al. 2012) profitably combines individual Dscore and MEscore, measuring distance from protein centre and residue propensities in the microenvironment of a given residue respectively, producing a MEDscore. The performance of this measure was comparable to a conservation-based CONscore with which it could be further combined to produce a best-performing CMEDscore. The recently introduced EXIA2 server (Lu et al. 2014) achieves excellent performance by combining a novel metric, based on the tendency of catalytic residue side chains to point towards the centre of the catalytic site, with conservation information (Chien and Huang 2012). The DISCERN server uses INTREPID phylogenomic data (see Sect. 10.2.3), catalytic site residue propensities, characteristics of neighbouring sites, location in a pocket (calculated with LIGSITE$^{CSC}$; see Sect. 10.3.3), centrality, solvent accessibility, rigidity (inferred from B-factors of crystal structures) and secondary structure (Sankararaman et al. 2009). These factors, weighted after benchmarking, are combined in a linear fashion using logical regression. The THEMATICS protocol is now bundled with pocket detection and optional conservation analysis in the POOL server (Somarowthu et al. 2011). Combination of these distinct characteristics using a Machine Learning approach boosts performance over THEMATICS alone to levels above comparable predictors.

## 10.5 Protein-Protein Interfaces

Many protein functions depend on interactions, either obligate or transitory, between two or more proteins. For the purposes of this chapter it is useful to distinguish protein-protein interactions, between two folded protein domains, and protein-peptide interactions where a linear motif in one partner interacts with a folded domain in another. Although both types of interface would be expected to be detectable using sequence conservation, in other ways they differ. Protein-protein interfaces are known to be relatively flat, somewhat more hydrophobic than protein surface in general, and to have distinct residue propensities such as a relative over-representation of aromatic amino-acids (Jones and Thornton 1997). In contrast, linear motifs typically bind to a cavity on their partner, with the interface producing more hydrogen bonds than protein-protein interfaces, and showing a clear over-representation of Ile and Leu, as well as aromatic residues (London et al. 2010). Indeed, peptides tend to bind to the largest pocket on the surface of their partner (London et al. 2010) so that many of the cavity detection methods mentioned above can be expected to work well for detecting putative interfaces. The geometry-independent method STP (Mehio et al. 2010) has also been shown to predict protein-peptide interfaces.

For protein-protein interactions, a number of predictors are available to try to predict the position and composition of interfaces from structural information for one partner (recently reviewed by Esmaielbeiki et al. 2016). Typically they integrate a selection of characteristics each of which helps differentiate interfaces in some way from the remaining protein surface. These include sequence conservation, solvent accessibility, hydrophobicity, amino-acid propensities and shape. The cons-PPISP server (Zhou and Shan 2001), for example, uses position-specific sequence profiles to represent residue propensities and sequence conservation, and solvent accessibility data. Information for a given residue is considered along with similar data for its sequential neighbours. The meta-PPISP (Qin and Zhou 2007) is a metaserver that takes cons-PPISP predictions, along with those from PINUP and Promate methods, to produce better joint predictions. CPORT is another consensus predictor that performed better than its six component contributions (de Vries and Bonvin 2011). Its predictions were used to inform data-driven 3D docking of protein structures (see Chap. 8) with HADDOCK; results comparing well to ab initio docking predictions. A different approach is taken by the PRISE server (Jordan et al. 2012) which defines a ‘structural element’ for each surface residue of a protein, considering atomic composition, residue type and solvent exposure of the central residue plus its neighbours on the surface. ‘Structural elements’ are compared to a database of known binding sites to predict whether they are likely to form part of an interface on the protein of interest. The VORFFIP method (Segura et al. 2011) again considers residues in their surface environment (defined using Voronoi diagrams) and integrates a variety of metrics from structural features (solvent accessibility, hydrophobicity and so on), energy predictions, sequence conservation

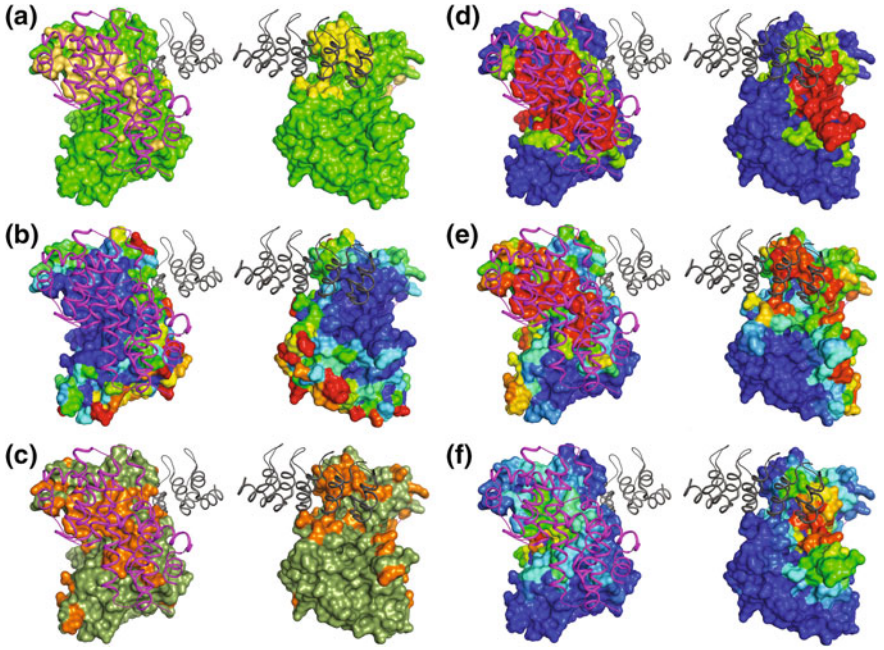



**Fig. 10.4** Predictions of binding sites for human cyclin-dependent kinase 6 in complex with cyclin (*magenta cartoon*) and the p18 kinase inhibitor (*grey cartoon*). In each component two views are shown, the *left* facing the cyclin binding site, the *right* focusing on the inhibitor site. **a** Interface residues at the two sites coloured shades of *yellow*. **b** Results of ConSurf analysis (Ashkenazy et al. 2010; Goldenberg et al. 2009). The surface is coloured using a spectrum from *blue*, most conserved, to *red*, least conserved. The *large blue* areas distinct from the protein interfaces contain the ATP-binding site and catalytic residues. **c** Residues predicted as protein interface by PRISE (Jordan et al. 2012) are coloured *orange*. **d** CPORT analysis (de Vries and Bonvin 2011) with predicted interface residues *red* and neighbouring possible interface residues *green*. Non-interface residues are coloured *blue*. **e** and **f** show results from VORFFIP (Segura et al. 2011) and meta-PPISP (Qin and Zhou 2007), respectively, with surface residues according to predicted interface propensity—*red*, strongly predicted to form a protein interface, to *blue*, not predicted

and crystallographic B-factors. Example results for several servers for protein-protein interface prediction are shown in Fig. 10.4.

## 10.6 Other Specialised Binding Site Predictors

As well as the broad range of general methods already discussed, certain ligands, through their importance or distinct features, have inspired the development of specialised predictors. First among these ligands is duplex DNA (Ding et al. 2010) and, to a lesser extent, RNA (Puton et al. 2012; Zhao et al. 2013). These nucleic

acids share a highly negatively charged backbone and so identifying patches with a distinct positive charge is often the starting point for predicting DNA- or RNA-binding sites (Jones et al. 2003). Currently available servers adopt distinct strategies. DNA-binding proteins, for example, finds conserved patches and then predicts DNA binding capacity using patch- and whole protein-derived measures of conservation, electrostatics and secondary structure (Nimrod et al. 2009). The recent BindUP server (Paz et al. 2016) finds patches, both positive and negative, feeding information on the largest of these and other structural features to a Support Vector Machine classifier. DISPLAR uses residue interface propensities, conservation information and solvent accessibility, supplementing these data for a given residue with those figures for its near spatial neighbours (Tjong and Zhou 2007) (see Fig. 10.2c for an example). DR-bind considers solvent exposure to target surface patches but not pockets, subsequently employing both electrostatics and conservation information (Chen et al. 2012b) (see Fig. 10.2c). The DP-dock server predicts which residues of a DNA-binding protein contact DNA through explicit docking of its structure with B-form DNA (Gao and Skolnick 2009), a more compute-intensive approach that somewhat outperforms DISPLAR. For RNA, the KYG server for prediction of RNA-binding interfaces uses propensities for different amino-acids to be found at known interfaces—calculated both for individual amino-acids and for doublets—as well as PSI-BLAST derived information on conservation (Kim et al. 2006). The recently introduced aaRNA predictor (Li et al. 2014b) uses structure-based data derived from geometry, conservation and residue composition. Unusually, it also integrates sequence-based predictive methods and can apply structure-based predictions from homology models when experimental structural data are not available. The very recent NABind server (Sun et al. 2016) detects RNA binding sites using electrostatic potential and interface triplet propensities (see also 10.2.4). Finally, the RBscore server (Miao and Westhof 2015), designed for RNA binding site prediction but also demonstrated to work well for DNA-binding sites, first assesses individual residues using electrostatic potential, solvent accessibility and sequence conservation. Residue binding probability is then predicted using neighbouring network based scoring.

Carbohydrate-binding sites are known to be enriched in certain amino-acids (Malik and Ahmad 2007; Taroni et al. 2000). Aromatic residues interact with the relatively flat surface of carbohydrate monomers while other common residues such as Arg and Asp can form bidentate hydrogen bonds with the hydroxyl groups of the carbohydrate ligand. These propensities were used to predict surface patches as candidate carbohydrate binding sites (Taroni et al. 2000). The InCa-SiteFinder algorithm (Kulharia et al. 2009) used these propensities in conjunction with Q-SiteFinder prediction (see Sect. 10.3.3). Most recently, 3D probability densities of interacting atoms have been used to obtain excellent results (Tsai et al. 2012) by one of the few methods available as a server (see Table 10.1 and Fig. 10.3f). This methodology, first applied by the authors to predict protein-protein interfaces (Chen et al. 2012a) (see also Sect. 10.5) has also found use for prediction of binding sites for other ligands—fatty acids (Mahalingam et al. 2014a) and flavin mononucleotide (Mahalingam et al. 2014b). A similarly flexible framework for discovery of binding

sites for particular classes of ligands is offered by EasyMIFS and SiteHound (Ghersi and Sanchez 2009), the former producing interaction energy maps for a given probe over the protein surface, the latter clustering the results to derive predicted binding sites as top ranking clusters. A particular application was the prediction of binding sites for phosphorylated ligands, including phosphorylated peptides, sugar phosphates and ATP (Ghersi and Sanchez 2009). For most ligand classes EasyMIFS in combination with SiteHound provided the best predictions. Example results for SiteHound are shown in Fig. 10.3e.

## 10.7 Medicinal Applications

As mentioned above, the properties of drugs differ somewhat to those of natural products (Feher and Schmidt 2003). Accordingly, some methods have been optimised for the distinct purpose of detecting pockets with suitable properties for binding drug-like molecules. The Fpocket method mentioned above has been extended to include a druggability score (Schmidtke and Barril 2010) based on hydrophobic and polarity measures of the pockets identified. Fpocket druggability scores are reported in the pocket PDB files downloaded from the server with scores greater than 0.5 indicating predicted druggability. The DoGSiteScorer server (Volkamer et al. 2012) automatically detects pockets in a protein structure provided and scores druggability using a support vector machine to process both global and local descriptors of the pocket. In the former category are characteristics such as size, shape and hydrophobicity, but local features such as frequency of interactions of certain residue types are shown to enhance predictive value. Also unusually, the server can decompose pockets into subpockets automatically. These methods appear to have comparable predictive value. Another druggability method, FTMAP, operates as an in silico analog of experimental methods for finding druggable pockets by exposure of a protein to small organic molecules (Ngan et al. 2012). Small probe molecules are rigid body docked to the protein surface, energy minimised and scored. Clusters of probes represent predictions of druggable loci on the protein surface.

Protein-protein interfaces, being more planar and larger than most surface pockets are often considered less druggable (Arkin and Wells 2004). However, the realisation that a few key residues—the ‘hot spots’—often contribute most of the interaction energy (Bogan and Thorn 1998) encourages optimism that these specific regions can be targeted by small molecules. Some methods take an experimentally determined interface and predict which residues contribute most to the interaction. Here, only the few methods for the distinct and more challenging task of predicting hot spots for a single structure, in the absence of its partner, are considered. Similar clustering of docked small molecules as with FTMAP has recently been shown not only to find druggable hot spots within interfaces, but also in fact to be useful for the preceding step of interface identification (Li et al. 2014a) (see above). Recently, unsuspected allosteric sites have emerged as loci for structure-based drug design

e.g. (De Smet et al. 2014). Since protein molecules are dynamic, and druggable pockets can transiently form and disappear, more compute-intensive Molecular Dynamics-based methods have also been devised. One protocol studied the interaction of hydrophobic solvent probes mixed with water for a series of test cases. Druggable sites were identified first by finding regions of the protein enriched in probe binding vs a reference simulation and then by clustering the resulting probe interaction spots (Bakan et al. 2012). Interesting recent work has discovered that allosteric sites can be identified through analysing calculated energy differences between residue pair interactions, in Molecular Dynamics simulations (see Chap. 12) of paired holo and apo structures (Ma et al. 2014). Categorising these differences in terms of their magnitude and comparing the sizes of the resulting categories allowed a predictive metric for allosteric sites to be produced. Based on this finding, a pipeline that finds cavities, scores their druggability and then compares residue-residue interactions in the two allosteric states was used to predict druggable allosteric sites, with data supporting the proposed functionality of some already being available (Ma et al. 2014). An obvious current limitation of the method is the requirement for reliable structural information on both allosteric states.

It is worth noting that pocket detection and comparison as discussed here has other potential roles in drug design and discovery. Drug promiscuity, binding to more than one target, is known to depend more on binding site similarity than any properties of the compound in question (Haupt et al. 2013). This promiscuity can be a problem—the undesirable binding of a drug to sites other than that originally targeted—but may also be advantageous: it underlies efforts to repurpose existing drugs for new diseases with considerable advantages compared to discovery and testing of novel compounds (Novac 2013). Drug promiscuity can be rationalised and predicted through binding site pocket comparison (Vulpetti et al. 2012). Interesting recent work using molecular interaction fields implemented in the FLAP software, could explain off-target binding of oestrogen receptor modulators to an ion channel ATPase, for example, as well as the polypharmacology of Nilotinib for oxidoreductase NQO2 alongside the originally targeted kinase (Siragusa et al. 2015). In two papers focusing on large-scale comparison of pockets and drug-target networks in the *Mycobacterium tuberculosis* 'pocketome', known polypharmacology of certain drugs could again be correlated to the statistically significant similarity observed between the respective binding sites (Anand and Chandra 2014; Kinnings et al. 2010).

## 10.8 Conclusions

This chapter has tried to cover up-to-date and readily available methods that help understand and predict protein function from analysis of the surface of a protein structure. Two themes emerge: first that a remarkable variety of properties can be analysed, understood and used predictively; second that individual factors can

frequently be advantageously combined to improve performance. Binding sites were here divided into surface patches and pockets but a clear classification is not necessarily possible or even desirable. Of course, many characteristics such as evolutionary conservation and electrostatic potential can be equally informative for both patches and pockets, as discussed. Furthermore, recent consensus methods such as COACH (Yang et al. 2013) conveniently predict both flatter sites for binding of DNA, for example, as well as pockets for binding small molecules. The geometry-independent STP method (Mehio et al. 2010) is another flexible tool. Encouragingly, several papers have shown that homology models, even those based on relatively distant relationships, can be profitably analysed by the methods described here (Lukk et al. 2012; Skolnick et al. 2013; Yang et al. 2013). Although the PDB continues to expand rapidly, with the total number of deposits having surpassed 100,000, this applicability to in silico models dramatically extends the value of structure-based function annotation methods.

## References

Adamian L, Naveed H, Liang J (2011) Lipid-binding surfaces of membrane proteins: evidence from evolutionary and structural analysis. Biochim Biophys Acta 1808(4):1092–1102

Anand P, Chandra N (2014) Characterizing the pocketome of *Mycobacterium tuberculosis* and application in rationalizing polypharmacological target selection. Sci Rep 4:6356

Arkin MR, Wells JA (2004) Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. Nat Rev Drug Discov 3(4):301–317

Ashkenazy H, Erez E, Martz E et al (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. Nucleic Acids Res 38(Web Server issue): W529–W533

Ashkenazy H, Abadi S, Martz E et al (2016) ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. Nucleic Acids Res 44(W1): W344–W350

Bakan A, Nevins N, Lakdawala AS et al (2012) Druggability assessment of allosteric proteins by dynamics simulations in the presence of probe molecules. J Chem Theory Comput 8(7): 2435–2447

Baker NA, Sept D, Joseph S et al (2001) Electrostatics of nanosystems: application to microtubules and the ribosome. Proc Natl Acad Sci U S A 98(18):10037–10041

Bartlett GJ, Porter CT, Borkakoti N et al (2002) Analysis of catalytic residues in enzyme active sites. J Mol Biol 324(1):105–121

Bateman A, Coggill P, Finn RD (2010) DUFs: families in search of function. Acta Crystallogr, Sect F: Struct Biol Cryst Commun 66(Pt 10):1148–1152

Beadle BM, Shoichet BK (2002) Structural bases of stability-function tradeoffs in enzymes. J Mol Biol 321(2):285–296

Ben-Shimon A, Eisenstein M (2005) Looking at enzymes from the inside out: the proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme-ligand interfaces. J Mol Biol 351(2):309–326

Berka K, Hanak O, Sehnal D et al (2012) MOLEonline 2.0: interactive web-based analysis of biomacromolecular channels. Nucleic Acids Res 40(Web Server issue):W222-W227

Bianchi V, Mangone I, Ferre F et al (2013) webPDBinder: a server for the identification of ligand binding sites on protein structures. Nucleic Acids Res 41(Web Server issue):W308-W313

Bogan AA, Thorn KS (1998) Anatomy of hot spots in protein interfaces. J Mol Biol 280(1):1–9

Brady GP Jr, Stouten PF (2000) Fast prediction and visualization of protein binding pockets with PASS. J Comput Aided Mol Des 14(4):383–401
Brylinski M (2014) eMatchSite: sequence order-independent structure alignments of ligand binding pockets in protein models. PLoS Comput Biol 10(9):e1003829
Brylinski M, Skolnick J (2009) FINDSITE: a threading-based approach to ligand homology modeling. PLoS Comput Biol 5(6):e1000405
Burgoyne NJ, Jackson RM (2009) Predicting protein function from surface properties. In: Rigden DJ (ed) From protein structure to function with bioinformatics, 1st edn. Springer, Berlin, pp 167–186
Capra JA, Laskowski RA, Thornton JM et al (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5(12): e1000585
Chagoyen M, Garcia-Martin JA, Pazos F (2016) Practical analysis of specificity-determining residues in protein families. Brief Bioinform 17(2):255–261
Chartier M, Najmanovich R (2015) Detection of binding site molecular interaction field similarities. J Chem Inf Model 55(8):1600–1615
Chartier M, Adriansen E, Najmanovich R (2016) IsoMIF finder: online detection of binding site molecular interaction field similarities. Bioinformatics 32(4):621–623
Chen BY (2014) VASP-E: specificity annotation with a volumetric analysis of electrostatic isopotentials. PLoS Comput Biol 10(8):e1003792
Chen CT, Peng HP, Jian JW et al (2012a) Protein-protein interaction site predictions with three-dimensional probability distributions of interacting atoms on protein surfaces. PLoS ONE 7(6):e37706
Chen YC, Wright JD, Lim C (2012b) DR_bind: a web server for predicting DNA-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res 40(Web Server issue):W249–W256
Chien YT, Huang SW (2012) Accurate prediction of protein catalytic residues by side chain orientation and residue contact density. PLoS ONE 7(10):e47951
Chovancova E, Pavelka A, Benes P et al (2012) CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol 8(10):e1002708
Connolly ML (1983) Analytical molecular surface calculation. J Appl Cryst 16:548–558
De Smet F, Christopoulos A, Carmeliet P (2014) Allosteric targeting of receptor tyrosine kinases. Nat Biotechnol 32(11):1113–1120
de Vries SJ, Bonvin AM (2011) CPORT: a consensus interface predictor and its performance in prediction-driven docking with HADDOCK. PLoS ONE 6(3):e17695
del Sol A, Fujihashi H, Amoros D et al (2006) Residue centrality, functionally important residues, and active site shape: analysis of enzyme and non-enzyme families. Protein Sci 15(9): 2120–2128
Ding XM, Pan XY, Xu C et al (2010) Computational prediction of DNA-protein interactions: a review. Curr Comput Aided Drug Des 6(3):197–206
Dundas J, Ouyang Z, Tseng J et al (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34(Web Server issue):W116–W118
Elcock AH (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. J Mol Biol 312(4):885–896
Esmaielbeiki R, Krawczyk K, Knapp B et al (2016) Progress and challenges in predicting protein interfaces. Brief Bioinform 17(1):117–131
Feher M, Schmidt JM (2003) Property distributions: differences between drugs, natural products, and molecules from combinatorial chemistry. J Chem Inf Comput Sci 43(1):218–227
Fersht A (1985) Enzyme structure and mechanism. Freeman, New York
Friesner RA, Banks JL, Murphy RB et al (2004) Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. J Med Chem 47(7):1739–1749
Gabdoulline RR, Stein M, Wade RC (2007) qPIPSA: relating enzymatic kinetic parameters and interaction fields. BMC Bioinformatics 8:373

Gao M, Skolnick J (2009) From nonspecific DNA-protein encounter complexes to the prediction of DNA-protein interactions. PLoS Comput Biol 5(3):e1000341
Gao M, Skolnick J (2013) A comprehensive survey of small-molecule binding pockets in proteins. PLoS Comput Biol 9(10):e1003302
Ghersi D, Sanchez R (2009) EasyMIFS and SiteHound: a toolkit for the identification of ligand-binding sites in protein structures. Bioinformatics 25(23):3185–3186
Goldenberg O, Erez E, Nimrod G et al (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res 37(Database issue):D323–D327
Han L, Zhang YJ, Song J et al (2012) Identification of catalytic residues using a novel feature that integrates the microenvironment and geometrical location properties of residues. PLoS ONE 7 (7):e41370
Haupt VJ, Daminelli S, Schroeder M (2013) Drug promiscuity in PDB: protein binding site similarity is key. PLoS ONE 8(6):e65894
Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15(6):359–363, 389
Heo L, Shin WH, Lee MS et al (2014) GalaxySite: ligand-binding-site prediction by using molecular docking. Nucleic Acids Res 42(Web Server issue):W210–W214
Hermann JC, Ghanem E, Li Y et al (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc 128(49):15882–15891
Hermann JC, Marti-Arbona R, Fedorov AA et al (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448(7155):775–779
Hernandez M, Ghersi D, Sanchez R (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. Nucleic Acids Res 37(Web Server issue):W413–W416
Huang B, Schroeder M (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6:19
Huang YF, Golding GB (2015) FuncPatch: a web server for the fast Bayesian inference of conserved functional patches in protein 3D structures. Bioinformatics 31:523–531
Jalencas X, Mestres J (2013) Identification of similar binding sites to detect distant polypharmacology. Mol Inform 32:976–990
Jambon M, Imberty A, Deleage G et al (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins 52(2):137–145
Jones S, Thornton JM (1997) Analysis of protein-protein interaction sites using surface patches. J Mol Biol 272(1):121–132
Jones S, Shanahan HP, Berman HM et al (2003) Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucleic Acids Res 31(24):7189–7198
Jordan RA, El-Manzalawy Y, Dobbs D et al (2012) Predicting protein-protein interface residues using local surface structural similarity. BMC Bioinformatics 13:41-2105-13-41
Kahraman A, Morris RJ, Laskowski RA et al (2010) On the diversity of physicochemical environments experienced by identical ligands in binding pockets of unrelated proteins. Proteins 78(5):1120–1136
Kalinina OV, Gelfand MS, Russell RB (2009) Combining specificity determining and conserved residues improves functional site prediction. BMC Bioinformatics 10:174-2105-10-174
Kanehisa M (2002) The KEGG database. Novartis Found Symp 247:91–101; discussion 101–103, 119–128, 244–252
Kawabata T (2010) Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins 78(5):1195–1211
Kawabata T, Go N (2007) Detection of pockets on protein surfaces using small and large probe spheres to find putative ligand binding sites. Proteins 68(2):516–529
Kim OT, Yura K, Go N (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. Nucleic Acids Res 34(22):6450–6460
Kim J-, Cho Y, Lee M et al (2015) BetaCavityWeb: a webserver for molecular voids and tunnels. Nucleic Acids Res (in press)
Kinnings SL, Xie L, Fung KH et al (2010) The *Mycobacterium tuberculosis* drugome and its polypharmacological implications. PLoS Comput Biol 6(11):e1000976

Kinoshita K, Nakamura H (2004) eF-site and PDBjViewer: database and viewer for protein functional sites. Bioinformatics 20(8):1329–1330

Kinoshita K, Sadanami K, Kidera A et al (1999) Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Eng 12(1):11–14

Konc J, Janezic D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26(9):1160–1168

Konc J, Janezic D (2014) ProBiS-ligands: a web server for prediction of ligands by examination of protein binding sites. Nucleic Acids Res 42(Web Server issue):W215–W220

Kozlikova B, Sebestova E, Sustr V et al (2014) CAVER analyst 1.0: graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. Bioinformatics 30(18): 2684–2685

Kulharia M, Bridgett SJ, Goody RS et al (2009) InCa-SiteFinder: a method for structure-based prediction of inositol and carbohydrate binding sites on proteins. J Mol Graph Model 28(3): 297–303

Lang PT, Brozell SR, Mukherjee S et al (2009) DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 15(6):1219–1230

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–330, 307–308

Laskowski RA, Luscombe NM, Swindells MB et al (1996) Protein clefts in molecular recognition and function. Protein Sci 5(12):2438–2452

Laurie AT, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21(9):1908–1916

Le Guilloux V, Schmidtke P, Tuffery P (2009) Fpocket: an open source platform for ligand pocket detection. BMC Bioinformatics 10:168-2105-10-168

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3):379–400

Lee TW, Yang AS, Brittain T et al (2015) An analysis approach to identify specific functional sites in orthologous proteins using sequence and structural information: application to neuroserpin reveals regions that differentially regulate inhibitory activity. Proteins 83(1):135–152

Levitt DG, Banaszak LJ (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10(4):229–234

Li H, Kasam V, Tautermann CS et al (2014a) Computational method to identify druggable binding sites that target protein-protein interactions. J Chem Inf Model 54(5):1391–1400

Li S, Yamashita K, Amada KM et al (2014b) Quantifying sequence and structural features of protein-RNA interactions. Nucleic Acids Res 42(15):10086–10098

Liang J, Edelsbrunner H, Fu P et al (1998) Analytical shape computation of macromolecules: II. Inaccessible cavities in proteins. Proteins 33(1):18–29

Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257(2):342–358

Lijnzaad P, Berendsen HJ, Argos P (1996) A method for detecting hydrophobic patches on protein surfaces. Proteins 26(2):192–203

London N, Movshovitz-Attias D, Schueler-Furman O (2010) The structural basis of peptide-protein binding strategies. Structure 18(2):188–199

Lu CH, Yu CS, Chien YT et al (2014) EXIA2: web server of accurate and rapid protein catalytic residue prediction. Biomed Res Int 2014:807839

Lukk T, Sakai A, Kalyanaraman C et al (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. Proc Natl Acad Sci U S A 109(11):4122–4127

Ma X, Qi Y, Lai L (2014) Allosteric sites can be identified based on the residue-residue interaction energy difference. Proteins

Mahalingam R, Peng HP, Yang AS (2014a) Prediction of fatty acid-binding residues on protein surfaces with three-dimensional probability distributions of interacting atoms. Biophys Chem 192:10–19

Mahalingam R, Peng HP, Yang AS (2014b) Prediction of FMN-binding residues with three-dimensional probability distributions of interacting atoms on protein surfaces. J Theor Biol 343:154–161
Malik A, Ahmad S (2007) Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct Biol 7:1
Mehio W, Kemp GJ, Taylor P et al (2010) Identification of protein binding surfaces using surface triplet propensities. Bioinformatics 26(20):2549–2555
Miao Z, Westhof E (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. Nucleic Acids Res 43(11):5340–5351
Morgan DH, Kristensen DM, Mittelman D et al (2006) ET viewer: an application for predicting and visualizing functional sites in protein structures. Bioinformatics 22(16):2049–2050
Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3(5):418–426
Nemoto W, Saito A, Oikawa H (2013) Recent advances in functional region prediction by using structural and evolutionary information—remaining problems and future extensions. Comput Struct Biotechnol J 8:e201308007
Ngan CH, Bohnuud T, Mottarella SE et al (2012) FTMAP: extended protein mapping with user-selected probe molecules. Nucleic Acids Res 40(Web Server issue):W271–W275
Nimrod G, Szilagyi A, Leslie C et al (2009) Identification of DNA-binding proteins using structural, electrostatic and evolutionary features. J Mol Biol 387(4):1040–1053
Novac N (2013) Challenges and opportunities of drug repositioning. Trends Pharmacol Sci 34(5): 267–272
Ohlendorf DH, Matthew JB (1985) Electrostatics and flexibility in protein-DNA interactions. Adv Biophys 20:137–151
Oliveira SH, Ferraz FA, Honorato RV et al (2014) KVFinder: steered identification of protein cavities as a PyMOL plugin. BMC Bioinformatics 15:197-2105-15-197
Ondrechen MJ, Clifton JG, Ringe D (2001) THEMATICS: a simple computational predictor of enzyme function from structure. Proc Natl Acad Sci U S A 98(22):12473–12478
Pandya MJ, Sessions RB, Williams PB et al (2000) Structural characterization of a methionine-rich, emulsifying protein from sunflower seed. Proteins 38(3):341–349
Paz I, Kligun E, Bengad B et al (2016) BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. Nucleic Acids Res 44(W1):W568–W574
Pellegrini-Calace M, Maiwald T, Thornton JM (2009) PoreWalker: a novel tool for the identification and characterization of channels in transmembrane proteins from their three-dimensional structure. PLoS Comput Biol 5(7):e1000440
Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605–1612
Pettit FK, Bare E, Tsai A et al (2007) HotPatch: a statistical approach to finding biologically relevant features on protein surfaces. J Mol Biol 369(3):863–879
Pravda L, Berka K, Svobodova Va Ekova R et al (2014) Anatomy of enzyme channels. BMC Bioinformatics 15(1):379
Puton T, Kozlowski L, Tuszynska I et al (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol 179(3):261–268
Qin S, Zhou HX (2007) meta-PPISP: a meta web server for protein-protein interaction site prediction. Bioinformatics 23(24):3386–3387
Ren J, Xie L, Li WW et al (2010) SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. Nucleic Acids Res 38(Web Server issue):W441–W444
Richter S, Wenzel A, Stein M et al (2008) webPIPSA: a web server for the comparison of protein interaction properties. Nucleic Acids Res 36(Web Server issue):W276–W280
Rocchia W, Sridharan S, Nicholls A et al (2002) Rapid grid-based construction of the molecular surface and the use of induced surface charge to calculate reaction field energies: applications to the molecular systems and geometric objects. J Comput Chem 23(1):128–137
Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(Web Server issue):W471–W477

Sacquin-Mora S, Laforet E, Lavery R (2007) Locating the active sites of enzymes using mechanical properties. Proteins 67(2):350–359
Sael L, Kihara D (2012) Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. Proteins 80(4):1177–1195
Sammond DW, Yarbrough JM, Mansfield E et al (2014) Predicting enzyme adsorption to lignin films by calculating enzyme surface hydrophobicity. J Biol Chem 289(30):20960–20969
Sankararaman S, Sjolander K (2008) INTREPID–INformation-theoretic TREe traversal for protein functional site IDentification. Bioinformatics 24(21):2445–2452
Sankararaman S, Kolaczkowski B, Sjolander K (2009) INTREPID: a web server for prediction of functionally important residues by evolutionary analysis. Nucleic Acids Res 37(Web Server issue):W390–W395
Sankararaman S, Sha F, Kirsch JF et al (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26(5):617–624
Schmidtke P, Barril X (2010) Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem 53(15):5858–5867
Schmidtke P, Le Guilloux V, Maupetit J et al (2010) Fpocket: online tools for protein ensemble pocket detection and tracking. Nucleic Acids Res 38(Web Server issue):W582–W589
Schneider S, Zacharias M (2012) Combining geometric pocket detection and desolvation properties to detect putative ligand binding sites on proteins. J Struct Biol 180(3):546–550
Segura J, Jones PF, Fernandez-Fuentes N (2011) Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. BMC Bioinformatics 12:352-2105-12-352
Sehnal D, Svobodova Varekova R, Berka K et al (2013) MOLE 2.0: advanced approach for analysis of biomacromolecular channels. J Cheminform 5(1):39-2946-5-39
Shazman S, Celniker G, Haber O et al (2007) Patch Finder Plus (PFplus): a web server for extracting and displaying positive electrostatic patches on protein surfaces. Nucleic Acids Res 35(Web Server issue):W526–W530
Siragusa L, Cross S, Baroni M et al (2015) BioGPS: navigating biological space to predict polypharmacology, off-targeting, and selectivity. Proteins
Skolnick J, Zhou H, Gao M (2013) Are predicted protein structures of any value for binding site prediction and virtual ligand screening? Curr Opin Struct Biol 23(2):191–197
Somarowthu S, Yang H, Hildebrand DG et al (2011) High-performance prediction of functional residues in proteins with machine learning and computed input features. Biopolymers 95(6): 390–400
Sun M, Wang X, Zou C et al (2016) Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors. BMC Bioinformatics 17(1):231-016-1110-x
Suzuki Y (2004) Three-dimensional window analysis for detecting positive selection at structural regions of proteins. Mol Biol Evol 21(12):2352–2359
Tan KP, Nguyen TB, Patel S et al (2013) Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins. Nucleic Acids Res 41(Web Server issue):W314–W321
Taroni C, Jones S, Thornton JM (2000) Analysis and prediction of carbohydrate binding sites. Protein Eng 13(2):89–98
Tian B, Wallrapp F, Kalyanaraman C et al (2013) Predicting enzyme-substrate specificity with QM/MM methods: a case study of the stereospecificity of (D)-glucarate dehydratase. Biochemistry 52(33):5511–5513
Tian BX, Wallrapp FH, Holiday GL et al (2014) Predicting the functions and specificity of triterpenoid synthases: a mechanism-based multi-intermediate docking approach. PLoS Comput Biol 10(10):e1003874
Tjong H, Zhou HX (2007) DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces. Nucleic Acids Res 35(5):1465–1477
Tsai KC, Jian JW, Yang EW et al (2012) Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. PLoS ONE 7(7):e40846

Volkamer A, Kuhn D, Grombacher T et al (2012) Combining global and local measures for structure-based druggability predictions. J Chem Inf Model 52(2):360–372
Vulpetti A, Kalliokoski T, Milletti F (2012) Chemogenomics in drug discovery: computational methods based on the comparison of binding sites. Future Med Chem 4(15):1971–1979
Walsh I, Minervini G, Corazza A et al (2012) Bluues server: electrostatic properties of wild-type and mutated protein structures. Bioinformatics 28(16):2189–2190
Ward RM, Venner E, Daines B et al (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25(11): 1426–1427
Warwicker J (1986) Continuum dielectric modelling of the protein-solvent system, and calculation of the long-range electrostatic field of the enzyme phosphoglycerate mutase. J Theor Biol 121(2):199–210
Wilkins A, Erdin S, Lua R et al (2012) Evolutionary trace for prediction and redesign of protein functional sites. Methods Mol Biol 819:29–42
Xie ZR, Hwang MJ (2012) Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics 28(12):1579–1585
Xie L, Xie L, Bourne PE (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. Bioinformatics 25(12):i305–i312
Xie ZR, Liu CK, Hsiao FC et al (2013) LISE: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. Nucleic Acids Res 41(Web Server issue):W292–W296
Yaffe E, Fishelovitch D, Wolfson HJ et al (2008) MolAxis: a server for identification of channels in macromolecules. Nucleic Acids Res 36(Web Server issue):W210–W215
Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29(20):2588–2595
Yuan Z, Zhao J, Wang ZX (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. Protein Eng 16(2):109–114
Zhang Z, Tang Y-, Sheng Z- et al (2009) An overview of the de novo prediction of enzyme catalytic residues. Curr Bioinformatics 4:197–206
Zhang Z, Li Y, Lin B et al (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27(15):2083–2088
Zhao H, Yang Y, Zhou Y (2013) Prediction of RNA binding proteins comes of age from low resolution to high resolution. Mol BioSyst 9(10):2417–2425
Zhou HX, Shan Y (2001) Prediction of protein interaction sites from sequence profile and residue neighbor list. Proteins 44(3):336–343

# Chapter 11
# 3D Motifs

**Jerome P. Nilmeier, Elaine C. Meng, Benjamin J. Polacco and Patricia C. Babbitt**

**Abstract** Three-dimensional (3D) motifs are patterns of local structure associated with function, typically based on residues in binding or catalytic sites. Protein structures of unknown function can be annotated by comparing them to known 3D motifs. Many methods have been developed for identifying 3D motifs and for searching structures for their occurrence. Approaches vary in the type and amount of input evidence, how the motifs are described and matched, whether the results include a measure of statistical significance, and how the motifs relate to function. Compared to algorithm development, less progress has been made in providing publicly searchable databases of 3D motifs that are both functionally specific and cover a broad range of functions. A roadblock has been the difficulty of generating detailed structure-function classifications; instead, automated, large-scale studies have relied upon pre-existing classifications of either structure or function. Complementary to 3D motif methods are approaches focused on molecular surface descriptions, global structure (fold) comparisons, predicting interactions with other macromolecules, and identifying physiological substrates by docking databases of small molecules.

J.P. Nilmeier (✉)
Lawrence Livermore National Laboratory (LLNL) Division of Physical and Life Sciences Directorate, Biotechnology and Biosciences Division, 7000 East Avenue, Livermore, CA 94550-9234, USA
e-mail: nilmeier1@llnl.gov

E.C. Meng · B.J. Polacco · P.C. Babbitt
Department of Pharmaceutical Chemistry, University of California San Francisco (UCSF), 600 16th Street, San Francisco, CA 94158-2517, USA
e-mail: meng@cgl.ucsf.edu

B.J. Polacco
e-mail: benjamin.polacco@ucsf.edu

P.C. Babbitt
e-mail: babbitt@cgl.ucsf.edu

P.C. Babbitt
UCSF Department of Biopharmaceutical Sciences, 1700 4th Street, San Francisco, CA 94158-2330, USA



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_11

**Keywords** Clique detection · Geometric hashing · Functional annotation · Function prediction · Active site · Binding site · Functional residues · Catalytic residues · Structural motifs · Pattern discovery

**List of Abbreviations**

| | |
|---|---|
| 3D | Three-dimensional |
| CSA | Catalytic Site Atlas |
| DRESPAT | Detection of REcurring Sidechain PATterns |
| EC | Enzyme Commission |
| FFF | Fuzzy Functional Form |
| GASPS | Genetic Algorithm Search for Patterns in Structures |
| GO | Gene Ontology |
| HMM | Hidden Markov Model |
| nr-PDB | Non-redundant PDB |
| NP | Nonpolynomial (scaling) |
| NOE | Nuclear Overhauser Effect |
| PAR-3D | Protein Active site Residues using 3-Dimensional structural motifs |
| PDB | Protein Data Bank |
| PINTS | Patterns in Non-homologous Tertiary Structures |
| RMSD | Root-mean-square Deviation |
| S-BLEST | Structure-Based Local Environment Search Tool |
| SCOP | Structural Classification of Proteins |
| SOIPPA | Sequence Order-Independent Profile-Profile Alignment |
| SPASM | SPatial Arrangements of Sidechains and Mainchains |
| TESS | TEmplate Search and Superposition |

## 11.1 Background: Functional Annotation

The genomic approach to biology has resulted not only in copious amounts of new sequence and structure data, but also the prospect of obtaining a complete "parts list" for many organisms. However, a parts list is of little use without some understanding of what each part does. Even with entire genome sequences in hand, not all genes have been identified, and among identified genes, significant numbers have not been annotated with any function. The amount of sequence data far outweighs the available structures, so to a large extent, the assignment of functions, or ***functional annotation***, has been performed by large-scale sequence searching. In many cases, the function of an unknown sequence is inferred, or ***transferred***, through similarity to a sequence with a known function.

### 11.1.1 What Is Function?

Function can be described at many levels and from many perspectives (Radivojac et al. 2013). Objective classifications of function are needed for training and testing any method of functional annotation. The Gene Ontology (GO) system (Ashburner et al. 2000) is a hierarchical set of functional descriptors ranging from broad to specific in each of three categories: biological process, cellular component, and molecular function. For the specific molecular functions of enzymes, GO embeds the Enzyme Commission (EC) system (International Union of Biochemistry and Molecular Biology: Nomenclature Committee and Webb 1992) which is also hierarchical: catalysed reactions are described with four integers, where the first number refers to a broad class of reactions and the last number refers to a specific substrate. GO also includes molecular function terms for stable binding relationships (where binding is not functionally associated with membrane transport or catalytic activity). The KEGG annotation (Kanehisa and Goto 2000; Ogata et al. 1999), while used mostly for studying reaction pathways, can also be used to annotate enzyme function.

Other methods for classifying proteins, while less directly related to function, can be used to infer relationships related to function. These include Structural Classification of Proteins (SCOP) (Murzin et al. 1995; Conte et al. 2000; Andreeva et al. 2004, 2008) and Class, Architecture, Topology, and Homologous superfamily (CATH) (Orengo et al. 1997, 1999, 2003). Both methods are hierarchical classifications of protein substructures such as ***folds*** (Richardson 1981) or ***domains*** (Chothia and Lesk 1986; Rost 1997), that can be "mixed and matched" evolutionarily (Chothia et al. 2003). In SCOP, domains are classified into families, superfamilies, folds, and classes. Folds are, in general, only indirectly related to function (Babbitt and Gerlt 1997; Todd et al. 2001), but they can be very informative for many cases. The use of fold similarity for annotation transfer is discussed in Chap. 9.

The GO and EC annotations for functional annotation cover nearly all reactions found in biochemical systems. They do not, however, include details on enzymatic mechanism, or the role of the protein in the reaction (Babbitt 2003). Two enzymes that catalyze the same overall reaction would have the same EC number, even if their structures and catalytic intermediates are very different. Additionally, many enzymes are evolutionarily related because they share an intermediate step in the overall reaction, that is, a ***common partial reaction***. The EC and GO naming systems do not account for such similarities in any practical way, and yet such similarities are a defining feature for many protein superfamilies, with the enolase superfamily as the most notable example. Figure 11.1 illustrates the variety of reactions associated with the enolase superfamily.

### 11.1.2 Genomics and Functional Annotation

The progress in the genomics community in assigning functional annotations through sequence-based methods is impressive. Given that function is related

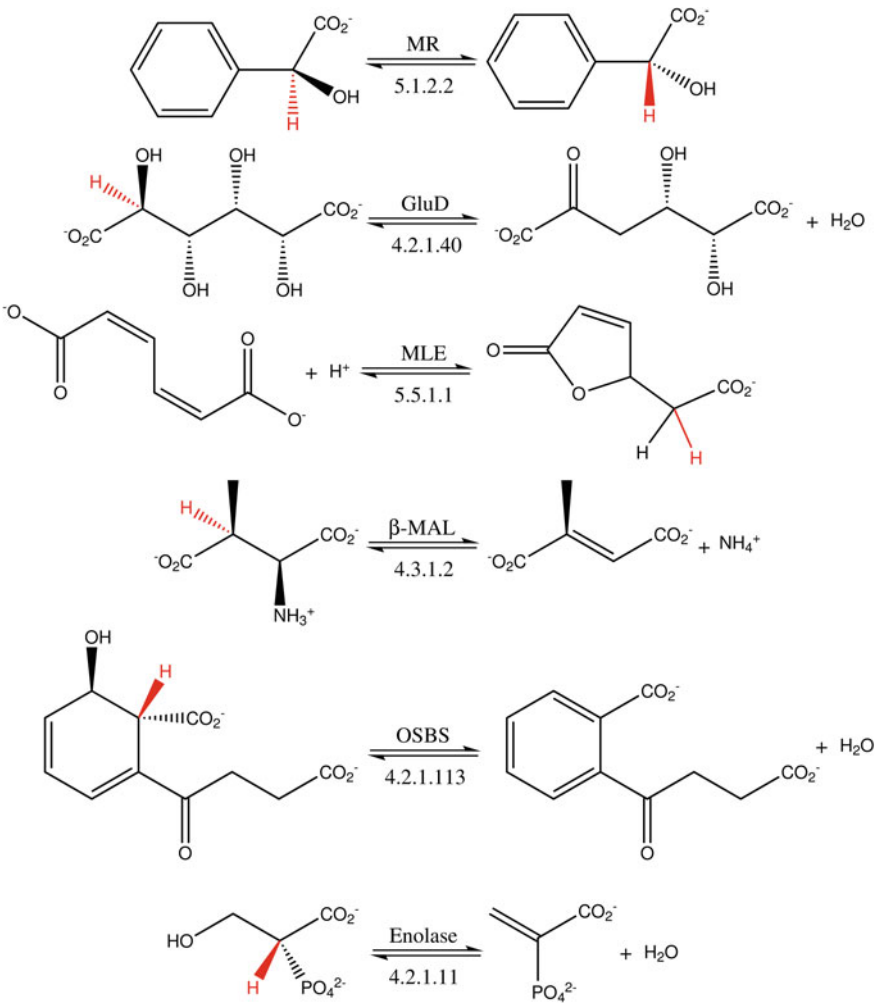



**Fig. 11.1** Illustration of the common partial reaction in the enolase superfamily. The extraordinary diversity of reactions shown in these enzymes share one step in common, which is the initial abstraction of a proton (indicated in *red*). Abbreviations are *MR* mandelate racemase, *GlucD* glucarate dehydratase, *MLE* muconate lactonizing enzyme, *β-MAL* β methylaspartate ammonia lyase, *OSBS* O-succinylbenzoate synthase

indirectly to sequence through a protein structure, however, it makes sense to consider methods that incorporate protein structure more directly in the inference of function.

Sequence alignment methods such as BLAST (Altschul et al. 1990) and CLUSTALW (Larkin et al. 2007; Thompson et al. 1994) have enjoyed wide success in inferring function when sequence similarity is greater than 40–60% (Tian and Skolnick 2003; Devos and Valencia 2001; Rost 2002). More sophisticated

methods, including Hidden Markov Model (HMM) methods (Krogh et al. 1994; Sjölander et al. 1996), and ancestry-based methods such as the Evolutionary Trace (Lichtarge et al. 1996), INTREPID (Sankararaman and Sjölander 2008), Phylofacts (Glanville et al. 2007; Krishnamurthy et al. 2006), Bayesian Monte Carlo inference from phylogenetic trees (Tseng and Liang 2006) and EFICAz (Arakaki et al. 2009; Tian et al. 2004) combine sequence alignment procedures and machine learning techniques to specifically assign function to a sequence.

### *11.1.3 The Need for Structure-Based Methods*

Protein structures, however, may reveal important similarities or possible evolutionary relationships that are not evident from their sequences alone. The natural analogue to a global sequence alignment is a global structure alignment. Methods like LGA (Zemla 2003), PINTS (Stark and Russell 2003) and CE (Shindyalov and Bourne 1998, 2001) can accomplish this alignment in various ways and sometimes reveal more significant relationships in the alignments.

Other approaches use combinations of sequence and structural information, such as SOIPPA (Xie and Bourne 2008, 2009; Ren et al. 2010), DISCERN (Sankararaman et al. 2010), and PevoSOAR (Tseng et al. 2009), and can provide improvements to sequence based methods alone. Additionally, methods like the FFF approach that are essentially structural in nature benefit from addition of sequence information (Cammer et al. 2003). The success of any of these global similarity-based techniques depends largely on the ability to distinguish conservation patterns that correspond to the actual functional or catalytic portions of a protein sequence or structure.

Related proteins may have diverged so far that global sequence or structure alignments are challenging. Conversely, proteins with highly similar folds can perform different functions (Babbitt and Gerlt 1997; Todd et al. 2001). This observation points to the need for a more fundamental definition of a structural unit, or ***3D motif*** which more specifically defines the functional aspects of a given protein structure.

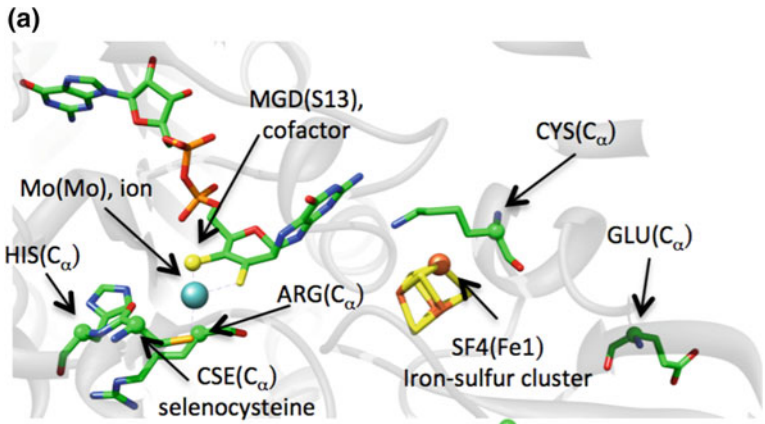
Structural genomics efforts have long recognized the fact that structural data is much more informative than sequence data alone. This data is used not only for annotation, but for homology modelling and in silico drug design. On principal driving idea behind this effort is to crystallize structures that are underrepresented in sequence space, so that more sequences can be more directly represented in structural forms (Berman et al. 2000; Baker and Sali 2001). The number of structures in the PDB from these initiatives has continued to grow at an increasing rate, and many target structures were previously completely unannotated, or annotated incorrectly using automated sequence-based methods.

