Focus on Structural Biology 8

Irena Roterman-Konieczna
Editor

Identification of Ligand Binding Site and Protein-Protein Interaction Area

Springer

Identification of Ligand Binding Site and Protein-Protein Interaction Area

FOCUS ON STRUCTURAL BIOLOGY

Volume 8

Series Editor

ROBERT KAPTEIN

Bijvoet Center for Biomolecular Research,
Utrecht University, The Netherlands

For further volumes:
http://www.springer.com/series/5990

Irena Roterman-Konieczna
Editor

Identification of Ligand Binding Site and Protein-Protein Interaction Area

Springer

Editor
Irena Roterman-Konieczna
Department of Bioinformatics and Telemedicine
Jagiellonian University
Medical College
Cracow, Poland

ISSN 1571-4853
ISBN 978-94-007-5284-9 ISBN 978-94-007-5285-6 (eBook)
DOI 10.1007/978-94-007-5285-6
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012951181

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword

The successful conclusion of the Human Genome Sequencing project, along with rapid progress in the development of analytical methods and high-performance computing solutions, has given rise to numerous biological databases of ever increasing volumes. Huge datasets, which nevertheless remain publicly accessible and affordable, are a crucial element of modern science. On the one hand, the ease with which research can be conducted is a great boon; on the other hand, however, one may feel somewhat overwhelmed by the immense quantity of available data. Such data is usually quite precise and detailed in nature, to the extent that modern scientific equipment and measuring devices allow. Information systems which assist in processing such data appear adequate, and their storage and processing capabilities – sufficient to meet the needs of modern researchers. Even so, further scientific breakthroughs are hindered by the relative lack of analysis methods targeted at large-scale datasets. This problem is particularly acute in analytical science, where it manifests itself as a general dearth of broad-scope methods with which to derive information (in the form of generalized models) approximating natural phenomena.

The above issue is the principal challenge in systems biology – a discipline which aims to develop comprehensive methods for simulating living organisms, so as to enable *in silico* experimentation on such organisms. A suitable system, properly reflecting the interactions and interdependencies observed in biological constructs, would support further research on specific anomalies, pathologies and diseases, well known to any clinician.

Before such a system can be designed and implemented, a fundamental biological axiom has to be addressed – namely, the relation between genetic information (genome) and the broad spectrum of active proteins, each of which facilitates a biological process, which, together, combine to form the extremely complex structure known as the organism.

Achieving this goal requires modeling three-dimensional structures of active proteins on the basis of their aminoacid sequences. The challenge lies not so much in predicting the structure itself, but rather in proposing a mechanism which leads to the generation of such structures. Another important issue, still waiting to be addressed, is the challenge of determining the biological function of a given protein. We would expect numerical methods (capable of predicting ligand binding sites or catalytic centers, where reaction substrates are processed) to also suggest the means by which such “active” sites are generated.

This handbook presents a review of numerical techniques used to identify ligand binding and protein complexation sites. It should be noted that there are many other theoretical studies devoted to predicting the activity of specific proteins and that useful protein data can be found in numerous databases. The aim of advanced computational techniques is to identify the active sites in specific proteins and moreover to suggest a generalized mechanism by which such protein-ligand (or protein-protein) interaction can be effected.

The project EFI similar to CASP and CAPRI has been initiated in regard to enzymatic active site recognition (http://enzymefunction.org).

Developing such tools is not an easy task – it requires extensive expertise in the area of molecular biology as well as a firm grasp of numerical modeling methods. Thus, it is often viewed as a prime candidate for interdisciplinary research. Gatenby R.A. and Maini P.K. (2003) postulate the creation of an entirely new branch of science called “mathematical ontology” (see “Cancer summed up”, *Nature*, 421, p. 321), which would bring together representatives of both – seemingly unconnected – disciplines. It is hoped that such close collaboration would lead to new systems enabling scientists to better simulate the properties and functioning of living organisms.

Cracow, 2012 Irena Roterman-Konieczna

Contents

Contributors

Paweł Alejster Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Cracow, Poland

Mateusz Banach Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Cracow, Poland

Adam S.Z. Belloum The Informatics Institue, University of Amsterdam, Amsterdam, The Netherlands

Emmmanuel Bettler Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Marian Bubak AGH University of Science and Technology Krakow, Poland and the Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Julie-Anne Chemelle Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Christophe Combet Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Reginald Cushing The Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Gilbert Deléage Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Christophe Geourjon Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Victor Guevara-Masis The Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Bingding Huang Systems Biology Division, Zhejiang-California International NanoSystems Institute, Zhejiang University, Hangzhou, China

Bioinformatics Group, Biotechnology Center, Technical University of Dresden, Dresden, Germany

Joël Janin IBBMC, Université Paris-Sud, Orsay, France

Wiktor Jurkowski Computational Biology Group, Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Esch-Belval, Luxembourg

Leszek Konieczny Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Cracow, Poland

Vladimir Korkhov Faculty of Applied Math and Control Processes, St. Petersburg State University , Saint Petersburg , Russia