Functional assignment to these proteins remains as a frontier challenge for structural genomics, and 3D motif-based methods are likely to play a prominent role for proteins where current methods fall short.

## 11.2 3D Motif Matching Techniques

### *11.2.1 What Is a 3D Motif?*

***3D motifs*** are spatial patterns of points based on a few residues (generally under a dozen) associated with some protein function or classification of interest. They are sometimes called ***active site templates***, since the residues may contribute to a




| | ARG | CYS | GLU | HIS | LYS | MGD | SF4 |
|---|---|---|---|---|---|---|---|
| ARG 333.A | | | | | | | |
| CYS 140.A | 5.54 | | | | | | |
| GLU 30.A | 24.15 | 27.56 | | | | | |
| HIS 141.A | 6.37 | 3.76 | 30.18 | | | | |
| LYS 44.A | 17.05 | 19.15 | 11.22 | 22.21 | | | |
| MGD 802.A | 5.80 | 4.24 | 25.04 | 7.11 | 15.79 | | |
| MO 803.A | 6.57 | 3.72 | 25.05 | 7.38 | 15.97 | 2.47 | |
| SF4 800.A | 13.65 | 17.00 | 11.71 | 19.50 | 5.15 | 13.93 | 14.51 |

**Fig. 11.2** Example of a catalytic template constructed from a Catalytic Site Atlas (CSA) entry, which has a corresponding EC number along with a list of residues that comprise the site. Each residue has a centroid associated with it, which is labelled in parentheses and shown as spheres in (**a**) and (**b**). Cofactors, ions, and residues can often have either a single centroid or many centroids associated with them (see Fig. 11.3). In this example, Cα coordinates are used as the residue centroids, but centroids may be computed in other ways. For this templating approach, a graphical representation of the template is used, with nodes associated with the centroid identity, and edges defined by the interatomic distances. The template is stored as a distance matrix, shown in (**b**). The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

binding or catalytic pocket, or ***structural templates***. The positions of one or a few atoms per residue are used, and the points are labelled with additional information, such as atom and residue type, used in matching. The residues are often strictly positioned in space but not necessarily in sequence. Figure 11.2 describes a typical binding site found in the Catalytic Site Atlas (CSA), and one way to represent it in a reduced form. In this example, the Cα atoms are used as pseudoatoms, but many approaches use atomic coordinates from the sidechains, or a centroid using clusters of atoms in the pseudoatom positions as well (Oldfield 2002), as is the case for the templates in Fig. 11.3.

3D motifs represent highly conserved patterns of local structure. Often the residues are conserved to sub-angstrom resolution, and the absence of one residue in the motif can completely eliminate its function. The remainder of the protein, however, can often vary substantially. Ideally, a 3D motif will describe exactly these function-critical structural components and serve as a sensitive and specific signature of the function.

Since such a motif can often be the only evolutionary constraint, many different structures can be present with the same motif, and there is no restriction on the

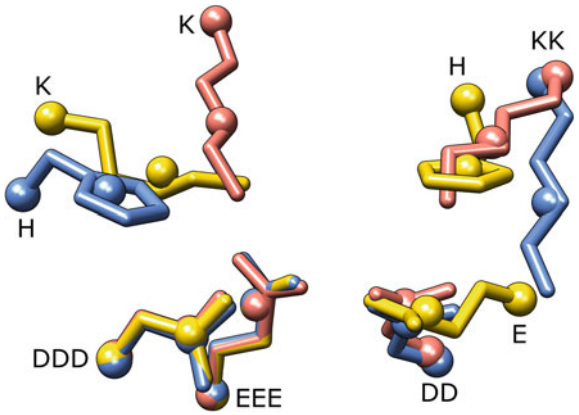



**Fig. 11.3** Active site residues from members of the enolase superfamily, illustrating aspects of motif representation and specificity. The superimposed side chains of two basic and three acidic residues are shown from each of the following: mandelate racemase (*yellow*, PDB 2mnr), enolase (*salmon*, PDB 4enl), and methylaspartate ammonia lyase (*blue*, PDB 1kcz). Balls indicate alpha-carbon (Cα) and side chain centroid locations. Single-letter codes near the alpha-carbons indicate residue types: *H* for histidine, *K* for lysine, *D* for aspartic acid, and *E* for glutamic acid. While the acidic residues at the two lower left positions are highly conserved in type and conformation, variations in the sites include: *1* differing (albeit similar) residue types at the other three positions; *2* different side chain conformations, exemplified by the two lysines on the right; *3* different locations in primary sequence, where the basic residue on the upper left is C-terminal to the others in enolase but N-terminal in the sequences of the other two proteins. Using side chain centroids rather than the positions of functional atoms generally allows for more variety in backbone conformations, assuming the sidechain positions are well conserved across templates (Todd et al. 2002). The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

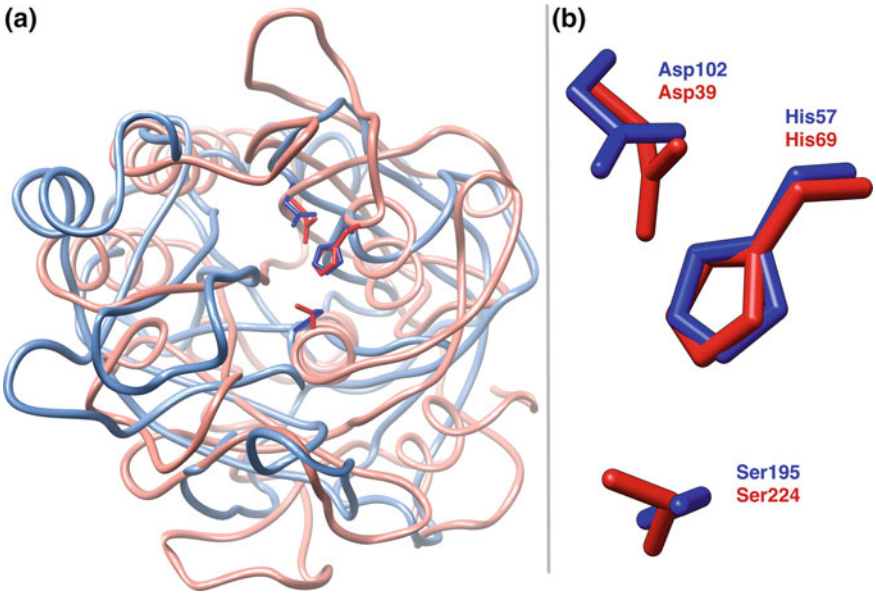
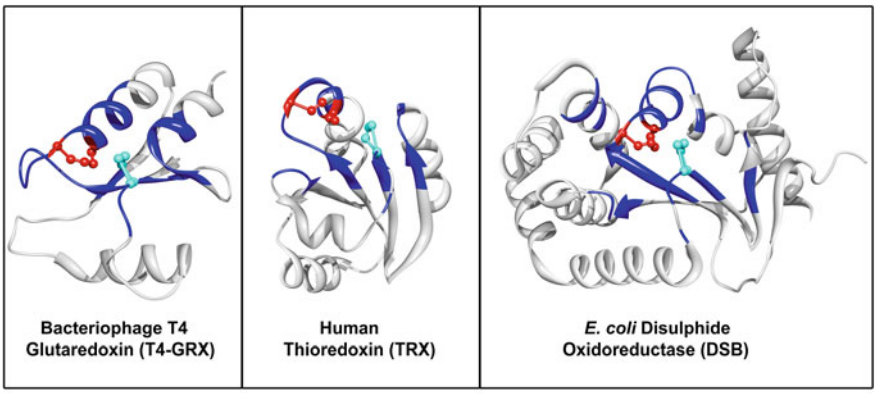



**Fig. 11.4** Two serine proteases superimposed at their catalytic triads reveals the close similarity of residues in the active sites despite different overall folds. **a** Ribbon diagrams of trypsin (*blue*/*light blue*, PDB 1sgt) and proteinase *K*, a homolog of subtilisin, (*red*/*salmon*, PDB 2pkc) show that the two proteins have different folds with no corresponding secondary structure elements, yet their catalytic triads (displayed in stick representation) overlap. They are considered to have no common ancestor. **b** The sidechains of the catalytic triads are shown enlarged to display the similar orientations of the catalytic triad residues (1sgt: Asp102, His57, Ser195; and 2pkc: Asp39, His69, Ser224). The similarity of the catalytic triad in these non-homologous structures demonstrates the ability of 3D motifs to detect similar functions in a pair of proteins where homology-based methods will fail. The image was created with UCSF Chimera (Pettersen et al. 2004) (http://www.cgl.ucsf.edu/chimera)

location or relative order of residues in the sequence. Figure 11.4 shows a case of convergent evolution in the serine protease Asp-His-Ser catalytic triad. While the catalytic triad is highly conserved structurally, the remaining structural elements display noticeable variations. This particular catalytic triad was, historically, the first to be thought of as a 'motif' based on these observations. Variations in structure relative to a motif are even more pronounced in other more recent examples, including the disulfide oxidoreductase site shown in Fig. 11.5, which is taken from an example of a Fuzzy Functional Form (FFF) template (Fetrow and Skolnick 1998; Di Gennaro et al. 2001).




```
T4-GRX  ...................MFKVYGYDSNIHKCVYCDNAKRLLTVKKQPF....EFINIMPEKGVFDDEKIAEL
DSB     AQYEDGKQYTTLEKPVAGAPQVLEFFSFFCPHCYQFEEVLHISDNVKKKLPEGVKMTKYHVNFMGGDLGKDLTQ
TRX     KQIESKTAFQEALDAAGDKLVVVDFSATWCGPCKMIKPFFHSLSE...KYSN.VIFLEVDVD...........D

T4-GRX  LTKLGRDTQIGLTMPQVFAPDGSHIGGFDQLREYFK.....................................
DSB     AWAVAMALGVEDKVTVPLFEGVQKTQTIRSASDIRDVFINAGIKGEEYDAAWNSFVVKSLVAQQEKAAADVQLR
TRX     CQDVASECEVKCTPTFQFFKKGQKVGE..........FSGA.NKEKLEATINELV..................

T4-GRX  ........................................
DSB     GVPAMFVNGKYQLNPQGMDTSNMDVFVQQYADTVKYLSEKK
TRX     ........................................
```

**Fig. 11.5** The FFF motif for the disulfide oxidoreductase active site is found in many proteins. Illustrated are T4 glutaredoxin, 1aaz, chain A (*left*), human thioredoxin, 4trx (*middle*) and proline disulfide isomerase, 1dsb, chain A. The three key residues which define this FFF are two cysteines (*red side* chains) and a proline (*cyan side* chain). The active site structure of these proteins is conserved, although the rest of the protein structures exhibit some differences. Using these three key residues, the active site signature for each protein was identified (fragments shown as *blue ribbons* in each protein). Global sequence alignment, produced using ClustalW, of these three proteins shows the location of the key residues (*red* and *cyan*, underlined) and the active site signature fragments (*blue*) within the whole sequence. The alignment illustrates the lack of overall sequence similarity between the three proteins, even though the active site structure itself is highly conserved

### *11.2.2 Historical Development of Motif Matching Methods*

Early ideas about catalytic motifs were based on observation, and were not algorithmic in nature. The most widely studied motif is the Ser-His-Asp catalytic triad mentioned above, first recognized in serine proteases (Blow et al. 1969; Wright et al. 1969) and later in other hydrolases such as esterases and lipases. The catalytic triad occurs in different folds, and thus it encompasses cases of both divergent and convergent evolution (Fig. 11.4). Early discoveries of the catalytic triad found it present in entirely different folds of subtilisins, (Fischer et al. 1994). The Thornton group, studying triads in detail, formulated a more careful description of the site, based on the observation that only the relative positions of serine and aspartate

oxygens and the histidine ring were preserved across many examples (Wallace et al. 1996).

During this time, the concept of a 3D motif began to emerge in an algorithmic context, which is generally described as ***template matching*** or ***motif matching***.

Artymiuk et al. (1994) appear to be the first to apply such a procedure, which they called ASSAM, to enzymatic site detection. Their work used the ***subgraph isomorphism*** procedure, which is a graph theoretic method for finding a motif graph in a larger structure graph. The method, originally proposed by Ullmann (1976), is described in Sect. 11.2.1. Later work by Artymiuk et al. expanded the approach beyond catalytic sites to other structural applications, such as the identification of tertiary structures (Mitchell et al. 1990; Spriggs et al. 2003). In this work, many careful choices were made with regard to which atoms to use as part of the template, and particular attention was paid to reliable detection of residue triads, given the importance of catalytic triads as an archetypal motif.

During this period, Kleywegt also developed a site-matching procedure originally designed to identify patterns in distance matrices determined by Nuclear Overhauser Effect (Radivojac et al.) measurements (Kleywegt et al. 1989). Later Kleywegt introduced a program called DEJAVU that detects protein motifs (Kleywegt and Jones 1997). A technique based on DEJAVU was later generalized to identify enzymatic sites with a method called SPASM, along with a complementary approach, known as RIGOR (Kleywegt 1999), used to search a list of motifs for similarity to a given structure. Early work with this method focused on triad motifs as well. A notable example from the Kleywegt study (Kleywegt 1999) was the discovery of a family of glucanases.

A related set of approaches to the template matching problem uses a procedure known as ***geometric hashing*** (Wolfson and Rigoutsos 1997; Brakoulias and Jackson 2004). The main difference between the geometric hashing procedure and graph-based procedures is that geometric hashing uses a Cartesian grid (with a suitable coordinate system) to bin similar coordinates. It is used widely in image processing, and has been successfully adapted to structural approaches. It is dependent on the frame of reference, however, and additional overhead is required to accomplish optimal translations and rotations for comparison. The Thornton group proposed a template-matching procedure, named TESS (Wallace et al. 1997), built on such an approach. A later iteration, known as JESS (Barker and Thornton 2003), incorporated recursive ideas and threshold constraints to improve searching procedures. More recently, the Kavraki group developed a series of procedures built on a match augmentation method, MASH, that iteratively grows a template match from pairwise matches obtained through geometric hashing (Chen et al. 2007a). Later developments from this group include the addition of residue hash matching, the LabelHash algorithm (Moll et al. 2010; Moll and Kavraki 2008), along with impressive optimizations at the hardware and software level to improve performance. Other geometric hashing approaches include SitesBase (Gold and Jackson 2006a, b), and GIRAF (Kinjo and Nakamura 2007).

Success of template-matching methods, within the Thornton group and elsewhere, led to the important recognition that a high quality curated database of enzymatic sites was needed. This recognition led directly to the development of the Catalytic Site Atlas (CSA) (Porter et al. 2004), which is a manually curated table of enzymes and binding site residues, as well as tabulated Enzyme Commission (EC) numbers (Bairoch 1994). The CSA is somewhat limited in coverage, however, and the scale of such a database will always be strictly limited by the capacity of expert manual curators. As a result, many approaches have been developed which attempt to automatically locate structural features that may be used as templates. These approaches include physics-based approaches (Halgren 2007, 2009) and statistical modelling of measures (Liang et al. 2003; Brylinski and Skolnick 2008; Skolnick and Brylinski 2009). Methods that consider protein dynamics (Yang and Bahar 2005; Glazer et al. 2008) represent a promising direction as computational capabilities improve (see also Chap. 12).

Other valuable resources related to this effort, including the MACiE database (Holliday et al. 2007), and the ProFunc server (Laskowski et al. 2005), as well as metaservers like ProKnow (Whisstock and Lesk 2003), resulted from the success and utility of structure-based approaches to understanding function. Table 11.1 lists some of the database resources that have resulted from efforts in this field.

The Babbitt and Gerlt groups have gone beyond matching of catalytic residues and matched enzymes by their chemical mechanism. They established the concept of a mechanistically diverse superfamily, where the similarity among members is governed by the conservation of partial reactions within the protein family, rather than by sequence or structure conservation alone (Galperin et al. 1998; Gerlt and Babbitt 2001; Gerlt et al. 2012). This approach is in contrast to a sequence-based approach, which relies on global sequence similarity with the expectation that conservation patterns can point to residues of functional interest. It also presents an alternative to the Enzyme Commission (EC) classification scheme (Webb 1992), which builds a hierarchy based on the substrate reaction chemistry. This alternative approach to classification, with its emphasis on binding site architecture and conservation of partial reactivity, led to the development of the Structure-Function Linkage Database SFLD (Pegg et al. 2005, 2006). These ideas led to the larger Enzyme Function Initiative (Gerlt et al. 2011), which has the goal of large-scale enzyme characterization and classification based on experimental and computational work (Gerlt et al. 2012; Kalyanaraman et al. 2008; Song et al. 2007). Template-matching procedures using superfamily template libraries were applied (Meng et al. 2004), and led to a procedure known as GASPS and the database GASPSdb. GASPS is designed to develop new template libraries based on any classification of structures into those with and without a function (or other property) of interest (Polacco and Babbitt 2006).

**Table 11.1** Servers and other web resources for 3D motif searching and comparison

| Server name and citation | Server URL |
|---|---|
| Description of resource | Motif database description |
| Catalytic site atlas (CSA) (Furnham et al. 2014) | http://www.ebi.ac.uk/thornton-srv/databases/CSA/ |
| Basic interface to motif database (CSA) | Cα and Cβ functional atom motifs for 147 well-characterized enzyme families. Database freely available for download |
| ProFunc (Laskowski et al. 2005) | http://www.ebi.ac.uk/thornton-srv/databases/profunc/ |
| Multi-search including motif search with JESS: whole structure query vs. motif database, fragment query versus whole chains | CSA motifs, 13,057 ligand-binding and 1200 DNA-binding modes from PDB. Motifs contain both sidechain and backbone atoms |
| Catalytic site identification (Kirshner et al. 2013) | http://catsid.llnl.gov/ |
| Finds matches to motifs with user defined target and/or protein databank. Uses subgraph isomorphism and machine learning | 2244 motifs, including modified CSA and enolase superfamily templates. User can also search for unannotated structures by EC number |
| Uppsala Software Factory (Kleywegt 1999) | http://xray.bmc.uu.se/usf/ |
| Software is available for download. SPASM compares a query motif to a database of targets. RIGOR compares a query structure to a database of motifs | RIGOR database contains 73,164 motifs from PDB. 57,719 motifs have residue type labels. The remaining are unlabelled (engineerable) |
| ProKnow (Pal and Eisenberg 2005) | http://proknow.mbi.ucla.edu/ |
| Multi-search, including RIGOR motif searches. GO annotations included in output | 10,230 motifs with GO annotations from their source structures, 7819 if electronic annotations are excluded |
| GASPSdb (Polacco and Babbitt 2006) | http://gaspsdb.rbvi.ucsf.edu/ |
| Browse database of 3D motifs representing SCOP families and superfamilies | Motifs have Cα and side chain coordinates. RIGOR-formatted database files are available for download |
| funClust (Ausiello et al. 2008) | http://pdbfun.uniroma2.it/funclust/ |
| Uses Query3D to identify motifs shared by groups of 3–20 structures | User supplied structures for consensus motif |
| pdbFun (Ausiello et al. 2005b) | http://pdbfun.uniroma2.it/ |
| Compares specified probe and target residue sets using Query3D | >12 M individual residues. Subsets are defined with Boolean descriptors combinations |
| ProBIS (Konc and Janežič 2012) | http://probis.cmm.ki.si/ |
| Detects similar binding sites using a clique detection algorithm (ProBIS) | Database contains pre-calculated matches for non-redundant (95%) pdb |
| The LabelHash server (Moll et al. 2011) | http://labelhash.kavrakilab.org/ |
| Compares motifs with PDB or user structures using LabelHash algorithm | 17 predefined motifs derived from CSA. User defined motifs are allowed |
| WebFEATURE (Liang et al. 2003; Buturovic et al. 2014) | http://feature.stanford.edu/webfeature/ |

(continued)

**Table 11.1** (continued)

| Server name and citation | Server URL |
|---|---|
| Uses radially symmetric patterns as motifs | Motifs are derived from PROSITE v20.81, and are available for individual download |
| PAR-3D (Goyal et al. 2007) | http://sunserver.cdfd.org.in:8080/protease/PAR_3D/access.html |
| Compares query to motifs expressed as distance and angle ranges | Cα and Cβ motifs for 6 protease classes and 10 glycolytic enzymes. Metal chelating sites have sidechain centroids as well |
| PDBSiteScan (Ivanisenko et al. 2004) | http://wwwmgs.bionet.nsc.ru/mgs/gnw/pdbsitescan/ |
| Compares query to all or a subset of motifs in the PDBSite database | 36,273 backbone-atom motifs from SITE annotations. Also includes interfaces with DNA, RNA, or other proteins |
| PINTS (Stark and Russell 2003) | http://www.russelllab.org/cgi-bin/tools/pints.pl |
| Compares query structure to motif database, query motif to PDB, or two proteins to each other | Ligand-binding and SITE-annotated motifs consisting of side chain points from polar residues |
| SuMo (Jambon et al. 2005) | http://sumo-pbil.ibcp.fr/ |
| Compares query structure, chain, or ligand-binding site to database | Database contains 34,210 ligand-binding sites, and also whole structures. Motifs are built from functional groups |
| S-BLEST (Schmitt et al.) (Mooney et al. 2005) | http://www.sblest.org/ |
| Queries residue-centred patterns against nr-PDB. Returns best-matching chains and annotations | Searches for similarity to uploaded structure only |
| SiteEngine (Shulman-Peleg et al. 2005) | http://bioinfo3d.cs.tau.ac.il/SiteEngine/ |
| Compares the binding site of a ligand-bound structure to the entire surface region of another structure | Linux executable for non-commercial use only |
| Nestor3D (Nebel et al. 2007) | http://staffnet.kingston.ac.uk/~ku33185/Nestor3D.html |
| Generates a consensus motif with input structures and structure alignments | User supplies input structures for comparison. Software is available for download |

## 11.3 Algorithmic Approaches to Motif Matching

The historical development of motif matching methods and current methods suggest the following categorization of these methods.

### 11.3.1 Methods Using 3D Motifs

Many elements can make up the definition of a motif, but the majority of approaches consider a motif as a constellation of labelled points derived directly from an important subset of atomic coordinates of a structure or set of structures. A side chain centroid, for example, is simply a pseudoatom at the average position of the atoms in the side chain. Up to a few points are used per residue in the motif, and the points are labelled with additional information such as atom type, residue type, or physicochemical characteristics.

Searching can be computationally intensive, especially considering that thousands of structures may be compared to thousands of motifs; 3D motif searching has relied on the development of efficient algorithms, often involving one or more of the following:

- **Geometric hashing**. Hashing is a broad term for reducing complex data to a simpler form that can be compared more rapidly. In its most basic form, a geometric hash can be a lookup table of Cartesian coordinate points (Fischer et al. 1994) and pseudoatom identities as well as many other properties, including distances, angles, and other residue features (Shulman-Peleg et al. 2004). In general, hash comparisons are very fast, especially compared to the time required to align the coordinates (Pennec and Ayache 1998). Hashing or preprocessing the data takes time, but only needs to be done once per structure and can greatly speed up comparisons.
- **Graph Theoretic Methods**. A graph consists of vertices (Kaminski et al. 2001) and edges (lines that connect pairs of vertices). A molecular structure or 3D motif can be treated as a labelled graph. Figure 11.2 shows how a catalytic site might be represented as a group of labelled vertices with interatomic distances used as edges. ***Subgraph isomorphism*** algorithms look for the occurrence of a subgraph (the 3D motif) in a larger graph (the structure). While the subgraph isomorphism is formally treated as a method for identical matches, many modifications to this basic approach are used for imperfect matches, including a variety of distance tolerances, as well as allowances for substitutions (Nilmeier et al. 2013). ***Clique detection*** (Schmitt et al. 2002) is essentially a similar algorithm, but the graph in this case describes the geometries of both structures together. A vertex in the graph represents a pair of atoms or pseudoatoms, one from structure A and one from structure B (where "structure" could be a 3D motif). Only atoms with matching types are allowed to pair. Two vertices are connected by an edge if the distance between the two atoms in A matches the distance between the two atoms in B within a specified tolerance. *A clique is a graph in which every vertex is connected to every other vertex*. Thus, clique detection identifies a set of atoms from A with internal distances completely consistent with those among a paired set of atoms from B.

### 11.3.2 Efficiency Considerations for 3D Motifs

Motif matching algorithms can be very fast for perfect matches. A challenge in the design of these algorithms, however, is that the extension to imperfect matches can lead to exponential scaling—sometimes referred to as nonpolynomial (Larkin et al. 2007) scaling—with respect to template and structure size, with concomitant losses in speed and efficiency.

To address this challenge variations of branch and bound approaches are used. These approaches leverage combinations of ***breadth-first*** and ***depth-first*** searches, and usually build a series of partial templates for comparison. In template matching algorithms, a breadth-first search typically refers to a method whereby a partially constructed template with few vertices and a 'breadth' of candidate edges are compared for fitness. The best candidates are then selected for the next iteration. Alternatively, a depth-first search builds a 'deeper' partial template with many vertices and fewer edges before iterating to the next comparison step. While described graphically, these ideas can be used in the geometric hashing comparisons as well.

During the buildup procedure, the list of candidates in the search is usually pruned using a heuristic similarity cutoff that can be highly specific to the algorithms and templates that are used. This buildup procedure is discussed in some of the isomorphism searches (Nilmeier et al. 2013), and in variants of the geometric hashing technique (Chen et al. 2007b).

Care must be applied in determining these cutoffs, especially in the time-intensive search portions. If the similarity cutoffs are relaxed, false positives may be obtained. More importantly, however, the scaling can rapidly become unmanageable, since each list is carried into the next iteration. On the other hand, if the cutoffs are too strict, then good matches are discarded. In addition to the pruning criterion, other measures are applied to restrict the search space. For example, in graph comparison algorithms the default description of the resulting graph would contain all distances, resulting in a large, fully connected (clique) structure graph. Nearly all of these edges are unnecessary when comparing the graphs, so careful construction of the graphs beforehand will vastly improve performance.

Application of similarity thresholds can be a nontrivial effort, and very specific to the templates under consideration. Consider the residues in the lower right hand corner of Fig. 11.3. The active site residues are represented as Cα and side chain centroids (Oldfield 2002). In this case, centroid position is highly conserved, but the Cα position is not, and the residue identity is also different (Asp $\longrightarrow$ Glu). The choice of which constraints to apply and which to relax in this case would require detailed knowledge about the significant elements involved (in this case, the proton abstraction residue).

### *11.3.3 Methods with Nonstandard Motif Information*

It is not always straightforward to differentiate between methods that use 'standard' 3D motifs from methods that incorporate additional information. For example, many techniques have multiple stages. In these techniques, a fast template matching algorithm is used to generate an initial candidate list, followed by a more complex scoring procedure to refine results (Laskowski et al. 2005; Kirshner et al. 2013; Nilmeier et al. 2013). While the second stage scoring procedure may incorporate more complex representations of the catalytic site, the core search algorithm uses the classic definition of a motif.

Other methods, however, incorporate a fundamentally different definition of a motif in the primary search machinery. For example, hybrid point-surface and single-point-centred descriptions of local structure do not fall under our working definition of a 3D motif approaches, but they do share many similarities. Methods primarily based on surface descriptions are covered in Chap. 10.

- **Single-Point-Centred Descriptions**. The program FEATURE (Bagley and Altman 1995) describes local structure as a set of properties in concentric shells emanating from a single point. The properties include descriptors of atoms, functional groups, residues, secondary structure, and simple biophysical characteristics. Because values are summed over spherical shells, however, directional information is lost. Both the WebFEATURE server (Liang et al. 2003; Buturovic et al. 2014) and the Structure-Based Local Environment Search Tool S-BLEST web server (Mooney et al. 2005; Peters et al. 2006) use FEATURE templates, and each provide their own results, along with enhanced annotations (Table 11.1).
- **Hybrid (Point-Surface) Descriptions**. Cavbase (Schmitt et al. 2002; Kuhn et al. 2006) and SiteEngine (Shulman-Peleg et al. 2004) describe binding sites as collections of pseudoatoms and their associated surface patches. The pseudoatoms represent surface-exposed functional groups of various types, such as a hydrogen bond donor or acceptor. Comparisons involve finding geometrically and physicochemically consistent sets of pseudoatoms, superimposing structures based on those matches, and then scoring based on surface patch overlap and physicochemical similarity. Surface points typically far outnumber the pseudoatoms, so scoring is relatively computationally demanding. The SiteEngine web server (Shulman-Peleg et al. 2005) performs pairwise comparisons but not database searches(Table 11.1). Other surface-based methods include eF-site (Kinoshita and Nakamura 2003), SuMo (Jambon et al. 2003), SiteEngine (Shulman-Peleg et al. 2004), and Query3D (Ausiello et al. 2005a)

### 11.3.4 Interpretation of Results

The previous sections have discussed the technical challenge of finding a given motif in a structure. However, there are still questions that must be answered when applying these methods. What can be said about the function of the structure if a positive match is found? What constitutes a positive match, and how reliable is it?

Several issues must be considered when deciding what a positive match means. The ideal case is when the motif perfectly defines the residues for a particular annotated function. In these cases, the interpretation of the match is straightforward: the structure has the annotated function that the matching motif has. Developing a motif library with these desirable properties is a challenge in itself, and is discussed in Sect. 11.3. This simple mapping of function from a motif to the structure is not always straightforward, as motifs may be only indirectly associated with a specific function. For example, if a motif is derived from a SCOP superfamily, a match may only imply some function which is commonly found in the SCOP structure.

Any given motif-to-structure comparison is an NP-hard challenge, and even an efficient procedure may still yield several different candidate matches. Additionally, motif libraries can number on the order of thousands, while the PDB has tens of thousands of structures. A comparison of the full set of possibilities can quickly lead to an intractable problem unless sensible cutoffs to candidate matches are applied during the evaluation steps.

It is even more important to be able to report a manageable list of matches that can be easily interpreted and understood by users. This list will likely contain trivial matches of nearly exact motifs found in proteins with very similar global structure. The more interesting matches in the list should include somewhat distant but still plausible relationships; possibly with residue substitutions, or noticeable differences in global structures.

Basic measures of structural similarity are usually the starting point for scoring. The root mean square deviation, or RMSD, is one very common measure. It has many limitations, however. Most notably, it is not a useful measure when comparing matches to motifs of different sizes. Many other nuances begin to become apparent, including substitution allowances as well as subtle geometric relationships that may not be properly represented by the reduced geometric form of the motif.

To account for these issues and provide a better ranking of hits, some groups apply a multistage method. The fast, coarse search method will generate a large candidate list that is then subjected to a more rigorous scoring procedure. Sometimes the scoring procedure is intended to have a direct statistical interpretation, much like a p-value or other probabilistic score (Barker and Thornton 2003; Nilmeier et al. 2013; Kirshner et al. 2013). The determination of the cutoff score, which indicates whether the candidate is a positive match, can often be heuristic. There are, however, classic machine learning techniques that can be applied to determine appropriate cutoffs.

The ability of a procedure to identify true positives, measured by the true positive rate (TPR) or ***sensitivity***, while also minimizing the false positive rate (FPR) is usually the measure of performance of many of these techniques. One technique that is used frequently is the Receiver Operating Characteristic (Bairoch), which is simply a plot of these values as the cutoff is adjusted, in which the Area Under the Curve (AUC) indicates a quality measure of the prediction procedure. This is only one of many techniques to identify good cutoff values, but is widely used in the motif matching literature and elsewhere.

Another approach to interpretation is to take the predictions of multiple methods into consideration. This can often prove to be more useful than relying on any one particular method. Some servers provide predictions from multiple sources, leaving the final determination to the user. Notable examples include the ProKnow server (Pal and Eisenberg 2005) and ProFunc (Laskowski et al. 2005) servers, and are listed in Table 11.1. These servers are also discussed in detail in Chap. 13.

Finally, common sense must be applied. Many confounding factors will still present themselves, even in the most carefully constructed procedures. For example, a motif may be correctly located in a structure, but there is no actual binding cavity to accommodate the substrate. It is prudent, if not essential, to inspect matches visually and to evaluate them using biologically relevant criteria when inferring the function from a match. Many of the most useful servers and software have some visualization process as an integral part of the procedure for studying matches, simply because expert evaluation of the matches is still the best way to determine if algorithms are working as expected.

## 11.4 Methods for Deriving Motifs

Most of the effort in motif matching approaches is invested in locating a motif in a protein structure. This challenge, however, assumes that the motif is available as a ground truth. Sometimes the methods allow the user to supply a motif, while other methods use a library of motifs. How, then, are these motifs generated in the first place?

Ideally, for ***motif discovery***, the set of positive examples should be as diverse as possible while retaining the common feature, and the negative examples should be as similar as possible to the positive examples while lacking that feature. In practice, the positive and negative sets may not be ideal, and part or all of a derived 3D motif could still reflect common ancestry or coincidence rather than shared function.

Others treat motif discovery or generation of motif libraries essentially as the primary goal of their method.

### 11.4.1 Literature Search and Manual Curation

Perhaps the most reliable approach to motif discovery is to mine the published literature for experimental evidence. For 3D motifs, the focus is on residues that provide a specific binding or catalytic capability.

The Catalytic Site Atlas (CSA) (Table 11.1) contains several hundred families of enzymes, each comprised of a structure with catalytic residue annotations from the literature (Barker and Thornton 2003; Porter et al. 2004; Torrance et al. 2005). The atlas library also includes structures related through sequence homology. Representative structural templates (3D motifs) are based on side-chain functional atoms, alpha carbons (Cα) and beta carbons (Cβ). In all, more than 2200 unique motifs were generated, whose function is verified through literature values, which often include experimental verification of the function.

The generation of this dataset was a fundamental advance in the field. Other servers rely on this dataset, including multiservers like ProFunc, (Laskowski et al. 2005), and groups who have curated or modified this Atlas and incorporated it in their own servers (Moll et al. 2011; Kirshner et al. 2013; Nilmeier et al. 2013).

### 11.4.2 Annotated Sites in PDB Structures

Another approach is to use the annotations given to the crystallographic structures in the PDB. In practice, this means looking at the SITE records of a given protein databank file, or at residues around molecules labelled as LIGAND. Sometimes even the residues around nonspecific heteroatoms (HET) or analysis of the residues of macromolecular interfaces can give some clue as to what portions of a protein may be involved in catalysis. This is not always informative, as these annotations are not guaranteed to point to the catalytic site of the protein of interest. It is often a very good starting point, however, and can provide new hypothesis for motifs.

Several databases of 3D motifs have been generated using only information from each source structure individually. For example, binding site motifs can be collected by taking residues within a cutoff distance of ligands, nucleic acids, or even other protein chains. The PINTS (Patterns in Non-homologous Tertiary Structures) server (Stark and Russell 2003) derives its database from binding sites defined as residues within three angstroms of a ligand as well as motifs annotated in the PDB as a SITE record (Russell 1998), along with careful statistical models (Stark et al. 2003, 2004) that estimate the statistical significance of matches. The PDBSite database (Ivanisenko et al. 2005) (Table 11.1) includes SITE records, along with interfacial reaction sites with other proteins, RNA, and DNA. Residues with at least three atoms within five angstroms of the other chain are included in an interaction site. The search machinery is called PDBSiteScan (Ivanisenko et al. 2004) (Table 11.1). The pdbFun web server (Ausiello et al. 2005b) uses sites defined as residues within 3.5 angstroms of a ligand (Ausiello et al. 2005a).

### *11.4.3 Mining for Emergent Properties*

When groups of structures are studied, local structural features shared among proteins may be taken as a 3D motif. The process of identifying these common features may be described as the ***mining*** step. It is helpful to separately identify the grouping methods as either ***undirected*** or ***directed***. In general, undirected (unsupervised) mining methods do not specifically use labels or annotations in the grouping step, while directed (supervised) mining methods tend to use labelled structures. Each approach will be discussed in the following sections.

In some cases investigators provide a mining toolset for the user. The technology is focused on mining the pattern or motif from a group, rather than in how the groups are defined. The *applications* of these methods are, in general, directed mining approaches. At the heart of these techniques is a search for a ***clique*** that is common to the grouping that can be interpreted as a functionally important motif. Methods such as the common structural cliques method (Milik et al. 2003), the maximum common clique (ProBIS) algorithm (Konc and Janežič 2010), as well as the Detection of REcurring Sidechain PATterns (DRESPAT), (Kar et al. 2012) are all designed to locate maximal cliques among sets of structures.

In other cases, the approaches for determining a motif are more dependent on the nature of the groupings: these are discussed in the next sections.

#### 11.4.3.1 Undirected Mining

Undirected mining refers to finding common patterns in unannotated, or ***unlabelled*** structures. The undirected mining approaches have elements of what is usually considered ***unsupervised learning***. For example, many of these approaches make ***all-to-all*** similarity comparisons (Russell 1998), which has some analogy to the notion of a ***distance matrix*** as seen in traditional clustering methods. Structures with sufficiently similar measures are grouped as a cluster. Other methods count motifs that appear with relatively high frequency (Oldfield 2002), and consider the structures having those motifs as a grouping.

Mining techniques apply to both unlabelled and labelled groupings, as well as cases where the distinction between unlabelled and labelled is not always straightforward. For example, a study that used groups of structures with similarity to sites with hypothesized function (Ausiello et al. 2007) was able to detect and propose new motifs. The reference structures were based on sequence similarity, proximity to a co-crystallized ligand, or contact with a cavity, but did not have a specific functional annotation.

#### 11.4.3.2 Directed Mining

In directed mining, the focus is on the use of ***labelled*** examples to suggest geometric features (residues) that are common, both within the labelled dataset and other structures that may be deemed similar to the labelled dataset. Directed mining may also be considered to be more of a targeted search for motifs and themes within a given group.

In general, only ***positive examples*** are used for motif discovery. Positive examples are those structures whose labels indicate a positive membership in the functional set. The motif discovery process is then to find what essential features define that set. The use of ***negative examples*** is not as frequent in the motif discovery process. It does, however, appear in the validation of the models. One notable exception to this approach is the GASPS method (Polacco and Babbitt 2006), which uses both positive and negative examples in the motif discovery process, and is discussed in the next section.

It is often more practical to develop motifs from crystallographic structures where the ligand is present. Studies of this sort tend to be more specific to the ligand types of interest. For example, one of the early approaches was developed for adenine mononucleotide sites, based on the fact that there were over 100 structures available for comparison at the time (Kobayashi and Go 1997). A high similarity was found between structures of different folds, which is a hallmark of a good motif. Later, after many more structures had become available, a similar approach was used to generate consensus binding-site motifs (Nebel et al. 2007), and the study was expanded to study mono-, di-, and tri-phosphate complexes as well, resulting in 13 high quality motifs. The same group developed motifs specifically for porphyrin-binding sites (Nebel 2006). Another study used phosphate groups as the ligand in protein-nucleotide complexes, and applied a clique detection algorithm to discover motifs (Brakoulias and Jackson 2004).

Other methods use more standard template-matching programs, but on smaller motifs, with emergent motifs built from the smaller ones. The funClust server (Ausiello et al. 2008) (Table 11.1) identifies 3D motifs shared by up to 20 input structures. The structures are then filtered by sequence identity and other geometric filters, and the comparison is made with Query3D (Ausiello et al. 2005a). Another method, the PAR-3D (Protein Active site Residues using 3-Dimensional structural motifs) server (Goyal et al. 2007) (Table 11.1) compares a structure to motifs for proteases, glycolysis enzymes, and metalloenzyme sites with only three or four residues (Goyal and Mande 2008) that are common to the broadly defined functions. The motifs returned are given as allowed ranges of interatomic distances to the library of motifs. Another approach, termed Geometric Sieving, starts with an existing motif or list of putatively important residues (Chen et al. 2007b), and develops candidate motifs by comparing them to a representative sample of structures. It is assumed that the low-RMSD tail in a distribution represents true positives.

#### 11.4.3.3 Directed Mining with Positive and Negative Examples

In most of the approaches listed above, only sets with known positives are used to discover the emergent features of a binding-site. Sometimes, however, it is important to know not just consensus features of a catalytic site, but the *essential* features. For this more subtle delineation, negative examples are needed to more precisely define what is an outlier.

For example, a simple mutation from Asp to Glu in a set of binding site residues may still preserve function, while a mutation from Asp to Asn may remove function completely if the residue needs to be protonated at some point in the catalysis. If, however, the residue only needs to be polar, then the Asp to Asn mutation might still be allowable.

These types of differences may not be easily seen by consensus methods, but some very carefully chosen negative examples can reveal these more subtle differences. The use of negative training examples is well understood in machine learning approaches with linear models. Here, the goal is to discover geometric features, rather than to apply a fitting procedure to determine parameters for a linear model. This presents a fundamentally different optimization problem.

One very successful approach to this problem is GASPS (Genetic Algorithm Search for Patterns in Structures), which finds patterns of residues that best separate the two groups (Polacco and Babbitt 2006). No prior residues list is required, and how the positive/negative groups are defined is independent of the method. The underlying search tool is SPASM (Kleywegt 1999), with residues represented by alpha-carbons and side chain centroids and only identical residue types allowed to match. To limit the search space, GASPS considers only the 100 most conserved residues in a structure chain, based on an automatically constructed sequence alignment. An initial candidate motif is constructed by picking one residue randomly and then choosing four more, also randomly except in the vicinity of the first. Each of 50 initial candidates is scored on how well it separates the positive and negative structures in terms of best match RMSD values. In each round of the genetic algorithm, the 16 highest-scoring motifs are used as the parents of 36 new motifs, and the top-scoring motif after 50 rounds is declared the winner. Motifs are allowed to contain from three to ten residues. Sensitive and specific motifs were obtained for diverse superfamilies (Babbitt and Gerlt 2000) and serine proteases. Most of the residues in the motifs were functionally important, but in some cases, residues with no known functional role were found to be equally predictive (Polacco and Babbitt 2006).

The GASPSdb database (Table 11.1) allows browsing and downloading 3D motifs previously generated by GASPS for SCOP families and superfamilies.

## 11.5 Molecular Docking for Functional Annotation

Ultimately, the ligand specificity and catalytic capabilities of a protein depend on the arrangement of atoms in its binding or active site(s). The use of 3D motifs can be seen as informatics approaches that are informed by the chemistry of the protein. These methods are limited in that they can only associate function to known motifs. There are many cases, however, where a high resolution target structure is available (either experimentally or through homology modelling), but there is no identifiable motif in the structure. For these cases, a more fundamental physical approach can fill in gaps in knowledge that the informatics approaches do not provide.

A computational method known as ***ligand docking*** can provide a different perspective on the problem of functional annotation (Jacobson et al. 2014). This technique (also mentioned in Chap. 10) uses molecular mechanics forcefields to directly estimate ligand protein energetics and complementarities. The field of docking is vast, and we list only a few examples for reference (Meng et al. 1992; Wang et al. 2003). As the name suggests, the molecule is 'docked' into the target protein, and the quality of the resulting pose is evaluated for fitness. Figure 11.6 illustrates a typical workflow that uses docking as a method for functional assignments. In general, the target is held rigid, but more recent approaches also allow for sidechain flexibility (Sherman et al. 2006). Since it is based on molecular interaction energies, this technique can conceivably predict molecular binding modes that are novel, but still physically reasonable.

Traditionally, database docking, or **in silico screening** has been applied to the ***lead discovery*** phase of drug design pipelines. As such, the technique is highly automated, and designed to dock large libraries of small molecules to selected targets (on the order of a million of compounds or more in some cases). While most ligand docking studies are focused on finding *inhibitors* to the target, the functional annotation effort seeks to find the *native metabolite* that is catalysed in the target. Many of the technical challenges in ligand docking are common to both goals, however, such as the need to distinguish true positives from false positives, or ***decoys*** (Huang et al. 2006). These studies highlight the need not only for high quality poses, but also for scoring procedures that will properly rank ligand affinities. Metabolite docking can be distinct from inhibitor docking, most notably due to the fact that most metabolites are highly charged (Song et al. 2007).

Despite these challenges, this approach has received considerable attention (Favia et al. 2008; Kalyanaraman et al. 2005; Macchiarulo et al. 2004; Paul et al. 2004; Tyagi and Pleiss 2006; Jacobson et al. 2014) In particular, studies of alpha-beta barrel enzymes (Song et al. 2007) and amidohydrolases (Hermann et al. 2007) have firmly established the capabilities of docking approaches as a supplement to approaches using sequence- and motif-based comparative approaches.

As these approaches have progressed, an emergent challenge for functional annotation is to not only generate comparative affinities for a particular target, but also to be able to compare affinities *across* targets. While inhibitor design is usually focused on a single target, the goal of functional annotation is to characterize entire

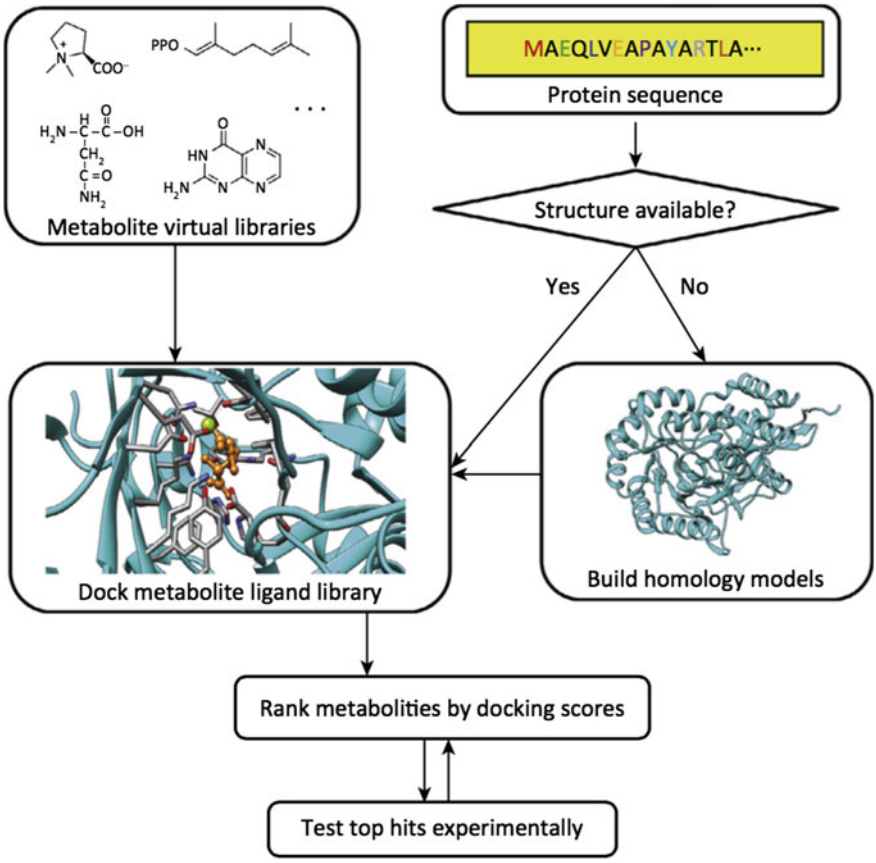



**Fig. 11.6** Structure-based virtual metabolite docking protocol for enzyme activity prediction. When no structure has been experimentally determined for a protein sequence, a model can be built using a variety of comparative modelling methods, if sequence identity is approximately 30% or more. Whether using a structure of a model, it is critical that active site metal ions and cofactors are present, and that catalytic residues are positioned appropriately for catalysis. Virtual metabolites libraries can be constructed and 'docked' against the putative active sites of structures or models using computational tools more commonly used in structure-based drug design (e.g., Glide or DOCK). The docking scoring functions can be used to rank the ligands according to their estimated relative binding affinities. Top-scoring metabolites are typically inspected for plausibility and then selected for in vitro testing. (This Figure was reprinted from Jacobson et al. (2014) with permission from Elsevier License #3624901501981)

synthetic pathways or proteomes in an automated fashion. Studying target groups for entire synthetic pathways provides a much larger perspective, as the molecules are related by a chain of incremental modifications, and the targets are often expressed from the same '***genome neighborhood***'. Applying these additional guidelines for self consistency, while also using homology modelling to construct missing targets, can allow for elucidation of complete pathways that were

previously unknown (Zhao et al. 2013), with potential applications to synthetic biology and other efforts that have not traditionally relied on structure-based techniques (Jacobson et al. 2014).

While molecular recognition techniques are significantly more computationally demanding than 3D motif matching, docking has the potential to extrapolate to functions not associated with previously characterized structures, and represents a frontier direction in the field for the most challenging of catalytic sites.

## 11.6 Discussion and Conclusions

The question of how best to describe the function of a protein with a meaningful language remains. While fold-based methods and ligand-based methods have been shown to be very useful, the use of a 3D motif as a signature for protein function has offered new perspectives on catalytic sites, and could ultimately form the foundation of a functional annotation language. Challenges remain on how to identify these motifs, and even with knowledge of the substrate and many examples, it can be nontrivial to identify the ideal 3D motif that uniquely and completely defines function for a given enzyme.

What, then, is the most natural classification of protein function, if we choose 3D motifs as a basis for classification? In enzymes, individual residues or functional groups play different roles in the course of a reaction: substrate recognition, catalysis of a particular step in the reaction, stabilization of an intermediate, or some combination of these. As proteins evolve to perform new functions, they can make use of existing pieces of catalytic machinery that carry out a ***common partial reaction*** (Babbitt and Gerlt 2000; Bartlett et al. 2003). This explains in part why members of a homologous but diverse group of enzymes often make use of the same configuration of a small number of amino acids, despite catalysing different overall reactions. It may well be that these subunits (which are 3D motifs) will form the basic building blocks of all enzymes, and a functional classification scheme should include these basic units in its language.

**Acknowledgements** We acknowledge support from NIH GM60595 and NSF DBI-0234768. Molecular graphics were produced with the UCSF Chimera package from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41-GM103311). We thank Jacquelyn Fetrow and Stacy Knutson (Wake Forest University) for providing Fig. 11.5 as an example of a result from their FFF/DASP/PASS motif analysis software. We gratefully acknowledge Dan Kirshner for enlightening discussions and a critical reading of the manuscript.

## References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410

Andreeva A, Howorth D, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2004) SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 32(suppl 1):D226–D229

Andreeva A, Howorth D, Chandonia J-M, Brenner SE, Hubbard TJ, Chothia C, Murzin AG (2008) Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 36 (suppl 1):D419–D425

Arakaki A, Huang Y, Skolnick J (2009) EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinform 10(1):107

Artymiuk PJ, Poirrette AR, Grindley HM, Rice DW, Willett P (1994) A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J Mol Biol 243(2):327–344

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29

Ausiello G, Gherardini PF, Marcatili P, Tramontano A, Via A, Helmer-Citterich M (2008) FunClust: a web server for the identification of structural motifs in a set of non-homologous protein structures. BMC Bioinform 9(Suppl 2):S2

Ausiello G, Peluso D, Via A, Helmer-Citterich M (2007) Local comparison of protein structures highlights cases of convergent evolution in analogous functional sites. BMC Bioinform 8 (Suppl 1):S24

Ausiello G, Via A, Helmer-Citterich M (2005a) Query3D: a new method for high-throughput analysis of functional residues in protein structures. BMC Bioinform 6(Suppl 4):S5

Ausiello G, Zanzoni A, Peluso D, Via A, Helmer-Citterich M (2005b) pdbFun: mass selection and fast comparison of annotated PDB residues. Nucleic Acids Res 33 (Web Server issue):W133–137

Babbitt PC (2003) Definitions of enzyme function for the structural genomics era. Curr Opin Chem Biol 7(2):230–237

Babbitt PC, Gerlt JA (1997) Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities. J Biol Chem 272(49):30591–30594

Babbitt PC, Gerlt JA (2000) New functions from old scaffolds: how nature reengineers enzymes for new functions. Adv Protein Chem 55:1–28

Bagley SC, Altman RB (1995) Characterizing the microenvironment surrounding protein sites. Protein Sci 4(4):622–635

Bairoch A (1994) The ENZYME data bank. Nucleic Acids Res 22(17):3626–3627

Baker D, Sali A (2001) Protein structure prediction and structural genomics. Science 294 (5540):93–96

Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19(13):1644–1649

Bartlett GJ, Borkakoti N, Thornton JM (2003) Catalysing new reactions during evolution: economy of residues and mechanism. J Mol Biol 331(4):829–860

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28(1):235–242

Blow DM, Birktoft JJ, Hartley BS (1969) Role of a buried acid group in the mechanism of action of chymotrypsin. Nature 221(5178):337–340

Brakoulias A, Jackson RM (2004) Towards a structural classification of phosphate binding sites in protein–nucleotide complexes: an automated all-against-all structural comparison using geometric matching. Proteins Struct Funct Bioinf 56(2):250–260

Brylinski M, Skolnick J (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. Proc Natl Acad Sci 105(1):129

Buturovic L, Wong M, Tang GW, Altman RB, Petkovic D (2014) High precision prediction of functional sites in protein structures. Publ Libr Sci One 9(3):e91240

Cammer SA, Hoffman BT, Speir JA, Canady MA, Nelson MR, Knutson S, Gallina M, Baxter SM, Fetrow JS (2003) Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 334(3):387–401

Chen BY, Bryant DH, Cruess AE, Bylund JH, Fofanov VY, Kristensen DM, Kimmel M, Lichtarge O, Kavraki LE (2007a) Composite motifs integrating multiple protein structures increase sensitivity for function prediction. Comput Syst Bioinform Conf 6:343–355

Chen BY, Fofanov VY, Bryant DH, Dodson BD, Kristensen DM, Lisewski AM, Kimmel M, Lichtarge O, Kavraki LE (2007b) The MASH pipeline for protein function prediction and an algorithm for the geometric refinement of 3D motifs. J Comput Biol 14(6):791–816

Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. Science 300(5626):1701–1703

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826

Conte LL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C (2000) SCOP: a structural classification of proteins database. Nucleic Acids Res 28(1):257–259

Devos D, Valencia A (2001) Intrinsic errors in genome annotation. Trends Genet 17(8):429–431

Di Gennaro JA, Siew N, Hoffman BT, Zhang L, Skolnick J, Neilson LI, Fetrow JS (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. J Struct Biol 134(2–3):232–245

Favia AD, Nobeli I, Glaser F, Thornton JM (2008) Molecular docking for substrate identification: the short-chain dehydrogenases/reductases. J Mol Biol 375(3):855–874

Fetrow JS, Skolnick J (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 281(5):949–968

Fischer D, Wolfson H, Lin SL, Nussinov R (1994) Three-dimensional, sequence order-independent structural comparison of a serine protease against the crystallographic database reveals active site similarities: potential implications to evolution and to protein folding. Protein Sci 3(5):769–778

Furnham N, Holliday GL, de Beer TA, Jacobsen JO, Pearson WR, Thornton JM (2014) The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 42 (D1):D485–D489

Galperin MY, Walker DR, Koonin EV (1998) Analogous enzymes: independent inventions in enzyme evolution. Genome Res 8(8):779–790

Gerlt JA, Allen KN, Almo SC, Armstrong RN, Babbitt PC, Cronan JE, Dunaway-Mariano D, Imker HJ, Jacobson MP, Minor W (2011) The enzyme function initiative. Biochem

Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 70(1):209–246

Gerlt JA, Babbitt PC, Jacobson MP, Almo SC (2012) Divergent evolution in enolase superfamily: strategies for assigning functions. J Biol Chem 287(1):29–34

Glanville JG, Kirshner D, Krishnamurthy N, Sjölander K (2007) Berkeley phylogenomics group web servers: resources for structural phylogenomic analysis. Nucleic Acids Res 35(suppl 2): W27–W32

Glazer DS, Radmer RJ, Altman RB Combining molecular dynamics and machine learning to improve protein function recognition. In: Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, 2008. NIH Public Access, p 332

Gold ND, Jackson RM (2006a) Fold independent structural comparisons of protein-ligand binding sites for exploring functional relationships. J Mol Biol 355(5):1112–1124

Gold ND, Jackson RM (2006b) SitesBase: a database for structure-based protein–ligand binding site comparisons. Nucleic Acids Res 34(suppl 1):D231–D234

Goyal K, Mande SC (2008) Exploiting 3D structural templates for detection of metal-binding sites in protein structures. Proteins 70(4):1206–1218

Goyal K, Mohanty D, Mande SC (2007) PAR-3D: a server to predict protein active site residues. Nucleic Acids Res 35 (Web Server issue):W503–505

Halgren T (2007) New method for fast and accurate binding-site Identification and analysis. Chem Biol Drug Des 69(2):146–148

Halgren TA (2009) Identifying and characterizing binding sites and assessing druggability. J Chem Inf Model 49(2):377–389

Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448 (7155):775–779

Holliday GL, Almonacid DE, Bartlett GJ, O'Boyle NM, Torrance JW, Murray-Rust P, Mitchell JBO, Thornton JM (2007) MACiE (mechanism, annotation and classification in enzymes): novel tools for searching catalytic mechanisms. Nucleic Acids Res 35(suppl 1): D515–D520

Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking sets for molecular docking. J Med Chem 49(23):6789–6801

International Union of Biochemistry and Molecular Biology: Nomenclature Committee, Webb EC (1992) Enzyme nomenclature 1992: recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes. Academic Press, San Diego

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2004) PDBSiteScan: a program for searching for active, binding and posttranslational modification sites in the 3D structures of proteins. Nucleic Acids Res 32(Web Server issue):W549–554

Ivanisenko VA, Pintus SS, Grigorovich DA, Kolchanov NA (2005) PDBSite: a database of the 3D structure of protein functional sites. Nucleic Acids Res 33(Database issue):D183–187

Jacobson MP, Kalyanaraman C, Zhao S, Tian B (2014) Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. Trends Biochem Sci 39(8):363–371

Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C (2005) The SuMo server: 3D search for protein functional sites. Bioinformatics 21(20):3929–3930

Jambon M, Imberty A, Deléage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Proteins Struct Funct Bioinf 52(2):137–145

Kalyanaraman C, Bernacki K, Jacobson MP (2005) Virtual screening against highly charged active sites: identifying substrates of alpha-beta barrel enzymes. Biochemistry 44(6):2059–2071

Kalyanaraman C, Imker HJ, Fedorov AA, Fedorov EV, Glasner ME, Babbitt PC, Almo SC, Gerlt JA, Jacobson MP (2008) Discovery of a dipeptide epimerase enzymatic function guided by homology modeling and virtual screening. Structure 16(11):1668–1677

Kaminski GA, Friesner RA, Tirado-Rives J, Jorgensen WL (2001) Evaluation and reparametrization of the OPLS-AA force field for proteins via comparison with accurate quantum chemical calculations on peptides. J Phys Chem B 105(28):6474–6487

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28(1):27–30

Kar S, Vijayakeerthi D, Tendulkar AV, Ravindran B Functional site prediction by exploiting correlations between labels of interacting residues. In: Proceedings of the ACM conference on bioinformatics, computational biology and biomedicine, 2012. ACM, pp 76–83

Kinjo AR, Nakamura H (2007) Similarity search for local protein structures at atomic resolution by exploiting a database management system. Biophysics 3:75–84

Kinoshita K, Nakamura H (2003) Identification of protein biochemical functions by similarity search using the molecular surface database eF-site. Protein Sci 12(8):1589–1595

Kirshner DA, Nilmeier JP, Lightstone FC (2013) Catalytic site identification—a web server to identify catalytic site structural matches throughout PDB. Nucleic Acids Res 41 (W1):W256–W265

Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285 (4):1887–1897

Kleywegt GJ, Jones TA (1997) Detecting folding motifs and similarities in protein structures. Methods Enzymol 277:525–545
Kleywegt GJ, Lamerichs RMJN, Boelens R, Kaptein R (1989) Toward automatic assignment of protein 1H NMR spectra. J Magn Reson 85(1):186–197
Kobayashi N, Go N (1997) A method to search for similar protein local structures at ligand binding sites and its application to adenine recognition. Eur Biophys J 26(2):135–144
Konc J, Janežič D (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26(9):1160–1168
Konc J, Janežič D (2012) ProBiS-2012: web server and web services for detection of structurally similar binding sites in proteins. Nucleic Acids Res 40(W1):W214–W221
Krishnamurthy N, Brown DP, Kirshner D, Sjölander K (2006) PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification. Genome Biol 7(9):R83
Krogh A, Brown M, Mian IS, Sjolander K, Haussler D (1994) Hidden Markov models in computational biology: applications to protein modeling. J Mol Biol 235(5):1501–1531
Kuhn D, Weskamp N, Schmitt S, Hullermeier E, Klebe G (2006) From the similarity analysis of protein cavities to the functional classification of protein families using cavbase. J Mol Biol 359(4):1023–1044
Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23 (21):2947–2948
Laskowski RA, Watson JD, Thornton JM (2005) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33(Web Server issue):W89–W93
Liang MP, Banatao DR, Klein TE, Brutlag DL, Altman RB (2003) WebFEATURE: an interactive web tool for identifying and visualizing functional sites on macromolecular structures. Nucleic Acids Res 31(13):3324–3327
Lichtarge O, Bourne HR, Cohen FE (1996) An evolutionary trace method defines binding surfaces common to protein families. J Mol Biol 257(2):342–358
Macchiarulo A, Nobeli I, Thornton JM (2004) Ligand selectivity and competition between enzymes in silico. Nat Biotechnol 22(8):1039–1045
Meng EC, Polacco BJ, Babbitt PC (2004) Superfamily active site templates. Proteins Struct Funct Bioinf 55(4):962–976
Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comput Chem 13(4):505–524
Milik M, Szalma S, Olszewski KA (2003) Common structural cliques: a tool for protein structure and function analysis. Protein Eng 16(8):543–552
Mitchell EM, Artymiuk PJ, Rice DW, Willett P (1990) Use of techniques derived from graph theory to compare secondary structure motifs in proteins. J Mol Biol 212(1):151–166
Moll M, Bryant DH, Kavraki LE (2010) The LabelHash algorithm for substructure matching. BMC Bioinform 11(1):555
Moll M, Bryant DH, Kavraki LE (2011) The LabelHash server and tools for substructure-based functional annotation. Bioinformatics 27(15):2161–2162
Moll M, Kavraki LE (2008) LabelHash: a flexible and extensible method for matching structural motifs. Available from Nature Precedings. http://dx.doi.org/10.1038/npre.2008.2199.1
Mooney SD, Liang MH, DeConde R, Altman RB (2005) Structural characterization of proteins using residue environments. Proteins 61(4):741–747
Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540
Nebel JC (2006) Generation of 3D templates of active sites of proteins with rigid prosthetic groups. Bioinformatics 22(10):1183–1189
Nebel JC, Herzyk P, Gilbert DR (2007) Automatic generation of 3D motifs for classification of protein binding sites. BMC Bioinformatics 8(1):321
Nilmeier JP, Kirshner DA, Wong SE, Lightstone FC (2013) Rapid catalytic template searching as an enzyme function prediction procedure. Publ Libr Sci One 8(5):e62535

Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M (1999) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 27(1):29–34

Oldfield TJ (2002) Data mining the protein data bank: residue interactions. Proteins 49(4):510–528

Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) CATH–a hierarchic classification of protein domain structures. Structure 5(8):1093–1108

Orengo CA, Pearl FMG, Bray JE, Todd AE, Martin A, Conte LL, Thornton JM (1999) The CATH Database provides insights into protein structure/function relationships. Nucleic Acids Res 27 (1):275–279

Orengo CA, Pearl FMG, Thornton JM (2003) The CATH domain structure database. Struct Bioinform 249–271

Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13 (1):121–130

Paul N, Kellenberger E, Bret G, Muller P, Rognan D (2004) Recovering the true targets of specific ligands by virtual screening of the protein data bank. Proteins 54(4):671–680

Pegg SC, Brown S, Ojha S, Huang CC, Ferrin TE, Babbitt PC (2005) Representing structure-function relationships in mechanistically diverse enzyme superfamilies. Pac Symp Biocomput 358–369

Pegg SCH, Brown SD, Ojha S, Seffernick J, Meng EC, Morris JH, Chang PJ, Huang CC, Ferrin TE, Babbitt PC (2006) Leveraging enzyme structure-function relationships for functional inference and experimental design: the structure-function linkage database. Biochemistry 45(8):2545–2555

Pennec X, Ayache N (1998) A geometric algorithm to find small but highly similar 3D substructures in proteins. Bioinformatics 14(6):516–522

Peters B, Moad C, Youn E, Buffington K, Heiland R, Mooney S (2006) Identification of similar regions of protein structures using integrated sequence and structure analysis tools. BMC Struct Biol 6:4

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) UCSF Chimera–a visualization system for exploratory research and analysis. Journal of Computaional Chemistry 25(13):1605–1612

Polacco BJ, Babbitt PC (2006) Automated discovery of 3D motifs for protein function annotation. Bioinformatics 22(6):723–730

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32(suppl 1):D129–D133

Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A (2013) A large-scale evaluation of computational protein function prediction. Nat Methods 10(3):221–227

Ren J, Xie L, Li WW, Bourne PE (2010) SMAP-WS: a parallel web service for structural proteome-wide ligand-binding site comparison. Nucleic Acids Res 38(suppl 2):W441–W444

Richardson JS (1981) The anatomy and taxonomy of protein structure. Adv Protein Chem 34:167–339

Rost B (1997) Protein structures sustain evolutionary drift. Fold Des 2(3):S19–S24

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2):595–608

Russell RB (1998) Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution. J Mol Biol 279(5):1211–1227

Sankararaman S, Sha F, Kirsch JF, Jordan MI, Sjölander K (2010) Active site prediction using evolutionary and structural information. Bioinformatics 26(5):617–624

Sankararaman S, Sjölander K (2008) INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification. Bioinformatics 24(21):2445–2452

Schmitt S, Kuhn D, Klebe G (2002) A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 323(2):387–406

Sherman W, Day T, Jacobson MP, Friesner RA, Farid R (2006) Novel procedure for modeling ligand/receptor induced fit effects. J Med Chem 49(2):534–553

Shindyalov IN, Bourne PE (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11(9):739–747
Shindyalov IN, Bourne PE (2001) A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. Nucleic Acids Res 29(1):228–229
Shulman-Peleg A, Nussinov R, Wolfson HJ (2004) Recognition of functional sites in protein structures. J Mol Biol 339(3):607–633
Shulman-Peleg A, Nussinov R, Wolfson HJ (2005) SiteEngines: recognition and comparison of binding sites and protein-protein interfaces. Nucleic Acids Res 33(Web Server issue):W337–W341
Sjölander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D (1996) Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci CABIOS 12(4):327–345
Skolnick J, Brylinski M (2009) FINDSITE: a combined evolution/structure-based approach to protein function prediction. Briefings Bioinform 10(4):378–391
Song L, Kalyanaraman C, Fedorov AA, Fedorov EV, Glasner ME, Brown S, Imker HJ, Babbitt PC, Almo SC, Jacobson MP (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 3(8):486–491
Spriggs RV, Artymiuk PJ, Willett P (2003) Searching for patterns of amino acids in 3D protein structures. J Chem Inf Comput Sci 43(2):412–421
Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 31(13):3341–3344
Stark A, Shkumatov A, Russell RB (2004) Finding functional sites in structural genomics proteins. Structure 12(8):1405–1412
Stark A, Sunyaev S, Russell RB (2003) A model for statistical significance of local similarities in structure. J Mol Biol 326(5):1307–1316
Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673–4680
Tian W, Arakaki AK, Skolnick J (2004) EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 32(21):6226–6239
Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333(4):863–882
Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143
Todd AE, Orengo CA, Thornton JM (2002) Plasticity of enzyme active sites. Trends Biochem Sci 27(8):419–426
Torrance JW, Bartlett GJ, Porter CT, Thornton JM (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. J Mol Biol 347 (3):565–581
Tseng YY, Dundas J, Liang J (2009) Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J Mol Biol 387(2):451–464
Tseng YY, Liang J (2006) Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. Mol Biol Evol 23(2):421–436
Tyagi S, Pleiss J (2006) Biochemical profiling in silico–predicting substrate specificities of large enzyme families. J Biotechnol 124(1):108–116
Ullmann JR (1976) An algorithm for subgraph isomorphism. J ACM (JACM) 23(1):31–42
Wallace AC, Borkakoti N, Thornton JM (1997) TESS: a geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. Protein Sci 6:2308–2323
Wallace AC, Laskowski RA, Thornton JM (1996) Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci 5(6):1001–1013

Wang R, Lu Y, Wang S (2003) Comparative evaluation of 11 scoring functions for molecular docking. J Med Chem 46(12):2287–2303

Webb EC (1992) Enzyme nomenclature 1992. In: Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes, vol Ed. 6. Academic Press

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36(03):307–340

Wolfson HJ, Rigoutsos I (1997) Geometric hashing: An overview. Comput Sci Eng IEEE 4(4):10–21

Wright CS, Alden RA, Kraut J (1969) Structure of subtilisin BPN' at 2.5 angstrom resolution. Nature 221(5177):235–242

Xie L, Bourne PE (2008) Detecting evolutionary relationships across existing fold space, using sequence order-independent profile–profile alignments. Proc Natl Acad Sci 105(14):5441

Xie L, Bourne PE (2009) A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. Bioinformatics 25(12):i305–i312

Yang LW, Bahar I (2005) Coupling between catalytic site and collective dynamics: a requirement for mechanochemical activity of enzymes. Structure 13(6):893–904

Zemla A (2003) LGA: a method for finding 3D similarities in protein structures. Nucleic Acids Res 31(13):3370–3374

Zhao S, Kumar R, Sakai A, Vetting MW, Wood BM, Brown S, Bonanno JB, Hillerich BS, Seidel RD, Babbitt PC (2013) Discovery of new enzymes and metabolic pathways by using structure and genome context. Nature 502(7473):698–702

# Chapter 12
# Protein Dynamics: From Structure to Function

**Marcus B. Kubitzki, Bert L. de Groot and Daniel Seeliger**

**Abstract** Understanding protein function requires detailed knowledge about protein dynamics, i.e. the different conformational states the system can adopt. Despite substantial experimental progress, simulation techniques such as molecular dynamics (MD) currently provide the only routine means to obtain dynamical information at an atomic level on timescales of nano- to microseconds. Even with the current development of computational power, sampling techniques beyond MD are necessary to enhance conformational sampling of large proteins and assemblies thereof. The use of collective coordinates has proven to be a promising means in this respect, either as a tool for analysis or as part of new sampling algorithms. Starting from MD simulations, several enhanced sampling algorithms for biomolecular simulations are reviewed in this chapter. Examples are given throughout illustrating how consideration of the dynamic properties of a protein sheds light on its function.

**Keywords** Protein dynamics · Molecular dynamics · Conformational sampling · Collective coordinates · Collective degrees of freedom · Enhanced sampling · Replica exchange · Principal component analysis/PCA · Essential dynamics · TEE-REX · CONCOORD/tCONCOORD · Geometrical constraints

## 12.1 Molecular Dynamics Simulations

Over the last decades, experimental techniques have made substantial progress in revealing the three-dimensional structure of proteins, in particular X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy and cryo-electron

M.B. Kubitzki · B.L. de Groot
Computational Biomolecular Dynamics Group, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Goettingen, Germany

D. Seeliger (✉)
Medicinal Chemistry, Boehringer Ingelheim Pharma GmbH & Co KG, Birkendorferstrasse 65, 88397 Biberach, Germany
e-mail: daniel.seeliger@boehringer-ingelheim.com



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_12

microscopy. Going beyond the static picture of single protein structures has proven to be more challenging, although, a number of techniques such as NMR relaxation, fluorescence spectroscopy or time-resolved X-ray crystallography have emerged (Kempf and Loria 2003; Weiss 1999; Moffat 2003; Schotte et al. 2003), yielding information about the inherent conformational flexibility of proteins. Despite this enormous variety, experimental techniques having spatio-temporal resolution in the nano- to microsecond as well as the nanometre regime are not routinely available, and thus information on the conformational space accessible to proteins in vivo often remains obscure. In particular, details on the pathways between different known conformations, frequently essential for protein function, are usually unknown. Here, computer simulation techniques provide an attractive possibility to obtain dynamic information on proteins at atomic resolution in the microsecond time range. Of all ways to simulate protein motions (Adcock and McCammon 2006), molecular dynamics (MD) techniques are among the most popular.

Since the first report of MD simulations of a protein some 30 years ago (McCammon et al. 1977), MD has become an established tool in the study of biomolecules. Like all computational branches of science, the MD field benefits from the ever increasing improvements in computational power. This progression also allowed for advancements in simulation methodology that have led to a large number of algorithms for such diverse problems as cellular transport, signal transduction, allostery, cellular recognition, ligand-docking, the simulation of atomic force microscopy and enzymatic catalysis.

### *12.1.1 Principles and Approximations*

Despite substantial algorithmic advances, the basic theory behind MD simulations is fairly simple. For biomolecular systems having N particles, the numerical solution of the time-dependent Schrödinger equation

$$i\hbar \frac{\partial}{\partial t}\psi(r,t) = H\psi(r,t)$$

for the N-particle wave function $\psi(r, t)$ of the system is prohibitive. Several approximations are therefore required to allow the simulation of solvated biomolecules at timescales on the order of microseconds. The first of these relates to positions of nuclei and electrons: due to the much lower mass and consequently much higher velocity of the electrons compared to the nuclei, electrons can often be assumed to instantaneously follow the motion of the nuclei. Thus, within the Born-Oppenheimer approximation, only the nuclear motion has to be considered, with the electronic degrees of freedom influencing the dynamics of the nuclei in the form of a potential energy surface V(r).

The second essential approximation used in MD is to describe nuclear motion classically by Newton's equations of motion

$$m_i \frac{d^2 r_i}{dt^2} = -\nabla_i V(r_1, \ldots, r_N),$$

where $m_i$ and $r_i$ are the mass and the position of the i-th nucleus. With the nuclear motion described classically, the Schrödinger equation for the electronic degrees of freedom has to be solved to obtain the potential energy V(r). However, due to the large number of electrons involved, a further simplification is necessary. A semi-empirical force field is introduced which approximates V(r) by a large number of functionally simple energy terms for bonded and non-bonded interactions. In its general form

$$\begin{aligned} V(r) &= V_{bonds} + V_{angles} + V_{dihedrals} + V_{improper} + V_{Coul} + V_{LJ} \\ &= \sum_{bonds} \frac{1}{2} k_i^l (l_i - l_{i,0})^2 + \sum_{angles} \frac{1}{2} k_i^\theta (\theta_i - \theta_{i,0})^2 \\ &+ \sum_{dihedrals} \frac{V_n}{2} (1 + \cos(n\phi - \delta)) + \sum_{improper} \frac{1}{2} k_\xi (\xi_{ijkl} - \xi_0)^2 \\ &+ \sum_{i,j;i\neq j} \frac{q_i q_j}{4\pi\varepsilon_0 \varepsilon_r r_{ij}} + \sum_{i,j;i\neq j} 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right]. \end{aligned}$$

The simple terms are often harmonic (e.g. $V_{bonds}$, $V_{angles}$, $V_{improper}$) or motivated by physical laws (e.g. Coulomb $V_{Coul}$, Lennard-Jones $V_{LJ}$). They are defined by their functional form and a small number of parameters, e.g. an atomic radius for van der Waals interactions. All parameters are determined using either ab initio quantum chemical calculations or comparisons of structural or thermodynamical data with suitable averages of small molecule MD ensembles. Between different force fields (Brooks et al. 1983; Weiner et al. 1986; van Gunsteren and Berendsen 1987; Jorgensen et al. 1996) the number of energy terms, their functional form and their individual parameters can vary considerably.

Given the above description of proteins as point masses (positions $r_i$, velocities $v_i$) moving in a classical potential under external forces $F_i$, a standard MD simulation integrates Newton's equations of motion in discrete timesteps $\Delta t$ on the femtosecond timescale by some numerical scheme, e.g. the leap-frog algorithm (Hockney et al. 1973):

$$\begin{aligned} v_i(t + \frac{\Delta t}{2}) &= v_i(t - \frac{\Delta t}{2}) + \frac{F_i(t)}{m_i} \Delta t \\ r_i(t + \Delta t) &= r_i(t) + v_i(t + \frac{\Delta t}{2}) \Delta t. \end{aligned}$$

Besides interactions with membranes and other macromolecules, water is the principal natural environment for proteins. For a simulation of a model system that matches the in vivo system as close as possible, water molecules and ions in

physiological concentration are added to the system in order to solvate the protein. Having a simulation box filled with solvent and solute, artefacts due to the boundaries of the system may arise, such as evaporation, high pressure due to surface tension and preferred orientations of solvent molecules on the surface. To avoid such artefacts, periodic boundary conditions are often applied. In this way, the simulation system does not have any surface. This, however, may lead to new artefacts if the molecules artificially interact with their periodic images due to e.g. long-range electrostatic interactions. These periodicity artefacts are minimized by increasing the size of the simulation box. Different choices of unit cells, e.g., cubic, dodecahedral or truncated octahedral allow an optimal fit to the shape of the protein, and, therefore, permit a suitable compromise between the number of solvent molecules while simultaneously keeping the crucial protein-protein distance high.

As the solvent environment strongly affects the structure and dynamics of proteins, water must be described accurately. Besides the introduction of implicit solvent models, where water molecules are represented as a continuous medium instead of individual "explicit" solvent molecules (Still et al. 1990; Gosh et al. 1998; Jean-Charles et al. 1991; Luo et al. 2002), a variety of explicit solvent models are used these days (e.g. Jorgensen et al. 1983). These models differ in the number of particles used to represent a water molecule and the assigned static partial charges, reflecting the polarity and, effectively, in most force fields, polarization. Because these charges are kept constant during the simulation, explicit polarization effects are thereby excluded. Nowadays, several polarizable water models (and force fields) exist, see Warshel et al. (2007) and Huang et al. (2014) for reviews.

In solving Newton's equations of motion, the total energy of the system is conserved, resulting in a microcanonical NVE ensemble having constant particle number N, volume V and energy E. However, real biological subsystems of the size studied in simulations constantly exchange energy with their surrounding. Furthermore, a constant pressure P of usually 1 bar is present. To account for these features, algorithms are introduced which couple the system to a temperature and pressure bath (Anderson 1980; Nose 1984; Berendsen et al. 1984), leading to a canonical NPT ensemble.

### *12.1.2 Applications*

Molecular Dynamics simulations have become a standard technique in protein science and are routinely applied to a wide range of problems. Conformational dynamics of proteins, however, is still a demanding task for MD simulations since functional conformational transitions often occur at timescales of microseconds to seconds which are not routinely accessible with current algorithms and computer power.

#### 12.1.2.1 Nuclear Transport Receptors

Despite their computational demands, MD simulations have been successfully applied to study functional modes of proteins. As an illustration, we will discuss in some detail the work of Zachariae et al. (Zachariae and Grubmüller 2006) that revealed a strikingly fast conformational transition of the exportin CAS (Cse1p in yeast) from the open to the closed state. CAS/Cse1p is a nuclear transport receptor consisting of 960 amino acids that binds importin-α and RanGTP in the nucleus. The heterotrimeric complex (Fig. 12.1) can cross nuclear pores and dissociates by catalyzed GTP hydrolysis in the cytoplasm and, thus, represents an important part of the nucleocytoplasmic transport cycle in cells.

For the function of the importin-α/CAS system it is essential that, after dissociation of the complex in the cytoplasm, CAS/Cse1p undergoes a large conformational change that prevents reassociation of the complex. X-ray structures of Cse1p show that the cargo bound conformation adopts a superhelical structure with curls around the bound RanGTP (Fig. 12.2 left), whereas the cytoplasmic form exhibits a closed ring conformation that leads to occlusion of the RanGTP binding site (Fig. 12.2 right). In order to understand the mechanism of this conformational switch, Zachariae et al. carried out MD simulations of Cse1p starting from the cargo bound conformation. They found that, mainly driven by electrostatic interactions, the structure of Cse1p spontaneously collapses and adopts a conformation close to the experimentally determined cytoplasmic form within a relatively short timescale of 10 ns. Simulations of mutants with different electrostatic surface potentials did not reveal a significant conformational change but remained in an open conformation which is in good agreement with experimental findings (Cook et al. 2005). This example shows that functionally relevant conformational changes that occur on short time scales can be studied by MD simulations. However, in this particular case the simulation has—due to the removal of importin-α and RanGTP—not been

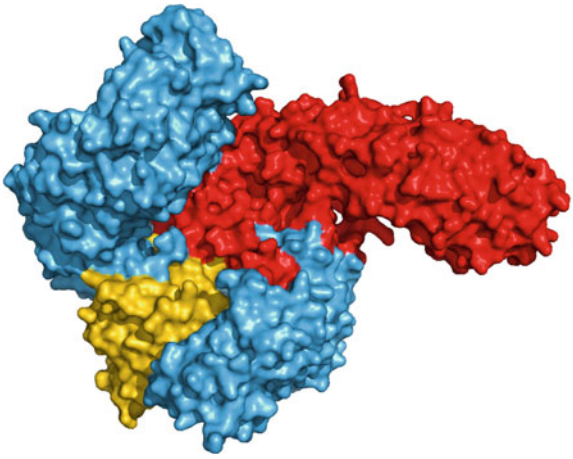


**Fig. 12.1** Heterotrimeric complex of Cse1p (*blue*), RanGTP (*yellow*) and importin-α (*red*). Cse1p adopts a superhelical structure and binds RanGTP and importin-α. The complex can cross nuclear pores and dissociates by catalyzed GTP hydrolysis in the cytoplasm

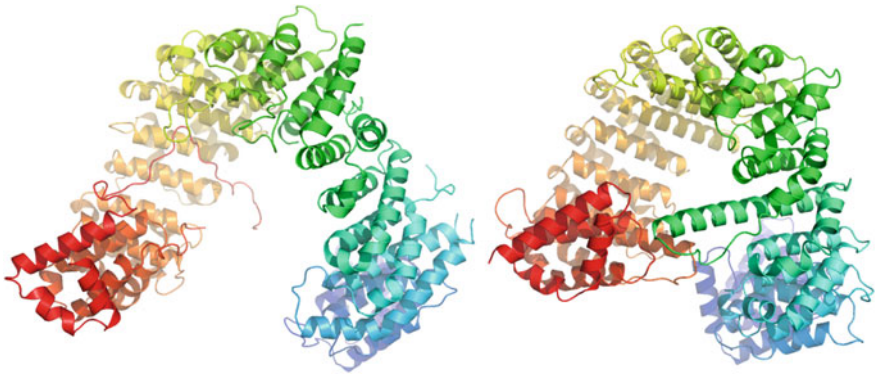


**Fig. 12.2** Nucleoplasmic (*left*) and cytoplasmic (*right*) form of Cse1p. In the nucleoplasmic form, Cse1p is bound to RanGTP and importin-α (both not shown) and adopts a superhelical structure. After dissociation in the cytoplasm, Cse1p undergoes a large conformational change and forms a ring conformation that occludes the RanGTP binding site and prevents reassociation of the complex. The structures are coloured in a spectrum from *blue* (N-terminus) to *red* (C-terminus)

started from an equilibrium conformation and thus, presumably, no significant energy barrier had to be overcome to reach the closed conformation. When simulations are started from a free energy minimum, which is usually the case, the accessible time scales are often too short to overcome higher energy barriers and, thus, to observe functionally relevant conformational transitions. This is known as the "sampling problem" and is a general problem for MD simulations.

#### 12.1.2.2 Lysozyme

MD simulations of bacteriophage T4-lysozyme (T4L), an enzyme which is six times smaller than Cse1p, impressively illustrate this sampling problem for relatively long MD trajectories. T4L has been extensively studied with X-ray crystallography (Faber and Matthews 1990; Kuroki et al. 1993) and, since it has been crystallized in many different conformations, represents one of the rare cases where information about functionally relevant modes can be directly obtained at atomic resolution from experimental data (Zhang et al. 1995; de Groot et al. 1998). The domain character of this enzyme is very pronounced (Matthews and Remington 1974) and from the differences between crystallographic structures of various mutants of T4L it has been suggested that a hinge-bending mode of T4L (Fig. 12.3) is an intrinsic property of the molecule (Dixon et al. 1992). Moreover, the domain fluctuations are predicted to be essential for the function of the enzyme, allowing the substrate to enter and the products to leave the active site in the open configuration, with the closed state presumably required for catalysis.

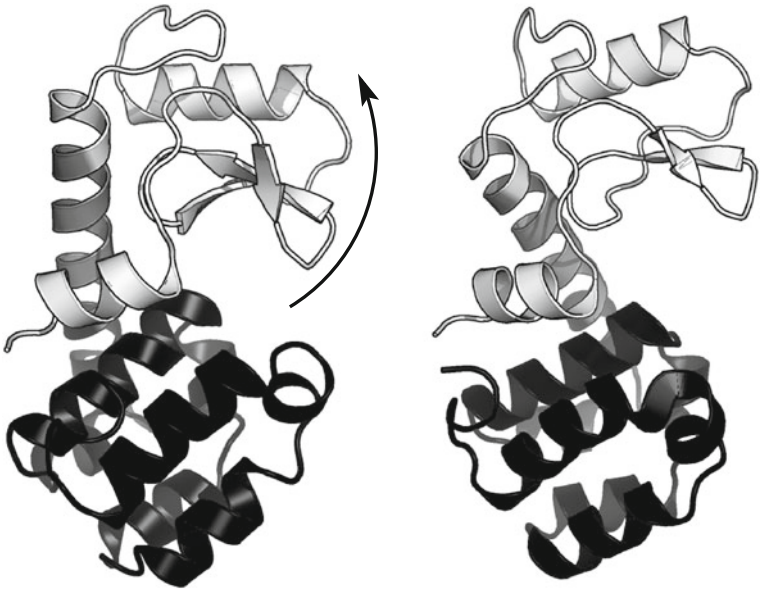


**Fig. 12.3** Hinge-bending motion in bacteriophage T4-lysozyme. Domain fluctuations (domains are coloured differently) are essential for enzyme function, allowing the substrate to enter and the products to leave the active site

The wealth of experimental data also provides the opportunity to assess the reliability and sampling performance of simulation methods. Two MD simulations have been carried out using a closed (simulation 1) and an open conformation (simulation 2) as starting points, respectively. In order to assess the sampling efficiency a principal components analysis (PCA, see Sect. 12.2 below) has been carried out on the ensemble of experimentally determined structures and the X-ray ensemble and the two MD trajectories have been projected onto the first two eigenvectors. The first eigenvector represents the hinge-bending motion, whereas the second eigenvector represents a twist of the two domains of T4L. The projections are shown in Fig. 12.4. The X-ray ensemble is represented by dots, each dot representing a single conformation. Movement along the first eigenvector (x-axis) describes a collective motion from the closed to the open state. It can be seen that neither of the individual the MD trajectories, represented by lines, fully samples the entire conformational space covered by the X-ray ensemble, although the simulation times (184 ns for simulation 1 and 117 ns for simulation 2) are one order of magnitude larger than in the previously discussed Cse1p simulation. From the phase space density one can assume that an energy barrier exists between the closed and the open state and neither simulation achieves a full transition, from the closed to the open state, or vice versa.

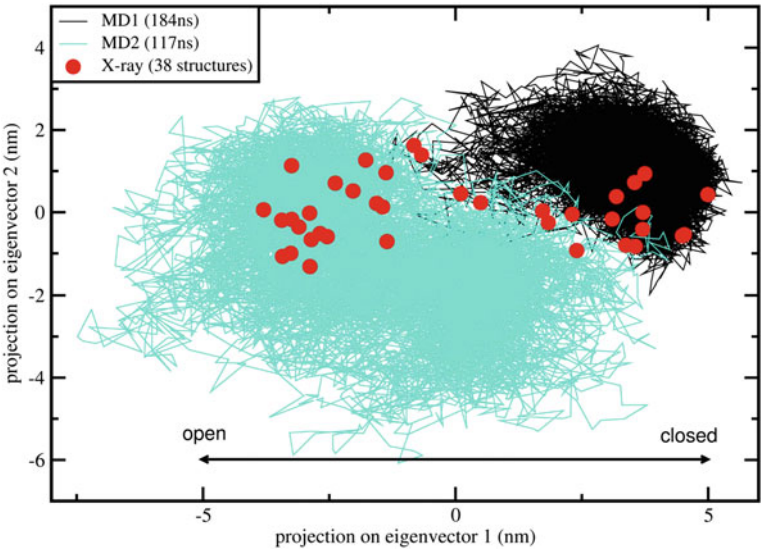



**Fig. 12.4** Principal components analysis of bacteriophage T4-lysozyme. The X-ray ensemble is represented by *dots*, MD trajectories by *lines*. A movement along the first eigenvector (*x-axis*) represents a collective motion from the open to the closed state. Neither simulation *1* started from a closed conformation—, nor simulation *2* started from an open conformation—show a full transition due to an energy barrier that separates the conformational states

#### 12.1.2.3 Aquaporins

Aquaporins present a prime example of how MD simulations have contributed to the understanding of protein function both in terms of dynamics and energetics. Aquaporins facilitate efficient and selective permeation of water across biological membranes. Related aquaglyceroporins in addition also permeate small neutral solutes like glycerol. Available high-resolution structures provided invaluable insights in the molecular mechanisms acting in aquaporins (Fu et al. 2000; Murata et al. 2000; de Groot et al. 2001; Sui et al. 2001). However, mostly static information is available from such structures and we can therefore not directly observe aquaporins "at work". So far, there is no experimental method that offers sufficient spatial and time resolution to monitor permeation through aquaporins on a molecular level. MD simulations therefore complement experiments by providing the progression of the biomolecular system at atomic resolution. As permeation is known to take place on the nanosecond timescale, spontaneous permeation can be expected to take place in multi-nanosecond simulations, allowing a direct observation of the functional dynamics. Hence, such simulations have been termed "real-time simulations" (de Groot and Grubmüller 2001).

Indeed, spontaneous permeation events were observed in MD simulations of aquaporin-1 and the aquaglyceroporin GlpF. These simulations identified that the efficiency of water permeation is accomplished by providing a hydrogen bond

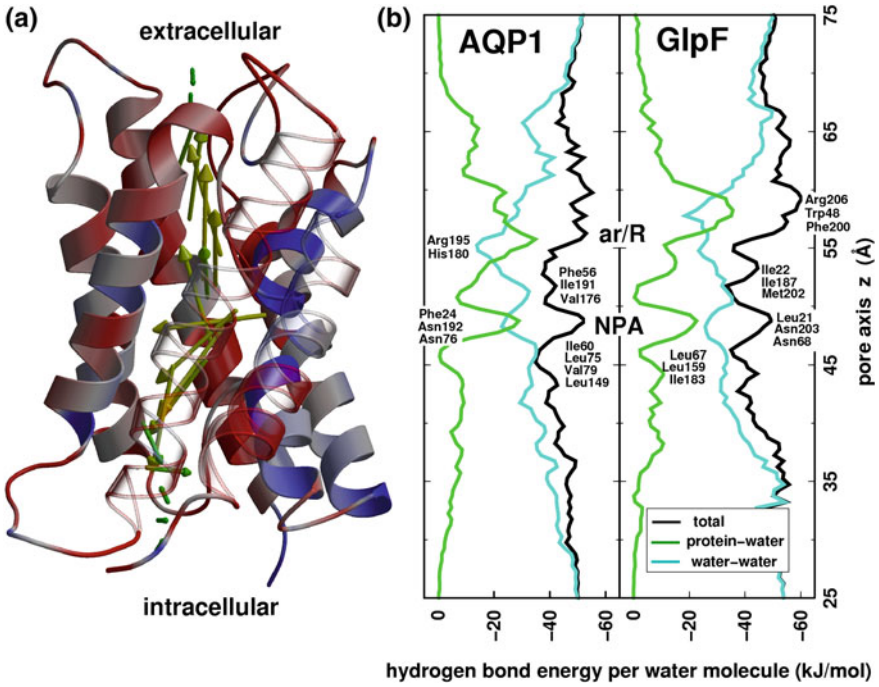



**Fig. 12.5** **a** Water molecules are strongly aligned inside the aquaporin-1 channel, with their dipoles pointing away from the central NPA region (de Groot and Grubmüller 2001). The water dipoles (*yellow arrows*) rotate by approx. 180° while permeating though the AQP1 pore. The *red* and *blue colours* indicate local electrostatic potential, negative and positive, respectively. **b** Hydrogen bond energies per water molecule (*solid black lines*) in AQP1 (*left*) and GlpF (*right*). Protein-water hydrogen bonds (*green*) compensate for the loss of water-water hydrogen bonds (*cyan*). The main protein-water interaction sites are the ar/R region and the NPA site

complementarity inside the channel comparable to bulk water, thereby establishing a low permeation barrier (de Groot and Grubmüller 2001; Tajkhorshid et al. 2002). The simulations furthermore identified that the selectivity in these channels is accomplished by a two-stage filter. The first stage of the filter is located in the central part of the channel at the conserved asparagine/proline/alanine (NPA) region; the second stage is located on the extracellular face of the channel in the aromatic/arginine (ar/R) constriction region (Fig. 12.5). As water permeation takes place on the nanosecond timescale, permeation coefficients can be directly computed from the simulations, and compared to experiment. Quantitative agreement was found between permeation coefficients from experiment and simulation, thereby validating the simulations.

A long standing question in aquaporin research has been the mechanism by which protons are excluded from the aqueous pores. The MD simulations addressing water permeation revealed a pronounced water dipole orientation pattern across the channel, with the NPA region as its symmetry center (de Groot and

Grubmüller 2001). In the simulations, the water molecules were found to rotate by 180° on their path through the pore (Fig. 12.5a). In a series of simulations addressing the mechanism of proton exclusion it was found that the pronounced water orientation is due to an electric field in the channel centred at the NPA region (de Groot et al. 2003; Chakrabarti et al. 2004; Ilan et al. 2004). Electrostatic effects therefore form the structural basis of proton exclusion. A debate continues about the origin of the electrostatic barrier, where both direct electrostatic effects caused by helix dipoles has been suggested (de Groot et al. 2003; Chakrabarti et al. 2004), as well as a specific desolvation effects (Burykin and Warshel 2003). Some results suggest that both effects contribute approximately equally (Chen et al. 2006).

Recently, MD simulations allowed for the elucidation of the mechanism of selectivity of neutral solutes in aquaporins and aquaglyceroporins. Aquaporins were found to be permeated solely by small polar molecules like water, and to some extent also ammonia, whereas aquaglyceroporins are also permeated by apolar molecules like $CO_2$ and larger molecules like glycerol, but not urea (Hub and de Groot 2008). For aquaporins, an inverse relation was observed between permeability and solute hydrophobicity—solutes competing with permeating water molecules for hydrogen bonds with the channel determine the permeation barrier. A combination of size exclusion and hydrophobicity therefore underlies the selectivity in aquaporins and aquaglyceroporins.

### *12.1.3 Limitations—Enhanced Sampling Algorithms*

Although molecular dynamics simulations have become an integral part of structural biology and provided numerous invaluable insights into biological processes at the atomic level, limitations occur due to both methodological restrictions and limited computer power. Methodological limitations arise from the classical description of atoms and the approximation of interactions by simple energy terms instead of the Schrödinger equation. This means that chemical reactions (bond breaking and formation) can not be described. Also polarization effects and proton tunnelling lie out of the scope of classical MD simulations.

The second class of limitations arises from the computational demands of MD simulations. Although bonds are usually treated as constraints thereby eliminating the highest frequency motions, the timestep length in MD simulations usually cannot be chosen longer than 4 fs. Hence, a nanosecond simulation requires 250,000 force calculations and integration steps. Despite the rapid progress in algorithm techniques, the development of special purpose hardware and the utilization of graphics processing units (GPUs) as calculation engines, the simulation of timescales at which biological phenomena occur are not routinely accessible.

Biologically relevant protein motions like large conformational transitions, folding and unfolding usually take place on the micro- to (milli)second timescale. Thus it becomes evident that, despite ever increasing computer power, which roughly grows by a factor of 100 per decade, MD simulations will not solve the

"sampling problem" anytime soon by just waiting for faster computers. Therefore, alternative methods—partly based on MD—have been developed to specifically address the problem of conformational sampling and to predict functionally relevant protein motions.

Reducing the number of particles is one approach. Since proteins are usually studied in solution most of the simulation system consists of water molecules. The development of implicit solvent models is therefore a promising means to reduce computational demands (Still et al. 1990; Gosh et al. 1998; Jean-Charles et al. 1991; Luo et al. 2002). Another possibility to reduce the number of particles is the use of so-called coarse-grained models (Bond et al. 2007; Saunders and Voth 2013). In these models, atoms are grouped together, for instance typically four water molecules are treated as one pseudo-particle (bead). These groupings have two effects. First, the number of particles is reduced and, second, the timestep, depending on the fastest motions in the system, can be increased. However, coarse-graining is not restricted to water molecules. Representations of several atoms up to complete amino acids by a single bead are nowadays used. This allows for a drastic reduction of computational demands, thereby enabling the simulation of large macromolecular aggregates on micro- to millisecond timescales. An even coarser model, collapsing entire protein domains into single interaction sites, has recently been introduced to study intermolecular interactions in solutions of antibodies. Here viscosity at high concentrations poses a severe challenge on antibody development and reliable viscosity predictions would be a major step forward in the field of computational biotechnology (Chaudhri et al. 2012).

The gain in efficiency due to coarse graining, however, comes with an inherent reduction of accuracy compared to all-atom descriptions of proteins, restricting current models to semi-quantitative statements. Essential for the success of coarse-grained simulations are the parameterizations of force fields that are both accurate and transferable, i.e. force fields capable of describing the general dynamics of systems having different compositions and configurations. As the graining becomes coarser, this process becomes increasingly difficult, since more specific interactions must effectively be included in fewer parameters and functional forms. This has led to a variety of models for proteins, lipids and water, representing different compromises between accuracy and transferability (see e.g. Marrink et al. 2004).

Other MD based enhanced sampling methods, which retain the atomistic description, include replica exchange molecular dynamics (REMD) and essential dynamics (ED) which are discussed in subsequent sections. Moreover, a number of non-MD based methods are discussed that aim towards the prediction of functional modes of proteins.

#### 12.1.3.1 Replica Exchange

The aim of most computer simulations of biomolecular systems is to calculate macroscopic behaviour from microscopic interactions. Following equilibrium

statistical mechanics, any observable that can be connected to macroscopic experiments is defined as an ensemble average over all possible realizations of the system. However, given current computer hardware, a fully converged sampling of all possible conformational states with the respective Boltzmann weight is only attainable for simple systems comprising a small number of amino acids (see e.g. Kubitzki and de Groot 2007). For proteins, consisting of hundreds to thousands of amino acids, conventional MD simulations often do not converge and reliable estimates of experimental quantities can not be calculated.

This inefficiency in sampling is a result of the ruggedness of the systems' free energy landscape, a concept put forward by Frauenfelder (Frauenfelder et al. 1991; Frauenfelder and Leeson 1998). The global shape is supposed to be funnel-like, with the native state populating the global free energy minimum (Anfinsen 1973). Looking in more detail, the complex high-dimensional free energy landscape is characterized by a multitude of almost iso-energetic minima, separated from each other by energy barriers of various heights. Each of these minima corresponds to one particular conformational substate, with neighboring minima corresponding to similar conformations. Within this picture, structural transitions are barrier crossings, with the transition rate depending on the height of the barrier. For MD simulations at room temperature, only those barriers are easily overcome that are smaller than or comparable to the thermal energy $k_BT$ and the observed structural changes are small, e.g. side chain rearrangements. Therefore the system will spend most of its time in locally stable states (*kinetic trapping*) instead of exploring different conformational states. This wider exploration is of greater interest, due to its connection to biological function, but requires that the system be able to overcome large energy barriers. Unfortunately, since MD simulations are mostly restricted to the nanosecond timescale, functionally relevant conformational transitions are rarely observed.