Spiros Koulouzis The Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Damian Marchewka Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Cracow, Poland

Astronomy and Applied Computer Science, Jagiellonian University, Cracow, Poland

Irena Roterman-Konieczna Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College, Cracow, Poland

Raphaël Terreux Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France

Dmitry Vasunin Faculty of Applied Math and Control Processes, St. Petersburg State University , Saint Petersburg , Russia

Zhiming Zhao The Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Chapter 1
SuMo: A Tool for Protein Function Inference Based on 3D Structures Comparisons

Julie-Anne Chemelle, Emmmanuel Bettler, Christophe Combet, Raphaël Terreux, Christophe Geourjon, and Gilbert Deléage

Abstract The prediction of important residues for binding/recognition sites in protein 3D structures is still a matter of challenge. Indeed, binding sites recognition is generally based on geometry often combined with physico-chemical properties of the site since the conformation, size and chemical composition of the protein surface are all relevant for the interaction with a specific ligand. In our group, we designed an innovative bioinformatics method called SuMo in order to detect similar 3-dimensional (3D) sites in proteins (Jambon et al. Protein-Struct Funct Genet 52:137–145, 2003). This approach allowed the comparison of protein structures or substructures, and detected local spatial similarities: the main advantage of the method is its independence for both amino acid sequences and backbone structures. In contrast to already existing tools, the basis for this method is a representation of the protein structure by a set of stereo chemical groups that are defined independently from the notion of amino acid. An efficient heuristics for finding similarities has been developed which uses graphs of triangles of chemical groups to represent the protein structures. The SuMo (Surfing the Molecules) program allows the dynamic definition of chemical groups, the selection of sites in the proteins, and the management and screening of databases. The basic principle of SuMo has been used in several recent studies (Sperandio et al. J Cheml Inf Model 47:1097–1110, 2007)

J.-A. Chemelle • E. Bettler • C. Combet • R. Terreux • C. Geourjon • G. Deléage (✉)
Université Lyon 1, CNRS, UMR 5086; Bases Moléculaires et Structurales des Systèmes Infectieux, Lyon, France
e-mail: gilbert.deleage@ibcp.fr

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_1,

(Doppelt-Azeroual et al. Protein Sci 19:847–867, 2010). In order to give access to the SuMo tool, we proposed a web server (Jambon et al. Bioinformatics 21:3929–3930, 2005) reachable at http://sumo-pbil.ibcp.fr. This chapter will describe the main rationale we initially took for designing the first release of SuMo. In addition, we propose a completely new set of parameters best suitable for proteins and finally, we illustrate its power with several biological examples. Two of them dealing with serine proteases and lectins are given for a comparison purpose. The first two examples illustrate the capability of SuMo to deal with completely opposite modes of evolution i.e. convergence and divergence. A new biological application dealing with betalactame binding protein PBB molecules is also presented.

Keywords Proteins • Structural bioinformatics • 3D structure • Physico-chemical groups • 3D sites • Annotation • Triangle form • Delta-plus • Delta-minus • Glycine polar • Hydrophobic aliphatic • Carbon alpha • Hydrophobic aromatic • Target structure • Objects • Proteases • Isomerases • Lectins • Betalactam • Penicillin drug • Cephalosporin drug • Ceftazidime • Serine proteases • Protein-protein interaction • SuMo

1.1 Introduction

Understanding and predicting the function of proteins using bioinformatics traditionally falls into three levels of knowledge: amino acid sequence, backbone structure (also called fold comparison/recognition) and local arrangement of atoms (sites detection). At the sequence level, similarity based methods such as FASTA (Pearson 1991) or BLAST (Altschul et al. 1997) are commonly used by molecular biologists for the retrieval of similar (or homologous) sequences. These methods are suitable for finding homologous proteins that share similar folds. Still at the sequence level, other tools exist that used the recognition of given patterns into protein sequences (Hulo et al. 2008; Sigrist et al. 2010). These methods can be used in non-homologous proteins as their sequences can share common similar sub sequences exhibiting common functions without the necessity to be homologues. However, receptor-ligand interaction can be conserved in functionally equivalent proteins even in the absence of sequence homology. If the 3D structure is available, the backbone level based comparison methods mainly rely on RMSD calculation after structure superimposition (Holm and Sander 1997). Alternatively, other methods rely on surface alignment (Guerra et al. 2010) or surface matching (Via et al. 2000) algorithms. Although very useful for a global fold comparison, this superimposition strategy is not suitable for sites detection in arbitrary 3D protein structures (Jambon et al. 2003). Although of large interest in drug design or in deciphering function from 3D structures, methods based on 3D sites prediction have been lately developed. Most of them have been designed for the recognition of protein-ligand binding sites and the comparison of protein-protein interfaces or protein pockets (Reisen et al. 2010). They include hashing techniques (Wallace et al. 1997; Shulman-Peleg et al. 2004), evolutionary trace