A plethora of enhanced sampling methods have been developed to tackle this multi-minima problem (see e.g. van Gunsteren and Berendsen 1990; Tai 2004; Adcock and McCammon 2006 and references therein). Among them, generalized ensemble algorithms have been widely used in recent years (for a review, see e.g. Mitsutake et al. 2001; Iba 2001). Generalized ensemble algorithms sample an artificial ensemble that is either constructed from combinations or alterations of the original ensemble. Algorithms of the second category (e.g. Berg and Neuhaus 1991) basically modify the original bell-shaped potential energy distribution p(V) of the system by introducing a so-called multicanonical weight factor w(V), such that the resulting distribution is uniform, $p(V)w(V) = \text{const}$. This flat distribution can then be sampled extensively by MD or Monte-Carlo techniques because potential energy barriers are no longer present. Due to the modifications introduced, estimates for canonical ensemble averages of physical quantities need to be obtained by reweighting techniques (Kumar et al. 1992; Chodera et al. 2007). The main problem with these algorithms, however, is the non-trivial determination of the different multicanonical weight factors by an iterative process involving short trial simulations. For complex systems this procedure can be very tedious and attempts have been made to accelerate convergence of the iterative process

(Berg and Celik 1992; Kumar et al. 1996; Smith and Bruce 1996; Hansmann 1997; Bartels and Karplus 1998).

The replica exchange (REX) algorithm, developed as an extension of simulated tempering (Marinari and Parisi 1992), removes the problem of finding correct weight factors. It belongs to the first category of algorithms where a generalized ensemble, built from several instances of the original ensemble, is sampled. Due to its simplicity and ease of implementation, it has been widely used in recent years. Most often, the standard temperature formulation of REX is employed (Sugita and Okamoto 1999), with the general Hamiltonian REX framework gaining increasing attention (Fukunishi et al. 2002; Liu et al. 2005; Sugita et al. 2000; Affentranger et al. 2006; Christen and van Gunsteren 2006; Lyman and Zuckerman 2006).

In standard temperature REX MD (Sugita and Okamoto 1999), a generalized ensemble is constructed from M + 1 non-interacting copies, or "replicas", of the system at a range of temperatures $\{T_0, \ldots, T_M\}$ ($T_m \leq T_{m+1}$; m = 0, …, M), e.g. by distributing the simulation over M + 1 nodes of a parallel computer (Fig. 12.6 left). A state of this generalized ensemble is characterized by $S = \{\ldots, s_m, \ldots\}$, where $s_m$ represents the state of replica m having temperature $T_m$. The algorithm now consists of two consecutive steps: (a), independent constant-temperature simulations of each replica, and (b), exchange of two replicas $S = \{\ldots, s_m, \ldots, s_n, \ldots\} \rightarrow S' = \{\ldots, s_n', \ldots, s_m', \ldots\}$ according to a Metropolis-like criterion. The exchange acceptance probability is thereby given by

$$P(S \rightarrow S') = \min\{1, \exp\{(\beta_m - \beta_n)[V_m - V_n]\}\} \tag{1.1}$$

with $V_m$ being the potential energy and $\beta_m^{-1} = k_B T_m$. Iterating steps a and b, the trajectories of the generalized ensemble perform a random walk in temperature space, which in turn induces a random walk in energy space. This facilitates an efficient and statistically correct conformational sampling of the energy landscape of the system, even in the presence of multiple local minima.

The choice of temperatures is crucial for an optimal performance of the algorithm. Replica temperatures have to be chosen such that (a) the lowest temperature is small enough to sufficiently sample low-energy states, (b) the highest temperature is large enough to overcome energy barriers of the system of interest, and (c) the acceptance probability $P(S \rightarrow S')$ is sufficiently high, requiring adequate overlap of potential energy distributions for neighboring replicas. For larger systems simulated with explicit solvent the latter condition presents the main bottleneck. A simple estimate (Cheng et al. 2005; Fukunishi et al. 2002) shows that the potential energy difference $\Delta V \sim N_{df} \Delta T$ is dominated by the contribution from the solvent degrees of freedom $N_{df}^{sol}$, constituting the largest fraction of the total number of degrees of freedom $N_{df}$ of the system. Obtaining a reasonable acceptance probability therefore relies on keeping the temperature gaps $\Delta T = T_{m+1} - T_m$ small (typically only a few Kelvin) which drastically increases computational demands for systems having more than a few thousand particles. Despite this severe limitation, REX methods have become an established tool for the study of peptide folding/unfolding

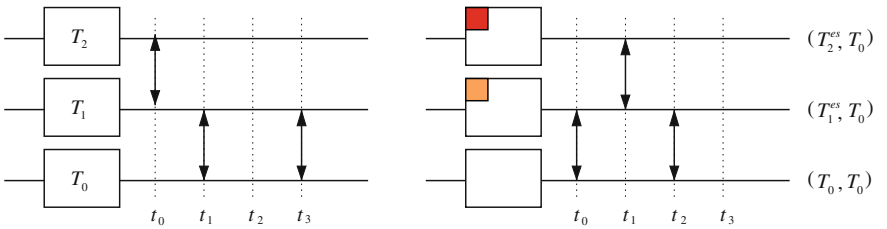



**Fig. 12.6** Schematic comparison of standard temperature REX (*left panel*) and the TEE-REX algorithm (*right panel*) for a three-replica simulation. Temperatures are sorted in increasing order, $T_{i+1} > T_i$. Exchanges (↔) are attempted (…) with frequency $\nu_{ex}$. Unlike REX, only an essential subspace {es} (*red boxes*) containing a few collective modes is excited within each TEE-REX replica. Reference replica ($T_0$, $T_0$), containing an approximate Boltzmann ensemble, is used for analysis

(Zhou et al. 2001; Rao and Caflisch 2003; García and Onuchic 2003; Pitera and Swope 2003; Seibert et al. 2005); structure prediction (Fukunishi et al. 2002; Kokubo and Okamoto 2004), phase transitions (Berg and Neuhaus 1991) and free energy calculations (Sugita et al. 2000; Lou and Cukier 2006).

Going beyond conventional MD, another class of enhanced sampling algorithms is successfully applied to the task of elucidating protein function. These algorithms make use of the fact that fluctuations in proteins are generally correlated. Extracting such collective modes of motion and their application in new sampling algorithms will be the focus of the following two sections.

## 12.2 Principal Component Analysis

Principal component analysis (PCA) is a well-established technique to obtain a low-dimensional description of high-dimensional data. Its applications include data compression, image processing, data visualization, exploratory data analysis, pattern recognition and time series prediction (Duda et al. 2001). In the context of biomolecular simulations PCA has become an important tool in the extraction and classification of relevant information about large conformational changes from an ensemble of protein structures, generated either experimentally or theoretically (García 1992; Gō et al. 1983; Amadei et al. 1993). Besides PCA, a number of similar techniques are nowadays used, most notably normal mode analysis (NMA) (Brooks and Karplus 1983; Gō et al. 1983; Levitt et al. 1983), quasi-harmonic analysis (Karplus and Kushick 1981; Levy et al. 1984a, b; Teeter and Case 1990) and singular-value decomposition (Romo et al. 1995; Bahar et al. 1997).

PCA is based on the notion that by far the largest fraction of positional fluctuations in proteins occurs along only a small subset of collective degrees of freedom. This was first realized from NMA of a small protein (Brooks and Karplus

1983; Gō et al. 1983; Levitt et al. 1983). In NMA (see Sect. 12.4.1), the potential energy surface is assumed to be harmonic and collective variables are obtained by diagonalization of the Hessian[1] matrix in a local energy minimum. Quasi-harmonic analysis, PCA and singular-value decomposition of MD trajectories of proteins that do not assume harmonicity of the dynamics, have shown that indeed protein dynamics is dominated by a limited number of collective coordinates, even though the major modes are frequently found to be largely anharmonic. These methods identify those collective degrees of freedom that best approximate the total amount of fluctuation. The subset of largest-amplitude variables form a set of generalized internal coordinates that can be used to effectively describe the dynamics of a protein. Often, a small subset of 5–10% of the total number of degrees of freedom yields a remarkably accurate approximation. As opposed to torsion angles as internal coordinates, these collective internal coordinates are not known beforehand but must be defined either using experimental structures or an ensemble of simulated structures. Once an approximation of the collective degrees of freedom has been obtained, this information can be used for the analysis of simulations as well as in simulation protocols designed to enhance conformational sampling (Grubmüller 1995; Zhang et al. 2003; He et al. 2003; Amadei et al. 1996).

In essence, a principal component analysis is a multi-dimensional linear least squares fit procedure in configuration space. The structure ensemble of a molecule, having N particles, can be represented in 3N-dimensional configuration space as a distribution of points with each configuration represented by a single point. For this cloud, always one axis can be defined along which the maximal fluctuation takes place. As illustrated for a two-dimensional example (Fig. 12.7), if such a line fits the data well, all data points can be approximated by only the projection onto that axis, allowing a reasonable approximation of the position even when neglecting the position in all directions orthogonal to it. If this axis is chosen as coordinate axis, the position of a point can be represented by a single coordinate. The procedure in the general 3N-dimensional case works similarly. Given the first axis that best describes the data, successive directions orthogonal to the previous set are chosen such as to fit the data second-best, third-best, and so on (the *principal components*). Together, these directions span a 3N-dimensional space. Mathematically, these directions are given by the eigenvectors $\mu_i$ of the covariance matrix of atomic fluctuations

$C = \left\langle (x(t) - \langle x \rangle)(x(t) - \langle x \rangle)^T \right\rangle$,with the angle brackets $\langle \cdot \rangle$ representing an ensemble average. The eigenvalues $\lambda_i$ correspond to the mean square positional fluctuation along the respective eigenvector, and therefore contain the contribution of each principal component to the total fluctuation (Fig. 12.8). Applications of such a multidimensional fit procedure on protein configurations from MD simulations of several proteins have proven that typically the first ten to twenty principal components are responsible for 90% of the fluctuations of a protein

[1]second derivative $(\partial^2 V)/(\partial x_i \partial x_j)$ of the potential energy.

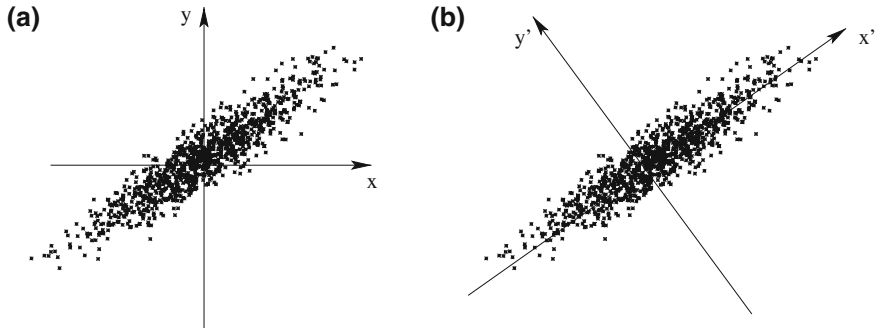



**Fig. 12.7** Illustration of PCA in two dimensions. Two coordinates (x, y) are required to identify a point in the ensemble in panel (**a**), whereas one coordinate x′ approximately identifies a point in panel (**b**)

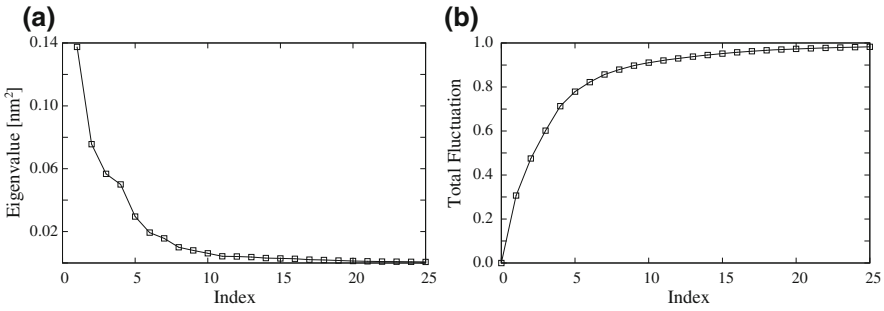



**Fig. 12.8** Typical PCA eigenvalue spectrum (MD ensemble of guanylin backbone structures). The first five eigenvectors (panel **a**) cover 80% of all observed fluctuations (panel **b**)

(Kitao et al. 1991; García 1992; Amadei et al. 1993). These principal components correspond to collective coordinates, containing contributions from every atom of the protein. In a number of cases these principal modes were shown to be involved in the functional dynamics of the studied proteins (Amadei et al. 1993; van Aalten et al. 1995a, b; de Groot et al. 1998). Hence, the subspace responsible for the majority of all fluctuations has been referred to as the *essential subspace* (Amadei et al. 1993).

The fact that a small subset of the total number of degrees of freedom (essential subspace) dominates the molecular dynamics of proteins originates from the presence of a large number of internal constraints and restrictions defined by the atomic interactions present in a biomolecule. These interactions range from strong covalent bonds to weak non-bonded interactions, whereas the restrictions are given by the dense packing of atoms in native-state structures.

Overall, protein dynamics at physiological temperatures has been described as diffusion among multiple minima (Kitao et al. 1998; Amadei et al. 1999; Kitao and Gō 1999). The dynamics on short timescales is dominated by fluctuations within a local minimum, corresponding to eigenvectors having low eigenvalues. On longer

timescales large fluctuations are dominated by a largely anharmonic diffusion between multiple wells. These slow dynamical transitions are usually represented by the largest-amplitude modes of a PCA. In contrast to normal mode analysis, PCA of a MD simulation trajectory does not rest on the assumption of a harmonic potential. In fact, PCA can be used to study the degree of anharmonicity in the molecular dynamics of the simulated system. For proteins, it was shown that at physiological temperatures, especially the major modes of collective fluctuation that are frequently functionally relevant, are dominated by anharmonic fluctuations (Amadei et al. 1993; Hayward et al. 1995).

## 12.3 Collective Coordinate Sampling Algorithms

Analyzing MD simulations in terms of collective coordinates (obtained e.g. by PCA or NMA) reveals that only a small subset of the total number of degrees of freedom dominates the molecular dynamics of biomolecules. As protein function could in many cases been linked to these essential subspace modes (e.g. Brooks and Karplus 1983; Gō et al. 1983; Levitt et al. 1983), the dynamics within this low-dimensional space was termed "essential dynamics" (ED). This not only aids the analysis and interpretation of MD trajectories but also opens the way to enhanced sampling algorithms that search the essential subspace in either a systematic or exploratory fashion (Grubmüller 1995; Amadei et al. 1996).

### *12.3.1 Essential Dynamics*

The first attempts in this direction were aimed at a simulation scheme in which the equations of motion were solely integrated along a selection of primary principal modes, thereby drastically reducing the number of degrees of freedom (Amadei et al. 1993). However, these attempts proved problematic because of non-trivial couplings between high- and low-amplitude modes, even though after diagonalization the modes are linearly independent (orthogonal). Therefore, instead, a series of techniques has prevailed that take into account the full-dimensional simulation system and enhance the motion along a selection of principal modes. The most common of these techniques are conformational flooding (Grubmüller 1995) and ED sampling (Amadei et al. 1996; de Groot et al. 1996a, b). In conformational flooding, an additional potential energy term that stimulates the simulated system to explore new regions of phase space is introduced on a selection of principal modes, whereas in ED sampling a similar goal is achieved by geometrical constraints along a selection of principal modes. With these techniques a sampling efficiency enhancement of up to an order of magnitude can be achieved, provided that a reasonable approximation of the principal modes has been obtained from a conventional simulation. However, due to the applied structural or energetic bias on the

system, the ensemble generated by ED sampling and conformational flooding is not canonical, restricting analysis to structural questions.

### *12.3.2 TEE-REX*

Enhanced sampling methods such as ED (Amadei et al. 1996) achieve their sampling power (Amadei et al. 1996; de Groot et al. 1996a, b) primarily from the fact that a small number of internal collective degrees of freedom dominate the configurational dynamics of proteins. Yet, systems simulated with such methods are always in a non-equilibrium state, rendering it difficult to extract thermodynamic, i.e. equilibrium properties of the system from such simulations. On the other hand, generalized ensemble algorithms such as REX not only enhance sampling but yield correct statistical ensembles necessary for the calculation of equilibrium properties which can be subjected to experimental verification. However, REX quickly becomes computationally prohibitive for systems of more than a few thousand particles, limiting its current applicability to smaller peptides (Pitera and Swope 2003; Cecchini et al. 2004; Nguyen et al. 2005; Liu et al. 2005; Seibert et al. 2005). The newly developed Temperature Enhanced Essential dynamics Replica EXchange (TEE-REX) algorithm (Kubitzki and de Groot 2007) combines the favorable properties of REX with those resulting from a *specific* excitation of functionally relevant modes, while at the same time avoiding the drawbacks of both approaches.

Figure 12.6 shows a schematic comparison of standard temperature REX (left) and the TEE-REX algorithm (right). TEE-REX builds upon the REX framework, i.e. a number of replicas of the system are simulated independently in parallel with periodic exchange attempts between neighbouring replicas. In contrast to REX, in each but the reference replica, only those degrees of freedom are thermally stimulated that contribute significantly to the total fluctuations of the system (essential subspace {es}). This way, several benefits are combined and drawbacks avoided. In contrast to standard REX, the specific excitation of collective coordinates promotes sampling along these often functionally relevant modes of motion, i.e. the advantages of ED are used. To counterbalance the disadvantages associated with such a specific excitation, i.e. the construction of biased ensembles, the scheme is embedded within the REX protocol. Thereby ensembles are obtained having approximate Boltzmann statistics and the enhanced sampling properties of REX are utilized. The exchange probability (1.1) between two replicas crucially depends on the excited number of degree of freedom of the system. Since the stimulated number of degrees of freedom makes up only a minute fraction of the total number of degrees of freedom of the system, the bottleneck of low exchange probabilities in all-atom REX simulations is bypassed. For given exchange probabilities, large temperature differences $\Delta T$ can thus be used, such that only a few replicas are required.

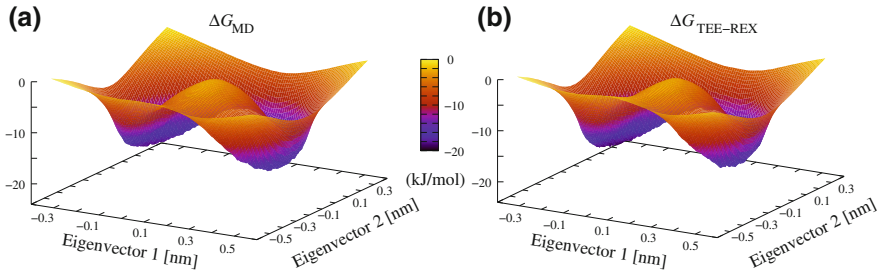



**Fig. 12.9** Comparison of two-dimensional relative free energy surfaces (in units of kJ/mol) of dialanine generated by MD (panel **a**) and TEE-REX (panel **b**). Deviations $\Delta G_{TEE\text{-}REX} - G_{MD}$ are commensurate with the statistical errors of $\sim 0.1\ k_BT$

Figure 12.9 shows a two-dimensional projection of the free energy landscape of dialanine, calculated with MD (panel A) and TEE-REX (panel B). The thermodynamic behaviour of a system is completely known once a thermodynamic potential such as the relative Gibbs free energy ΔG is available. Comparing free energies thus enables us to decide to which degree ensembles created by different simulation methods coincide. In doing so, ensemble convergence is an absolute necessity. For the dialanine test case, this requirement is met. A detailed analysis of the shape of the free energy surfaces generated by MD and TEE-REX shows that the maximum absolute deviations of 1.5 kJ/mol $\cong 0.6 k_BT$ from the ideal case $\Delta G_{TEE\text{-}REX} - G_{MD} = 0$, commensurate with the maximum statistical errors of $0.15 k_BT$ found for each method. The small deviations found for the TEE-REX ensemble are presumably due to the exchange of non-equilibrium structures into the TEE-REX reference ensemble.

The sampling efficiency of the TEE-REX algorithm compared to MD was evaluated for guanylin, a small 13 amino-acid peptide hormone (Currie et al. 1992). Trajectories generated with both methods-using the same computational effort-were projected into (ϕ, ψ)-space as well as different two-dimensional subspaces spanned by PCA modes calculated from an MD ensemble of guanylin structures. From these projections, the time evolution of sampled configuration space volume was measured. Overall, the sampling performance of MD is quite limited compared to TEE-REX, the latter outperforming MD on average by a factor of 2.5, depending on the subspace used for projecting.

#### 12.3.2.1 Applications: Finding Transition Pathways in Adenylate Kinase

Understanding the functional basis for many protein functions (Gerstein et al. 1994; Berg et al. 2002; Karplus and Gao 2004; Xu et al. 1997) requires detailed knowledge of transitions between functionally relevant conformations. Over the last years X-ray crystallography and NMR spectroscopy have provided mostly static

pictures of different conformational states of proteins, leaving questions related to the underlying transition pathway unanswered. For atomistic MD simulations, elucidating the pathways and mechanisms of protein conformational dynamics poses a challenge due to the long timescales involved. In this respect, *E. coli* adenylate kinase (ADK) is a prime example. ADK is a monomeric enzyme that plays a key role in energy maintenance within the cell, controlling cellular ATP levels by catalyzing the reaction $Mg^{2+}$: ATP + AMP ↔ $Mg^{2+}$: ADP + ADP. Structurally, the enzyme consists of three domains (Fig. 12.10): the large central "CORE" domain (light grey), an AMP binding domain referred to as "AMPbd" (black), and a lid-shaped ATP-binding domain termed "LID" (dark grey), which covers the phosphate groups at the active centre (Müller et al. 1996). In an unligated structure of ADK the LID and AMPbd adopt an open conformation, whereas they assume a closed conformation in a structure crystallized with the transition state inhibitor $Ap_5A$ (Müller and Schulz 1992). Here, the ligands are contained in a highly specific environment required for catalysis. $^{15}N$ nuclear magnetic resonance spin relaxation studies (Shapiro and Meirovitch 2006) have shown the existence of catalytic domain motions in the flexible AMPbd and LID domains on the nanosecond time scale, while the relaxation in the CORE domain is on the picosecond time scale (Tugarinov et al. 2002; Shapiro et al. 2002). For ADK, several computational studies have addressed its conformational flexibility (Temiz et al. 2004; Maragakis and Karplus 2005; Lou and Cukier 2006; Whitford et al. 2007; Snow et al. 2007). However, due to the magnitude and timescales involved, spontaneous transitions between the open and closed conformations have not been achieved until now by all-atom MD simulations. Using TEE-REX, spontaneous transitions between the open and closed structures of ADK are facilitated, and a fully atomistic description of the transition pathway and its underlying mechanics could be achieved (Kubitzki and de Groot 2008). To this end, different essential subspaces {es} were constructed from short MD simulations of either conformation as well as from a combined ensemble holding structures from both the open and closed conformation. In the latter case, {es} modes were excited containing the difference X-ray mode connecting the open and closed experimental structures.

The observed transition pathway can be characterized by two phases. Starting from the closed conformation (Fig. 12.10 left), the LID remains essentially closed while the AMPbd, comprising helices $\alpha_2$ and $\alpha_3$, assumes a half-open conformation. In doing so, $\alpha_2$ bends towards helix $\alpha_4$ of the CORE by 15° with respect to $\alpha_3$. This opening of the AMP binding cleft could facilitate an efficient release of the formed product. For the second phase, a partially correlated opening of the LID domain together with the AMPbd is observed. Compared to coarse-grained approaches, all-atom TEE-REX simulations allow detailed analyses of inter-residue interactions. For ADK, a highly stable salt bridge between residues Asp118 and Lys136 forms during phase one, connecting the LID and CORE domains. Estimating the total non-bonded interaction between LID and CORE, it was found that this salt-bridge contributes substantially to the interaction of the two domains. Breaking this salt bridge via mutation, e.g. Asp118Ala, should thus decrease the

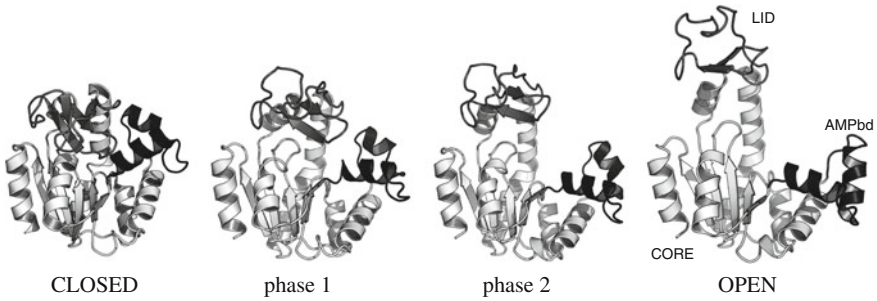



**Fig. 12.10** Closed (*left*) and open (*right*) crystal structures of *E. coli* adenylate kinase (ADK) together with intermediate structures characterizing the two phases of the closed-open transition. ADK has domains CORE (*light grey*), AMPbd (*black*) and LID (*dark grey*). The transition state inhibitor $Ap_5A$ is removed in the closed crystal structure (*left*)

stability of the open state. From a comparison of fourteen Protein Data Bank (PDB) structures from yeast, maize, human and bacterial ADK, eleven structures feature such a salt-bridge motif at the LID-CORE interface.

Alternative transition pathways seem possible, however an analysis of all TEE-REX simulations suggests a high free energy barrier obstructing the full opening of the AMPbd after the LID has opened. Together with the observed larger fluctuations in secondary structure elements, indicating high internal strain energies, the enthalpic penalty along this route possibly renders it unfavourable as a transition pathway of ADK.

## 12.4 Methods for Functional Mode Prediction

As discussed in the previous section, functional modes in proteins are usually those with the lowest frequencies. Apart from molecular dynamics based techniques, there are several alternative methods that focus on the prediction of these essential degrees of freedom based on a single input structure.

### *12.4.1 Normal Mode Analysis*

Normal mode analysis (NMA) is one of the major simulation techniques used to probe the large-scale, shape-changing motions in biological molecules (Gō et al. 1983; Brooks and Karplus 1983; Levitt et al. 1983). These motions are often coupled to function and a consequence of binding other molecules like substrates, drugs or other proteins. In NMA studies it is implicitly assumed that the normal modes with the largest fluctuation (lowest frequency modes) are the ones that are

functionally relevant, because, like function they exist by evolutionary "design" rather than by chance.

Normal mode analysis is a harmonic analysis. The underlying assumption is that the conformational energy surface can be approximated by a parabola, despite the fact that functional modes at physiological temperatures are highly anharmonic (Brooks and Karplus 1983; Austin et al. 1975). To perform a normal mode analysis one needs a set of coordinates, a force field describing the interactions between constituent atoms, and software to perform the required calculations. The performance of a normal mode analysis in Cartesian coordinate space requires three main calculation steps.

1. Minimization of the conformational potential energy as a function of the atomic coordinates.
2. The calculation of the so-called "Hessian" matrix

$$H = \frac{\partial^2 V}{\partial x_i \partial x_j}$$

   which is the matrix of second derivatives of the potential energy with respect to the mass-weighted atomic coordinates.
3. The diagonalization of the Hessian matrix. This final steps yields eigenvalues and eigenvectors (the "normal modes").

Energy minimization can require quite a lot of CPU time. Furthermore, as the Hessian matrix is a 3N × 3N matrix, where N is the number of atoms, the last step can be computationally demanding.

### *12.4.2 Elastic Network Models*

Elastic or Gaussian network models (Tirion 1996) (ENM) are basically a simplification of normal mode analysis. Usually, instead of an all atom representation, only $C_\alpha$ atoms are taken into account. This means a ten-fold reduction of the number of particles which decreases the computational effort dramatically. Moreover, as the input coordinates are taken as representing the ground state, no energy minimization is required.

The potential energy is calculated according to

$$V = \frac{\gamma}{2} \sum_{|r_{ij}^0| < R_C} (r_{ij} - r_{ij}^0)^2$$

where $\gamma$ denotes the spring constant and $R_C$ the cut-off distance. Regarding the drastic assumptions inherent in normal mode analysis, these simplifications do not mean a severe loss of quality. This, together with the relatively low computational

costs, explains the current popularity of elastic network models. ENM calculations are also offered on web servers such as ElNemo (Suhre and Sanejouand 2004a, b) (http://www.igs.cnrs-mrs.fr/elnemo/) and AD-ENM (Zheng and Doniach 2003; Zheng and Brooks 2005) (http://enm.lobos.nih.gov/).

### *12.4.3 CONCOORD*

CONCOORD (de Groot et al. 1997) uses a geometry-based approach to predict protein flexibility. The three-dimensional structure of a protein is determined by various interactions such as covalent bonds, hydrogen bonds and non-polar interactions. Most of these interactions remain intact during functionally relevant conformational changes. This notion lies at the heart of the CONCOORD simulation method: based on an input structure, *alternative* structures are generated that share the large majority of interactions found in the original configuration. To this end, in the first step of a CONCOORD simulation (Fig. 12.11) interactions in a single input structure are analyzed and turned into geometrical constraints, mainly distance constraints with upper and lower bounds for atomic distances but also angle constraints and information about planar and chiral groups. This geometrical description of the structure can be compared to a construction plan of the protein. In the second step, starting from random atomic coordinates, the structure is iteratively rebuilt based on the predefined construction plan, commonly several hundreds of times. As each run starts from random coordinates, the method does not suffer from sampling problems like MD simulations and the resulting ensemble covers the whole conformational space that is available within the predefined constraints. However, the method does not provide information about the path between two conformational substates or about timescales and energies (Fig. 12.12).

#### 12.4.3.1 Applications

CONCOORD and the extension tCONCOORD (t stands for transition) (Seeliger et al. 2007) have been applied to diverse proteins. Adenylate kinase displays a distinct domain-closing motion upon binding to its substrate (ATP/AMP) or an inhibitor (see Fig. 12.13 left) with a $C_\alpha$-RMSD of 7.6 Å between the ligand-bound and the ligand-free conformation. Two tCONCOORD simulations were carried out using a closed conformation (PDB 1AKE) as input. In one simulation the ligand ($Ap_5A$) was removed. Figure 12.13 (right) shows the result of a principal components analysis (PCA) applied to the experimental structures. The first eigenvector (x-axis) corresponds to the domain-opening motion indicated by the arrow in Fig. 12.13 (left). Every dot in the plot represents a single structure. Red dots represent the ensemble that has been generated using the closed conformation of adenylate kinase without ligand as input. Green dots represent the ensemble that has been generated using the ligand-bound structure as input. Whereas the simulation

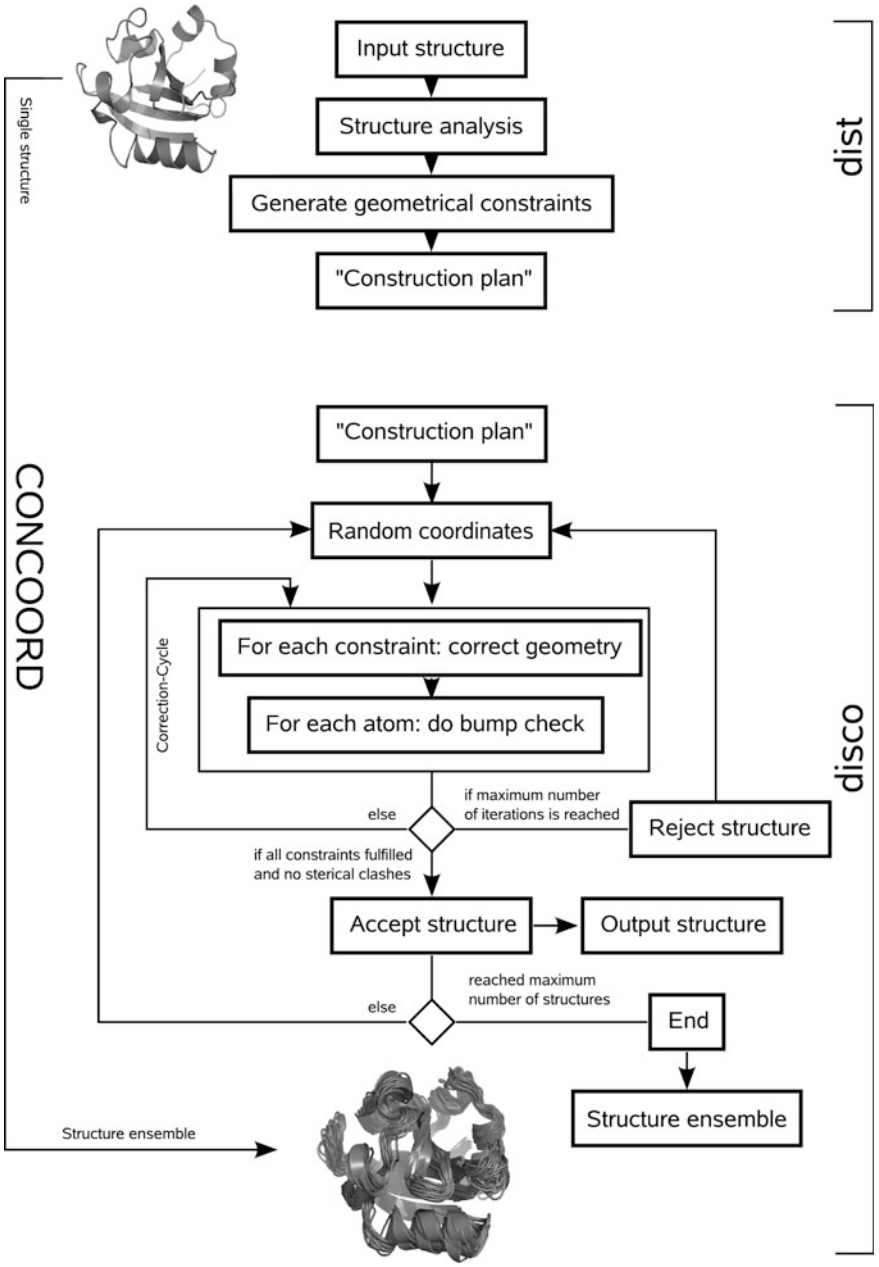



**Fig. 12.11** Schematic representation of the CONCOORD method for generating structure ensembles from a single input structure. In a first step (program *dist*) a single input structure is analyzed and turned into a geometric description of the protein. In a second step (program *disco*) the structure is rebuilt based on the predefined constraints, starting from random coordinates

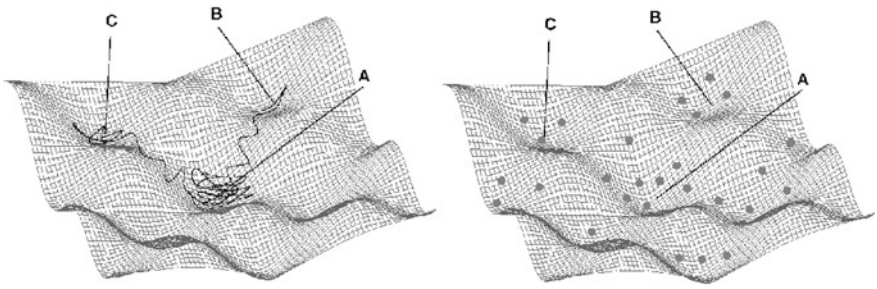



**Fig. 12.12** Comparison of the sampling properties of Molecular Dynamics and CONCOORD on hypothetic energy landscapes. A MD-trajectory (*left*) "walks" on the energy landscape, thereby providing information about timescales and paths between conformational substates. The (non-deterministic) CONCOORD-ensemble (*right*) "jumps" on the energy landscape, thereby offering better sampling of the conformational space

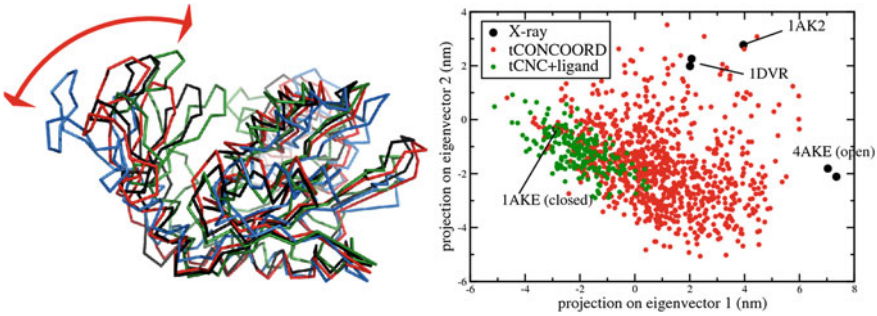



**Fig. 12.13** *Left* Overlay of X-ray structures of adenylate kinase. *Right* principal component analysis. Two tCONCOORD ensembles are projected onto the first two eigenvectors of a PCA carried out on an ensemble of X-ray structures. The ensemble represented by *red dots* has been started from a closed conformation (1AKE) with removed inhibitor. The generated ensemble samples both, closed and open conformations. The ensemble represented with *green dots* has also been started from a closed conformation (1AKE) but with inhibitor present. The generated ensemble only samples closed conformations around the ligand bound conformation

with inhibitor basically samples closed conformations around the ligand-bound state, the ligand-free simulation samples both, open and closed conformations, thereby reaching the experimentally determined open conformations with RMSD's of 2.4, 2.6, and 3.1 Å for 1DVR, 1AK2, and 4AKE, respectively. In structure-based drug design, often the reverse problem, predicting ligand-bound structures from unbound conformations, needs to be addressed. A tCONCOORD simulation starting with an open conformation (4AKE) as input produced structures that approach the closed conformations with RMSD's of 2.5, 2.9, and 3.3 Å for 1DVR, 1AK2, and 1AKE, respectively. Thus, the functional domain-opening motion has been predicted in both cases, when using a closed, ligand-bound conformation as input and when using an open, ligand-free conformation as input.

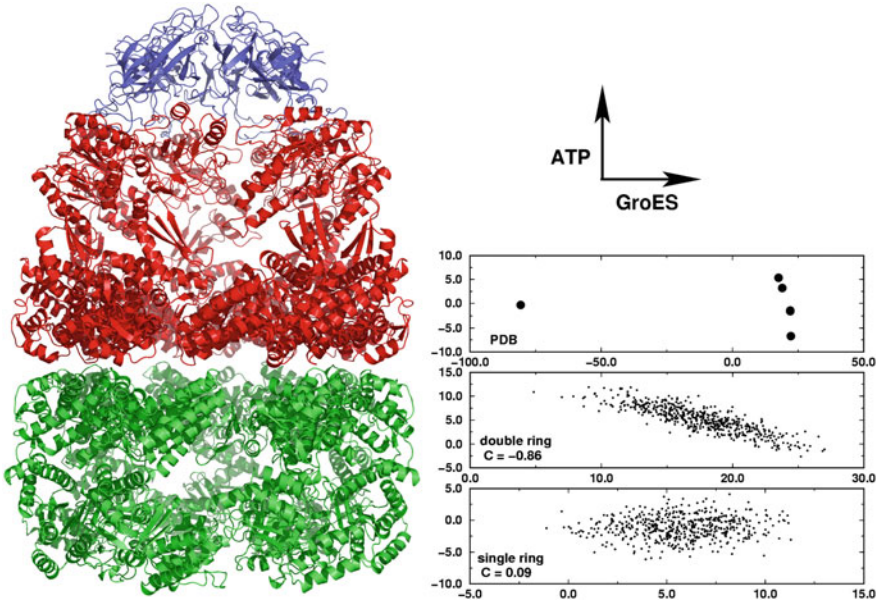



**Fig. 12.14** Asymmetric GroEL-GroES complex (*left*), together with CONCOORD simulation results (*right*). The GroEL-GroES complex consists of the co-chaperonin GroES (*blue*), the trans-ring of GroEL, bound to GroES (*red*), and the cis-ring (*green*). A principal component analysis revealed two main structural transitions per GroEL ring, upon nucleotide binding (*vertical axis* in the *right panels*) and GroES binding (*horizontal axis*), respectively. In simulations of the double ring, but not in a single ring, these modes were found to be coupled, suggesting a coupling between intra-ring and inter-ring cooperativity

Because of its computational efficiency, CONCOORD can be routinely applied to extract functionally relevant modes of flexibility for molecular systems that are beyond the size limitations of other atomistic simulation techniques like molecular dynamics simulations. An application to the chaperonin GroEL-GroES complex that contains more than 8000 amino acids revealed a novel form of coupling between intra-ring and inter-ring cooperativity (de Groot et al. 1999). Each GroEL ring displays two main modes of collective motion: the main conformational transition upon binding of the co-chaperonin GroES, and a secondary transition upon ATP binding (Fig. 12.14 upper right panel). CONCOORD simulations of a single GroEL ring did not show any coupling between these modes, whereas simulations of the double ring system showed a strict correlation between the two modes, thereby providing an explanation for how nucleotide binding is coupled to GroES affinity in the double ring, but not in a single ring.

## 12.5 Summary and Outlook

Computational methods gain growing recognition in structural biology and protein research. Protein function is usually a dynamic process involving structural rearrangements and conformational transitions between stable states. Since such dynamic processes are difficult to study experimentally, in silico methods can significantly contribute to the understanding of protein function at atomic resolution. The most prominent method to study protein dynamics is molecular dynamics (MD), where atoms are treated as classical particles and their interactions are approximated by an empirical force field. Newton's equations of motion are solved at discrete time steps, leading to a trajectory that describes the dynamical behaviour of the system. Despite their growing popularity the scope of application for MD simulations is limited by computational demands. Within the next 10 years the accessible timescales for the simulation of average sized proteins will, in all likelihood, not exceed the low microsecond range for most biomolecular systems. However, since functionally relevant protein dynamics is usually represented by collective, low-frequency motions taking place on the micro- to millisecond timescale, standard MD simulations are ill-suited to be routinely applied to study conformational dynamics of large biomolecules.

Different methodologies have been developed to alleviate this sampling problem that standard MD suffers from. One approach is to reduce the number of particles, either by fusing groups of atoms into pseudo-atoms (coarse-graining), or by replacing explicit solvent molecules with an implicit solvent continuum model. In both cases the number of particles is significantly reduced, facilitating much longer time scales than in all-atom simulations using explicit solvent. However, the loss of "resolution" inherent to both methods may limit their accuracy and hence, their applicability. Other approaches retain the atomistic description and pursue different sampling strategies.

Generalized ensemble algorithms such as Replica Exchange (REX) make use of the fact that conformational transitions occur more frequently at higher temperatures. In standard temperature REX, several copies (replicas) of the system are simulated with MD at different temperatures, with frequent exchanges between replicas. Thereby, low-temperature replicas utilize the enhanced barrier-crossing capabilities of high-temperature replicas. Although dynamical information gets lost in this setup, each replica still represents a Boltzmann ensemble at its corresponding temperature, providing valuable information about thermodynamics and thus the stability of different conformational substates. Although often used in the context of protein folding, REX simulations at full atomic resolution quickly become computationally very demanding for systems comprising more than a few thousand atoms.

Whereas REX is an unbiased sampling method, several other methods exist that bias the system in order to enhance sampling predominantly along certain collective degrees of freedom. Functionally relevant protein motions often correspond to those eigenvectors of the covariance matrix of atomic fluctuations having the largest

eigenvalues. If these eigenvectors are known from a principal component analysis (PCA), either using experimental data or previous simulations, they can be used in simulation protocols like Conformational Flooding or Essential Dynamics (ED). However, in both methods the enhancement in sampling is paid for by losing the canonical properties of the resulting trajectory.

The recently developed TEE-REX protocol combines the favourable properties of REX with those resulting from a specific excitation of functionally relevant modes (as e.g. in ED), while at the same time avoiding the aforementioned drawbacks of each method. In particular, approximate canonical integrity of the reference ensemble is maintained and sampling along the main collective modes of motion is significantly enhanced. The resulting reference ensemble can thus be used to calculate equilibrium properties of the system which allows comparison with experimental data.

Although significant progress has been made in the development of enhanced sampling methods, computational demands of MD based methods are still substantial. For many questions in structural biology it is already beneficial to have an idea about possible protein conformations and functional modes without the need to get detailed information about energetics and timescales. In this respect, elastic network models offer a cheap way to get an estimate of possible functional protein motions. Although drastic assumptions are made and no atomistic picture is obtained the predicted collective motions are often in qualitatively good agreement with experimental results. Another computational efficient way which retains the atomistic description of protein structures is the CONCOORD method where a protein is described with geometrical constraints. Based on a construction plan derived from a single input structure, an ensemble of structures is generated which represents an exhaustive sampling of conformational space that is available within the predefined constraints. However, no information about timescales or energies is obtained.

Right now there is no single method that is routinely applicable to predict functionally relevant protein motions from a given three-dimensional structure. However, there are a large number of methods available, capturing different aspects of the problem and contributing to our understanding of protein function. Thus, combinations of existing methods will presumably be the most straightforward way of enhancing the predictive power of in silico methods.

## References

Adcock SA, McCammon JA (2006) Molecular dynamics: survey of methods for simulating the activity of proteins. Chem Rev 106:1589–1615

Affentranger R, Tavernelli I, di Iorio E (2006) A novel Hamiltonian replica exchange MD protocol to enhance protein conformational space sampling. J Chem Theory Comput 2:217–228

Amadei A, Linssen ABM, Berendsen HJC (1993) Essential dynamics of proteins. Proteins 17:412–425

Amadei A, Linssen ABM, de Groot BL et al (1996) An efficient method for sampling the essential subspace of proteins. J Biomol Struct Dyn 13:615–626
Amadei A, de Groot BL, Ceruso M-A et al (1999) A kinetic model for the internal motions of proteins: Diffusion between multiple harmonic wells. Proteins 35:283–292
Anderson HC (1980) Molecular dynamics simulations at constant pressureand/or temperature. J Chem Phys 72:2384–2393
Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230
Austin RH, Beeson KW, Eisenstein L et al (1975) Dynamics of ligand binding to myoglobin. Biochemistry 14(24):5355–5373
Bahar I, Erman B, Haliloglu T et al (1997) Efficient characterization of collective motions and interresidue correlations in proteins by low-resolution simulations. Biochemistry 36:13512–13523
Bartels C, Karplus M (1998) Probability distributions for complex systems: adaptive umbrella sampling of the potential energy. J Phys Chem B 102:865–880
Berendsen HJC, Postma JPM, di Nola A et al (1984) Molecular dynamics with coupling to an external bath. J Chem Phys 81:3684–3690
Berg BA, Celik T (1992) New approach to spin-glass simulations. Phys Rev Lett 69:2292–2295
Berg BA, Neuhaus T (1991) Multicanonical algorithms for first-order phase transitions. Phys Lett 267:249–253
Berg JM, Tymoczko JL, Stryer L (2002) Biochemistry, 5th edn. WH Freeman and Co., New York
Bond PJ, Holyoake J, Ivetac A et al (2007) Coarse-grained molecular dynamics simulations of membrane proteins and peptides. J Struct Biol 157:593–605
Brooks B, Karplus M (1983) Harmonic dynamics of proteins: normal modes and fluctuations in bovine pancreatic trypsin inhibitor. Proc Natl Acad Sci 80:6571–6575
Brooks BR, Bruccoleri RE, Olafson BD et al (1983) CHARMM: a program for macromolecular energy minimization and dynamics calculations. J Comp Chem 4:187–217
Burykin A, Warshel A (2003) What really prevents proton transport through aquaporin? Charge self-energy versus proton wire proposals. Biophys J 85:3696–3706
Cecchini M, Rao F, Seeber M et al (2004) Replica exchange molecular dynamics simulations of amyloid peptide aggregation. J Chem Phys 121:10748–10756
Chakrabarti N, Tajkhorshid E, Roux B et al (2004) Molecular basis of proton blockage in aquaporins. Structure 12:65–74
Chaudhri A, Zarranga IE, Kamerzell TJ et al (2012) Coarse-grained modeling of the self-association of therapeutic monoclonal antibodies. J Phys Chem B 116:8045–8057
Chen H, Wu Y, Voth GA (2006) Origins of proton transport behavior from selectivity domain mutations of the aquaporin-1 channel. Biophys J 90:L73–L75
Cheng X, Cui G, Hornak V et al (2005) Modified replica exchange simulation for local structure refinement. J Phys Chem B 109:8220–8230
Chodera JD, Swope WC, Pitera JW et al (2007) Use of the weighted histogram analysis method for the analysis of simulated and parallel tempering simulations. J Chem Theory Comput 3:26–41
Christen M, van Gunsteren WF (2006) Multigraining: an algorithm for simultaneous fine-grained and coarse-grained simulation of molecular systems. J Chem Phys 124:154106
Cook A, Fernandez E, Lindner D et al (2005) The structure of the nuclear export receptor Cse1 in its cytosolic state reveals a closed conformation incompatible with cargo binding. Mol Cell 18:355–357
Currie MG, Fok KF, Kato J et al (1992) Guanylin: an endogenous activator of intestinal guanylate cyclise. Proc Natl Acad Sci 89:947–951
de Groot BL, Grubmüller H (2001) Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and GlpF. Science 294:2353–2357
de Groot BL, Amadei A, Scheek RM et al (1996a) An extended sampling of the configurational space of HPr from *E. coli*. Proteins 26:314–322
de Groot BL, Amadei A, van Aalten DMF et al (1996b) Towards an exhaustive sampling of the configurational spaces of the two forms of the peptide hormone guanylin. J Biomol Str Dyn 13:741–751

de Groot BL, van Aalten DMF, Scheek RM et al (1997) Prediction of protein conformational freedom from distance constraints. Proteins 29:240–251
de Groot BL, Hayward S, van Aalten DMF et al (1998) Domain motions in bacteriophage T4 lysozyme: a comparison between molecular dynamics and crystallographic data. Proteins 31:116–127
de Groot BL, Vriend G, Berendsen HJC (1999) Conformational changes in the chaperonin GroEL: new insights into the allosteric mechanism. J Mol Biol 286:1241–1249
de Groot BL, Engel A, Grubmüller H (2001) A refined structure of human aquaporin-1. FEBS Lett 504:206–211
de Groot BL, Frigato T, Helms V et al (2003) The mechanism of proton exclusion in the aquaporin-1 water channel. J Mol Biol 333:279–293
Dixon MM, Nicholson H, Shewchuk L et al (1992) Structure of a hinge-bending bacteriophage T4 lysozyme mutant Ile3 → Pro. J Mol Biol 227:917–933
Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley, New York
Faber HR, Matthews BW (1990) A mutant T4 lysozyme displays five different crystal conformations. Nature 348:263–266
Frauenfelder H, Leeson DT (1998) The energy landscape in non-biological and biological molecules. Nat Struct Biol 5:757–759
Frauenfelder H, Sligar SG, Wolynes PG (1991) The energy landscapes and motions of proteins. Science 254:1598–1603
Fu D, Libson A, Miercke LJ et al (2000) Structure of a glycerol-conducting channel and the basis for its selectivity. Science 290:481–486
Fukunishi H, Watanabe O, Takada S (2002) On the Hamiltonian replica exchange method for efficient sampling of biomolecular systems: application to protein structure prediction. J Chem Phys 116:9058–9067
García AE (1992) Large-amplitude nonlinear motions in proteins. Phys Rev Lett 68:2696–2699
García AE, Onuchic JN (2003) Folding a protein in a computer: an atomic description of the folding/unfolding of protein A. Proc Natl Acad Sci 100:13898–13903
Gō N, Noguti T, Nishikawa T (1983) Dynamics of a small globular protein in terms of low-frequency vibrational modes. Proc Natl Acad Sci 80:3696–3700
Gerstein M, Lesk AM, Chothia C (1994) Structural mechanisms for domain movements in proteins. Biochemistry 33:6739–6749
Gosh A, Rapp CS, Friesner RA (1998) Generalized Born model based on a surface integral formulation. J Phys Chem B 102:10983–10990
Grubmüller H (1995) Predicting slow structural transitions in macromolecular systems: conformational flooding. Phys Rev E 52:2893–2906
Hansmann UHE (1997) Effective way for determination of multicanonical weights. Phys Rev E 56:6200–6203
Hayward S, Kitao A, Gō N (1995) Harmonicity and anharmonicity in protein dynamics: a normal mode analysis and principal component analysis. Proteins 23:177–186
He J, Zhang Z, Shi Y et al (2003) Efficiently explore the energy landscape of proteins in molecular dynamics simulations by amplifying collective motions. J Chem Phys 119:4005–4017
Hockney RW, Goel SP, Eastwood JW (1973) 10000 particle molecular dynamics model with long-range forces. Chem Phys Lett 21:589–591
Huang J, Lopes PEM, Roux B et al (2014) Recent advances in polarizable force fields for macromolecules: microsecond simulations of proteins using the classical Drude oscillator model. J Phys Chem Lett 5:3144–3150
Hub JS, de Groot BL (2008) Mechanism of selectivity in aquaporins and aquaglyceroporins. Proc Natl Acad Sci 105:1198–1203
Iba Y (2001) Extended ensemble Monte Carlo. Int J Mod Phys C 12:623–656
Ilan B, Tajkhorshid E, Schulten K et al (2004) The mechanism of proton exclusion in aquaporin channels. Proteins 55:223–228

Jean-Charles A, Nicholls A, Sharp K et al (1991) Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations. J Am Chem Soc 113:1454–1455

Jorgensen WL, Chandrasekhar J, Madura JD et al (1983) Comparison of simple potential functions for simulating liquid water. J Chem Phys 79:926–935

Jorgensen WL, Maxwell DS, Tirado-Rives J (1996) Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 118:11225–11236

Karplus M, Gao YQ (2004) Biomolecular motors: the F1-ATPase paradigm. Curr Opin Struct Biol 14:250–259

Karplus M, Kushick JN (1981) Method for estimating the configurational entropy of macromolecules. Macromolecules 14:325–332

Kempf JG, Loria JP (2003) Protein dynamics from solution NMR theory and applications. Cell Biochem Biophys 37:187–211

Kitao A, Gō N (1999) Investigating protein dynamics in collective coordinate space. Curr Opin Struct Biol 9:143–281

Kitao A, Hirata F, Gō N (1991) The effects of solvent on the conformation and the collective motions of proteins—normal mode analysis and molecular-dynamics simulations of melittin in water and vacuum. Chem Phys 158:447–472

Kitao A, Hayward S, Gō N (1998) Energy landscape of a native protein: jumping-among-minima model. Proteins 33:496–517

Kokubo H, Okamoto Y (2004) Prediction of membrane protein structures by replica-exchange Monte Carlo simulations: case of two helices. J Chem Phys 120:10837–10847

Kubitzki MB, de Groot BL (2007) Molecular dynamics simulations using temperature-enhanced essential dynamics replica exchange. Biophys J 92:4262–4270

Kubitzki MB, de Groot BL (2008) The atomistic mechanism of conformational transition in adenylate kinase: a TEE-REX molecular dynamics study. Structure 16(8):1175–1182

Kumar S, Bouzida D, Swendsen RH et al (1992) The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method. J Comput Chem 13:1011–1021

Kumar S, Payne PW, Vásquez M (1996) Method for free-energy calculations using iterative techniques. J Comput Chem 17:1269–1275

Kuroki R, Weaver LH, Matthews BW (1993) A covalent enzyme-substrate intermediate with saccharide distortion in a mutant T4 lysozyme. Science 262:2030–2033

Levitt M, Sander C, Stern PS (1983) Normal-mode dynamics of a protein: bovine pancreatic trypsin inhibitor. Int J Quant Chem Quant Biol Symp 10:181–199

Levy RM, Karplus M, Kushick J et al (1984a) Evaluation of the configurational entropy for proteins: application to molecular dynamics of an α-helix. Macromolecules 17:1370–1374

Levy RM, Srinivasan AR, Olsen WK et al (1984b) Quasi-harmonic method for studying very low frequency modes in proteins. Biopolymers 23:1099–1112

Liu P, Kim B, Friesner RA et al (2005) Replica exchange with solute tempering: a method for sampling biological systems in explicit water. Proc Natl Acad Sci 102:13749–13754

Lou H, Cukier RI (2006) Molecular dynamics of apo-adenylate kinase: a distance replica exchange method for the free energy of conformational fluctuations. J Phys Chem B 110:24121–24137

Luo R, David L, Gilson ML (2002) Accelerated Poisson-Boltzmann calculations for static and dynamic systems. J Comput Chem 23:1244–1253

Lyman E, Zuckerman DM (2006) Ensemble-based convergence analysis of biomolecular trajectories. Biophys J 91:164–172

Maragakis P, Karplus M (2005) Large amplitude conformational change in proteins explored with a plastic network model: adenylate kinase. J Mol Biol 352:807–822

Marinari E, Parisi G (1992) Simulated tempering: a new Monte Carlo scheme. Europhys Lett 19:451–458

Marrink SJ, de Vries AH, Mark AE (2004) Coarse grained model for semiquantitative lipid simulations. J Phys Chem B 108:750–760

Matthews BW, Remington SJ (1974) The three dimensional structure of the lysozyme from bacteriophage T4. Proc Natl Acad Sci 71:4178–4182
McCammon JA, Gelin BR, Karplus M (1977) Dynamics of folded proteins. Nature 267:585–590
Mitsutake A, Sugita Y, Okamoto Y (2001) Generalized-ensemble algorithms for molecular simulations of biopolymers. Biopolymers 60:96–123
Moffat K (2003) The frontiers of time-resolved macromolecular crystallography: movies and chirped X-ray pulses. Faraday Discuss 122:65–77
Murata K, Mitsuoka K, Walz T et al (2000) Structural determinants of water permeation through Aquaporin-1. Nature 407:599–605
Müller CW, Schulz GE (1992) Structure of the complex between adenylate kinase from *Eschericia coli* and the inhibitor $Ap_5A$ refined at 19 Å resolution: a model for a catalytic transition state. J Mol Biol 224:159–177
Müller CW, Schlauderer G, Reinstein J et al (1996) Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. Structure 4:147–156
Nguyen PH, Mu Y, Stock G (2005) Structure and energy landscape of a photoswitchable peptide: a replica exchange molecular dynamics study. Proteins 60:485–494
Nose S (1984) A unified formulation of the constant temperature molecular dynamics method. J Chem Phys 81:511–519
Pitera JW, Swope W (2003) Understanding folding and design: replica-exchange simulations of "Trp-cage" miniproteins. Proc Natl Acad Sci 100:7587–7592
Rao F, Caflisch A (2003) Replica exchange molecular dynamics simulations of reversible folding. J Chem Phys 119:4035–4042
Romo TD, Clarage JB, Sorensen DC et al (1995) Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. Proteins 22:311–321
Saunders MG, Voth GA (2013) Coarse-graining methods for computational biology. Ann Rev Biophys 42:73–93
Schotte F, Lim M, Jackson TA et al (2003) Watching a protein as it functions with 150 ps time-resolved X-ray crystallography. Science 300:1944–1947
Seeliger D, Haas J, de Groot BL (2007) Geometry-based sampling of conformational transitions in proteins. Structure 15:1482–1492
Seibert MM, Patriksson A, Hess B et al (2005) Reproducible polypeptide folding and structure prediction using molecular dynamics simulations. J Mol Biol 354:173–183
Shapiro YE, Meirovitch E (2006) Activation energy of catalysis-related domain motion in *E. coli* adenylate kinase. J Phys Chem B 110:11519–11524
Shapiro YE, Kahana E, Tugarinov V et al (2002) Domain flexibility in ligand-free and inhibitor bound *Eschericia coli* adenylate kinase based on a mode-coupling analysis of $^{15}N$ spin relaxation. Biochemistry 41:6271–6281
Smith GR, Bruce AD (1996) Multicanonical Monte Carlo study of solid-solid phase coexistence in a model colloid. Phys Rev E 53:6530–6543
Snow C, Qi G, Hayward S (2007) Essential dynamics sampling study of adenylate kinase: comparison to citrate synthase and implication for the hinge and shear mechanisms of domain motion. Proteins 67:325–337
Still WC, Tempczyk A, Hawley RC et al (1990) Semianalytical treatment of solvation for molecular mechanics and dynamics. J Am Chem Soc 112:6127–6129
Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314:141–151
Sugita Y, Kitao A, Okamoto Y (2000) Multidimensional replica-exchange method for free-energy calculations. J Chem Phys 113:6042–6051
Suhre K, Sanejouand YH (2004a) ElNemo: a normal mode web-server for protein movement analysis and the generation of templates for molecular replacement. Nucl Acids Res 32:610–614
Suhre K, Sanejouand YH (2004b) On the potential of normal mode analysis for solving difficult molecular replacement problems. Act Cryst D 60:796–799

Sui H, Han B-G, Lee JK et al (2001) Structural basis of water-specific transport through the AQP1 water channel. Nature 414:872–878

Tai K (2004) Conformational sampling for the impatient. Biophys Chem 107:213–220

Tajkhorshid E, Nollert P, Jensen MØ et al (2002) Control of the selectivity of the aquaporin water channel family by global orientational tuning. Science 296:525–530

Teeter MM, Case DA (1990) Harmonic and quasi harmonic descriptions of crambin. J Phys Chem 94:8091–8097

Temiz NA, Meirovitch E, Bahar I (2004) *Eschericia coli* adenylate kinase dynamics: comparison of elastic network model modes with mode-coupling $^{15}$N-NMR relaxation data. Proteins 57:468–480

Tirion MM (1996) Large amplitude elastic motions in proteins from a single-parameter atomic analysis. Phys Rev Lett 77:186–195

Tugarinov V, Shapiro YE, Liang Z et al (2002) A novel view of domain flexibility in E coli adenylate kinase based on structural mode-coupling $^{15}$N NMR spin relaxation. J Mol Biol 315:155–170

Van Aalten DMF, Amadei A, Vriend G et al (1995a) The essential dynamics of thermolysin—confirmation of hinge-bending motion and comparison of simulations in vacuum and water. Prot Eng 8:1129–1136

Van Aalten DMF, Findlay JBC, Amadei A et al (1995b) Essential dynamics of the cellular retinol binding protein—evidence for ligand induced conformational changes. Prot Eng 8:1129–1136

Van Gunsteren WF, Berendsen HJC (1987) Groningen molecular simulation (GROMOS) library manual. Biomos, Groningen

Van Gunsteren WF, Berendsen HJC (1990) Computer-simulation of molecular-dynamics—methodology, applications, and perspectives in chemistry. Angew Chem Int Edit Engl 29:992–1023

Warshel A, Kato M, Pisliakov AV (2007) Polarizable force fields: history test cases and prospects. J Chem Theory Comput 3:2034–2045

Weiner SJ, Kollman PA, Nguyen DT et al (1986) An all atom force field for simulations of proteins and nucleic acids. J Comp Chem 7:230–252

Weiss S (1999) Fluorescence spectroscopy of single biomolecules. Science 283:1676–1683

Whitford PC, Miyashita O, Levy Y et al (2007) Conformational transitions of adenylate kinase: switching by cracking. J Mol Biol 366:1661–1671

Xu Z, Horwich AL, Sigler PB (1997) The crystal structure of the asymmetric Gro-EL-GroES-$(ADP)_7$ chaperonin complex. Nature 388:741–750

Zachariae U, Grubmüller H (2006) A highly strained nuclear conformation of the exportin Cse1p revealed by molecular dynamics simulations. Structure 14:1469–1478

Zhang X-J, Wozniak JA, Matthews BW (1995) Protein flexibility and adaptability seen in 25 crystal forms of T4 lysozyme. J Mol Biol 250:527–552

Zhang Z, Shi Y, Liu H (2003) Molecular dynamics simulations of peptides, and proteins with amplified collective motions. Biophys J 84:3583–3593

Zheng W, Brooks BR (2005) Probing the local dynamics of nucleotide-binding pocket coupled to the global dynamics: myosin versus kinesin. Biophys J 89(1):167–178

Zheng W, Doniach S (2003) A comparative study of motor-protein motions by using a simple elastic-network model. Proc Natl Acad Sci 100(23):13253–13258

Zhou R, Berne BJ, Germain R (2001) The free energy landscape for β-hairpin folding in explicit water. Proc Natl Acad Sci 98:14931–14936

# Chapter 13
# Integrated Servers for Structure-Informed Function Prediction

**Roman A. Laskowski**

**Abstract** No single method for predicting a protein's function from its three-dimensional structure is perfect; some methods work well in some cases, whereas other methods perform better in others. Consequently, it makes sense to apply a number of different predictive methods to a given protein structure and obtain either a consensus or the most likely prediction from them all. In this chapter we describe two web servers, ProKnow (http://proknow.mbi.ucla.edu) and ProFunc (http://www.ebi.ac.uk/profunc), that use a cocktail of methods for predicting function from 3D structure.

**Keywords** Function prediction · ProKnow · ProFunc · 3D structure · Motifs · Homology · Fold-matching · Protein interactions · 3D templates

## 13.1 Introduction

For a protein of unknown function, can knowledge of its 3D structure help identify its function? The structure undoubtedly holds clues to what the protein does, but the problem is how to identify those clues, discarding any red herrings, and arrive at a reliable prediction of the protein's biochemical, or even biological, function.

This topic became particularly significant in the early 2000s with the birth of the various Structural Genomics (SG) initiatives (Burley 2000; Blundell and Mizuguchi 2000; Chandonia and Brenner 2006; Norvell and Berg 2007). Before then, experimentalists would already know much about their protein before embarking on determination of its 3D structure and would have selected it for its biological interest. Much of the point of solving the structure was to gain an insight into how the protein achieved its biological function at the atomic level. The motivation of the SG groups, with their high-throughput structure determination methods, differed

R.A. Laskowski (✉)
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
e-mail: roman@ebi.ac.uk



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_13

markedly. Now a protein would be solved if it belonged to a family with no structural representatives, or was expected to have a novel fold, or was relevant to some disease. Knowledge of its function no longer came into it.

Consequently, many structures started to emerge of proteins of unknown function. Indeed, about a third of the SG structural models were of proteins of unknown or uncertain function (Lee et al. 2011). This rather limited their usefulness. No longer did the models explain how the protein's function is achieved as the protein's function was not in fact known.

Although the thrust of the SG projects has largely shifted away from solving structures merely to plug missing gaps in structural knowledge, the prediction of protein function from structure remains of interest for annotation purposes. A recent review (Nadzirin and Firdaus-Raih 2012) found that 2549 non-redundant entries in the PDB were categorized as of "unknown function", although in truth only 1084 (42.5%) of them were genuinely unknown, the remainder being cases where their PDB entries had not been updated with newer functional information.

So, to what extent can function be determined from protein structure? The history of structural biology demonstrates that 3D structure can explain function, in terms of a protein's interactions, catalytic residues, trans-membrane regions, etc. Indeed, virtually every structure solved before the advent of SG helped explain some biological or biochemical process. So, given a structure, out should pop the function.

Sadly, few things in life—or in bioinformatics—are that simple. Structure may explain a function, but only if the function is known already. Despite the availability of the many diverse methods discussed earlier in this book, it is surprisingly difficult to determine the function from structure alone.

### *13.1.1 The Problem of Predicting Function from Structure*

Why is it difficult to get function from structure? Firstly, if one has a protein of unknown function it means that, not only is there no experimental information about its function, but also that the standard sequence methods for functional annotation have failed. These methods, particularly the various profile methods such as the Hidden Markov Model (HMM) methods, have become quite sophisticated in recent years and can detect similarity of function at quite low levels of sequence identity (Soding 2005). So if these methods have failed we really are relying on the 3D structure alone.

The structure can provide clues to function at various levels and in varying degrees of reliability as the preceding chapters have described. Chap. 9 showed how, at the global level, a protein's fold can very often give a clue to its function as some folds are strongly associated with certain functions. So the first step in identifying function from structure is invariably to find a protein of known function

with a similar fold. There are a large number of fold comparison servers on the web that will do this, and these have been compared in several reviews (Sierk and Pearson 2004; Novotny et al. 2004). However, you need to bear in mind that a similarity of fold does not necessarily imply a similarity of function. For example, the so-called *superfolds* (Orengo et al. 1994), such as the TIM-barrel family, can support large numbers of different functions (Nagano et al. 1999; Anantharaman et al. 2003). Indeed, even proteins with highly similar sequences can perform different functions if key functional residues have evolved for a different purpose. Furthermore, if the protein has a completely novel fold—formerly a successful outcome in the eyes of many SG projects—there will be no fold match at all.

More locally, the surface of the protein, particularly its clefts and pockets, can hold important clues to function (Chap. 10) as can specific local arrangements of residues, such as those involved in catalysis, DNA recognition, etc. (Chap. 11). So you may be able to identify, say, a possible ATP-binding site. This would be an important clue to function, but not the full story. Of course, the sensitivity of predictions from these so called "functional motifs" depends on how well the sequence or structure of these motifs are evidenced in sequence alignment or substructural clustering. Using structure-derived sequence signatures offers another way of enhancing the predictions and using newer motifs provides clues to function (Das et al. 2014).

Plus there are various spanners that can jam the works. Firstly, it is often difficult to solve the whole intact protein. In these cases the best one can get is a structural model for part of the protein—say, just a single domain. On its own this domain may say little about the protein's function. Secondly, even if the whole protein is solved, it may be just one component of a multi-protein complex. Again, the structure gives only part of the story. More dastardly still are the so-called moonlighting proteins which can actually have more than one function, depending on their context: cellular location, environment, and so on (Jeffery 1999, 2009). And some proteins can alter their function according to which alternatively spliced variant is expressed at any given time (Stamm et al. 2005).

Another problem with function prediction is the difficulty of assessing the success or failure of a given prediction method and, indeed, even defining what is meant by *function*. Function can be described at many levels, ranging from biochemical function through biological processes and pathways, all the way up to the organ or organism level (Shrager 2003). Consequently, a given protein may be annotated at several different levels of functional specificity: for example, *ubiquitin-like domain*, *signalling protein*, *predicted serine hydrolase*, *probable eukaryotic D-amino acid tRNA deacetylase*, and so on. Thus it is difficult to judge the accuracy of any such assignment, especially if the assignment is one of the more vague ones.

A common strategy for assessing function prediction methods is to use the Gene Ontology (GO) (Ashburner et al. 2000; Gene Ontology Consortium 2015). This is an open source scheme for functional annotation of protein sequences. It is a machine-readable ontology based on a controlled vocabulary of functional

descriptors and many function prediction methods couch their results in terms of GO codes. Although not strictly hierarchical, the GO functional descriptors range from the truly unspecific (e.g. enzyme) down to the highly precise (e.g. 1-pyrroline-4-hydroxy-2-carboxylate deaminase).

More recently, there has been an effort to move beyond static information (i.e. sequence and structure snapshots) by using long time-scale coarse grained structural dynamics to obtain clues to protein function (Bhadra and Pal 2014). Highly mobile protein segments identified from dynamics can be matched for dynamics-function correlation using a combination of inputs from root-mean-square fluctuations and auto-correlation vector profiles. The method is able to identify moonlighting functions of proteins and match function between proteins that have poor sequence identity.

### *13.1.2 Structure-Function Prediction Methods*

As the previous chapters show, there are very many different methods for predicting function from structure. Several reviews have described them and considered their usefulness (Kim et al. 2003; Watson et al. 2005; Rigden 2006; Lee et al. 2007). None of the methods is perfect and none can hope to be successful in all situations. For example, some methods are only suitable for enzymes—and so cannot help at all if the protein in question is not an enzyme. Other methods rely strongly on some match—whether of the fold, or a motif, or a binding site, etc.—to a protein of known structure. So if no match can be found, or the match is merely to another hypothetical protein, such a method effectively returns a blank.

Consequently, a sensible approach is to throw a large number of these predictive methods at the protein structure and see what drops out. The two servers described in this chapter do just that. They are ProKnow from UCLA at http://proknow.mbi.ucla.edu, and ProFunc from the European Bioinformatics Institute (EBI) at http://www.ebi.ac.uk/profunc. Both use a cocktail of sequence-based as well as structure-based predictions and are largely automated: the user uploads a PDB-format file and waits patiently for the results.

To illustrate the two methods we use as an example the structure of a putative acetyltransferase from *Vibrio cholerae,* solved in 2005 by the Midwest Center for Structural Genomics (MCSG). It was released by the PDB as entry 2fck on 28 February 2006 (Cuff et al. 2007). The function of this protein was only tentatively known at the time; its sequence had over 50% identity to a ribosomal-protein-serine acetyltransferase and contained several sequence motifs characteristic of acetyltransferase activity. Once its structure was known, these tentative functional assignments were greatly strengthened as it revealed strong structural similarities, both global and local, to other—distantly related—acetyltransferases. The strongest similarities occurred at the putative binding site where coenzyme A (coA) is likely to bind. Some of these similarities will be illustrated below.

## 13.2 ProKnow

The first of the two integrated servers is ProKnow (Pal and Eisenberg 2005; Medrano-Soto et al. 2008) at UCLA (http://proknow.mbi.ucla.edu). The current version of ProKnow runs six principal prediction methods on any uploaded 3D structure (Fig. 13.1). In fact, the server can also accept just a protein sequence; in which case, one of the six methods is dropped. The six features examined include the protein's overall fold, various 3D structural motifs (omitted for sequence-only submissions), sequence similarities, sequence motifs, and functional linkages from the Database of Interacting Proteins (DIP) and the Prolinks Database. Each method may provide one or more functional clues, with varying degrees of confidence. These clues are weighted using Bayes's theorem and combined to give the most likely overall function, expressed as GO terms and measures of confidence for each. A map showing the relationship between the top GO predictions is returned (Fig. 13.2), allowing the user to more confidently interpret the predictions. Also given are the detailed hits and their scores. The top results for our example structure, 2fck, are shown in Fig. 13.3. Here, essentially only one hit of significance was returned: N-acetyltransferase, which is very confidently predicted and agrees with the protein's putative function.

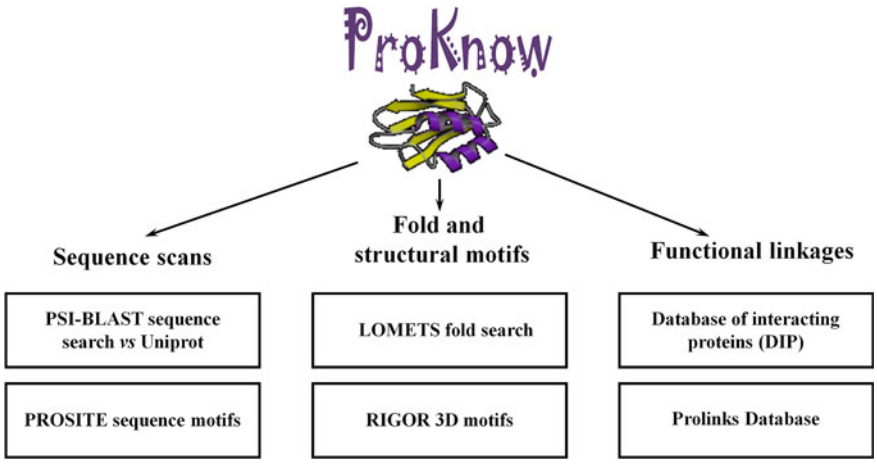



**Fig. 13.1** Schematic diagram of the sequence- and structure-based methods applied to any protein 3D structure submitted to the ProKnow function prediction server. The sequence-based methods are PSI-BLAST (Altschul et al. 1997) and PROSITE (Hulo et al. 2004). The structure-based methods are the LOMETS fold search (Wu and Zhang 2007) and RIGOR structural motif search (Kleywegt 1999). The final two methods use DIP, the Database of Interacting Proteins (Xenarios et al. 2002), and the Prolinks Database (Bowers et al. 2004) to identify any interesting functional linkages for each of the PSI-BLAST hits. The Gene Ontology (GO) functional annotations are obtained from all the results and combined using Bayesian weighting to arrive at a set of functional prediction and associated reliability estimates

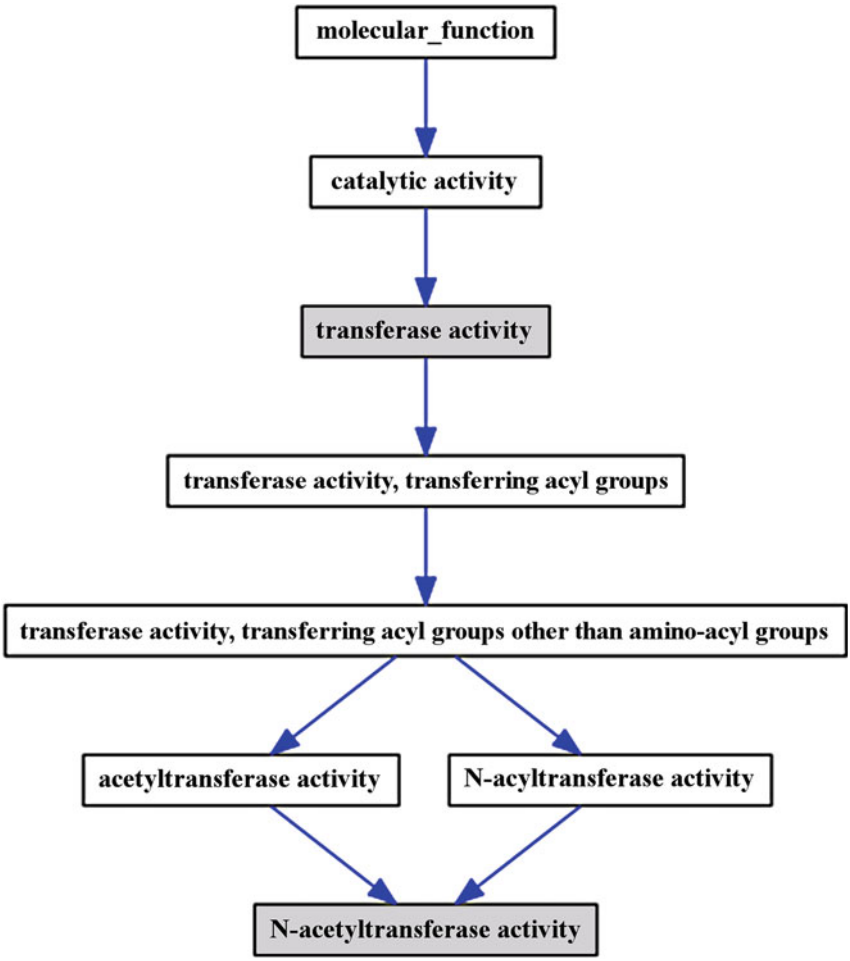



**Fig. 13.2** Gene Ontology map generated for PDB entry 2fck showing the hierarchy of functional terms, from the general to the specific. Where ProKnow identifies more than one functional prediction the map returned will show a network of possibilities, each linked to any others that are similar, with the connections colour-coded by the similarity of the terms

### *13.2.1 Fold Matching*

The first stage in ProKnow is the identification of other protein structures having the same, or most similar, fold to that of the query protein. In fact, this part is not automated and requires the user to first run the Dali fold-matching program (Holm and Sander 1998) and then upload the results, in FSSP format, to ProKnow. The matches obtained from Dali provide the first set of clues used by ProKnow about the protein's function.

**(a)**

**Prediction Result Summary**

| Type | GO Term | Evidence | Rank | Clues | Description |
|---|---|---|---|---|---|
| Function | 0008080 | 0.5340 | 4.3 | 6 | "Catalysis of the transfer of an acetyl group to a nitrogen atom on the acceptor molecule." [GOC:ai] |
| Function | 0016740 | 0.4660 | 4.3 | 6 | "Catalysis of the transfer of a group, e.g. a methyl group, glycosyl group, acyl group,... |
| Process | 0008152 | 1.0000 | 4.3 | 6 | "The chemical reactions and pathways, including anabolism and catabolism, by which living... |

**(b)**

**Prediction Results**

**Prediction results for your job are below.**
Questions how to interpret these results? Click here for information.

Function & Process | Gene Ontology Graphs | Clue Details

| | Clues | | | | | | | | | | Evidence | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | BLAST | | RIGOR | | LOMETS/DALI | | DIP | | PROSITE | PROLINKS | BLAST | | RIGOR | | LOMETS/DALI | | DIP | | PROSITE | PROLINKS |
| GO Code | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 0008080 | 57 | 0 | 2 | 4438 | 36 | 567 | 0 | 0 | 0 | 0 | 6.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 0.00 | 0.00 | 0.00 |
| 0016740 | 47 | 0 | 2 | 4438 | 36 | 600 | 0 | 0 | 0 | 0 | 6.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 0.00 | 0.00 | 0.00 |
| 0008152 | 57 | 0 | 2 | 4438 | 36 | 343 | 0 | 0 | 0 | 0 | 6.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 5.00 | 0.00 | 0.00 | 0.00 |

**Fig. 13.3** **a** The top ProKnow functional predictions for PDB entry 2fck. The top hit predicts, with high confidence, the protein to have N-acetyltransferase activity. **b** Master table of clues used in each GO term prediction for 2fck. Clicking on any of the numbers in the table shows the details for the given clue

Interestingly, if only the sequence is submitted, then ProKnow does all the work itself: it identifies a fold compatible with the sequence and uses that for clues about function. The potential matching folds are identified using LOMETS (Wu and Zhang 2007), which itself is a metaserver that runs nine different algorithms, namely: FUGUE (Shi et al. 2001), HHSEARCH (Soding et al. 2005), PROSPECT2 (Xu and Xu 2000), SAM-T02 (Karplus et al. 2003), SPARKS2 (Zhou and Zhou 2004), SP3 (Zhou and Zhou 2005), and PAINT, PPA-I, PPA-II (Wu and Zhang 2007). Each of the algorithms are run locally on the server and their predicted templates are used as input for clues to function.

One thing that needs to be remembered when relying on the results of any fold-recognition, or *threading*, method is that these methods are something of a Black Art, and require careful interpretation. Occasionally, they can give approximately the right answer—usually for small, single-domain proteins where a topologically near-correct model is obtained (Moult 2005); but, in general, accuracy varies widely. If the sequence is a very long one the chances of success are even smaller as the protein almost certainly comprises several structural domains, the boundaries of which would ideally be manually identified. Each domain's fold has then to be recognized. Even if these stages are successful, the 3D arrangement of the domains may be crucial for the protein's function, and although methods for predicting domain packing exist (Xu et al. 2014) they will not succeed in all cases. Nevertheless, using more information for arriving at any answer has a higher likelihood to succeed compared to a single approach.

### 13.2.2 3D Motifs

After the fold-matching stage, the protein's 3D structure is scanned against the RIGOR database of automatically generated 3D motifs (Kleywegt 1999). The motifs consist of 'interesting' arrangement of residues. RIGOR has three rules for distinguishing interesting arrangements from uninteresting ones: (*a*) the protein contains *n* sequential residues of the same type (*e.g.* four consecutive arginine residues), (*b*) a set of neighbouring residues are all hydrophobic, or all polar/charged, or a mixture of hydrophobic and polar/charged, and (*c*) residues that all make contact to a single hetero compound. ProKnow uses over 10,000 RIGOR motifs; associated with each are the GO terms of the corresponding protein chain.

### 13.2.3 Sequence Homology

The PSI-BLAST program (Altschul et al. 1997) is used to scan the UniProtKB sequence database (UniProt Consortium 2014) for proteins homologous to the target protein. Any matches with GO annotations add their functional clues to the pot.

### 13.2.4 Sequence Motifs

The target protein's sequence is then scanned for sequence motifs using the PROSITE database of functionally-associated motifs (Sigrist et al. 2012). Again, each motif has a set of associated GO codes.

### 13.2.5 Protein Interactions

The final set of features extracted by ProKnow relate to protein-protein interactions taken from the Database of Interacting Proteins, DIP (Salwinski et al. 2004), and functional annotations from the Prolinks Database (Bowers et al. 2004). Any sequence matched by the PSI-BLAST search can return a functional linkage if present in either DIP or Prolinks.

### 13.2.6 Combining the Predictions

Once all processes have completed, the functions (i.e. GO terms) associated with any extracted features that reoccur are combined using Bayes's Theorem weighting. This provides an estimate of the significance of any predicted GO term. Only terms relating to molecular function and biological process are considered—i.e. terms relating to cellular component are ignored. The significance of any predicted GO term is reflected by three numbers. The first is the Bayesian weight which represents the probability, 0.0–1.0, of the predicted GO term being correct. The second is the evidence rank and relates to how reliable a particular GO assignment is deemed to be in the first place. GO assignments come from various sources: inferred by the curator, inferred from direct assay, inferred from sequence or structural similarity, and so on. These have a range of reliabilities, the most reliable being any that have direct experimental evidence to support them. The source of the annotation is recorded by the *evidence code* in the GO data. In ProKnow, each type of evidence code is assigned a *rank* to quantify its reliability, and the ranks from several predictions are averaged to give the evidence rank. The third measure of significance is the clue count which is the number of weights used to calculate the Bayesian weight and is related to how many of the ProKnow sequence and structure methods contributed to a given GO prediction.

### 13.2.7 Prediction Success

Figure 13.3 shows some of the output on our example structure, 2fck. Figure 13.3a lists the top GO codes returned by the predictions, with the code for N-acetyltransferase (0008080) on top. Figure 13.3b shows the 'clues' (i.e. methods) that gave rise to these predictions and the 'evidence score' for each. In this example, the DIP, PROSITE and Prolinks searches returned nothing of significance. So the strongest prediction was the one that appears to be the correct one; namely, that the protein is an acetyltransferase.

In general, ProKnow performs quite well. Its authors tested it on a non-redundant data set of proteins of known function and found that around 70% of the functional annotations were correct (Pal and Eisenberg 2005). Less specific predictions (e.g. hydrolase) tended to be more accurate than more specific ones (e.g. leucyl aminopeptidase). The prediction accuracy has been increased slightly by the recent inclusion of Prolinks, not present in the original version, and should improve more as the coverage of Prolinks increases.

## 13.3 ProFunc

The second integrated server described here is ProFunc (Laskowski et al. 2005a) at the European Bioinformatics Institute (EBI), http://www.ebi.ac.uk/profunc developed as part of a collaboration with the Midwest Center for Structural Genomics (MCSG) . ProFunc allows the user to either upload a protein structural model or to enter the PDB code of a structure already in the Protein Data Bank. In the latter case, if ProFunc has already been run on that PDB entry, the results will be displayed immediately.

When ProFunc runs it applies a number of sequence- and structure-based methods to the structure, as shown in Fig. 13.4. A processor farm is used, with different methods sent to different processors. Several of the compute intensive

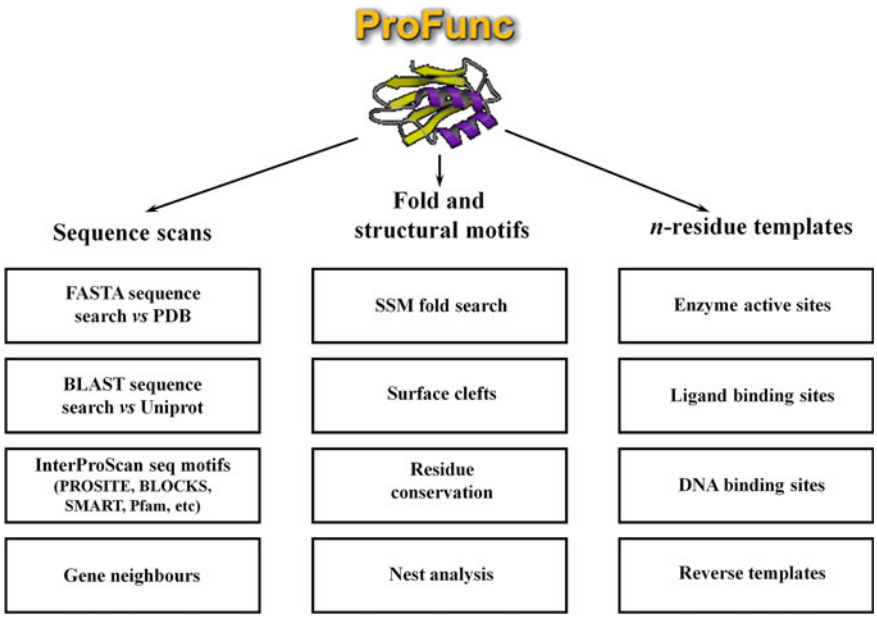



**Fig. 13.4** Schematic diagram of the sequence- and structure-based methods used in ProFunc. The sequence-scans in the *left-hand* column include searches against the protein sequences in the PDB and UniProt databases. The InterProScan search returns any sequence motifs present in the query protein's sequence. For each UniProt BLAST hit the Gene Neighbours search locates the gene in its genome, if available, and identifies all neighbouring genes. The first of the structure-based searches in the middle column uses SSM to identify structures with the most similar overall fold to that of the query protein. Surface clefts are computed and can be visualized coloured by residue type or residue conservation. Nests, which are often found at functionally important locations, are identified, and finally, in the *right*-hand column, are the various template methods that find local 3D matches to known protein structures

method are themselves subdivided to run in parallel on multiple processors. Processing is usually complete within about an hour.

The results of each method are then summarized, with further details available for each method. However, the results are not combined in any sophisticated way, as is done in ProKnow. Rather, there is a summary at the top of the results page showing the most commonly occurring GO terms and protein names, but this is meant only as a quick guide. The primary aim of the server is to present the results in an easily accessible format to enable researchers to interpret them, using their own expertise and knowledge of the protein in question.

Now, although ProFunc does apply a number of sequence-based methods, using several well-known search techniques such as FASTA and InterProScan (Jones et al. 2014), we will only describe the structure-based methods here as most of them are unique to this server.

## *13.3.1 ProFunc's Structure-Based Methods*

### 13.3.1.1 Fold-Matching

The first of the structure-based methods is a search for proteins with similar fold as the target. The PDBeFold program is used, based on the Secondary Structure Matching program, SSM (Krissinel and Henrick 2004). It performs a fast graph-matching procedure to compare the secondary structure elements (SSEs) of the target structure against those of the structures in the database. Any strong matches are superposed and an r.m.s.d. for equivalent Cαs is calculated together with a *z*-score measure of significance and SSM's own significance measure, called the *Q*-score. In ProFunc the top ten hits, ordered by *Q*-score, are shown and any, or all, can be viewed superposed on the target structure using the molecular graphics viewer RasMol (Sayle and Milner-White 1995).

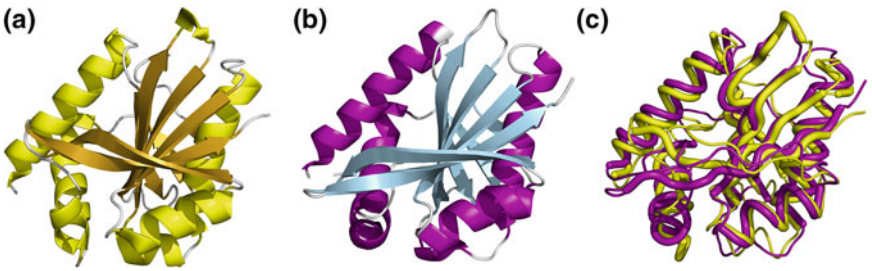



**Fig. 13.5** The closest fold to that of 2fck, found by the SSM's fold-matching program, is PDB entry 1z9u, an N-acetyltransferase from *Salmonella typhimurium*. **a** Overall 3D structure of 2fck and **b** overall structure of 1z9u in the same orientation. **c** The two structure superposed and each shown as a $C^{\alpha}$ trace, 2fck in *yellow* and 1z9u in *purple*. The matched regions are shown using a thicker representation of the trace in each structure

The top fold match to our example structure, 2fck, is shown in Fig. 13.5. The match is to PDB entry 1z9u, an N-acetyltransferase from *Salmonella typhimurium*. The protein forms a homodimer with a large trough at the dimer interface where the substrate binds. The protein also binds coenzyme A (coA).

#### 13.3.1.2 Surface Clefts

The second method computes all the clefts in the protein's surface, using the SURFNET program (Laskowski 1995). The clefts are ranked in order of size and can be viewed in RasMol. The viewing options allow for the cleft surfaces to be coloured by specific properties, such as cleft size, residue type or residue conservation score. Cleft size is important as the largest cleft in a protein's surface tends to be where the protein's active site is located (Laskowski et al. 1996). Also important is residue conservation, as clusters of highly conserved residues, particularly if located in a large pocket, are highly indicative of a functional site (Lichtarge and Sowa 2002; Madabushi et al. 2002; Glaser et al. 2003). Like nest analysis (below), study of the protein's surface clefts is of most use when the other methods have failed or have suggested only vague possibilities as it can identify the functionally important parts of the structure.

In our example structure, the largest cleft does indeed correspond to the protein's putative binding site, matching the location of the bound coA in the related structures identified from the fold-match above and the template methods to be described shortly.

#### 13.3.1.3 Nests

Next, any nest motifs in the structure are identified. These are frequently associated with functional sites. A *nest* is an anion or cation binding site formed by three or more amino acids in the sequence whose main-chain $\psi$-$\varphi$ dihedral angles alternate between the right- and left-handed helical ($\alpha$ and $\gamma$) regions of the Ramachandran plot (Watson and Milner-White 2002a, b). Again, a RasMol view shows the location of the nest in the context of the whole 3D structure. ProFunc assigns a score to each nest based on: the number of NH atoms that are accessible to solvent, the conservation scores of its constituent residues, and whether the nest occurs in one of the larger clefts on the surface.

The 2fck structure contains several nests, three of which score highly enough to indicate that they may be functionally significant. And indeed, the top-scoring nest is in the protein's likely substrate binding site (based on the similarity identified above to the 1z9u structure), while nests two and three are found at the entrance to the coA binding site.

#### 13.3.1.4 Template Methods

The final ProFunc methods involve four different types of residue template searches (Laskowski et al. 2005b). The templates are defined as specific 3D conformations of, typically, three amino acid residues. The template searches are carried out by a fast 3D search algorithm called Jess (Barker and Thornton 2003), performed in parallel on the processor farm.

Enzyme templates

The first group of templates are the enzyme active site templates which come from the manually compiled Catalytic Site Atlas, CSA (Porter et al. 2004). Here each template consists of two to five residues that are identified in the literature as being catalytic or are highly-conserved residues in the neighbourhood of the catalytic residues. A strong match (see below) to one of these templates can be highly suggestive of the protein's function.

Ligand- and DNA-binding templates

The next two groups of templates are the ligand- and DNA-binding templates. These are automatically generated once a week so as to be as up-to-date as the PDB. The ligand templates are generated by considering in turn every type of Het Group (as defined in the PDB's Het Group Dictionary) and retrieving a non-homologous list of structures in the PDB that contain this Het Group. Residues interacting with the Het Group in each selected structure are marked. Templates are compiled by selecting groups of 3 residues from each structure's marked residues. The selection criteria governing which groups of 3 residues can form a template are as follows: each of the residues must be within 5 Å of one of the others in the template, each template can have at most one hydrophobic residue (i.e. Ala, Phe, Ile, Leu, Met, Pro or Val) to bias the templates towards surface residues, and no two templates from the same structure can have more than one residue in common. The order in which the potential templates are considered is biased by their relative importance. Thus a template containing residues that make several hydrogen bonds to the given Het Group are more highly valued than those whose residues only interact with the Het Group via a small number of non-bonded contacts.

The DNA-binding templates are generated in exactly the same way except that all DNA and RNA molecules are treated as though they were a single Het Group. As of February 2017, there were 584 CSA templates, 94,055 ligand-binding templates and 5320 DNA-binding templates in these template databases.

Figure 13.6 shows a template match in 2fck to a coA ligand-binding template taken from PDB entry 1s7 l, a RimL N(α)-acetyltransferase from *Salmonella typhimurium*.

Reverse templates

The fourth group of templates are intended to find any matches that the first three sets may have missed. They are the *reverse templates* and are computed from the target structure itself. The rules for generating them are the same as for the ligand- and DNA-binding templates except that, firstly, the whole protein structure is considered, rather than merely the residues in contact with ligand or DNA, and secondly the weighting of each of the templates is by residue conservation (as

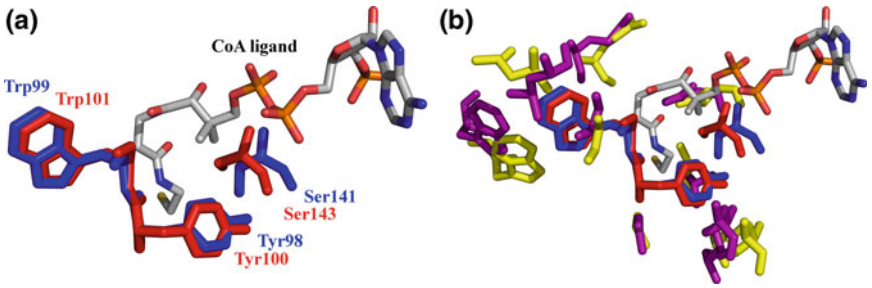



**Fig. 13.6** A match from ProFunc to a ligand-binding template for coenzyme A (CoA). The CoA is shown in the thinner sticks and is coloured by atom type (carbon *grey*, nitrogen *blue*, oxygen *red*, sulphur *yellow* and phosphorus *orange*). **a** The template is defined by the three *red* residues: Tyr100, Trp101 and Ser143 from PDB entry 1s7l, a RimL N(α)-acetyltransferase from *Salmonella typhimurium*. The three *blue* residues correspond to the residues in the query structure, PDB entry 2fck, that match the template residues. They are: Tyr98, Trp99 and Ser141, respectively. The rmsd of the 26 matched side chain atoms is 0.72 Å. **b** As in a, but with additional matching residues lying within 10 Å of the template's centre shown. These are residues of identical residue type which overlap when the query and template structures are superposed on the basis of the template match in **a**. The *purple* residues are from the template structure (1s7l) while the *yellow* ones are from the query structure (2fck)

obtained from a multiple alignment of the sequences returned by a BLAST search against the UniProtKB sequence database). The templates are selected so that, ideally, each residue in the protein is present in at least one template, although, if too many are generated, their number is capped at twice the number of residues in the sequence.

The top reverse template hit to 2fck is shown in Fig. 13.7. The match is to PDB entry 3r9f, a microcin c7 self-immunity acetyltransferase from *E. coli* (Agarwal et al. 2011).

Template searching and scoring

The template searches can return hundreds, thousands or even tens of thousands of matches, particularly in the case of the reverse templates. The problem, then, is to discard fortuitous matches and retain only significant matches, ranked in order of relevance. ProFunc does this by comparing the environment around the template residues in their parent structure with the environment around the residues that were matched. Residues within 10 Å of the template's geometrical centre in both structures are paired off according to their degree of similarity and overlap. Where alternative pairings are possible an optimization procedure is applied to maximize the numbers of paired identical or similar residues in equivalent 3D positions. The number of paired residues gives a crude measure of the local similarity of the matched sites in the two proteins (Figs. 13.6b and 13.7b). However, this crude measure still lets through too many false positives. Therefore the measure that is actually used takes into account the relative positions of the paired residues in their respective amino acid sequences. If the sets of paired residues appear in the same

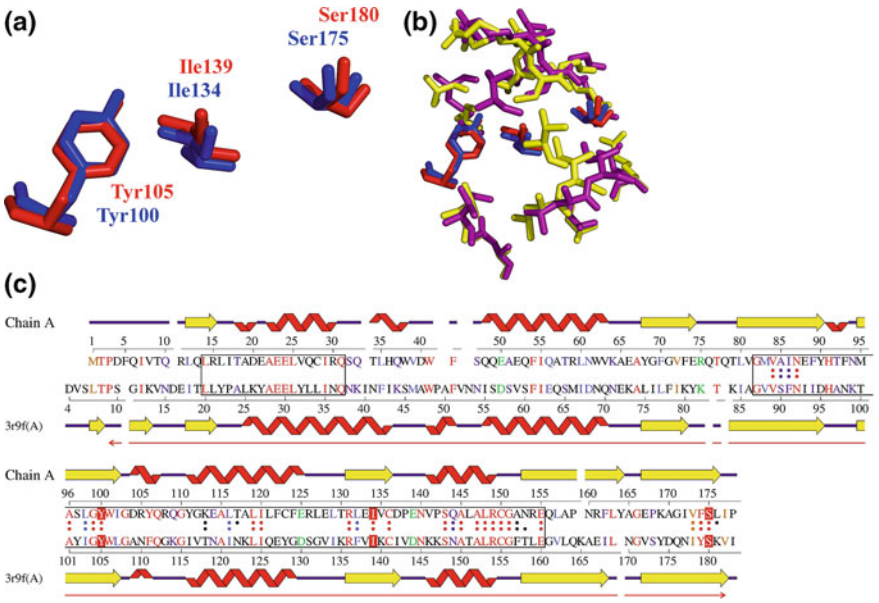



**Fig. 13.7** A reverse template match between 2fck and 3r9f, a microcin c7 self-immunity acetyltransferase from *E. coli*. **a** The *blue* are the template residues from 2fck (Tyr100, Ile134 and Ser175) which match to the *red* residues (Tyr105, Ile 139 and Ser180, respectively) in 3r9f with an rmsd of 0.55 Å for 18 matched atoms. **b** The equivalent residues of identical type within 10 Å of the template centre, *yellow* from 2fck and purple from 3r9f. There are 20 residues in all (out of 43 within 10 Å) giving a local similarity of 46.5%. A further 10 superposed residues (not shown) are of similar type (e.g. Ile for Val). **c** Sequence alignment obtained from the structural superposition. The *top row* shows the secondary structure of 2fck and the *bottom* shows that of 3r9f; any helices are represented by the jagged elements and β-strands by *arrows*. The three highlighted residues in the sequence alignment correspond to the template residues. Double *dots* between the two sequences identify the residues contained within the 10 Å sphere centred on the template and hence show which residues were used to drive the alignment. The boxed regions represent segments of the alignment where the sequence identity of the two sequences exceeds 35%. The long thin arrows at the bottom show structurally "fittable" regions; that is, segments from both proteins whose Cα atoms can be structurally superposed with an rmsd of less than 3.0 Å

order in both sequences then the likelihood of the sequences being homologues is high.

To see why this is so, consider two sequences descended from a common ancestor protein which have diverged so much that their relationship cannot be detected by sequence methods. However, if both have retained the same function, then the region that will have changed least is likely to be the active site as any significant changes here might have altered the function. The net result will be that the highest level of similarity between the two proteins will be among the residues in the vicinity of the active site. These residues will be close in 3D, but may be scattered along the lengths of the two sequences. That is why the similarity can be

detected in 3D, but may be virtually impossible to pick up from comparison of the sequences.

Figure 13.7c provides an illustration of this. It shows a sequence alignment between 2fck and its top reverse template hit, 3r9f. The alignment has been driven by the residues determined to be equivalent in the local matching procedure described above. The residues are marked by the double dots between the sequences. (The single dots correspond to residues that have lost their 3D-equivalent partners in the alignment). One can see that the paired residues, which lie in a compact region in 3D, are spread far apart in both sequences.

More interestingly, while the whole alignment gives a sequence identity of 25.9% between the two proteins, 20 of the 43 residues within 10 Å of the template centre are identical, giving a local sequence identity of 46.5%. As this region corresponds to a significant part of the coA binding site in the 3r9f structure it provides strong structural evidence that 2fck also binds coA. It also covers part of the putative substrate binding site, but not enough to suggest the substrates of both proteins are the same nor, indeed, that they perform the same function.

In addition to the local similarity score, various other statistics are quoted by ProFunc. One of these is an estimated *E*-value associated with the score. For the reverse templates these are calculated from the distribution of all scores obtained in a given search using the same procedure that FASTA uses for computing its *E*-values (Pearson 1998). For the other template searches the *E*-values are calculated using pre-computed parameters. The hits are ranked by *E*-value and categorized into four groups: certain matches ($E < 10^{-6}$), probable matches ($10^{-6} < E < 0.01$), possible matches ($0.01 < E < 0.1$) and long shots ($0.1 < E < 10.0$).

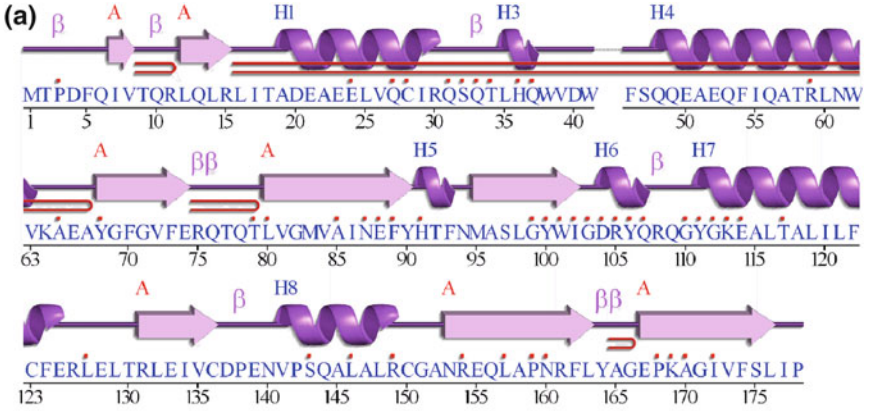
Also quoted is the overall structural similarity of the structures and the longest stretch of the two sequences that superposes with an rmsd of 3.0 Å on the Cα atoms. This latter can be particularly revealing when there is a long overlap, suggesting a significant structural match, even for the long shot cases.

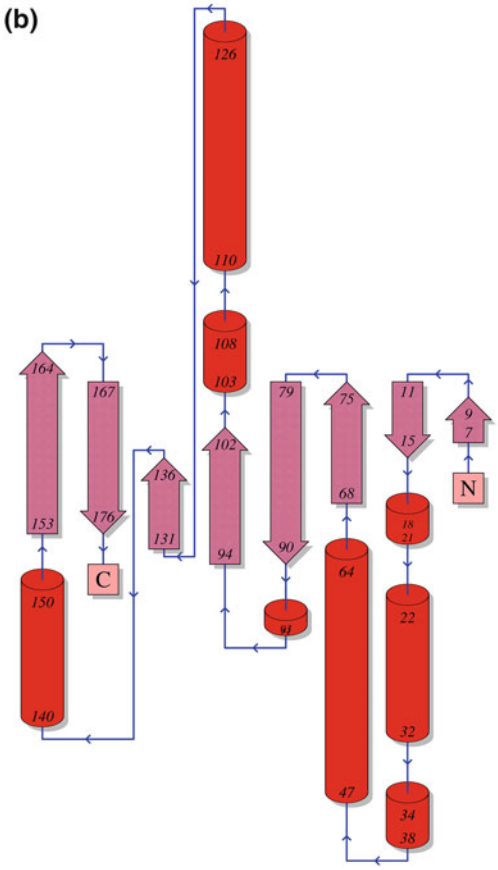
#### 13.3.1.5 PDBsum Structural Analyses

Although not strictly relevant to function prediction, a useful side effect of submitting a structure to ProFunc is that a set of PDBsum pages are also generated for it. PDBsum is a largely pictorial protein structure atlas at http://www.ebi.ac.uk/pdbsum that performs a number of structural analyses on the submitted protein and illustrates the results using various schematic diagrams (de Beer et al. 2014). A couple of examples are given in Fig. 13.8.

### *13.3.2 Assessment of the Structural Methods*

How good are the structural methods at predicting protein function? Attempts to answer this question are described in Chap. 14. Two studies (Watson et al. 2007;


(a)
β A β A H1 β H3 H4
MTPDFQIVTQRLQLRLITADEAEELVQCIRQSQTLHQWVDW FSQQEAEQFIQATRLNW
1 5 10 15 20 25 30 35 40 50 55 60
A ββ A H5 H6 β H7
VKAEAYGFGVFERQTQTLVGMVAINEFYHTFNMASLGYWIGDRYQRQGYGKEALTALILF
63 70 75 80 85 90 95 100 105 110 115 120
A β H8 A ββ A
CFERLELTRLEIVCDPENVPSQALALRCGANREQLAPNRFLYAGEPKAGIVFSLIP
123 130 135 140 145 150 155 160 165 170 175


(b)
126
110
108
103
102
94
164
153
167
176
C
136
131
150
140
79
90
91
75
68
64
47
11
15
18
21
22
32
34
38
9
7
N

◂**Fig. 13.8** Example analyses from the PDBsum pages generated when any structure is submitted to ProFunc. **a** A schematic diagram of the protein chain showing the protein's secondary structure elements (α-helices and β–sheets) together with various structural motifs such as β– and γ–turns, and β-hairpins. In this example residues interacting with bound ligands are indicated by the *red dots* above the single-letter amino acid code. In the 2fck structure the ligands are not particularly interesting or functionally informative, deriving from elements of the crystallization solution, and comprise 12 nitrate ions and one molecule of glycerol. **b** Topology diagram of the 2fck protein chain. The diagram illustrates how the β-strands, represented by the large *pink arrows*, join up, side-by-side, to form the domain's central β-sheet. The diagram also shows the relative locations of the α-helices, here represented by the *red cylinders*. The small blue arrows indicate the directionality of the protein chain, from the N- to the C-terminus. The numbers within the secondary structural elements correspond to the residue numbering given in the PDB file. The diagram is generated from the output of the Hera program (Hutchinson and Thornton 1990)

Lee et al. 2011) showed that the most successful of the structure based methods were the SSM fold comparison method and the reverse templates. Both had a success rate of 50–60%.

Of course, the only true way of validating a prediction is to confirm it experimentally. This is difficult, time-consuming and expensive although some progress has been made towards development of high-throughput functional assays (Yakunin et al. 2004; Proudfoot et al. 2008).

## 13.4 Conclusion

Here we have described ProKnow and ProFunc, two integrated servers that use a combination of structure- and sequence-based matching methods to try to predict the function of a protein from an uploaded 3D structural model. In most cases, they are able to offer some suggestions about possible function, although these may sometimes be rather vague (e.g. DNA-binding activity). In other cases, however, all their methods draw a blank and they fail completely. The most challenging cases are structures belonging to uncharacterized protein families possessing a novel fold. So, all one may be left with is the knowledge that the structure has an interesting-looking cleft in its surface, lined by highly conserved residues, but with no hint of what might bind there. For cases such as these, new methods need to be developed and incorporated into the servers. Of most utility would be methods that can predict what a given protein's likely substrate is from an analysis of the structure alone. That is, the methods should not rely on a match to an existing structure as, for novel folds, there is by definition no such match. At present, such methods are very compute-intensive and tend to commence with some idea of class of substrate at least—e.g. (Hermann et al. 2006). So, for the time being, prediction of a protein's function will continue to rely on clever sleuthing and deduction.

**Acknowledgements** The author would like to thank Debnath Pal for help with ProKnow and for his useful comments on this chapter.

## References


Agarwal V, Metlitskaya A, Severinov K, Nair SK (2011) Structural basis for microcin C7 inactivation by the MccE acetyltransferase. J Biol Chem 286(24):21295–21303. doi:10.1074/jbc.M111.226282

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Anantharaman V, Aravind L, Koonin EV (2003) Emergence of diverse biochemical activities in evolutionarily conserved structural scaffolds of proteins. Curr Opin Chem Biol 7(1):12–20

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25–29

Barker JA, Thornton JM (2003) An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. Bioinformatics 19(13):1644–1649

Bhadra P, Pal D (2014) De novo inference of protein function from coarse-grained dynamics. Proteins 82(10):2443–2454. doi:10.1002/prot.24609

Blundell TL, Mizuguchi K (2000) Structural genomics: an overview. Prog Biophys Mol Biol 73 (5):289–295

Bowers PM, Pellegrini M, Thompson MJ, Fierro J, Yeates TO, Eisenberg D (2004) Prolinks: a database of protein functional linkages derived from coevolution. Genome Biol 5(5):R35. doi:10.1186/gb-2004-5-5-r35

Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7(Suppl):932–934. doi:10.1038/80697

Chandonia JM, Brenner SE (2006) The impact of structural genomics: expectations and outcomes. Science 311(5759):347–351. doi:10.1126/science.1121018

Cuff ME, Li H, Moy S, Watson J, Cipriani A, Joachimiak A (2007) Crystal structure of an acetyltransferase protein from Vibrio cholerae strain N16961. Proteins 69(2):422–427. doi:10.1002/prot.21417

Das S, Ramakumar S, Pal D (2014) Identifying functionally important cis-peptide containing segments in proteins and their utility in molecular function annotation. FEBS J 281(24):5602–5621. doi:10.1111/febs.13100

de Beer TA, Berka K, Thornton JM, Laskowski RA (2014) PDBsum additions. Nucleic Acids Res 42 (Database issue):D292–296. doi:10.1093/nar/gkt940

Gene Ontology Consortium T (2015) Gene Ontology Consortium: going forward. Nucleic Acids Res 43 (Database issue): D1049–1056. doi:10.1093/nar/gku1179

Glaser F, Pupko T, Paz I, Bell RE, Bechor-Shental D, Martz E, Ben-Tal N (2003) ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic information. Bioinformatics 19(1):163–164

Hermann JC, Ghanem E, Li Y, Raushel FM, Irwin JJ, Shoichet BK (2006) Predicting substrates by docking high-energy intermediates to enzyme structures. J Am Chem Soc 128(49):15882–15891. doi:10.1021/ja065860f

Holm L, Sander C (1998) Touring protein fold space with Dali/FSSP. Nucleic Acids Res 26 (1):316–319

Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, Gattiker A, De Castro E, Bucher P, Bairoch A (2004) Recent improvements to the PROSITE database. Nucleic Acids Res 32 (Database issue): D134–137. doi:10.1093/nar/gkh04432/suppl_1/D134

Hutchinson EG, Thornton JM (1990) HERA–a program to draw schematic diagrams of protein secondary structures. Proteins 8(3):203–212. doi:10.1002/prot.340080303

Jeffery CJ (1999) Moonlighting proteins. Trends Biochem Sci 24(1):8–11

Jeffery CJ (2009) Moonlighting proteins–an update. Mol BioSyst 5(4):345–350. doi:10.1039/b900658n

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240. doi:10.1093/bioinformatics/btu031

Karplus K, Karchin R, Draper J, Casper J, Mandel-Gutfreund Y, Diekhans M, Hughey R (2003) Combining local-structure, fold-recognition, and new fold methods for protein structure prediction. Proteins 53(Suppl 6):491–496. doi:10.1002/prot.10540

Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R (2003) Structure-based functional inference in structural genomics. J Struct Funct Genomics 4(2–3):129–135

Kleywegt GJ (1999) Recognition of spatial motifs in protein structures. J Mol Biol 285(4):1887–1897. doi:10.1006/jmbi.1998.2393

Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1):2256–2268. doi:10.1107/S0907444904026460

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13 (5):323–330, 307–328.

Laskowski RA, Luscombe NM, Swindells MB, Thornton JM (1996) Protein clefts in molecular recognition and function. Protein Sci 5(12):2438–2452. doi:10.1002/pro.5560051206

Laskowski RA, Watson JD, Thornton JM (2005a) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33 (Web Server issue):W89–93. doi:10.1093/nar/gki41433/suppl_2/W89

Laskowski RA, Watson JD, Thornton JM (2005b) Protein function prediction using local 3D templates. J Mol Biol 351(3):614–626. doi:10.1016/j.jmb.2005.05.067

Lee D, de Beer TA, Laskowski RA, Thornton JM, Orengo CA (2011) 1,000 structures and more from the MCSG. BMC Struct Biol 11:2. doi:10.1186/1472-6807-11-2

Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8(12):995–1005

Lichtarge O, Sowa ME (2002) Evolutionary predictions of binding surfaces and interactions. Curr Opin Struct Biol 12(1):21–27

Madabushi S, Yao H, Marsh M, Kristensen DM, Philippi A, Sowa ME, Lichtarge O (2002) Structural clusters of evolutionary trace residues are statistically significant and common in proteins. J Mol Biol 316(1):139–154. doi:10.1006/jmbi.2001.5327

Medrano-Soto A, Pal D, Eisenberg D (2008) Inferring molecular function: contributions from functional linkages. Trends Genet 24(12):587–590

Moult J (2005) A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15(3):285–289. doi:10.1016/j.sbi.2005.05.011

Nadzirin N, Firdaus-Raih M (2012) Proteins of unknown function in the Protein Data Bank (PDB): an inventory of true uncharacterized proteins and computational tools for their analysis. Int J Mol Sci 13(10):12761–12772. doi:10.3390/ijms131012761

Nagano N, Hutchinson EG, Thornton JM (1999) Barrel structures in proteins: automatic identification and classification including a sequence analysis of TIM barrels. Protein Sci 8 (10):2072–2084. doi:10.1110/ps.8.10.2072

Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15(12):1519–1522. doi:10.1016/j.str.2007.11.004

Novotny M, Madsen D, Kleywegt GJ (2004) Evaluation of protein fold comparison servers. Proteins 54(2):260–270. doi:10.1002/prot.10553

Orengo CA, Jones DT, Thornton JM (1994) Protein superfamilies and domain superfolds. Nature 372(6507):631–634. doi:10.1038/372631a0

Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13 (1):121–130. doi:10.1016/j.str.2004.10.015

Pearson WR (1998) Empirical statistical estimates for sequence similarity searches. J Mol Biol 276 (1):71–84. doi:10.1006/jmbi.1997.1525

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32 (Database issue):D129–133. doi:10.1093/nar/gkh02832/suppl_1/D129

Proudfoot M, Kuznetsova E, Sanders SA, Gonzalez CF, Brown G, Edwards AM, Arrowsmith CH, Yakunin AF (2008) High throughput screening of purified proteins for enzymatic activity. Methods Mol Biol 426:331–341. doi:10.1007/978-1-60327-058-8_21

Rigden DJ (2006) Understanding the cell in terms of structure and function: insights from structural genomics. Curr Opin Biotechnol 17(5):457–464. doi:10.1016/j.copbio.2006.07.004

Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 32 (Database issue):D449–451. doi:10.1093/nar/gkh08632/suppl_1/D449

Sayle RA, Milner-White EJ (1995) RASMOL: biomolecular graphics for all. Trends Biochem Sci 20(9):374

Shi J, Blundell TL, Mizuguchi K (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310 (1):243–257. doi:10.1006/jmbi.2001.4762

Shrager J (2003) The fiction of function. Bioinformatics 19(15):1934–1936

Sierk ML, Pearson WR (2004) Sensitivity and selectivity in protein structure comparison. Protein Sci 13(3):773–785. doi:10.1110/ps.0332850413/3/773

Sigrist CJ, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, Bougueleret L, Xenarios I (2012) New and continuing developments at PROSITE. Nucleic Acids Res 41 (Database issue):D344–347. doi:10.1093/nar/gks1067

Soding J (2005) Protein homology detection by HMM-HMM comparison. Bioinformatics 21 (7):951–960. doi:10.1093/bioinformatics/bti125

Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33 (Web Server issue):W244–248. doi:10.1093/nar/gki40833/suppl_2/W244

Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, Thanaraj TA, Soreq H (2005) Function of alternative splicing. Gene 344:1–20. doi:10.1016/j.gene.2004.10.022

UniProt Consortium T (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Res 42 (Database issue):D191–198. doi:10.1093/nar/gkt1140 gkt1140 [pii]

Watson JD, Laskowski RA, Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15(3):275–284. doi:10.1016/j.sbi.2005.04.003

Watson JD, Milner-White EJ (2002a) The conformations of polypeptide chains where the main-chain parts of successive residues are enantiomeric. Their occurrence in cation and anion-binding regions of proteins. J Mol Biol 315(2):183–191. doi:10.1006/jmbi.2001.5228

Watson JD, Milner-White EJ (2002b) A novel main-chain anion-binding site in proteins: the nest. A particular combination of phi, psi values in successive residues gives rise to anion-binding sites that occur commonly and are found often at functionally important regions. J Mol Biol 315(2):171–182. doi:10.1006/jmbi.2001.5227

Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM (2007) Towards fully automated structure-based function prediction in structural genomics: a case study. J Mol Biol 367(5):1511–1522. doi:10.1016/j.jmb.2007.01.063

Wu S, Zhang Y (2007) LOMETS: a local meta-threading-server for protein structure prediction. Nucleic Acids Res 35(10):3375–3382. doi:10.1093/nar/gkm251

Xenarios I, Salwinski L, Duan XJ, Higney P, Kim SM, Eisenberg D (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30(1):303–305

Xu D, Jaroszewski L, Li Z, Godzik A (2014) AIDA: ab initio domain assembly server. Nucleic Acids Res 42 (Web Server issue):W308–313. doi:10.1093/nar/gku369

Xu Y, Xu D (2000) Protein threading using PROSPECT: design and evaluation. Proteins 40 (3):343–354

Yakunin AF, Yee AA, Savchenko A, Edwards AM, Arrowsmith CH (2004) Structural proteomics: a tool for genome annotation. Curr Opin Chem Biol 8(1):42–48. doi:10.1016/j.cbpa.2003.12.003

Zhou H, Zhou Y (2004) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 55 (4):1005–1013. doi:10.1002/prot.20007

Zhou H, Zhou Y (2005) Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. Proteins 58(2):321–328. doi:10.1002/prot.20308

# Chapter 14
# Case Studies: Function Predictions of Structural Genomics Results

**James D. Watson, Roman A. Laskowski and Janet M. Thornton**

**Abstract** The various Structural Genomics initiatives around the globe succeeded in solving several thousand protein structures, many of which were novel folds or structures of biological interest. Nevertheless, because of the high-throughput strategies employed, a significant proportion of the proteins were of unknown function, and remain so to this day. A number of computational methods have been developed to help ascertain protein function from three dimensional structure, the approaches ranging from large scale fold comparison to highly specific residue template matching. Each has its own advantages and disadvantages. Here we look at various analyses conducted to assess function prediction from structure, with specific examples of some of the success stories.

## 14.1 Introduction

Genome sequencing projects around the globe have already provided enormous amounts of data on the genes that are essential to a number of organisms, and this information is expanding rapidly with the large-scale metagenomics projects currently under way (Reddy et al. 2014). By comparison, the amount of protein structure data available lags far behind. The main aim of the Structural Genomics projects in the early 2000s was to bridge this gap by solving, in a high-throughput manner, a large number of novel structures that could be used to model a larger number of sequences (Fox et al. 2008; Service 2005). A consequence of this approach was the deposition of large numbers of structures with little or no functional annotation (Watson et al. 2007; Nadzirin and Firdaus-Raih 2012). This was in direct contrast to traditional structural biology where the function of a protein is often known in advance and the purpose of solving its 3D structure is to identify the biochemical mechanisms and unique subtleties of its action.

J.D. Watson · R.A. Laskowski · J.M. Thornton (✉)
European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
e-mail: thornton@ebi.ac.uk



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_14

Determination of a protein's function experimentally is a highly resource intensive process. So, faced with a large number of structures of unknown function, a major challenge is the accurate and automatic prediction of their function. A variety of computational methods now exist which aim to do this, many of which have been discussed in detail in previous chapters, but they effectively fall into two major categories: those which are predominantly sequence-based and those which are structure-based.

Sequence analysis is usually the first step in predicting a protein's function as significant sequence similarity is still the most reliable way of inferring function. A number of studies have investigated this and have shown that homologous proteins sharing over 40% sequence identity are likely to have conserved function (Todd et al. 2001). However, care must be taken as there are a number of exceptions where almost identical proteins have been shown to have different functions (Gerlt and Babbitt 2001; Rost 2002; Tian and Skolnick 2003), and, conversely, where proteins with almost undetectable sequence similarity have retained the same function (Whisstock and Lesk 2003). The development of powerful and sensitive profile- and pattern-based methods has increased our ability to infer functional similarities through the detection of increasingly distant sequence relationships. Other methods developed to help gain functional clues involve looking at residue conservation, phylogenetic profiles (i.e. groups of proteins which are jointly present/absent in different genomes) and gene location (e.g. within an operon of functionally related proteins) (Gabaldon and Huynen 2004; Gabaldon 2008).

When the sequence provides few clues to function, or there are no detectable homologues in the databases, a protein's structure can often provide further insight. As elements of a protein's structure are often conserved for functional reasons, structure-based approaches can identify more distant relationships than methods based on sequence (Chothia and Lesk 1986). The methods which have been developed range from large scale fold (Krissinel and Henrick 2004; Holm and Rosenstrom 2010) and biological assembly comparisons (Krissinel and Henrick 2007) (see also Chap. 9), down through localised pockets and clefts (Laskowski 1995; Binkowski et al. 2004; Glaser et al. 2006) (also discussed in Chap. 10), to highly specific three-dimensional clusters of functional residues (Stark and Russell 2003; Laskowski et al. 2005b; Kristensen et al. 2008; Wu et al. 2008; Roy et al. 2012) (see Chap. 11).

No single method is 100% successful and therefore a more prudent approach is to use as many methods as possible to try to gain functional clues: the more independent methods that agree on the same putative function, the more likely it is to be a correct prediction. As a result, a number of servers have been developed that utilise a range of methods to try to predict function. Some of these resources, such as the ProKnow server (Pal and Eisenberg 2005), try to make an overall consensus prediction, whereas others, like the ProFunc server (Laskowski et al. 2005a), present the results of a variety of methods for the user to interpret with their expert insight (see Chap. 13). The question that arises, however, is how successful have all of the attempts to predict function from structure actually been? In this chapter we shall review the various analyses that have been made of structure-based function

prediction, and the difficulties encountered, with reference to case studies from structural genomics projects.

## 14.2 Function Prediction Case Studies

There have been a number of attempts to assess the effectiveness of structure-based function prediction using structural genomics targets.

### *14.2.1 Teichman et al. (2001)*

An early review of 16 hypothetical proteins of known structure and their functional assignment (Teichmann et al. 2001) provided some glimpses of the quality of functional assignments that can be made from structure. The structures, in conjunction with alignments of homologous sequences, were used to find surface cavities and grooves in which conserved residues indicated an active site. Using knowledge of any bound co-factors in the structures, together with available experimental data for the proteins in question or their related sequences, assessments were made as to the depth of functional information that could be obtained. For the proteins examined, detailed functional information was obtained for a quarter, some functional information could be obtained for half, and no functional information could be obtained for the remaining quarter.

### *14.2.2 Kim et al. (2003)*

An analysis of eight protein structures, some solved at the Berkeley Center for Structural Genomics, others with collaborators, showed how the 3D structures provided functional or evolutionary insights (Kim et al. 2003). The examples were classed into one of five categories:

1. **Remote homologues**. Here function was inferred from structural similarity that could not be observed through sequence. An example was MJ0882: fold similarity suggested this to be a putative methyltransferase, and this was later verified experimentally (Huang et al. 2002). Figure 14.1 shows the structural match between the two proteins.
2. **Proteins with unexpected bound ligands**. Here function was inferred by the chance binding of a substrate or cofactor in the crystal structure. The first example, MJ0577 from *Methanococcus jannaschii*, had a bound ATP—suggesting an ATP hydrolysis function. More detailed analysis of the binding pocket identified motifs commonly found in nucleotide-binding proteins.

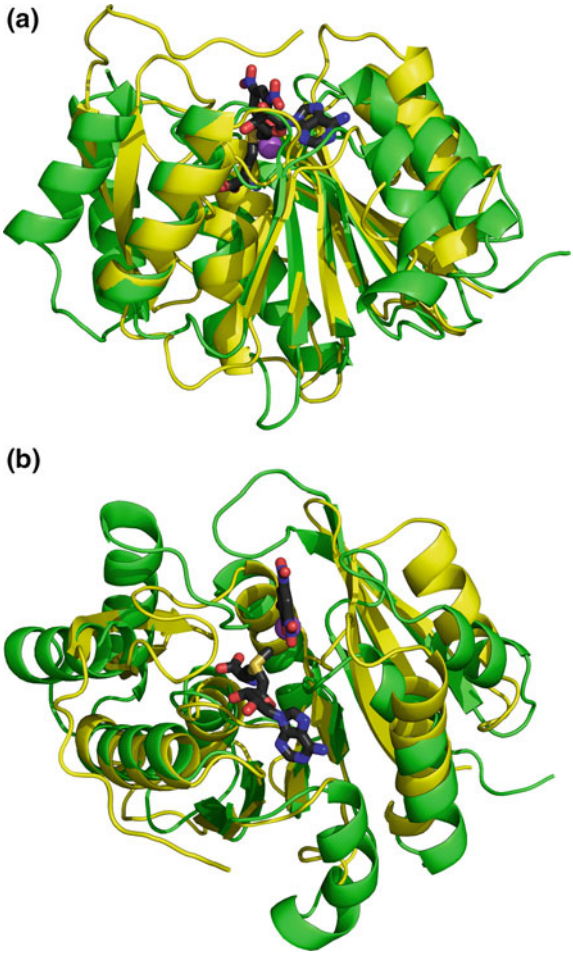



**Fig. 14.1** Structural superposition between a protein of unknown function, MJ0882 (PDB entry 1dus) in *green*, and a catechol O-methyltransferase (1vid) in *yellow*. The structures were superposed using PDBeFold, and fitted with a Cα r.m.s.d. of 2.57 Å over an alignment of 154 residues. The small molecules shown in *stick* representation are ligands from the 1vid structure and are S-adenosylmethionine and 3,5-dinitrocatechol. The 1vid structure also contained a bound magnesium ion, represented by the *purple sphere*. **a** Side view and **b** top view

The sequential arrangement of these motifs was unusual, which explained why motif-based methods had failed to detect them. Experimental assays confirmed an ATP hydrolysis function but only in the presence of cell extract, suggesting the protein is a molecular switch requiring one or more partner proteins to activate it.

The second example, TM841 from *Thermotoga maritima*, is a member of the large DegV Pfam family and belongs to the COG1307 group which has unknown function. The structure contained a bound palmitate molecule, suggesting a fatty acid binding function. Comparison with other members of the DegV and COG1307 families indicated a high degree of conservation where the head group of the palmitate was bound and, conversely, a higher variability where the tail was bound. This suggests that different members of the protein family bind different fatty acids with selectivity for tail length.

3. **"Twilight zone" proteins**. Here only weak matches were obtained for both the protein's sequence and structure. The structure of MJ0226 had a novel fold but showed weak similarity to nucleotide binding proteins (Hwang et al. 1999). Experimental analysis identified the biochemical function as a novel nucleotide triphosphatase. In conjunction with observed weak similarities to HAM1 protein (Noskov et al. 1996) the authors suggested a possible role in preventing mutations through removal of non-standard nucleotide triphosphates. This prediction was later confirmed through a complementation experiment (Stepchenkova et al. 2005).
4. **New molecular function for known cellular function**. Here the overall function of the protein was known but the biochemical details of its mechanism of action were revealed by the structure. In MJ0285, from *M. jannaschii*, the protein was annotated as being a small heat shock protein induced under cellular stress. The structure showed that 24 copies of the protein formed a hollow sphere with 8 triangular "windows" and 6 square "windows" (Kim et al. 1998). This raised the question of whether it works by trapping partially denatured proteins inside the sphere or by them becoming attached to its outside surface. Biochemical experiments strongly suggested the latter was the case, with the binding to the surface helping to prevent the proteins from aggregating and becoming inactivated.
   A second example was MPN625, a member of the OsmC domain family which exhibits a wide sequence variety but contains two highly conserved cysteine residues. The crystal structure revealed that these two cysteines lie in the cleft of a putative active site similar to that of the 2-cysteine peroxiredoxin family. This latter family inactivate reactive oxygen species (Schroder et al. 2000), which suggested that the OsmC family had a similar function, with the wide sequence variety being responsible for different substrate specificity. A subsequent experimental study demonstrated that OsmC proteins can indeed reduce both inorganic and organic peroxides and hence are involved in peroxide metabolism and protecting mycobacteria against oxidative stress (Saikolappan et al. 2011).
5. **Proteins where the function remains unknown**. Here the two examples quoted, Aq1575 from *Aquifex aeolicus* and MPN314 from *Mycoplasma pneumoniae*, are both hypothetical proteins which are members of Pfam domains of unknown function. In both cases there is evidence from conserved residues to suggest a putative active site, but searches of all the motif and functional databases failed to provide any clues as to their molecular function.

### *14.2.3 Watson et al. (2007)*

The first large-scale analysis of structure-to-function prediction (Watson et al. 2007) was performed to assess the effectiveness of the ProFunc server (see Chap. 13). First, all 319 structures solved by the Midwest Center for Structural Genomics

(MCSG) during the first stage of the NIH/NIGMS Protein Structure Initiative (PSI-1) were classified into those with known, putative or unknown function. Then the 93 proteins of known function were submitted to the ProFunc server and the top scoring matches from each of the server's sequence- and structure-based method were retrieved and stored. The results were then 'backdated' to each structure's deposition date by removing any more recent data. The aim was to get the same results as would have been obtained at the time of each structure's solution.

The top hit for each method was then manually compared with the known function and a judgement made as to whether the prediction was correct.

The results indicated that, of the ProFunc structure-based methods, the fold recognition and "reverse template" approaches were the most successful with approximately 60% of the known functions identified correctly. Both of these methods often identify the same function by matching to the same protein, but in some cases one method succeeded where the other failed. Differences occur because fold-matching looks for global similarity whereas the "reverse template" approach is a local comparison of residue locations.

A major drawback of the study was that it was unable to address the question of how structure-based approaches compare with sequence-based ones. This is a generic problem, not adequately addressed in the literature, caused by the difficulty of 'rolling back' to a particular date for the sequence databases and the databases of motifs, patterns and profiles derived from them. The results suggested that sequence methods can provide functional annotation in the majority of cases while structure can come to the rescue in a limited set of cases that are 'difficult', though not 'too difficult' (e.g. the protein has a completely novel fold).

An example is that of the BioH protein from *Escherichia coli* (Sanishvili et al. 2003) which was known to be involved in biotin synthesis but no biochemical function had been assigned to it. Its structure consisted of a Rossmann fold—a very common fold adopted by proteins performing a wide variety of functions. Indeed, fold comparison using Dali (Holm and Rosenstrom 2010) indicated structural similarity to many different enzymes: bromoperoxidase (EC 1.11.1.10), an aminopeptidase (EC 3.4.11.5), two epoxide hydrolases (EC 3.3.2.3), two haloalkane dehalogenases (EC 3.8.1.5), and a lyase (EC 4.2.1.39). The sequence identity of these hits was low, ranging from 15 to 25%. However, ProFunc found an enzyme template match for the Ser-His-Asp catalytic triad of the lipases, E.C.3.1.1.3 (with an rmsd of 0.28 Å). Experimental characterisation of the protein revealed it to be a novel carboxylesterase acting on short acyl chain substrates (Sanishvili et al. 2003). Figure 14.2 shows the enzyme template match.

A second example is hypothetical protein IsdG from *Staphylococcus aureus*. ProFunc's sequence-based methods revealed a variety of possible functions: antibiotic biosynthesis monooxygenase, cysteine peptidase, oxidoreductase, methyltransferase, epimerase, transportation, possible RNA binding, and others. The top fold matches found by SSM were all hypothetical proteins with no functional annotation. Further down the list, though, were a number of monooxygenases. Of the other structure-based methods, the reverse template scan returned a large number of matches, mostly to proteins of unknown function, but the first

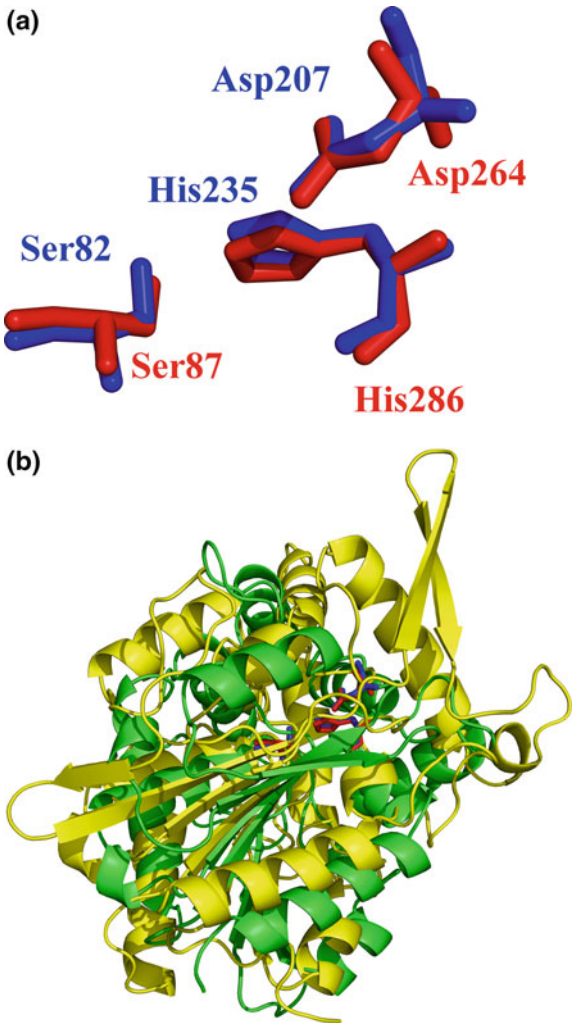



**Fig. 14.2** Enzyme template match for the BioH protein (PDB entry 1m33). **a** The match between the catalytic residues in *Pseudomonas* lipase (PDB entry 2lip), shown in *red*, and the three matching residues from BioH, in *blue*, **b** the corresponding superposition of the two structures showing their similar, yet far from identical, folds, with the BioH structure in *green* and 2lip in *yellow*

significant hit with an assigned function was to a monooxygenase from *Streptomyces coelicolor* (PDB entry 1lq9). Taken together these results were suggesting a monooxygenase function. Subsequent experimental analysis characterised the protein as a haem-degrading enzyme with structural similarity to monooxygenases (Wu et al. 2005). This example shows that structural knowledge can help tip the balance between several equally confident sequence-based predictions in favour of the correct functional assignment.

### 14.2.4 Lee et al. (2011)

A larger analysis (Lee et al. 2011) was carried out on 1165 structures, relating to 1118 protein targets solved by the MCSG during the first two stages of the Protein Structure Initiative (Norvell and Berg 2007). The proteins were first categorized according to whether their function was known (31%), putative (48%), possible (14%), or unknown (7%). The annotations were manually performed, using data from the Gene3D database (Lees et al. 2012) which contains all protein sequences in UniProtKB and most complete genomes. After running ProFunc on all the structures, the authors focused on the 78 (7%) that were of unknown function. ProFunc's reverse template method found a 'certain' match (i.e. *E*-value < $10^{-6}$) for one of these, and 'probable' matches ($10^{-6} \leq E < 10^{-2}$) for a further 17, showing that in cases where sequence methods cannot identify a protein's function, structural methods such as the reverse template method can help.

One example was for PDB entry 2aua, the structure of BC2332, an uncharacterized protein from *Bacillus cereus*. The strongest reverse template match was to a diphtheria toxin (PDB code 1f01), while two other matches gave an exotoxin (1xk9) and a cholix toxin (3ess). All matches were in the protein's known substrate binding site, strongly suggesting it, too, might function as a toxin; *Bacillus cereus* is known to produce enterotoxins that cause food poisoning (Granum and Lund 1997), so this could be another such toxin.

## 14.3 Some Specific Examples

Apart from the few large-scale studies described above, there have been many interesting analyses of individual proteins, or sets of proteins, where the 3D structure provided vital clues for some element of functional characterisation.

### 14.3.1 Adams et al. (2007)

In the first example, the functions of five hypothetical proteins from *E. coli* were deduced using a variety of methods, including structural information, operon prediction, related function from other operon members and catalytic residue conservation. Further information was gathered from co-crystallization trials and virtual ligand screening (Adams et al. 2007).

The first case involved a protein, ChuS, with a novel fold, thought to be involved in haem uptake and utilisation. Biochemical analysis suggested a haem oxygenase function—the first to be identified in *E. coli*. A multiple sequence alignment highlighted four conserved histidine residues which, when mapped onto the structure of ChuS, revealed that three are adjacent to, or point into, one of two large

clefts on opposite sides of the central core. Further structural studies with haem co-crystallization, and mutagenesis of the conserved histidines, identified a novel haem coordination, unlike any previously in haem degradation enzymes.

The second example, protein YgiN, showed a fold similarity to ActVA-Orf6, a monooxygenase from *S. coelicolor*. Members of this family are involved in the synthesis of large, polyketide compounds in the antibiotic biosynthetic pathways of Gram-positive bacteria. Their function is to tailor an antifungal compound, dihydrokalafungin, to confer its specific activity (Sciara et al. 2003). As *E. coli* was not known to produce this compound, it was expected that the natural substrate of YgiN would be different. Using information in the literature about previous studies, the authors were able to suggest that YgiN may be involved in menadione metabolism. Further experimental work resulted in the structure being co-crystallised with menadione.

In the third case, the protein YjjX showed a similar fold to a number of nucleotide binding proteins (including the MJ0226 protein—the third of the proteins discussed above from the paper by Kim et al. 2003). There was significant similarity in the active sites of these structural matches, with a number of residues either conserved or substituted with functionally equivalent residues. Biochemical analysis revealed YjjX to be a novel ITPase/XTPase that acts as a housekeeping enzyme in *E. coli* during oxidative stress to prevent the accumulation and subsequent incorporation into nucleic acids of non-canonical nucleotides.

The fourth case was of YhhW, a member of the cupin family. Its fold, as expected, had a similar core to known cupin structures, while the sequence pointed to its closest relatives being the pirins. A deep, charged pocket next to a metal binding site, also seen in one of its homologues, hPirin, suggested a possible active site. The pocket was found to be similar to that of quercetin 2,3-dioxygenase, and a biochemical assay confirmed quercetin 2,3-dioxygenase activity—the first enzymatic activity determined for any members of the pirin family. This example illustrates the problems faced when dealing with a large protein superfamily covering a diverse range of functions; a mere fold match is insufficient to identify a specific function, with local analysis required to pin it down.

The final example, z3393, was another member of the cupin superfamily. This time the sequence suggested a close relationship to the gentisate 1,2-dioxygenases, and this was supported by global structural and local molecular surface comparisons. The structure may help understand how the gentisate operon may be associated with pathogenic strains of *E. coli*.

### *14.3.2 AF0491 Protein*

The AF0491 protein from *A. fulgidus* was solved by the MCSG and deposited as PDB entry 1p9q (Savchenko et al. 2005). The protein is a homologue of the human Shwachman-Bodian-Diamond syndrome (SBDS) protein. SBDS is a rare autosomal recessive disorder caused by mutations in the SBDS gene on chromosome 7 and is

characterized by abnormal pancreatic exocrine function, skeletal defects, and haematological dysfunction (Boocock et al. 2003).

The structure revealed three domains (Fig. 14.3) although domain sequence databases indicate only two. The C-terminal domain is one found in many RNA- and DNA-binding proteins. The central domain has a winged helix-turn-helix (wHTH) fold that is commonly involved in DNA binding (Aravind et al. 2005) and has also been identified in RNA-binding proteins (Schade et al. 1999). However, in AF0491, the surface of this middle domain does not have a general basic character for DNA or RNA binding so is not expected to bind nucleic acids but be involved in protein-protein interactions instead.

The N-terminal domain was thought to be a novel fold and is where most of the disease-linked mutations are identified in SBDS patients. But it was noticed that the fold was similar to that of a protein that had been solved as one of the group's NMR structural genomics targets: yeast protein YHR087W. This protein was experimentally shown to be involved in RNA processing, although whether the same is true of the SBDS is not established. However, this example showed that the structural determination of a bacterial homologue of a human protein has identified additional homologues in yeast useful for experimental-based inference of function.

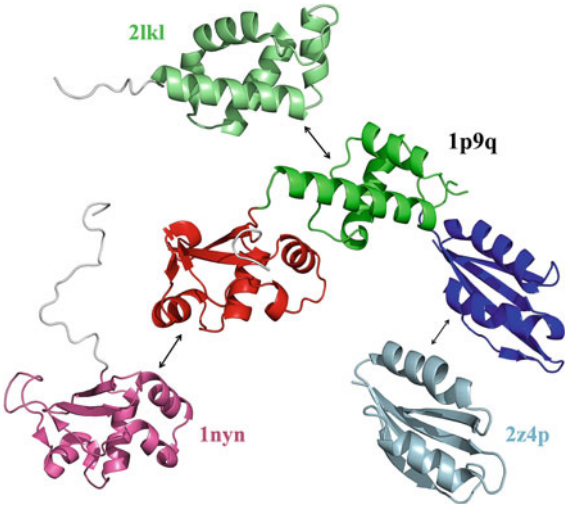



**Fig. 14.3** Monomer of AF0491 protein from *A. fulgidus*, a homologue of the human Shwachman-Bodian-Diamond syndrome (SBDS) protein clearly showing the protein consists of three structural domains. The domains are coloured *red*, *green* and *blue* in order from the N-terminus to the C-terminus. Matching domains from the PDB are coloured in pale versions of these *colours*. The C-terminal domain is commonly found in RNA- and DNA-binding proteins, the matching domain here being from feast/famine regulatory protein DM1 (PDB entry 2z4p). The central domain has a winged helix-turn-helix fold characteristic of DNA binding—as exemplified by PDB entry 2lkl, an erythrocyte membrane protein. The N-terminal domain's role is unknown, but could be involved in RNA processing. Its fold matches yeast protein YHR087W (PDB 2z4p)

### *14.3.3 The GxGYxYP Family*

The structure of BT2193 protein from *B. thetaiotaomicron* containing the GxGYxYP domain (PDB entry 3sgg) provided strong clues to this domain's function (Rigden et al. 2014). The domain is found in proteins that are over-represented in mammalian gut microbiomes, the *Bacteroides* genus making up to 30% of the microbiota. The domain occurs in Polysaccharide Utilization Loci (PULs) which code for different sets of enzymes and other proteins that can collectively digest a specific carbohydrate.

The BT2193 protein was found to contain two structural domains (Fig. 14.4a). The C-terminal domain, a 7-stranded beta barrel, exhibited many fold matches to other proteins in the PDB, yet no functional inferences could be drawn from these matches as the sequence identities were far too low. Structural alignments between BT2193 and its fold matches failed to identify any residues in BT2193 that might correspond to the catalytic residues in the PDB proteins. The N-terminal domain comprised 3 subdomains, structurally similar to one another, but not to any known structure.

So the domain structures did not point strongly at any function. However, the large cleft between the two domains held some functional clues. A 'nest motif', involving Asp331–Asp333, was identified within the cleft at a point of high sequence conservation. Additionally, a molecule of glycerol from the crystallization buffer was hydrogen-bonded to Asp333 and Glu272, suggesting a carbohydrate binding site. A search against catalytic residues from the Catalytic Site Atlas, CSA (Porter et al. 2004), identified that the three residues Asp333, Asp331 and Glu272 match the catalytic residues of a GH9 bacterial cellulase, PDB entry 1js4 (Fig. 14.4c). The residues are highly conserved, although have changed in some species, possibly indicating a loss of catalytic function there. Since the folds of the GxGYxYP domain (Fig. 14.4a) and the cellulase structure (Fig. 14.4b) are completely different, the similarity could be the result of convergent evolution (Fig. 14.4).

Additional support for the protein's glycoside hydrolase activity came from the fact that there were several solvent-exposed aromatic amino acids in the neighbourhood of the catalytic residues, as is found in other glycosidases. The structure was analyzed using the THEMATICS method (Wei et al. 2007) which computes the likely pKa perturbation of each amino acid and consequently identifies those most likely to have catalytic activity. The method ranked the putative catalytic residues, Asp333, Asp331 and Glu272, at positions 1, 3 and 5, respectively.

So, the evidence from several sources all seemed to indicate that the GxGYxYP domain is a new class of glycoside hydrolase. Further experimental work is required to identify likely substrates to help explain the role of the domain in the gut microbiome.

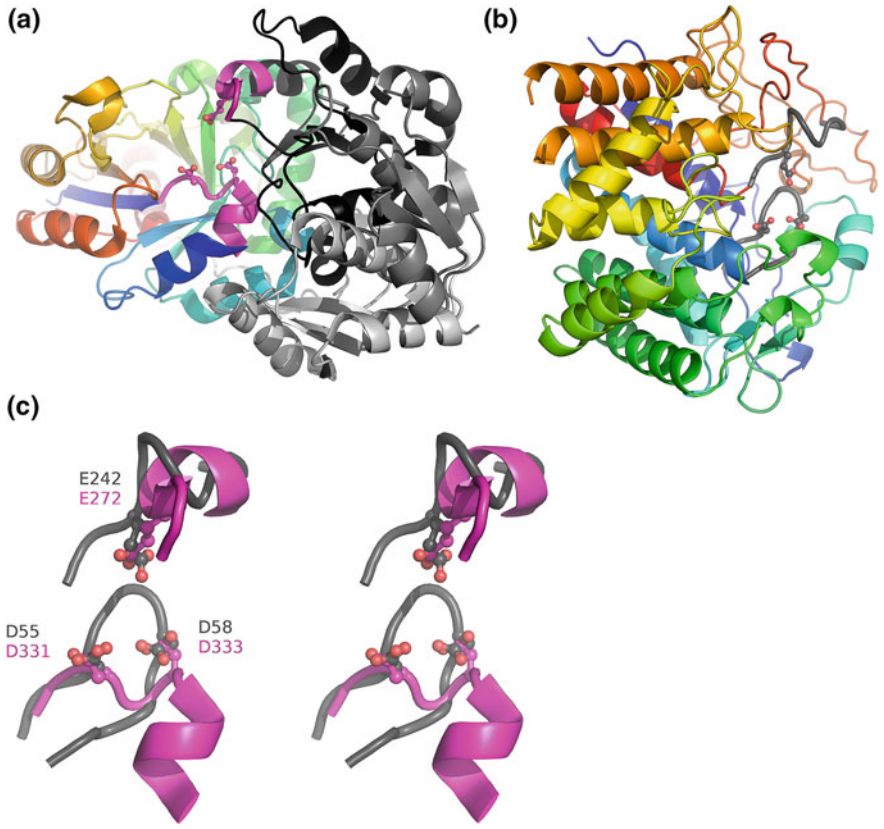



**Fig. 14.4** Comparison of overall structures (**a**, **b**) and (putative) catalytic residues (**c**; cross-eyed stereo) of BT2193 protein from *B. thetaiotaomicron* (PDB entry 3sgg; **a**) and a bacterial cellulase (PDB entry 1js4; **b**). Catalytic domains are coloured in a spectrum from *blue* at the N-terminus to *red* at the C-terminus. The three repeat domains of BT2193 are shown in different shades of *grey*. (Putative) catalytic residues are shown in a *ball*-and-*stick* representation along with some sequential context and are coloured distinctly as *magenta* (3sgg) or *grey* (1js4)

## 14.4 Community Annotation

One of the criticisms of the early SG projects was that, as the focus was on how many structures could be solved and deposited, their publication in the literature became a low priority. Indeed, for many structures, where the function remained unknown, there was little to report in any case. And, even where the desire to publish was strong, the process was something of a bottleneck in the high-throughput structure determination pipelines (Rigden 2006).

For functionally unknown proteins the experimental determination of function was a particular problem, requiring time, equipment and expertise. One solution was to collaborate with laboratories specialising in the particular protein being

studied. Another was to use high-throughput experimental screens for, say, enzymatic activities (Kuznetsova et al. 2005; Proudfoot et al. 2008; Simon and Cravatt 2010). A recent initiative, funded by the National Institute of General Medical Sciences (NIGMS), aims to tackle the problem using a community-wide approach. It is called COMBREX (COMputational BRidge to EXperiments) (Anton et al. 2013, 2014). It has compiled a database of experimentally determined functions of microbial proteins together with the functional predictions that can be inferred from them (http://combrex.bu.edu). It also encourages the experimental characterization of high priority targets, providing funding by way of assistance, to encourage experimental function determination to cope with the ever-increasing flood of newly sequenced genomes.

Other community-led approaches to annotation have included a number of wiki-based databases wherein scientists with expert knowledge in particular fields can annotate protein structures (Giles 2007; Mons et al. 2008).

One of the first such attempts was TOPSAN (The Open Protein Structure Annotation Network), initiated by the Joint Centre for Structural Genomics (JCSG) for the annotation of proteins solved by SG efforts (Ellrott et al. 2011). A wiki (http://www.topsan.org) allows registered users to annotate the pages of each structure. The pages themselves are initially filled with automatically generated data.

On a larger scale was PDBWiki (http://pdbwiki.org), created in August 2007 by the Structural Proteomics Group at the Max-Planck-Institute for Molecular Genetics (Stehr et al. 2010) and covering every structure deposited in the PDB. Sadly this went offline in January 2014, and it is not clear whether it will be reinstated.

A similar, PDB-based wiki, which is still running, is called Proteopedia at http://www.proteopedia.org (Prilusky et al. 2011). As in PDBWikiiP, each PDB entry has its own page automatically seeded, the information here coming from OCA (http://bip.weizmann.ac.il/oca) and other sources. Proteopedia provides a fully interactive Jmol view of each entry, plus anyone editing a page can create a 3D scene in Jmol to illustrate the point being made in the text. Every author is fully acknowledged, with no anonymous edits allowed. A nice feature is that any user can set up their own visible, but non-editable, areas in the system. This allows for the generation of topic-based or example pages that remain stable and can therefore be used as a teaching tool.

## 14.5 Conclusions

The SG initiatives resulted in a great number of structures solved and deposited in the PDB, contributing to our treasury of protein 3D structures. According to one estimate (Khafizov et al. 2014), the past ten years have seen the overall structural coverage of proteins, both experimentally solved and for which reliable homology models can be generated, increase from 30 to 40% (in terms of total number of protein residues). The contribution from SG efforts was ~50% of this new

structural coverage, despite coming from only ~10% of all new structures. However, due to the rapid data release required by these projects, a large proportion of structures still have little or no functional annotation. The ability to predict a protein's function from sequence and structure has been something of a Holy Grail for bioinformaticians and consequently a wide variety of methods have been developed over the years. Each of these methods has its own pros and cons and, currently, no single method shows a 100% success rate.

Thus, as the case studies here have demonstrated, one often needs the clues from several sources to arrive at a convincing case for a particular function. Experimental verification is the only way of checking whether the final prediction is correct, but this is a time-consuming business and requires specialist knowledge and equipment (and funding). There is an awareness that the difficult business of experimental determination of function needs to be rationalized, and use of community-wide initiatives like the COMBREX project seem a very promising start. Similarly the annotation of proteins can benefit greatly from a community-wide approach, enabling the experts in each field to contribute to building up the store of functional information on proteins and so allow deeper biological insights to be gained.

**Acknowledgements** The authors would like to thank Vicky Schneider for her useful comments on the first version of this chapter.

## References

Adams MA, Suits MD, Zheng J, Jia Z (2007) Piecing together the structure-function puzzle: experiences in structure-based functional annotation of hypothetical proteins. Proteomics 7 (16):2920–2932. doi:10.1002/pmic.200700099

Anton BP, Chang YC, Brown P, Choi HP, Faller LL, Guleria J, Hu Z, Klitgord N, Levy-Moonshine A, Maksad A, Mazumdar V, McGettrick M, Osmani L, Pokrzywa R, Rachlin J, Swaminathan R, Allen B, Housman G, Monahan C, Rochussen K, Tao K, Bhagwat AS, Brenner SE, Columbus L, de Crecy-Lagard V, Ferguson D, Fomenkov A, Gadda G, Morgan RD, Osterman AL, Rodionov DA, Rodionova IA, Rudd KE, Soll D, Spain J, Xu SY, Bateman A, Blumenthal RM, Bollinger JM, Chang WS, Ferrer M, Friedberg I, Galperin MY, Gobeill J, Haft D, Hunt J, Karp P, Klimke W, Krebs C, Macelis D, Madupu R, Martin MJ, Miller JH, O'Donovan C, Palsson B, Ruch P, Setterdahl A, Sutton G, Tate J, Yakunin A, Tchigvintsev D, Plata G, Hu J, Greiner R, Horn D, Sjolander K, Salzberg SL, Vitkup D, Letovsky S, Segre D, DeLisi C, Roberts RJ, Steffen M, Kasif S (2013) The COMBREX project: design, methodology, and initial results. PLoS Biol 11(8):e1001638. doi:10.1371/journal.pbio.1001638

Anton BP, Kasif S, Roberts RJ, Steffen M (2014) Objective: biochemical function. Front Genet 5:210. doi:10.3389/fgene.2014.00210

Aravind L, Anantharaman V, Balaji S, Babu MM, Iyer LM (2005) The many faces of the helix-turn-helix domain: transcription regulation and beyond. FEMS Microbiol Rev 29(2):231–262. doi:10.1016/j.femsre.2004.12.008

Binkowski TA, Freeman P, Liang J (2004) pvSOAR: detecting similar surface patterns of pocket and void surfaces of amino acid residues on proteins. Nucleic Acids Res 32 (Web Server issue):W555–558. doi:10.1093/nar/gkh390

Boocock GR, Morrison JA, Popovic M, Richards N, Ellis L, Durie PR, Rommens JM (2003) Mutations in SBDS are associated with Shwachman-Diamond syndrome. Nat Genet 33(1):97–101. doi:10.1038/ng1062

Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826

Ellrott K, Zmasek CM, Weekes D, Sri Krishna S, Bakolitsa C, Godzik A, Wooley J (2011) TOPSAN: a dynamic web database for structural genomics. Nucleic Acids Res 39 (Database issue):D494–496. doi:10.1093/nar/gkq902

Fox BG, Goulding C, Malkowski MG, Stewart L, Deacon A (2008) Structural genomics: from genes to structures with valuable materials and many questions in between. Nat Methods 5 (2):129–132. doi:10.1038/nmeth0208-129

Gabaldon T (2008) Comparative genomics-based prediction of protein function. Methods Mol Biol 439:387–401. doi:10.1007/978-1-59745-188-8_26

Gabaldon T, Huynen MA (2004) Prediction of protein function and pathways in the genome era. Cell Mol Life Sci 61(7–8):930–944. doi:10.1007/s00018-003-3387-y

Gerlt JA, Babbitt PC (2001) Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. Annu Rev Biochem 70:209–246. doi:10.1146/annurev.biochem.70.1.209

Giles J (2007) Key biology databases go wiki. Nature 445(7129):691. doi:10.1038/445691a

Glaser F, Morris RJ, Najmanovich RJ, Laskowski RA, Thornton JM (2006) A method for localizing ligand binding pockets in protein structures. Proteins 62(2):479–488. doi:10.1002/prot.20769

Granum PE, Lund T (1997) Bacillus cereus and its food poisoning toxins. FEMS Microbiol Lett 157(2):223–228

Holm L, Rosenstrom P (2010) Dali server: conservation mapping in 3D. Nucleic Acids Res 38 (Web Server issue):W545–549. doi:10.1093/nar/gkq366

Huang L, Hung L, Odell M, Yokota H, Kim R, Kim SH (2002) Structure-based experimental confirmation of biochemical function to a methyltransferase, MJ0882, from hyperthermophile Methanococcus jannaschii. J Struct Funct Genomics 2(3):121–127

Hwang KY, Chung JH, Kim SH, Han YS, Cho Y (1999) Structure-based identification of a novel NTPase from *Methanococcus jannaschii*. Nat Struct Biol 6(7):691–696. doi:10.1038/10745

Khafizov K, Madrid-Aliste C, Almo SC, Fiser A (2014) Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. Proc Natl Acad Sci U S A 111(10):3733–3738. doi:10.1073/pnas.1321614111

Kim KK, Kim R, Kim SH (1998) Crystal structure of a small heat-shock protein. Nature 394 (6693):595–599. doi:10.1038/29106

Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R (2003) Structure-based functional inference in structural genomics. J Struct Funct Genomics 4(2–3):129–135

Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1):2256–2268. doi:10.1107/S0907444904026460

Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372(3):774–797. doi:10.1016/j.jmb.2007.05.022

Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. BMC Bioinform 9:17. doi:10.1186/1471-2105-9-17

Kuznetsova E, Proudfoot M, Sanders SA, Reinking J, Savchenko A, Arrowsmith CH, Edwards AM, Yakunin AF (2005) Enzyme genomics: application of general enzymatic screens to discover new enzymes. FEMS Microbiol Rev 29(2):263–279. doi:10.1016/j.femsre.2004.12.006

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–330, 307–328.

Laskowski RA, Watson JD, Thornton JM (2005a) ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 33 (Web Server issue):W89–93. doi:10.1093/nar/gki414

Laskowski RA, Watson JD, Thornton JM (2005b) Protein function prediction using local 3D templates. J Mol Biol 351(3):614–626

Lee D, de Beer TA, Laskowski RA, Thornton JM, Orengo CA (2011) 1,000 structures and more from the MCSG. BMC Struct Biol 11:2. doi:10.1186/1472-6807-11-2

Lees J, Yeats C, Perkins J, Sillitoe I, Rentzsch R, Dessailly BH, Orengo C (2012) Gene3D: a domain-based resource for comparative genomics, functional annotation and protein network analysis. Nucleic Acids Res 40 (Database issue):D465–471. doi:10.1093/nar/gkr1181

Mons B, Ashburner M, Chichester C, van Mulligen E, Weeber M, den Dunnen J, van Ommen GJ, Musen M, Cockerill M, Hermjakob H, Mons A, Packer A, Pacheco R, Lewis S, Berkeley A, Melton W, Barris N, Wales J, Meijssen G, Moeller E, Roes PJ, Borner K, Bairoch A (2008) Calling on a million minds for community annotation in WikiProteins. Genome Biol 9(5):R89. doi:10.1186/gb-2008-9-5-r89

Nadzirin N, Firdaus-Raih M (2012) Proteins of unknown function in the Protein Data Bank (PDB): an inventory of true uncharacterized proteins and computational tools for their analysis. Int J Mol Sci 13(10):12761–12772. doi:10.3390/ijms131012761

Norvell JC, Berg JM (2007) Update on the protein structure initiative. Structure 15(12):1519–1522. doi:10.1016/j.str.2007.11.004

Noskov VN, Staak K, Shcherbakova PV, Kozmin SG, Negishi K, Ono BC, Hayatsu H, Pavlov YI (1996) HAM1, the gene controlling 6-N-hydroxylaminopurine sensitivity and mutagenesis in the yeast Saccharomyces cerevisiae. Yeast 12(1):17–29. doi:10.1002/(SICI)1097-0061(199601)12:1<17:AID-YEA875>3.0.CO;2-I

Pal D, Eisenberg D (2005) Inference of protein function from protein structure. Structure 13(1):121–130. doi:10.1016/j.str.2004.10.015

Porter CT, Bartlett GJ, Thornton JM (2004) The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucleic Acids Res 32 (Database issue):D129–133. doi:10.1093/nar/gkh028

Prilusky J, Hodis E, Canner D, Decatur WA, Oberholser K, Martz E, Berchanski A, Harel M, Sussman JL (2011) Proteopedia: a status report on the collaborative, 3D web-encyclopedia of proteins and other biomolecules. J Struct Biol 175(2):244–252. doi:10.1016/j.jsb.2011.04.011

Proudfoot M, Kuznetsova E, Sanders SA, Gonzalez CF, Brown G, Edwards AM, Arrowsmith CH, Yakunin AF (2008) High throughput screening of purified proteins for enzymatic activity. Methods Mol Biol 426:331–341. doi:10.1007/978-1-60327-058-8_21

Reddy TB, Thomas AD, Stamatis D, Bertsch J, Isbandi M, Jansson J, Mallajosyula J, Pagani I, Lobos EA, Kyrpides NC (2014) The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. Nucleic Acids Res 43 (Database issue):D1099–1106. doi:10.1093/nar/gku950

Rigden DJ (2006) Understanding the cell in terms of structure and function: insights from structural genomics. Curr Opin Biotechnol 17(5):457–464. doi:10.1016/j.copbio.2006.07.004

Rigden DJ, Eberhardt RY, Gilbert HJ, Xu Q, Chang Y, Godzik A (2014) Structure- and context-based analysis of the GxGYxYP family reveals a new putative class of glycoside hydrolase. BMC Bioinform 15:196. doi:10.1186/1471-2105-15-196

Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2):595–608. doi:10.1016/S0022-2836(02)00016-5

Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40 (Web Server issue): W471–477. doi:10.1093/nar/gks372

Saikolappan S, Das K, Sasindran SJ, Jagannath C, Dhandayuthapani S (2011) OsmC proteins of Mycobacterium tuberculosis and Mycobacterium smegmatis protect against organic hydroperoxide stress. Tuberculosis (Edinb) 91(Suppl 1):S119–127. doi:10.1016/j.tube.2011.10.021

Sanishvili R, Yakunin AF, Laskowski RA, Skarina T, Evdokimova E, Doherty-Kirby A, Lajoie GA, Thornton JM, Arrowsmith CH, Savchenko A, Joachimiak A, Edwards AM (2003) Integrating structure, bioinformatics, and enzymology to discover function: BioH, a new carboxylesterase from Escherichia coli. J Biol Chem 278(28):26039–26045. doi:10.1074/jbc.M303867200

Savchenko A, Krogan N, Cort JR, Evdokimova E, Lew JM, Yee AA, Sanchez-Pulido L, Andrade MA, Bochkarev A, Watson JD, Kennedy MA, Greenblatt J, Hughes T, Arrowsmith CH, Rommens JM, Edwards AM (2005) The Shwachman-Bodian-Diamond syndrome protein family is involved in RNA metabolism. J Biol Chem 280(19):19213–19220. doi:10.1074/jbc.M414421200

Schade M, Turner CJ, Lowenhaupt K, Rich A, Herbert A (1999) Structure-function analysis of the Z-DNA-binding domain Zalpha of dsRNA adenosine deaminase type I reveals similarity to the (alpha + beta) family of helix-turn-helix proteins. EMBO J 18(2):470–479. doi:10.1093/emboj/18.2.470

Schroder E, Littlechild JA, Lebedev AA, Errington N, Vagin AA, Isupov MN (2000) Crystal structure of decameric 2-Cys peroxiredoxin from human erythrocytes at 1.7 A resolution. Structure 8(6):605–615

Sciara G, Kendrew SG, Miele AE, Marsh NG, Federici L, Malatesta F, Schimperna G, Savino C, Vallone B (2003) The structure of ActVA-Orf6, a novel type of monooxygenase involved in actinorhodin biosynthesis. EMBO J 22(2):205–215. doi:10.1093/emboj/cdg031

Service R (2005) Structural biology. Structural genomics, round 2. Science 307(5715):1554–1558. doi:10.1126/science.307.5715.1554

Simon GM, Cravatt BF (2010) Activity-based proteomics of enzyme superfamilies: serine hydrolases as a case study. J Biol Chem 285(15):11051–11055. doi:10.1074/jbc.R109.097600

Stark A, Russell RB (2003) Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. Nucleic Acids Res 31(13):3341–3344

Stehr H, Duarte JM, Lappe M, Bhak J, Bolser DM (2010) PDBWiki: added value through community annotation of the Protein Data Bank. Database (Oxford) 2010:baq009. doi:10.1093/database/baq009

Stepchenkova EI, Kozmin SG, Alenin VV, Pavlov YI (2005) Genome-wide screening for genes whose deletions confer sensitivity to mutagenic purine base analogs in yeast. BMC Genet 6:31. doi:10.1186/1471-2156-6-31

Teichmann SA, Murzin AG, Chothia C (2001) Determination of protein function, evolution and interactions by structural genomics. Curr Opin Struct Biol 11(3):354–363

Tian W, Skolnick J (2003) How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 333(4):863–882

Todd AE, Orengo CA, Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4):1113–1143. doi:10.1006/jmbi.2001.4513

Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM (2007) Towards fully automated structure-based function prediction in structural genomics: a case study. J Mol Biol 367(5):1511–1522

Wei Y, Ko J, Murga LF, Ondrechen MJ (2007) Selective prediction of interaction sites in protein structures with THEMATICS. BMC Bioinform 8:119. doi:10.1186/1471-2105-8-119

Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36(3):307–340

Wu R, Skaar EP, Zhang R, Joachimiak G, Gornicki P, Schneewind O, Joachimiak A (2005) Staphylococcus aureus IsdG and IsdI, heme-degrading enzymes with structural similarity to monooxygenases. J Biol Chem 280(4):2840–2846. doi:10.1074/jbc.M409526200

Wu S, Liang MP, Altman RB (2008) The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. Genome Biol 9(1):R8. doi:10.1186/gb-2008-9-1-r8

# Chapter 15
# Prediction of Protein Function from Theoretical Models

**Daniel J. Rigden, Iwona A. Cymerman and Janusz M. Bujnicki**

**Abstract** Explicit 3D models can be obtained by comparative protein modelling, a mature and predictable technique, fragment assembly ab initio methods for smaller novel or unrecognisable folds and contact-based methods for large protein families. Each modelling method has limitations in model accuracy, which vary further according to the characteristics of the target: as a result, the performance of structure-based function prediction algorithms applied to models is variable. Nevertheless, with care, a wide variety of structure-based methods can be productively applied to protein models, frequently facilitating the planning and interpretation of experimental results. This chapter will first survey the literature on applicability of structure-based methods specifically to models, before discussing a selection of examples in more detail.

**Keywords** Model added value · Binding site prediction · Function prediction · Prediction of specificity · Comparative modelling · Ab initio modelling · Homology modelling · Protein model databases

## 15.1 Background

In this era of Big Data, biologists benefit from the exponential growth of both sequence (UniProt Consortium 2015) and protein structure databases (Rose et al. 2015), growth driven by technological innovations such as pyrosequencing and

D.J. Rigden (✉)
Institute of Integrative Biology, University of Liverpool,
Liverpool L69 7ZB, UK
e-mail: drigden@liverpool.ac.uk

I.A. Cymerman · J.M. Bujnicki
International Institute of Molecular and Cell Biology,
Trojdena 4, 02-109 Warsaw, Poland

J.M. Bujnicki
Faculty of Biology, Institute of Molecular Biology and Biotechnology,
Adam Mickiewicz University, Umultowska 89, 61-614 Poznań, Poland



D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3_15

robotised crystal growth and handling. However, the gap between the numbers of available protein sequences and protein structures remains huge. At the time of writing, the PDB had just breached 100,000 entries yet UniProt contained around 48 million entries. To bridge that gap the structure prediction methods featured in Section 1 of this book must play a major role. Currently, ab initio structure prediction (see Chap. 1) is limited to relatively small proteins while the exciting new area of contact-driven modelling (Chap. 2) requires large numbers of reasonably diverse homologous sequences to the target, limiting the number of protein families of currently unknown structure to which it can be applied (Hopf et al. 2012; Kamisetty et al. 2013). Thus, the burden falls largely on homology modelling (Chap. 4) which, where a suitable template can be found, is generally quick, easy and broadly applicable. Recognising the key role of comparative modelling, Structural Genomics consortia made a concerted push, especially in the PSI-2 era, to seed protein fold space with new structures. These structures were chosen to open up the possibility of homology modelling of large numbers of sequences that were hitherto intractable (Dessailly et al. 2009). The success of SG initiatives in providing structural novelty has recently been analysed (Khafizov et al. 2014). The proportion of protein sequence space that can be mapped to protein structures (in this case, meaning residues aligned by BLAST with e-value $<10^{-10}$) increased modestly from 2001 to 2011 from 30 to 40%. However, around half of this expansion was accounted for by SG-derived structures despite their only accounting for 10% of the new structures determined in the period. In recent times the interests of SG centres have diversified but for transmembrane proteins at least a strong interest in improving structural coverage of fold space is maintained (Pieper et al. 2013). Model organisms of particular interest tend to be better catered for than average. For example, recent analysis shows residue-level coverage of the human proteome approaching 60% (Schwede 2013), a figure demonstrating higher completeness than initially apparent since around 30% of the proteome is predicted to be intrinsically disordered.

Although the value of homology models for structure-based analysis is undisputed, the inevitable errors they contain can ultimately confound predictive methods. Thus it is important for the user to have a good understanding of the quality of their model both overall and in terms of local details. Model Quality Assessment Programs (MQAPs; Kryshtafovych et al. 2011) are available to consider various aspects of protein structure. Models available in databases and repositories will generally be accompanied by MQAP analyses and online modelling pipelines such as HHpred (Soding et al. 2005) often offer this type of analysis. A major source of error in comparative models lies in the backbone which will generally strongly resemble that of the template(s) used in model construction. An estimate of the backbone error of a homology model, sharing a certain % sequence identity with template(s), can therefore be obtained by considering the backbone root-mean-square deviation (rmsd) between crystal structures sharing the same % identity. This relationship was analysed by Chothia and Lesk (1986) showing that backbone rmsd values for common cores steadily increase from around 0.5 Å for crystal structures of the same protein to more than 2.0 Å for homologous pairs sharing less than 20% sequence identity. Of course, this is an optimistic estimate in

a modelling context since it assumes no alignment errors between target and template, this in fact being the largest source of error in homology models (Ginalski 2006). Furthermore, the shared cores analysed (Chothia and Lesk 1986) sometimes comprised only half the protein. Finally, outside the core, deletions and insertions will be likely less well-modelled since the template provides less information to guide their construction, and side chains will generally be more poorly modelled than the backbone, especially at higher divergences between target and template (Chung and Subbiah 1996). An objective assessment of a protein model, ideally employing multiple MQAPs, is an essential precursor to structure-based inference (Schwede 2013).

While the overall quality of homology models can be broadly estimated in advance based on target-template identity, the same cannot be said for fragment-based or contact-driven methods. By these techniques a single target will be modelled many times, perhaps thousands of times in the case of fragment assembly methods. Predicting which fragment assembly model is likely to best represent the unknown target structure, and with what degree of confidence that can be asserted, is done by clustering for low-resolution fold predictions (Shortle et al. 1998) and by energy funnel analysis for the much more time-consuming all-atom protocols (e.g. Barth et al. 2007) (see Chap. 1). The QUARK fragment assembly modelling server (Xu and Zhang 2012) offers a predicted TM-score (Zhang and Skolnick 2004) for its top returned model, and the same for the best of the top ten models, based on an assessment of the fragments available and how the models clustered (Xu and Zhang 2013). TM-scores range from 0 to 1 with 1 being a perfect match, 0.17 the average score for unrelated proteins and 0.5 broadly indicating that the overall fold has been predicted correctly. Bespoke scoring systems have been developed for contact-based models and will surely improve in the short term (Marks et al. 2011). In two studies the top-scoring models gave TM-scores of 0.25–0.70 (Cα rmsd values in the range 2.7–4.8 Å) for globular proteins (Marks et al. 2011) and 0.40–0.70 (2.8–5.1 Å Cα rmsd) for transmembrane proteins (Hopf et al. 2012).

## 15.2 Suitability of Protein 3D Models for Structure-Based Predictions

As mentioned above, models will vary considerably in their overall accuracy. In particular, the quality of comparative models depends on the degree of evolutionary divergence between the modelled target and the template that was used. Broadly, the higher the sequence identity shared between the two, the lower the expected structural difference between the two, and so the lower the error likely to be present in the model built using the template. Bearing this in mind, it is important to ask how reliably different structural and functional characteristics can be inferred from model structures of different predicted quality. This section therefore focuses on published work that explicitly considers the value of modelled structures with

respect to the methods covered earlier in Section 2 of this book. It is worth remembering that these studies generally used comparative single-template models. Thus, the results are representative of the models typically produced by large-scale fully automated methods. However, in some cases, more elaborate modelling procedures employing multiple templates and refinement may improve the accuracy of models and hence of structure-based properties inferred from them. Care must be taken, however, since use of multiple templates can also degrade model quality in some circumstances (Hasegawa and Funatsu 2012). Guidelines for choice of template in order to maximise the chance of a positive effect on model quality have been presented (Hasegawa and Funatsu 2012). The comparative modelling protocols of the Rosetta suite also now effectively combine information from multiple templates with impressive results (Song et al. 2013).

### *15.2.1 Surface Properties*

The functions of most proteins rely on intermolecular recognition, with ligands ranging from small molecules to multi-protein complexes, so surface properties (see also Chap. 10) are of particular interest. The accuracy of prediction of these properties was addressed in large-scale analyses of simple comparative models performed by Chakravarty et al. (2005). It was shown that the overall accuracy of all analyzed structural model-derived properties (SDPs) drops as a function of template-to-target sequence similarity, but that this decrease has different degrees of impact on the accuracy of different structural features (Table 15.1). For example, alignment errors have a negligible effect on the correctness of the prediction of accessible surface area (ASA) while the correctness of the electrostatic potential prediction is already affected when the sequence identity drops below 50% (Table 15.1). Knowing how reliable the different model-derived properties are, it is interesting to investigate what additional information (added value) they carry with reference to the template structures used for model building. Again, systematic analysis of the model added value was performed on the large scale only for single-template models (Chakravarty and Sanchez 2004), but it provides valuable guidelines as to which particular model-derived properties can be informative (Table 15.1).

In general, the greater the difference between target and template sequences, the more significant the added value becomes. This results from the fact that lower-similarity cases contain less information in the template about the size and physicochemical properties of particular residues in the target. However, not all structure-derived properties provide additional information with respect to the template. For SDPs that depend mostly on position of residues, such as exposure state, neighbourhood of buried residues and number of surface pockets, models do not provide added value. It is probably caused by the fact that buried residues are more conserved than exposed residues, comprising protein cores that are responsible for protein integrity. For other SDPs, such as neighbourhood of exposed residues

**Table 15.1** The accuracy and added value of structure-derived properties in single-template based comparative models (Chakravarty and Sanchez 2004; Chakravarty et al. 2005)

| Model-derived property | Accuracy | Added value |
|---|---|---|
| All | Increases with template-to-target identity | Increases when template-to-target identity drops |
| Residue exposure state | Decreases with the protein size; affected by alignment errors below 30% sequence identity | No added value |
| Buried residues neighbourhood | No clear dependence on protein size, higher than for exposed residues; affected by alignment errors below 30% sequence identity | No added value |
| Exposed residues neighbourhood | No clear dependence on protein size, lower than for buried residues; affected by alignment errors below 30% sequence identity | Moderate added value |
| Accessible surface area (ASA) | Error in total ASA increases with protein size, influence of misalignment is very small | Moderate added value |
| Surface pockets identification | Pocket artefacts; increased number of surface pockets in comparison to the template and target structure; alignment errors have no clear effect on the number of pockets | Negative added value |
| Surface pockets composition | | High added value |
| Electrostatic potential (EP) | Affected by alignment errors below 50% sequence identity | High added value |

and total accessible surface area (ASA), models show some added value. This is very important as residues accessible to the solvent are responsible for interactions with other molecules, thus determining the biological function of the protein. Finally, for properties that strongly depend on the physicochemical characteristics of the amino acids in the sequence, such as composition of pockets and electrostatic potential, models show large added value. The identification of charged regions is of large value as they may represent binding or active sites (see Chap. 10).

In summary, the studies performed by Chakravarty et al. demonstrated that, with the exception of the detection of pockets, most model-derived structural properties exhibit some level of added value. The more a given property depends on the sequence of the protein the more useful a model will be in estimating the value of that property. Encouragingly, depending on the feature, 25–40% sequence identity between target and template was sufficient to produce a SDP estimate of comparable accuracy to that available from an NMR structure.

A later study (Piedra et al. 2008) specifically examined the quality of modelling of surface clefts in comparative models, focusing on medium to low quality models (defined as those based on target-template alignments of 30–60% or <30%, respectively). Six metrics reporting the reproduction of known benchmark structure cleft structure by the models were analysed including measurements of rmsd,

protrusion index and accessible surface area. Some expected factors were confirmed as important: thus, cleft model quality improves as the sequence identity rises, improves with accurate alignment of target and template, and is degraded when residues contributing to the cleft in the model structure are not aligned with counterparts in the template. Most interestingly, the authors define a threshold of around 20% sequence identity between target and template, below which there is a steep drop in the quality of modelled clefts by various metrics (Piedra et al. 2008). Encouragingly, in the range of 20–30% identity protein clefts are generally sufficiently well-modelled to be of value, although a significant proportion of poor models are found in this category.

Further work (Zhao et al. 2011) was limited in scope to models based on target-template alignments of >30% sequence identity, but usefully added measures capturing the recall of atoms in the surface pockets, in particular the recall of 'signature', biologically important atoms i.e. whether key binding determinants in the native structure were correctly solvent exposed in the modelled structure. Importantly, the latter 'signature' atoms were much better modelled than cleft-lining atoms on average. Again, even lower quality models contained high value information: considering models containing clefts contributed by sequence segments down to only 45% identity between target and template, 77% of true binding pocket atoms were present in the modelled pocket (Zhao et al. 2011).

### *15.2.2 Functional Sites*

Early work proposed a multi-step procedure that enables identification of protein functional sites in low-to-moderate resolution models (Fetrow and Skolnick 1998). Based on the geometry, residue identity, their distances between alpha carbons and conformation, the active site residues become a three dimensional descriptor termed Fuzzy Functional Form (FFF) which could be used to screen homology models. The usefulness of the method was proved by the identification of the novel members of the disulphide glutaredoxin/thioredoxin protein family in the yeast (Fetrow and Skolnick 1998) and *E. coli* genomes (Fetrow et al. 1998), whose functions could not be identified by sequence comparison methods. The great advantage of FFF and related approaches is that the method distinguishes protein pairs with similar active sites from proteins pairs that may have similar folds, but not necessarily similar active sites. The FFF technology was further developed to the method called active site profiling (Cammer et al. 2003) and was successfully combined with experimental procedures to determine new serine hydrolases in yeast (Baxter et al. 2004). The main advantage of the method is that it does not rely on residue conservation across an entire family and the key functional residues are specifically identified regardless of overall global sequence similarity to any other protein exhibiting the same function. It could therefore be applicable to identification and annotation of different functional sites, including enzyme-active sites, regulatory and cofactor-binding sites.

A general approach to functional site detection is also implemented in advanced methods like FINDSITE (Brylinski and Skolnick 2009) and COFACTOR (Roy et al. 2012), sometimes termed Ligand Homology Modelling (LHM). Putative binding sites in a protein of interest along with their candidate ligands are first discovered using, for example, 3D motif matching (see also Chap. 11) or by matching to templates containing bound compounds. The candidates can be clustered, refined and scored with strong predictions revealing not only a proposed functional site but also suggested ligand. Importantly, the quality of predictions degrades only slowly with lower quality starting structures e.g. homology models of increasing evolutionary divergence from the best available template (Lee and Zhang 2012; Skolnick et al. 2013; Yang et al. 2013). This points the way to the use of LHM for ligand discovery in a pharmaceutical context (see also Sect. 15.2.4 below), not simply for annotating functional sites. For example, the complete human kinome was subject to a LHM study involving the construction of homology models for each kinase. That work concluded that performance at ligand ranking using modelled structures was at least as good as conventional virtual screening applied to crystal structures (Brylinski and Skolnick 2010). Indeed, the technique is viewed as having potential in a number of medicinal areas including the off-target binding responsible for drug side-effects (Skolnick et al. 2013).

### *15.2.3 Specific Binding Predictions*

Other studies have addressed whether more specific function predictions can be made as accurately for models as for experimental structures. For metal-binding sites, the results of the MetSite method that combines sequence and structure information were encouraging (Sodhi et al. 2004). Although performance with modelled structures was inferior to that with experimental structures, correct metal site predictions could be made for around half of reliable mGenTHREADER-derived models. Notably, these models are backbone-only so that performance would not be at all affected by errors in side-chain positioning.

Similarly, a method for predicting DNA-binding ability using sequence information, structural asymmetry in distribution of some amino acids and dipole moments, has been benchmarked against both experimental structures and models (Szilagyi and Skolnick 2006). The method uses Cα-only structures. Performance of this method vs that obtained for experimental structures, was found to decrease only very slightly for models of up to 6 Å rms deviation from native structure. Thus, it will be appropriate to use the method on model structures of all kinds, including the template-free and fold recognition-derived models for which lower accuracy would be expected.

A more challenging problem is not just distinguishing DNA-binding proteins from non-binding structures, but to accurately predict the mode of DNA binding. In one such study (Gao and Skolnick 2009), models were automatically built using TASSER excluding any templates that shared >30% sequence identity with the

target. Remarkably, for the easier cases where threading indicated the availability of better (though still <30% sequence identical) templates, performance in predicting DNA binding mode was broadly comparably to that achieved using the apo crystal structures of the proteins in question (Gao and Skolnick 2009).

Model structures have also been shown to be valuable in detecting proteins that bind RNA (Li et al. 2014). Even predictions based on the poorest quality models, built using templates sharing <30% sequence identity, outperformed those made using sequences alone. Furthermore, good quality models from >90% identical templates performed comparably to the experimental structures themselves (Li et al. 2014).

### *15.2.4 Small Molecule Binding*

One of the important practical applications of protein models is for in silico screening against small compound databases in order to pick out likely inhibitors for development into drug leads (Jacobson and Sali 2004). While not the focus of this book, such docking employs the same principles and programs as are increasingly used to predict natural ligands for proteins in a structure-based manner (see also Chap. 10) (Hermann et al. 2007; Song et al. 2007). It is therefore relevant here to mention studies that explore the performance of protein models, compared to experimental structures, in both small molecule docking scenarios.

In early work in the pharmaceutical context, McGovern and Shoichet (2003) compared enrichment of known ligands vs decoys in docking results for holo, apo and model structures of nine enzymes. Templates used for model construction shared 34–87% sequence identity with targets overall, and 45–100% identity in the region of the binding site. In the best enrichment class were results for eight holo structure, two apo structures and three models, confirming the general superiority of experimental structures. Nevertheless, modelled structures as a whole almost always gave better than random selection of active compounds. There was a tendency for models built using more closely-related templates to perform better, but small conformational changes in the binding site could sometimes lead to poor performance even in these cases. Later, Oshiro et al. (2004) compared the enrichment of known active compounds in docking results for compound databases of experimental structures and comparative models, several for each, of CDK2 and factor VIIa. The templates used for model construction shared 37–77% sequence identity in the vicinity of the binding site. Remarkably, where the local sequence identity of the model was higher than 50%, performance was similar to that obtained with an experimental structure. Below 50% binding site identity, performance was clearly degraded. Later work introduced the idea of ‘consensus’ enrichment where compounds are ranked according to their binding scores against multiple comparative models. Strikingly, this approach produced performance comparable to or even better than that achieved against X-ray structures of the target in question (Fan et al. 2009). Also importantly, even models built using

distantly homologous templates sharing as low as 25% sequence identity produce better enrichment of active compounds than the template itself, demonstrating the added value of the modelling exercise (Fan et al. 2009). This message was reinforced by very recent work using automatically generated homology models built with I-TASSER (Du et al. 2015). Even restricting the model building to templates sharing <30% sequence identity to the target, in a majority of cases compound enrichment approached that achieved with the crystal structure (Du et al. 2015). Taken together, these papers strongly encourage the use of models for docking studies where the obviously preferable experimental structures are unavailable.

A different perspective on the accuracy of small molecule docking to protein models was provided by Bordogna et al. (2011). They asked directly whether the known, experimentally observed mode of ligand binding was observed when the small molecule was docked in silico using AutoDock to a model of the protein receptor. A set of 21 protein-ligand complexes was used to benchmark performance with a total of 245 models constructed for the receptors using templates sharing a wide range of % sequence identity. A strong relationship between the receptor-template sequence identity and accuracy of the best-scoring docking pose was observed (Bordogna et al. 2011). Where the former was at least 80%, the top-scoring docking pose generally deviated by less than 2 Å rmsd from the experimentally observed binding mode. At lower similarities between receptor and template similarly accurate predictions were still sometimes seen, but the spread of accuracy became progressively larger and totally inaccurate predictions became increasingly common. Interestingly, the authors could reliably predict the accuracy with which retinol could be docked to models of its binding protein, a case not in their original dataset, using the results of their study (Bordogna et al. 2011).

Recent papers (Wallrapp et al. 2013) have amply demonstrated the value of homology models for structure-based assignment of function to proteins by the metabolite docking approach. A particular focus has been enzyme superfamilies in which substrate specificities of individual families can be confidently predicted. For example, in the enolase superfamily, homology models of different dipeptide epimerases built using a single template, docked and scored against all 400 possible dipeptide substrates, correctly predicted favoured substrates (Lukk et al. 2012). Interestingly, this work revealed both families with the same specificity as experimentally characterised groups, but different structural bases of substrate preference, as well as families with entirely novel specificity. Equally impressive was work predicting substrate chain length preferences in uncharacterised groups of prenyltransferases (Wallrapp et al. 2013). These enzymes synthesise linear allylic diphosphates with chain lengths ranging from C10 to C50 as building blocks for a wide variety of isoprenoid metabolites. In blind experiments, homology models were made, mostly based on templates sharing 30–60% sequence identity with the protein of interest, and docked with potential substrates containing 5–25 carbon atoms. Predicted binding energies were used to suggest substrate specificities which later experiment showed to be correct in 45% of cases (spanning essentially the full range of target-template sequence identities), differing by only one C5 unit in a

further 15% and profoundly wrong in only 5%. Subsequent crystal structures confirmed the broad features of substrate binding modes predicted using homology models (Wallrapp et al. 2013).

### *15.2.5 Protein-Protein Interactions*

The use of homology models to predict protein-protein interactions on a large scale has seen rapidly increasing interest recently (see also Chap. 8). Indeed, protein models have been termed the 'Grand Challenge' in the area of protein docking (Anishchenko et al. 2014). Two broad approaches can be distinguished. The simpler, template-free approach uses regular docking methods to seek favourable modes of interaction between two protein models. The second, template-based modelling (Szilagyi and Zhang 2014), is based on finding complex templates from the PDB that are proposed to represent the mode of interaction of the proteins of interest. Commonly, but not invariably (e.g. Tuncbag et al. 2012, Zhang et al. 2012), the two proteins of interest will be homologous to the two partners in the identified complex template. Thus, a crystal structure of proteins X and Y in complex might be used to model the interaction of proteins A and B where A is homologous to X and B is homologous to Y. The two approaches have been usefully compared (Vreven et al. 2014), highlighting strengths, weaknesses and potential synergies.

Encouragement for the use of homology models in template-free docking came from the results of a large scale exercise in which protein models of complexed pairs were systematically constructed with errors over a 1–10 Å rmsd range (Tovchigrechko et al. 2002). As expected, the accuracy of prediction of interface formation declined with decreasing model quality but, nevertheless, gross features of the interaction are frequently present in complexes containing models of up to around 6 Å rmsd from the experimental structure (Tovchigrechko et al. 2002). More recently, the accuracy of the binding sites in automatically produced protein models has been assessed (Kundrotas and Vakser 2010): the quality of these regions is obviously much more important for successful docking than that of non-binding regions. In that study only a low-resolution criterion of docking success was used but the authors concluded that interface regions in homology models were predicted well enough for low- or medium-resolution docking in around 50% of their cases (Kundrotas and Vakser 2010).

It is highly desirable to steer the template-free docking to the known interface regions where this information is available. A recent study of such information-driven docking (Rodrigues et al. 2013) demonstrated, most interestingly, that the quality of those restraints is more important for overall docking accuracy than that of the model that is being docked. Thus, where interface information was accurate, reasonable quality docking poses (<3 Å interface rmsd) were obtained for modelled structures built using templates with which they shared as little as 20% sequence identity. Another key finding was that the quality of the

docking results could be predicted based simply on the template sequence identity shared with the protein model being docked (Rodrigues et al. 2013).

Interest in template-based protein docking grew with the publication of a paper suggesting that templates suitable for modelling essentially all complexes between structurally characterised proteins are already available in the PDB (Kundrotas et al. 2012) and it has therefore been argued that a near-complete albeit low-resolution description of the interactome will be achieved soon (Vakser 2013). However, other work demonstrates that the availability of more interaction templates will be required to enable better quality modelling of protein interactions in the twilight zone where interactors share only low sequence identity with proteins in structurally characterised complexes (Negroni et al. 2014). An exercise incorporating template-based protein docking on a large scale to probe host-parasite interactions is discussed in detail below (Davis et al. 2007). More recently, genome-scale prediction of protein-protein interactions has been done (Zhang et al. 2012). Homology models were assembled into putative complexes using the template-based approach and the resulting interfaces assessed against expectations from known complexes. Overall confidence measures combined these structure-based scores in a Bayesian framework with other information such as co-expression. Notably, high confidence predictions were sometimes obtained in this way even for interactions that were considered low probability in structural terms alone (Zhang et al. 2012).

### *15.2.6 Protein Model Databases*

Although model databases are not a main focus of this chapter, we hope the foregoing discussion and the examples that follow will encourage the reader to explore the use of protein models for function prediction, and so it is worth mentioning that actually carrying out the modelling is not always essential: ready-made models may already be available. This is only generally the case for comparative models, but accessing these automatically generated models is very easily done through the Protein Model Portal (http://www.proteinmodelportal.org; PMP; Haas et al. 2013). The user is greeted by a single search box which accepts database accessions, sequences or free text queries. A single page of search results contains links out to two well-established large-scale automated modelling exercises—ModBase (Pieper et al. 2014b) and SWISS-MODEL Repository (Kiefer et al. 2009)—as well as to smaller scale resources allied to Structural Genomics Consortia or devoted specifically to G protein-coupled receptors (Vroling et al. 2011). Interestingly, where several models for the same protein region are available their structural variability can be analysed and illustrated. Such variability may result from genuine conformational properties—two models may have been built on different templates representing alternative allosteric states—or can be indicative of

localised uncertainty in the model. The results of specialised Model Quality Assessment Programs (MQAPs) typically accompany models linked to by the PMP, but MQAPs may also be run for any structure provided at the PMP. The use of a model from ModBase for function prediction is illustrated below in Sect. 15.3.5. Finally, the PMP now also encompasses The Model Archive (http://modelarchive.org) designed as a repository for community-generated models and accompanying data (Schwede et al. 2009).

## 15.3 Function Prediction Examples

Earlier chapters of this book show the many and diverse ways in which structures may be used to infer function. As outlined above, homology models—and sometimes models deriving from other methodologies—have proved to be eminently suitable for a variety of structure-based function inference techniques. Specific case studies are now presented to illustrate the application of many of these methods to structural information generated by template-free modelling, contact-based modelling, fold recognition or comparative modelling (Chaps. 1–4, respectively).

### *15.3.1 Fold Prediction with Fragment-Based Ab Initio Models*

Although impressive accuracy has occasionally been achieved using fragment-based ab initio or de novo modelling in favourable cases (Bradley et al. 2005), a more conservative objective for such modelling has been simply been prediction of the correct fold rather than highly accurate predictions (see Chap. 1). This has limited the range of function inference techniques that have be applied, and means that most predictions in the literature are based mainly on the protein fold predicted, and its functional correlations (discussed in Chap. 9). Exceptions to this trend involve functional annotation methods that are comparatively tolerant of model error such as electrostatic analyses.

In an early large-scale application of ROSETTA, Bonneau et al. (2002) produced models for 510 Pfam families with average length of less than 150 residues. These were of unknown structure at the time, but for some a function was known or suspected. Tentative predictions could be bolstered by the modelling results in several cases. For example, PF01938, the TRAM domain was suspected at the time to be a nucleic-acid binding protein, a prediction strongly supported by the resemblance of its de novo model to structures in a SCOP superfamily containing diverse nucleic acid binding proteins. The accuracy of the model was subsequently revealed by crystal structures of the RNA-binding protein RumA (Lee et al. 2004). An example of function prediction for a completely uncharacterised protein was

made for what was known at the time as Domain of Unknown Function 37 (PF01809). Its model matched the structure of NK-lysin a haemolytic protein expressed in natural killer T-cells. Although the structure of PF01809 proteins remains unknown, the Pfam database has renamed this entry as 'Haemolytic' on the basis of characterisation of a member from *Bacillus subtilis* as a haemolysin (Liu et al. 2009).

Interestingly, a de novo model need not match a known fold exactly in order to offer clues to function; the broad structural class of the model may sometimes be suggestive. An example of this is the model produced for a mucin-binding domain (Bumbaca et al. 2007). The favoured model contained a β-sandwich fold, of the kind strongly associated with carbohydrate binding. At the time of publication, half the families of carbohydrate-binding domains of known structure folded into β-sandwich structures of some kind. This would be consistent with the domain binding to the carbohydrate component, rather than a protein part, of its target, the highly glycosylated mucin. Subsequent determination of the crystal structure of the protein confirmed the existence of a β-sandwich fold, although the elongated immunoglobulin-like architecture revealed was of quite different proportions to the de novo model (Du et al. 2011).

Another example showed how function suggested by the fold of a de novo model could be supported by other analyses (Rigden and Galperin 2008). The SpoVS protein is known as being required for sporulation in sporulating bacteria, but in fact has a wider distribution. The phenotypic characterisation of SpoVS mutants says very little about its molecular role. However, the top models produced by both ROSETTA and I-TASSER matched well to the fold of the Alba archaeal chromatin protein (Fig. 15.1a). This fold is strongly associated with nucleic acid binding in various contexts and, furthermore, mapping electrostatic potential on to the models revealed the pronounced positively-charged region characteristic of nucleic acid-binding proteins (Fig. 15.1b; see also Chap. 10). Taken together these analyses suggested that SpoVS is a novel transcription factor that contributes to the control of intricate gene expression patterns involved in sporulation (Rigden and Galperin 2008). Subsequent determination of the crystal structure of *Thermus thermophilus* SpoVS confirms the accuracy of the fold prediction (Fig. 15.1c).

A large scale application of de novo modelling, in a pipeline also involving PSI-BLAST and threading-based structure predictions, analysed the yeast genome (Malmstrom et al. 2007). The authors used a novel strategy to use known functional information to help pick out correct putative structure-based matches of de novo models to SCOP superfamilies. To this end, in addition to structural comparisons, the overlap of Gene Ontology (GO) terms between the target protein and proteins of the superfamily in question was assessed. These complementary sources of information were combined using Bayesian statistics. Figure 15.2 shows an example of a prediction, that the structure of protein TRS20/YBR254C belongs in the SNARE superfamily of the SCOP database, that was later confirmed by the determination of an experimental structure. The match between model and crystal structure is partial and limited (Fig. 15.2), illustrating the value of including GO information for target

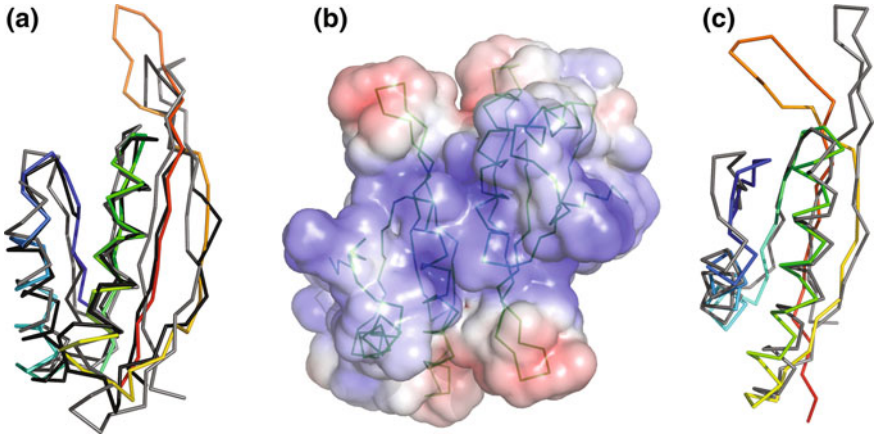



**Fig. 15.1** Analysis of template free models of SpoVS suggests a nucleic acid-binding function (Rigden and Galperin 2008). **a** Both ROSETTA (*grey*) and I-TASSER (*black*) models of SpoVS are strongly similar to the structure of Alba, an archaeal chromatin protein (PDB code 1nfj; *coloured* in a spectrum from *blue* N-terminus to *red* C-terminus). **b** The electrostatic potential of a putative SpoVS dimer, based on the ROSETTA model, with *blue* showing positive regions and *red* negative regions. **c** Comparison of the ROSETTA model (*grey*) with the unpublished structure of *Thermus thermophilus* SpoVS (PDB code 2eh1; *coloured* in a spectrum from *blue* N-terminus to *red* C-terminus)

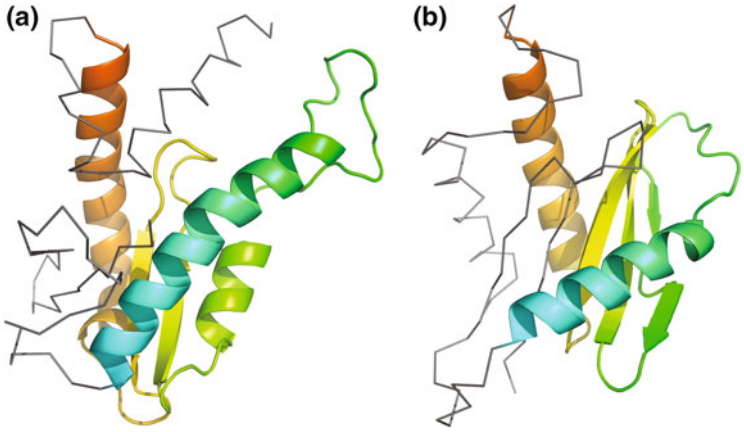



**Fig. 15.2** A confirmed structure prediction from Malmstrom et al. (2007). The model of TRS20/YBR254C (**a**) was matched to the SNARE superfamily in the SCOP database, an assignment later validated by a later experimental structure (PDB code 1h3q) of a related protein (**b**). *Colours* are used for structurally matched regions, *grey* elsewhere

and superfamilies of putatively matched structures. In this case, the target TRS20/YBR254C is one of the subunits of the transport protein particle (TRAPP) complex involved in vesicle docking and fusion. Its match with structures from the SNARE-like superfamily of SCOP was therefore strongly supported since vesicle trafficking is a strong theme of proteins in that superfamily.

More recently, proteins encoded by the *Escherichia coli* genome were addressed by a hybrid pipeline in which conventional template-based modelling for targets with recognisable folds was supplemented by QUARK ab initio modelling of the 'hard' sequences for which fold recognition (see Chap. 3) failed to identify close homologues (Xu and Zhang 2013). TM-scores vs the unknown native structures were estimated using a score that drew on quality measures of the original fragments used as well as the clustering behaviour of the eventual model set. These predicted TM-scores suggested that fold predictions of 72/495 hard targets were essentially correct (TM-score >0.5) and a further 321 were partially correct (TM-score >0.35). Again, comparison of predicted fold to the PDB was used to highlight cases where structural similarity might be string enough to indicate unexpected homology and thus, potentially, indicate a function for the target protein by comparison.

### *15.3.2 Fold Prediction with Contact-Based Models*

As a much more recently arrived technique, examples of application of contact-derived modelling to genuinely unknown protein folds are few: more typically methods have been developed and benchmarked against known structures. However, work applying the EVfold method (Marks et al. 2011) to helical transmembrane proteins (Hopf et al. 2012) included fold predictions for several families. In some cases, the inclusion of the target family in a Pfam clan in which other families had been structurally characterised would have offered a strong advance indication of the overall fold. This was not the case, however, for the human adiponectin receptor (Uniprot entry ADR1_HUMAN) which resides in the Pfam family 'Haemolysin-III related' (PF03006). Here the most similar PDB structures to the highest-ranked adiponectin receptor model produced bore a striking resemblance to 7-transmembrane proteins such as bacteriorhodopsin. The topologies of the model and bacteriorhodopsin are identical (Fig. 15.3) and the structures can be aligned over the majority of their length with a Cα-rmsd of around 4.5 Å. Although an unsuspected homology between the two is an obvious plausible explanation for the similarity, the authors also raise the possibility that the 7-helical fold may be particularly energetically favourable with each family independently converging on the fold. Although the examples tackled were all proteins of known function, it is clear that the performance of contact-based modelling in favourable cases will in the future allow function annotation by fold prediction for hypothetical proteins, Domains of Unknown Function and the like.

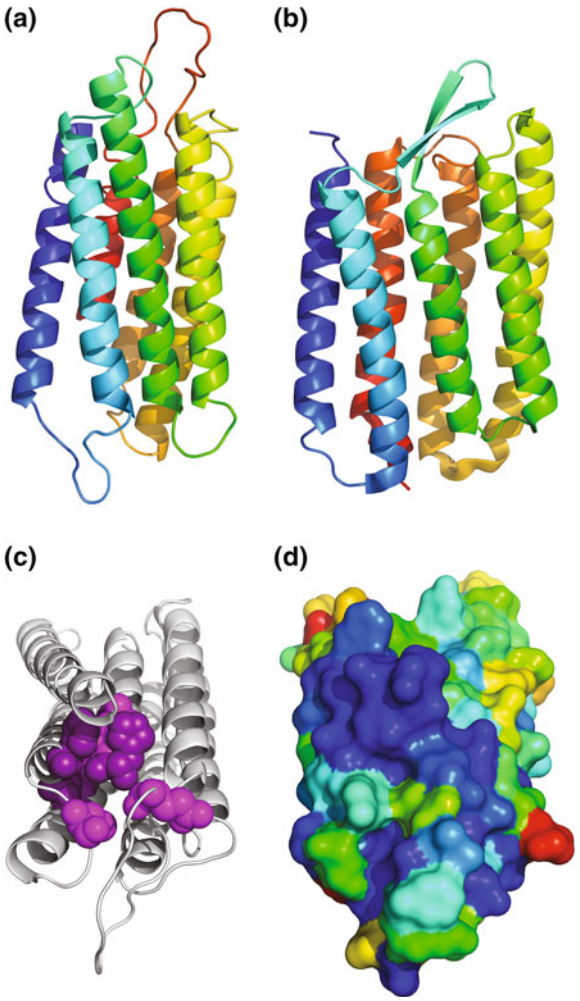



**Fig. 15.3** An EVfold model of the human adiponectin receptor **a** is predicted to have the 7-transmembrane topology seen in bacteriorhodopsin (**b**; PDB code 3hao). A functionally important region in the model is revealed by mapping of residues that are highly involved in evolutionary couplings (**c**; *purple* used for residues already suspected to be of functional importance, *magenta* for novel predictions) and by ConSurf mapping of sequence conservation (**d**; *coloured* on a scale from *blue*, highly conserved to *red*, unconserved)

As discussed in Chap. 2, modelling based on evolutionary constraints is considered to produce results in which functionally significant regions of the protein may be predicted as containing residues implicated in higher than average numbers of constraints i.e. with high 'coupling scores'. On the adiponectin receptor model mentioned above, such residues form a cluster on the cytoplasmic side which includes both residues already predicted to be involved in catalysis, as well as others nearby (Fig. 15.3c). Strong support for the significance of the cluster comes from the independent Consurf analysis mapping sequence conservation (not evolutionary coupling) onto the protein (Fig. 15.3d): the strongly coupled cluster overlaps a highly conserved area of protein surface.

### 15.3.3 Plasticity of Catalytic Site Residues

Despite the efforts undertaken by the Structural Genomics initiatives to cover the protein fold space by providing structural templates for all existing protein families, there are cases where the sequence similarity criterion is insufficient to assign any defined functionality to the analyzed family. In many cases, however, the protein structure can be inferred with the aid of protein fold-recognition methods, alone or in combination with de novo modelling (Kolinski and Bujnicki 2005) and then used to pinpoint the potential active site, suggesting a possible function. This can be exemplified by the published analysis (Feder and Bujnicki 2005) of family of sequences grouped together in the Clusters of Orthologous Groups (COG) database (Tatusov et al. 2003) as COG4636 and annotated as "uncharacterized protein conserved in Cyanobacteria". The detailed analysis of sequence conservation within COG4636 family combined with secondary structure prediction revealed a pattern of α-helices and β-strands associated with conserved carboxylate residues, which has been previously identified in the PD-(D/E)XK superfamily of nucleases (Bujnicki 2003).This similarity suggested that members of COG4636 may belong to the PD-(D/E)XK superfamily (Fig. 15.4a, b). However, the multiple sequence alignment revealed that only the "PD" half-motif is nearly perfectly conserved, while a critical Lys residue is missing from the second half-motif "(D/E)XK". Specifically, instead of the Lys residue most members of COG4636 possessed a hydrophobic amino-acid, such as Leu or Val. One possibility was therefore that this family was not related at all to PD-(D/E)XK proteins. Another possibility was that they are related to these nucleases, but they lost the active site residue and became catalytically inactive. A third possibility was that the function of the "missing" Lys residue was be taken over by another residue, but based only on the sequence alignment it was not possible to identify which of the other residues could fulfil this role. Could structure predictions enable the true function of COG4636 to be determined?

First, a fold-recognition analysis of COG4636 sequence supported the prediction that they are indeed related to PD-(D/E)XK enzymes. A comparative model was then built based on a structure of a bona-fide PD-(D/E)XK nuclease and analyzed for the presence of spatially adjacent conserved residues. Analysis of the model revealed that in COG4636 the missing Lys residue had been replaced by another Lys residue that has appeared in a distinct region in the sequence (Fig. 15.4c). The replacement Lys could place its functional group in the same spatial position as the catalytic Lys residue of the templates thereby allowing the completion of the PD-(D/E)XK motif in three dimensions, despite the lack of sequence conservation. This allowed for a strong prediction, unavailable from purely sequence analyses, that COG4636 indeed contained active nucleases. Later on the correctness of the prediction of unusual configuration of the active site was confirmed by crystallographic analysis of another member of the COG4636 family (Fig. 15.4d; PDB code 1wdj) as well as by identification of other bona fide nucleases with the same spatial rearrangement of the active site (Tamulaitiene et al. 2006).

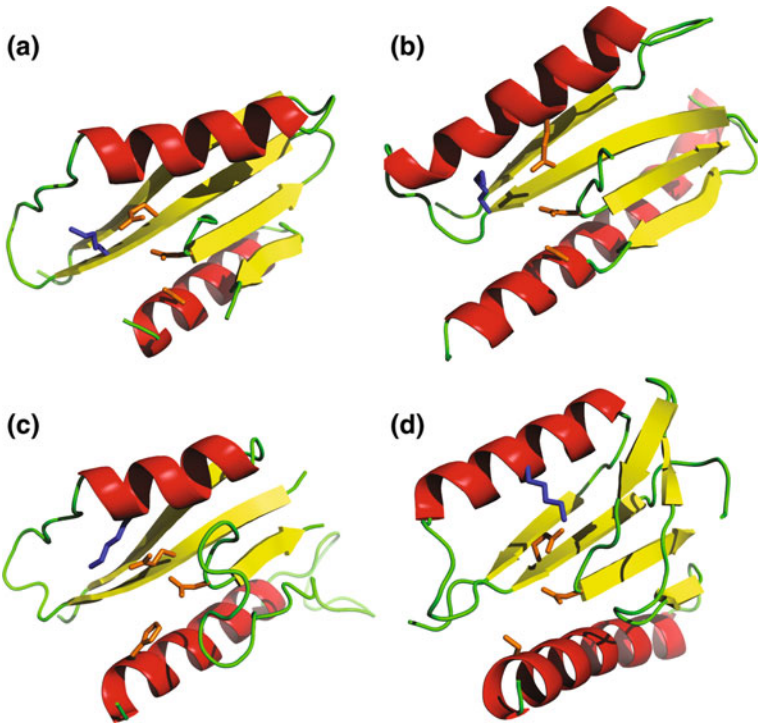



**Fig. 15.4** Spatial conservation of the PD-(D/E)XK active site. Only the structurally superimposed common cores are shown, terminal regions and insertions have been omitted for clarity of the presentation. The *upper panels* show bona fide PD-(D/E)XK nucleases—**a** Holliday junction resolvase Hje (PDB code 1ob8) and **b** REase Ngo-MIV (PDB code 1fiu). The *lower panels* present COG4636 structures—**c** a theoretical model (Feder and Bujnicki 2005) and **d** crystal structure of another member of the family (PDB code 1wdj). The side chains of the typical and variant PD-(D/E)xK active site are *coloured orange* except the Lys residues which are *coloured blue*

### *15.3.4 Prediction of Ligand Specificity*

One of the most basic function predictions that can be obtained from a protein model is ligand specificity. Frequently, if the structure of protein A bound to ligand X is known, it is of interest to predict whether protein B, homologous to A, shares the same specificity as X, or in fact binds a different ligand Y. These analyses rely on the assumption that the binding sites of A and B are similarly positioned. This is usually the case between homologous proteins and the presence of key catalytic residues nearby, in the case for enzymes, often offers confirmation. A comparative model is then made B, based on the structure of A. Examination of the modelled B structure, and in particular its comparison with the template A, should show whether the binding site appears to have changed. A reduction in size, for example, would lead to the prediction of a smaller ligand.

An early example of work in this area was modelling of brain lipid-binding protein (BLBP), based on the related fatty-acid binding protein structures (Xu et al. 1996). Interactions of known fatty acid ligands of BLBP were modelled in an effort to discover the molecular basis of the 20-fold tighter binding of docosahexaenoic acid relative to the shorter oleic and arachidonic acids. The model revealed that the two extra carbon atoms of the former fatty acid could be accommodated in the pocket of BLBP, making additional favourable hydrophobic interactions. The calculated additional binding free energy, based on the size of the additional hydrophobic contact area, of around 2 kcal/mol correlated nicely with the difference in affinity. With the model validated in this way, the authors were able to predict that still larger fatty acids would not be able to make additional contacts and would therefore not bind any more tightly.

The molecular bases of different specificities may sometimes be surprisingly simple. Such is the case with the phospho donors of some 6-phosphofructokinases (PFKs). PFK is a glycolytic enzyme catalysing the transfer of a phospho group from a donor, which may be ATP, ADP or inorganic pyrophosphate (PPi). The ATP- and PPi-dependent enzymes share an evolutionary relationship, while ADP-dependent PFKs belong to a different structural class. It was noticed early on that the ATP-dependent enzymes from trypanosomatids bore a closer relationship to certain PPi-dependent enzymes that they did to the better-characterised ATP-dependent enzymes from bacteria and mammals (Michels et al. 1997). Modelling later revealed that the basis for ATP or PPi specificity could be pinned down to a single amino-acid which was Gly in the ATP enzymes but Asp in the PPi enzymes (Lopez et al. 2002). As shown in Fig. 15.5, an Asp at this position clashes sterically and electrostatically with the α-phosphate of bound ADP or ATP, reducing the binding site to a size that can only accommodate PPi as phospho donor. The conversion of a PPi-dependent enzyme to an ATP-dependent one by the replacement of the Asp at this position with a Gly confirms the dramatically simple origin of specificity in this case (Chi and Kemp 2000).

### *15.3.5 Prediction of Cofactor Specificity Using an Entry from a Database of Models*

As mentioned above, databases such as Swiss-Model Repository (Kiefer et al. 2009) and ModBase (Pieper et al. 2014b) automatically calculate comparative models of protein sequences periodically. These models eliminate the effort required of the user even to use a web-service for comparative modelling, much less to carry out modelling on their own computer. The models are accompanied by quality indicators and colour coding providing the user with a rapid indication of their reliability. The user should obviously consider the age of the model: there is always the possibility that a better template structure, allowing for higher quality modelling, has emerged since the model was constructed. Nevertheless, the scale of

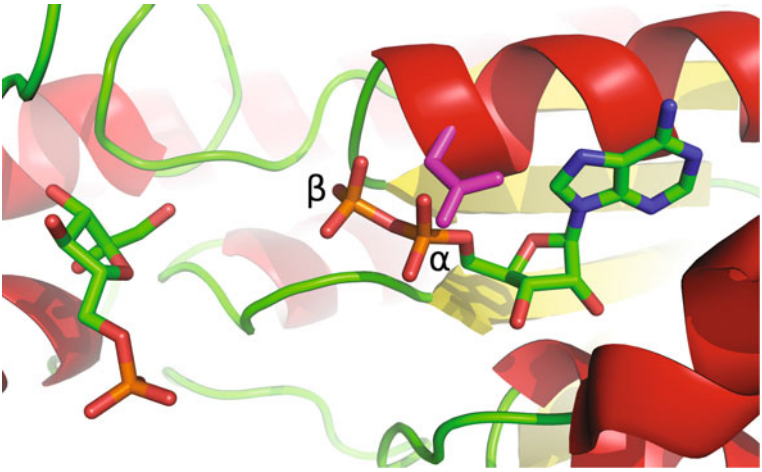



**Fig. 15.5** Catalytic site of *E. coli* 6-phosphofructokinase bound to fructose-6-phosphate (F6P) and adenosine diphosphate (ADP) (PDB code 4pfk). Ligands are shown as *coloured sticks* (F6P on left, ADP on right). ATP-dependent enzymes, like that from *E. coli*, have a Gly at the catalytic site (not shown). Modelling of an Asp residue at the same position (*magenta*), as found in PPi-dependent enzymes, shows that it is responsible for changing the specificity for phospho donor (see text)

these automatic modelling exercises and the ease with which the models are obtained means that their usefulness for function prediction should be seriously assessed.

As a test case, we consider here the new family of glucose-6-phosphate dehydrogenases (Glc6PDH) that have recently been characterised (Pickl and Schonheit 2015). They form part of a newly identified oxidative pentose phosphate pathway in archaea. Interestingly, a 6-phosphogluconate dehydrogenase (6PGDH) was readily identifiable and correctly annotated in the genomes studied, but the novel Glc6PDH was only identified after purification of the protein and peptide mass fingerprinting. This led to the identification of HVO_0511, until then misannotated as an epimerase/dehydratase of the SDR (short-chain dehydrogenase/reductase) superfamily, as the genome locus encoding the Glc6PDH enzyme in *Haloferax volcanii*. Conventional Glc6PDH (and 6PGDH) enzymes are $NADP^+$-dependent enzymes, but assay showed that the archaeal enzyme was specific for $NAD^+$ as a cofactor. The SDR superfamily encompasses both $NAD^+$ and $NADP^+$-dependent enzymes (Kavanagh et al. 2008). Could an automatically-generated model have predicted this preference?

The UniProt entry (UniProt Consortium 2015) for the *H. volcanii* enzyme (D4GS48_HALVD) contains a link to a ModBase (Pieper et al. 2014a) search which reveals the existence of an automatically generated model for residues 5–252 of the 262 residue protein. Colour-coded quality indicators are all green suggesting that the model would be good enough for structure-based function inference.

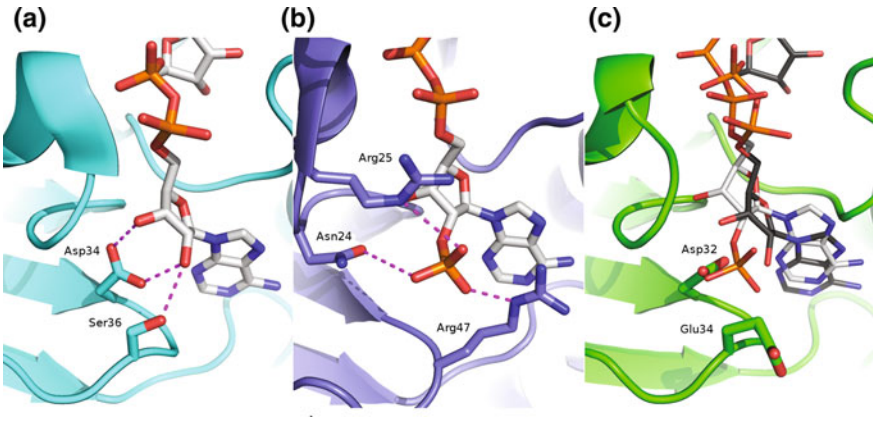



**Fig. 15.6** Predicting cofactor specificity of archaeal Glc6PDH enzymes using a ModBase model. The cofactor-binding sites are shown for **a** $NAD^+$-dependent *Agrobacterium tumefaciens* uronate dehydrogenase (PDB code 3rfv), **b** $NADP^+$-dependent *Mycobacterium tuberculosis* MabA reductase (PDB code 1uzn) and **c** the ModBase model of *H. volcanii* Glc6PDH (Uniprot identifier D4GS48_HALVD). Cofactors are shown as sticks along with key specificity-determining residues. In the model structure superimposed cofactors from both crystal structures are shown as thin sticks, along with residues predicted to be important in cofactor specificity

Figure 15.6 shows a comparison of the cofactor binding sites of $NAD^+$-dependent uronate dehydrogenase from *Agrobacterium tumefaciens* (Parkkinen et al. 2011; PDB code 3rfv; the template used to make the model was the corresponding cofactor-free form with PDB code 3rft), $NADP^+$-dependent *Mycobacterium tuberculosis* MabA reductase (Cohen-Gonsaud et al. 2002; PDB code 1uzn; the most similar structurally characterised $NADP^+$-dependent enzyme) and the ModBase model of *H. volcanii* Glc6PDH. Specific $NAD^+$ and $NADP^+$ recognition is encoded by hydrogen-bonding and electrostatic interactions between protein and, respectively, either the unmodified or phosphorylated 2′ OH group of the cofactor adenine moiety. In addition, the negatively charged Asp residue employed for recognition of the unmodified 2′ OH group of $NAD^+$ would electrostatically repel the phospho group borne by $NADP^+$. The model of *H. volcanii* Glc6PDH contains the key Asp32 to potentially recognise the unmodified 2′ OH and the phospho group, although its orientation is not suitable for ideal hydrogen bonding, presumably because cofactor was not present in the template structure. It appears not to be capable of making the additional hydrogen bond, as Ser36 does in uronate dehydrogenase, but Glu34 present at the corresponding position would provide additional electrostatic repulsion of $NADP^+$. Overall, the model resembles neither of the comparator structures exactly reinforcing the desirability of a 3D atomic view rather than relying on a purely sequence-based comparison. According to that structural analysis, the model's cofactor binding site would certainly be predicted to have a preference for $NAD^+$ over $NADP^+$, in accord with the experimental data.

### *15.3.6 Mutation Mapping*

Rare mutations in important proteins underlie many genetic diseases. Similarly, allelic variations in drug targets can lead to differential drug binding and hence to different drug responses by patients. Structural mapping of mutations, a key use of molecular models, is therefore useful for understanding molecular mechanisms of disease as well as predicting patient responses as a step towards personalised medicine.

ATP-sensitive potassium (KATP) channels play key roles in many tissues by linking cell metabolism to electrical activity. KATP channels are octameric complexes of two different proteins Kir6.2 and SUR. Binding of ATP or ADP to a KATP channel causes its inhibition. The identification of the number of mutations in Kir6.2 leading to reduced ATP sensitivity of the channel has turned out to be the cause of permanent neonatal diabetes (Hattersley and Ashcroft 2005). In pancreatic ß-cells the inhibited KATP channel causes membrane hyperpolarization which in turn leads to a reduction in insulin secretion and, consequently, diabetes. The diagnosis of the genetic etiology of the disease has revolutionized therapy for patients with neonatal diabetes resulting from Kir6.2 mutations, as those channels can still be closed by therapeutics such sulfonylureas and glinides and the insulin treatment could be limited or discontinued. The homology model of Kir6.2 subunit allowed for the spatial mapping of the residues mutated in the neonatal diabetes and thus illustrated the molecular mechanism underlying reduced KATP sensitivity (Hattersley and Ashcroft 2005). Patients carrying mutations in Kir6.2 exhibit spectrum of phenotypes that are directly correlated to the nature of mutation. For example patients with neurological symptoms carry the mutations which do not directly impair ATP binding but markedly bias the channel toward the open state and thus reduce the ability of ATP to block the channel (ATP stabilizes the closed state of the channel).

Studies showed that there is a group of patients with permanent neonatal diabetes, carrying the L164P mutation in Kir6.2, who are unresponsive to sulfonylurea therapy (Tammaro et al. 2008). Analysis of the spatial L164 position reveals that this residue lies deep within the structure, 35 Å away from the ATP-binding site. It is therefore unlikely that it acts by reducing ATP binding directly. Instead, L164P probably destabilizes the closed state of the channel, to which sulfonylureas preferentially bind, and which is rarely reached in channels with enhanced channel open probability. Taken together, these results show that the drug response is dependent of the nature of particular mutation, but that it can be predicted by detailed analysis of a protein model.

### 15.3.7 Protein Complexes

A full understanding of the complex networks of protein-protein interactions that exist in cells is essential if systems biology, whereby these and other large-scale datasets are integrated into a meaningful whole, is ever to become a success. There is therefore much interest in adding predictions from comparative modelling to the battery of experimental and computational methods for prediction of protein-protein interaction (Aloy and Russell 2006). The template-free and template-based approaches were introduced above in Sect. 15.2.5 (see also Chap. 8).

An interesting large-scale application, incorporating template-based docking into an integrative approach, has been reported (Davis et al. 2007). In this work prediction of interactions were made for human proteins with proteins from the genomes of ten pathogenic organisms responsible for neglected diseases. The pathogen and host genomes were first scanned for proteins homologous to those known to interact. The pipeline proceeded when structural information for the interaction is not available by employing simple sequence similarity scores. This approach produced few predictions, however, since strict criteria were necessary in order for confident interaction prediction by this approach. More interesting and powerful was the explicit comparative modelling of the potential interaction partners based on protein complex templates. These modelled complexes were assessed using a statistical potential with favourable interactions passed on to a further ingenious filter. This employed known information about (sub-) cellular localization and function in order to eliminate from consideration interactions which could not occur in vivo. Thus, only host proteins known to be expressed in skin, lymph node or lung were considered as possible interaction partners for *Mycobacterium leprae* proteins. Pathogen proteins were also required to pass specific biological criteria. For *M. leprae*, for example, a protein had to have a relevant GO annotation (e.g. pathogenesis) or be annotated as being extracellular or surface located. The number of filtered predictions varied from 0 to 1501 between the pathogens. Rather few known interactions were available with which to benchmark the technique, but the method predicted four of the 33 interactions demonstrated at the time. In the remaining cases there was no available template to model the interaction suggesting that this lack was consistently responsible for the low coverage of known interactions (Davis et al. 2007). Interestingly, one prediction was experimentally validated: the method predicted the interaction of falcipain-2 and cystatin (PDB code 1yvb) based on the earlier structure of cathepsin-H bound to stefin A (PDB code 1nb3) (Fig. 15.7). The two enzymes share around 24% sequence identity while the inhibitors are around 11% identical. The success of the prediction in the face of these sequence differences and considerable structural variation (Fig. 15.7) illustrates the power of the methodology.

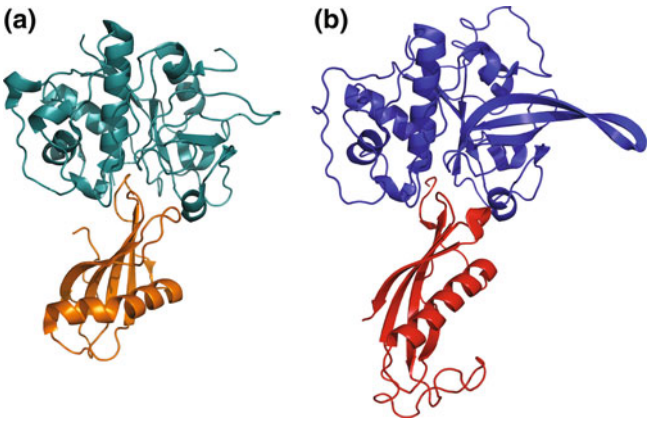



**Fig. 15.7** Modelling-based prediction of protein-protein interactions. A pipeline based on comparative modelling of protein complexes (Davis et al. 2006) was able to use the structure of cathepsin A in complex with stefin A (PDB code 1nb3 **a**) to infer a probable interaction between falcipain-2 and cystatin, as confirmed by crystallography (PDB code 1yvb **b**). Enzymes are shown above, and inhibitors below, in each panel

### *15.3.8 Structure Modelling of Alternatively Spliced Isoforms*

Many, if not most, eukaryotic genes are alternatively spliced, dramatically increasing the diversity of transcripts. It is often difficult to predict from the sequences of alternatively spliced transcripts whether function is retained or modified. Structure modelling, where possible, can shed light on the structure-function relationship among alternatively spliced transcripts from a single gene.

Early work by Furnham et al. (2004), involving 40 splice variant models of 14 proteins, showed that exon loss often involved loss of complete structural units rather than small regions. The authors showed that deletions were more reliably modelled, according to structure validation software, than insertions. For four proteins with biomedical implications the authors could correlate known function properties of splice variants with their modelled structures. Later Wang et al. (2005) showed that boundaries of splicing events tend to lie both in coil regions (rather than within elements of regular secondary structure) and at the protein surface. Splicing events were generally few in number for a particular gene, 1 or 2, and small in size, with 60% affecting 50 residues or fewer. These findings suggested that splicing tends to occur in positions and in ways that perturb only minimally the protein tertiary structure consistent with most alternative isoforms having folding properties similar to the original form and thus potential functionality. However, a later study (Tress et al. 2007), in which fewer transcripts were analysed in structural terms, revealed that many alternatively spliced isoforms would have to have dramatically different structures to determined structures of other isoforms. For fully 49 of 85 transcripts mapped onto homologous structures, the authors inferred that

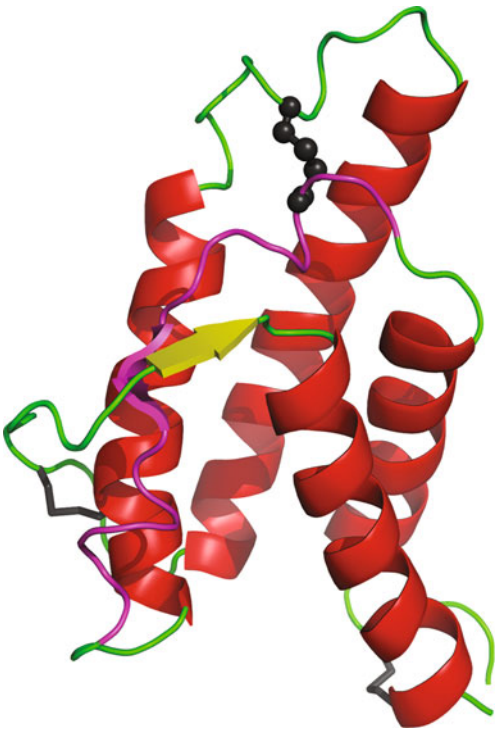


**Fig. 15.8** Structure of interleukin 4 showing the portion encoded by exon 2. The experimental structure (PDB core 1ilt) is shown as *coloured* cartoon, with exon 2-encoded protein *coloured magenta*. Disulphide bridges are shown as sticks, with the bridge contributed to by exon 2-encoded protein shown as ball-and-stick

isoform and principal sequences would adopt substantially different structures. An example, taken from Tress et al. (2007) and shown in Fig. 15.8, relates to an isoform of interleukin 4 lacking exon 2. The structural region encoded by that exon contributes to both the folding core of the protein and to a disulphide bridge, showing that the 3D structure of the isoform must be substantially different to the determined structure of the complete protein. As yet, we have only a very incomplete larger scale picture of the functional consequences of structural changes —minor and major—due to alternative splicing. For example, for only four of 214 loci could experimental data illustrating functional differences between splice isoforms be found by Tress et al. (2007).

### *15.3.9 From Broad Function to Molecular Details*

Protein function can be considered on different complexity levels—ranging from the involvement into the cellular processes to the knowledge of the mode of action on the molecular level. Lysosomal deoxyribonuclease II α (DNase IIα) was one of the earliest endonucleases identified (1947), with considerable biochemical characterization reported already in the 1960s. This enzyme is indispensable for the

organism development as it is responsible for DNA waste removal and auxiliary apoptotic DNA fragmentation in higher eukaryotes—the knockout of murine lysosomal DNase IIα turned out to be lethal. Despite the intensive research for over 50 years and unquestionable importance of DNase IIα no similarity to any other protein family could be detected, hampering function studies on the molecular level for this protein. No Fold Recognition method reported any target-template alignment with a score above the documented level of significance, but analysis of their results revealed that several of them reported a similarity to the phospholipase D (PLD) fold in the region comprising part of the active site—the so called HxK motif (Cymerman et al. 2005). Known members of the PLD superfamily possess a bilobed structure, with a single active site composed of two "HxK-Xn-N-Xn-(E/Q/D)" motifs located at the interface between two domains. Based on the alignments alone it was not possible to define the remaining residues of the active site. Analysis of the placement of particular residues in the structural model however, delivered the essential information and allowed for the selection of amino acids that potentially could serve for the formation of the catalytic centre (Fig. 15.9). The finding that DNase IIα is a remote relative of phospholipase D was later confirmed by experimental studies (Schafer et al. 2007) and explained unusual features of this nuclease, such as its resistance to EDTA. By similarity to PLD whose mechanism has been elucidated, it was also possible to infer that the reaction

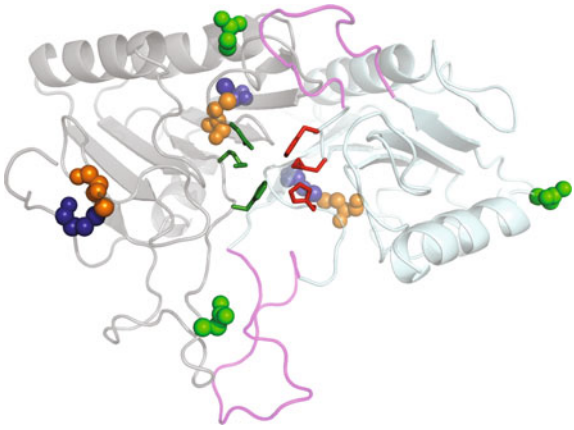


**Fig. 15.9** Structural model of human DNase IIα. The computational analyses enabled the assignment of DNase IIα as member of PLD family. The enzyme adopts a monomeric structure with a pseudodimeric architecture. The two HxK motifs in the N (*cartoon light blue* representation) and in the C-terminal (*cartoon grey* representation) domains contain the catalytically relevant amino acid residues (*red and green sticks*), which collectively form a single active site. In addition to the identities of putative catalytic residues, the structural model accounts for the proximal positions of cysteine residue disulphide bonds (*orange and dark blue balls*), and the exposed character of N-glycosylated residues (represented as *green balls*). Putative DNA-binding loops are shown in *magenta*. The identification of the functionally important residues in the theoretical model can greatly facilitate the process of enzyme engineering

of phosphodiester bond hydrolysis by DNase IIα will proceed by a covalently linked reaction intermediate. The case of DNase IIα exemplifies the bioinformatics can bypass some experimental limitations (DNase IIα is heavily glycosylated making the enzyme resistant to the crystallization) and thereby allow further exploration of the protein properties.

## 15.4 Conclusions

Using homology models to guide experiments or interpret existing biological data shows no sign of losing importance or popularity. Since 1985 when a single PubMed entry can be found with the phrases 'homology model(l)ing' or 'comparative model(l)ing' (albeit one in a high-profile journal; Greer 1985) annual records rise to around 270 at the time of the first edition of this text. Since then they have risen further, averaging around 450 per year recently. This reflects the generality of the method, the increasing availability of templates in an expanding PDB (Rose et al. 2015), and the increasing availability of automated methods of model construction which facilitate access of novice users to structure-based methods. Recent innovations under the ambit of the Protein Model Portal (Haas et al. 2013)—comprising links to model databases, modelling servers and MQAPs as well as entry level documentation—should encourage uptake still further. At the other end of the spectrum, we will no doubt see more examples of papers in which homology models, treated in an integrated manner with diverse experimental data, help yield insights into complex molecular systems (Bui et al. 2013). For the diminishing set of folds for which absence from the PDB precludes homology modelling, we can look forward to ab initio and especially contact-driven modelling filling the gap, as ever more inexpensive sequencing continues to drive expansion of the protein databases. A wider availability of protein models in coming years, combined with an improving understanding of their value but also their limitations, add up to a bright future for the area.

## References

Aloy P, Russell RB (2006) Structural systems biology: modelling protein interactions. Nat Rev Mol Cell Biol 7:188–197

Anishchenko I, Kundrotas PJ, Tuzikov AV et al (2014) Protein models: the grand challenge of protein docking. Proteins 82(2):278–287

Barth P, Schonbrun J, Baker D (2007) Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci U S A 104(40):15682–15687

Baxter SM, Rosenblum JS, Knutson S et al (2004) Synergistic computational and experimental proteomics approaches for more accurate detection of active serine hydrolases in yeast. Mol Cell Proteomics 3:209–225

Bonneau R, Strauss CE, Rohl CA et al (2002) *De novo* prediction of three-dimensional structures for major protein families. J Mol Biol 322:65–78
Bordogna A, Pandini A, Bonati L (2011) Predicting the accuracy of protein–ligand docking on homology models. J Comput Chem 32(1):81–98
Bradley P, Misura KM, Baker D (2005) Toward high-resolution *de novo* structure prediction for small proteins. Science 309:1868–1871
Brylinski M, Skolnick J (2009) FINDSITE: a threading-based approach to ligand homology modeling. PLoS Comput Biol 5(6):e1000405
Brylinski M, Skolnick J (2010) Comprehensive structural and functional characterization of the human kinome by protein structure modeling and ligand virtual screening. J Chem Inf Model 50(10):1839–1854
Bui KH, von Appen A, DiGuilio AL et al (2013) Integrated structural analysis of the human nuclear pore complex scaffold. Cell 155(6):1233–1243
Bujnicki JM (2003) Crystallographic and bioinformatic studies on restriction endonucleases: inference of evolutionary relationships in the "midnight zone" of homology. Curr Protein Pept Sci 4:327–337
Bumbaca D, Littlejohn JE, Nayakanti H et al (2007) Genome-based identification and characterization of a putative mucin-binding protein from the surface of Streptococcus pneumoniae. Proteins 66:547–558
Cammer SA, Hoffman BT, Speir JA et al (2003) Structure-based active site profiles for genome analysis and functional family subclassification. J Mol Biol 334:387–401
Chakravarty S, Sanchez R (2004) Systematic analysis of added-value in simple comparative models of protein structure. Structure 12:1461–1470
Chakravarty S, Wang L, Sanchez R (2005) Accuracy of structure-derived properties in simple comparative models of protein structures. Nucleic Acids Res 33:244–259
Chi A, Kemp RG (2000) The primordial high energy compound: ATP or inorganic pyrophosphate? J Biol Chem 275:35677–35679
Chothia C, Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4):823–826
Chung SY, Subbiah S (1996) A structural explanation for the twilight zone of protein sequence homology. Structure 4(10):1123–1127
Cohen-Gonsaud M, Ducasse S, Hoh F et al (2002) Crystal structure of MabA from Mycobacterium tuberculosis, a reductase involved in long-chain fatty acid biosynthesis. J Mol Biol 320(2):249–261
Cymerman IA, Meiss G, Bujnicki JM (2005) DNase II is a member of the phospholipase D superfamily. Bioinformatics 21:3959–3962
Davis FP, Braberg H, Shen MY et al (2006) Protein complex compositions predicted by structural similarity. Nucleic Acids Res 34:2943–2952
Davis FP, Barkan DT, Eswar N et al (2007) Host pathogen protein interactions predicted by comparative modeling. Protein Sci 16:2585–2596
Dessailly BH, Nair R, Jaroszewski L et al (2009) PSI-2: structural genomics to cover protein domain family space. Structure 17(6):869–881
Du Y, He YX, Zhang ZY et al (2011) Crystal structure of the mucin-binding domain of Spr1345 from Streptococcus pneumoniae. J Struct Biol 174(1):252–257
Du H, Brender JR, Zhang J et al (2015) Protein structure prediction provides comparable performance to crystallographic structures in docking-based virtual screening. Methods 71:77–84
Fan H, Irwin JJ, Webb BM et al (2009) Molecular docking screens using comparative models of proteins. J Chem Inf Model 49(11):2512–2527
Feder M, Bujnicki JM (2005) Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site. BMC Genom 6:21
Fetrow JS, Skolnick J (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J Mol Biol 281:949–968

Fetrow JS, Godzik A, Skolnick J (1998) Functional analysis of the Escherichia coli genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/thioredoxin disulfide oxidoreductase activity. J Mol Biol 282:703–711

Furnham N, Ruffle S, Southan C (2004) Splice variants: a homology modeling approach. Proteins 54:596–608

Gao M, Skolnick J (2009) From nonspecific DNA–protein encounter complexes to the prediction of DNA–protein interactions. PLoS Comput Biol 5(3):e1000341

Ginalski K (2006) Comparative modeling for protein structure prediction. Curr Opin Struct Biol 16 (2):172–177

Greer J (1985) Model structure for the inflammatory protein C5a. Science 228(4703):1055–1060

Haas J, Roth S, Arnold K, et al (2013) The protein model portal—a comprehensive resource for protein structure and model information. Database (Oxford) 2013:bat031

Hasegawa K, Funatsu K (2012) A new method for mapping the molecular surface of a protein structure using a spherical self-organizing map. Mol Inf 31(2):161–166

Hattersley AT, Ashcroft FM (2005) Activating mutations in Kir6.2 and neonatal diabetes: new clinical syndromes, new scientific insights, and new therapy. Diabetes 54:2503–2513

Hermann JC, Marti-Arbona R, Fedorov AA et al (2007) Structure-based activity prediction for an enzyme of unknown function. Nature 448:775–779

Hopf TA, Colwell LJ, Sheridan R et al (2012) Three-dimensional structures of membrane proteins from genomic sequencing. Cell 149(7):1607–1621

Jacobson M, Sali A (2004) Comparative protein structure modelling and its applications to drug discovery. Annu Rep Med Chem 39:259–274

Kamisetty H, Ovchinnikov S, Baker D (2013) Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. Proc Natl Acad Sci U S A 110(39):15674–15679

Kavanagh KL, Jornvall H, Persson B et al (2008) Medium- and short-chain dehydrogenase/reductase gene and protein families: the SDR superfamily: functional and structural diversity within a family of metabolic and regulatory enzymes. Cell Mol Life Sci 65 (24):3895–3906

Khafizov K, Madrid-Aliste C, Almo SC et al (2014) Trends in structural coverage of the protein universe and the impact of the protein structure initiative. Proc Natl Acad Sci U S A 111 (10):3733–3738

Kiefer F, Arnold K, Kunzli M, et al (2009) The SWISS-MODEL repository and associated resources. Nucleic Acids Res 37(Database issue):D387–D392

Kolinski A, Bujnicki JM (2005) Generalized protein structure prediction based on combination of fold-recognition with *de novo* folding and evaluation of models. Proteins 61(Suppl 7):84–90

Kryshtafovych A, Fidelis K, Tramontano A (2011) Evaluation of model quality predictions in CASP9. Proteins 79(Suppl 10):91–106

Kundrotas PJ, Vakser IA (2010) Accuracy of protein-protein binding sites in high-throughput template-based modeling. PLoS Comput Biol 6(4):e1000727

Kundrotas PJ, Zhu Z, Janin J et al (2012) Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A 109(24):9438–9441

Lee HS, Zhang Y (2012) BSP-SLIM: A blind low-resolution ligand-protein docking approach using predicted protein structures. Proteins Struct Funct Bioinf 80(1):93–110

Lee TT, Agarwalla S, Stroud RM (2004) Crystal structure of RumA, an iron-sulfur cluster containing E. coli ribosomal RNA 5-methyluridine methyltransferase. Structure 12(3):397–407

Li S, Yamashita K, Amada KM et al (2014) Quantifying sequence and structural features of protein-RNA interactions. Nucleic Acids Res 42(15):10086–10098

Liu J, Fang C, Jiang Y et al (2009) Characterization of a hemolysin gene ytjA from Bacillus subtilis. Curr Microbiol 58(6):642–647

Lopez C, Chevalier N, Hannaert V et al (2002) Leishmania donovani phosphofructokinase. Gene characterization, biochemical properties and structure-modeling studies. Eur J Biochem 269:3978–3989

Lukk T, Sakai A, Kalyanaraman C et al (2012) Homology models guide discovery of diverse enzyme specificities among dipeptide epimerases in the enolase superfamily. Proc Natl Acad Sci U S A 109(11):4122–4127

Malmstrom L, Riffle M, Strauss CE et al (2007) Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 5:e76

Marks DS, Colwell LJ, Sheridan R et al (2011) Protein 3D structure computed from evolutionary sequence variation. PLoS ONE 6(12):e28766

McGovern SL, Shoichet BK (2003) Information decay in molecular docking screens against holo, apo, and modeled conformations of enzymes. J Med Chem 46:2895–2907

Michels PA, Chevalier N, Opperdoes FR et al (1997) The glycosomal ATP-dependent phosphofructokinase of Trypanosoma brucei must have evolved from an ancestral pyrophosphate-dependent enzyme. Eur J Biochem 250:698–704

Negroni J, Mosca R, Aloy P (2014) Assessing the applicability of template-based protein docking in the twilight zone. Structure 22(9):1356–1362

Oshiro C, Bradley EK, Eksterowicz J et al (2004) Performance of 3D-database molecular docking studies into homology models. J Med Chem 47:764–767

Parkkinen T, Boer H, Janis J et al (2011) Crystal structure of uronate dehydrogenase from Agrobacterium tumefaciens. J Biol Chem 286(31):27294–27300

Pickl A, Schonheit P (2015) The oxidative pentose phosphate pathway in the haloarchaeon Haloferax volcanii involves a novel type of glucose-6-phosphate dehydrogenase—the archaeal Zwischenferment. FEBS Lett

Piedra D, Lois S, de la Cruz X (2008) Preservation of protein clefts in comparative models. BMC Struct Biol 8:2-6807-8-2

Pieper U, Schlessinger A, Kloppmann E et al (2013) Coordinating the impact of structural genomics on the human [alpha]-helical transmembrane proteome. Nat Struct Mol Biol 20 (2):135–138

Pieper U, Webb BM, Dong GQ, et al (2014a) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42(Database issue): D336–D346

Pieper U, Webb BM, Dong GQ, et al (2014b) ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res 42(Database issue): D336–D346

Rigden DJ, Galperin MY (2008) Sequence analysis of GerM and SpoVS, uncharacterised bacterial 'sporulation' proteins with widespread phylogenetic distribution. Bioinform. doi:10.1093/bioinformatics/btn314 (accepted)

Rodrigues J, Melquiond A, Karaca E et al (2013) Defining the limits of homology modeling in information-driven protein docking. Proteins Struct Funct Bioinf 81(12):2119–2128

Rose PW, Prlic A, Bi C, et al (2015) The RCSB protein data bank: views of structural biology for basic and applied research and education. Nucleic Acids Res 43(Database issue):D345–D356

Roy A, Yang J, Zhang Y (2012) COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 40(Web Server issue):W471–W477

Schafer P, Cymerman IA, Bujnicki JM et al (2007) Human lysosomal DNase IIalpha contains two requisite PLD-signature (HxK) motifs: evidence for a pseudodimeric structure of the active enzyme species. Protein Sci 16:82–91

Schwede T (2013) Protein modeling: what happened to the "protein structure gap"? Structure 21 (9):1531–1540

Schwede T, Sali A, Honig B et al (2009) Outcome of a workshop on applications of protein models in biomedical research. Structure 17(2):151–159
Shortle D, Simons KT, Baker D (1998) Clustering of low-energy conformations near the native structures of small proteins. Proc Natl Acad Sci U S A 95(19):11158–11162
Skolnick J, Zhou H, Gao M (2013) Are predicted protein structures of any value for binding site prediction and virtual ligand screening? Curr Opin Struct Biol 23(2):191–197
Sodhi JS, Bryson K, McGuffin LJ et al (2004) Predicting metal-binding site residues in low-resolution structural models. J Mol Biol 342:307–320
Soding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33(Web Server issue):W244–W248
Song L, Kalyanaraman C, Fedorov AA et al (2007) Prediction and assignment of function for a divergent N-succinyl amino acid racemase. Nat Chem Biol 3:486–491
Song Y, DiMaio F, Wang RY et al (2013) High-resolution comparative modeling with Rosetta CM. Structure 21(10):1735–1742
Szilagyi A, Skolnick J (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. J Mol Biol 358:922–933
Szilagyi A, Zhang Y (2014) Template-based structure modeling of protein-protein interactions. Curr Opin Struct Biol 24:10–23
Tammaro P, Flanagan SE, Zadek B et al (2008) A Kir6.2 mutation causing severe functional effects in vitro produces neonatal diabetes without the expected neurological complications. Diabetologia
Tamulaitiene G, Jakubauskas A, Urbanke C et al (2006) The crystal structure of the rare-cutting restriction enzyme SdaI reveals unexpected domain architecture. Structure 14:1389–1400
Tatusov RL, Fedorova ND, Jackson JD et al (2003) The COG database: an updated version includes eukaryotes. BMC Bioinform 4:41
Tovchigrechko A, Wells CA, Vakser IA (2002) Docking of protein models. Protein Sci 11 (8):1888–1896
Tress ML, Martelli PL, Frankish A et al (2007) The implications of alternative splicing in the ENCODE protein complement. Proc Natl Acad Sci U S A 104:5495–5500
Tuncbag N, Keskin O, Nussinov R et al (2012) Fast and accurate modeling of protein-protein interactions by combining template-interface-based docking with flexible refinement. Proteins 80(4):1239–1249
UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Res 43 (Database issue):D204–D212
Vakser IA (2013) Low-resolution structural modeling of protein interactome. Curr Opin Struct Biol 23(2):198–205
Vreven T, Hwang H, Pierce BG et al (2014) Evaluating template-based and template-free protein-protein complex structure prediction. Brief Bioinform 15(2):169–176
Vroling B, Sanders M, Baakman C, et al (2011) GPCRDB: information system for G protein-coupled receptors. Nucleic Acids Res 39(Database issue):D309–D319
Wallrapp FH, Pan JJ, Ramamoorthy G et al (2013) Prediction of function for the polyprenyl transferase subgroup in the isoprenoid synthase superfamily. Proc Natl Acad Sci U S A 110 (13):E1196–E1202
Wang P, Yan B, Guo JT et al (2005) Structural genomics analysis of alternative splicing and application to isoform structure modeling. Proc Natl Acad Sci U S A 102:18920–18925
Xu D, Zhang Y (2012) Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 80(7):1715–1735
Xu D, Zhang Y (2013) Ab initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. Sci Rep 3:1895
Xu LZ, Sanchez R, Sali A et al (1996) Ligand specificity of brain lipid-binding protein. J Biol Chem 271:24711–24719
Yang J, Roy A, Zhang Y (2013) Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29 (20):2588–2595

Zhang Y, Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57(4):702–710

Zhang QC, Petrey D, Deng L et al (2012) Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 490(7421):556–560

Zhao J, Dundas J, Kachalo S et al (2011) Accuracy of functional surfaces on comparatively modeled protein structures. J Struct Funct Genomics 12(2):97–107

# Index





D.J. Rigden (ed.), *From Protein Structure to Function with Bioinformatics*,
DOI 10.1007/978-94-024-1069-3