(Capra and Singh 2007; Kristensen et al. 2008; Capra et al. 2009; Ward et al. 2009; Erdin et al. 2010), graph theory (Jambon et al. 2003; Weskamp et al. 2004), several kinds of descriptors (Ballester and Richards 2007; Sael et al. 2008; Schalon et al. 2008; Venkatraman et al. 2009) or support vectors machines (Sonavane and Chakrabarti 2010). In 2003, we proposed the SuMo method (Jambon et al. 2003) that has been recently used or improved in several studies by other groups for ligand-based screening for efficient scaffold hopping (Sperandio et al. 2007), analyzing protein-ligand interactions exposed at the surface of a protein (Doppelt-Azeroual et al. 2010) or Fragments-Based Drug Design (Moriaud et al. 2009). A similar approach using clouds of atoms has been recently described (Hoffmann et al. 2010) and today a database of protein complexes suitable for a critical assessment of predicted interaction is available to check blind predictions (Janin et al. 2003; Janin 2010). From the biologist point of view, the methods should be either usable from a software or reachable from a web server (Jambon et al. 2005). This server offers the possibility to match a query structure against a database of active site templates.

1.2 Methods

1.2.1 SuMo General Methodology

The rationale behind the SuMO program is the comparison of 3D local structures, in order to identify possible common functions or to explain unexpected functional results within a protein family. The methodology is divided into two major steps. First, the PDB file containing the atomic coordinates for a protein structure is converted into a data structure suitable for fast comparison. This representation may be stored into a database dedicated to comparison. The 3D structures of the proteins have to be preformatted, i.e. converted into a representation that will be used by the comparison heuristics. Since this operation takes usually longer than the comparison itself, the preformatted data may be stored in a database. Four successive levels will be considered: (1) atoms, (2) groups of atoms, (3) triangles formed by chemical groups and (4) vertices in the final graph representing the molecule. Then SuMo performs the comparison itself by using preformatted data of this database. Proteins are described in terms of amino acids chemical functions. Thus, all amino acids are converted to local functions in space according to a dictionary of physico-chemical functions or "SuMo objects". Then, chemical groups are used to build triangles of chemical neighbours groups (cut-off 8 Å). The choice of triangles allows the description of surface patches rather than isolated points. The burial of each chemical group is estimated using a local atomic density function. The orientation of the triangle towards the rest of the molecule is also estimated. The final representation of the molecule is a set of connected adjacent triangles, i.e. triangles that share exactly two chemical groups, to make a graph in which each triangle forms a vertex. Details about algorithms and parameters can be found in the original paper (Jambon et al. 2003).

1.2.2 New Definition of SuMo Objects

The set of object was not modified since the first version of the software (Jambon et al. 2003) and since the original release, we have improved this set to increase the efficacy and the accuracy of the SuMo software. The strategy is to use a minimal number of different types of object and to have also fewer objects that the previous set to compare a complete surface versus a large database like the PDB. This calculation requires large computer resources and may take a long time to compute or requires the use of a super computer. That is why, we have tried to reduce the number of object and thus drastically decreasing the combinatorial number of superposition to score. The new set has 10 different types of objects versus 15 for the first one.

The first set (Jambon et al. 2003) described hydrogen bonds with an object called "delta plus" and this object is placed on the position of the receptor atom of the hydrogen bond. This position is computed by extension in the direction of the bond hetero atom – hydrogen of 1.8 Å. This description induces two problems, the first one is about the flexibility of the side chain, a small variation of the direction of the hydrogen can displaced the object to 5 Å and cannot match for calculation. Side chain of amino acid could generate a large number of hydrogen bonds and all bonds cannot be treated. We decided to not describe hydrogen bond, by this way, on the new set.

We decide to focus on the chemical function of amino acid and to homogenously treat all residues. The backbone is described by a new object called "carbon alpha" centered on the carbon alpha of each amino acid, instead of a "delta plus" extended from the N–H bond, one "delta minus" centred on the oxygen of the hydroxyl and just for glycine residue a object called "glycine polar" on the nitrogen of the backbone. For hydrophobic moiety, we put one object "hydrophobic aliphatic" in the centred of bond C–C of the side chain (carbon alpha is not included in this calculation). One exception is made for the alanine residue, one "hydrophobic aliphatic" is centred on the methyl. We define ten types of objects centered on chemical function like hydroxyl, carboxylic acid, amine. For hydrophobic or amine function we define the aliphatic or the aromatic object type, except for hydroxyl because the capacity for this function to make hydrogen bond is roughly the same for the two types. We also create the following objects : amide, thiol and carboxylic acid. The guanidinium moiety could be described with aromatic amine and hydrophobic, but we prefer to create a specific object for this function due to the particularity of the reactivity of the residue. The group of residue composed with aromatic cyclic was described with "hydrophobic aromatic" and "amine aromatic" objects. The hydrophobic object was positioned at the barycentre of the cycle. In the case of tryptophan residue one object was positioned was each cycle. Each nitrogen atoms of aromatic cycle were described by an aromatic amine object and centred on it. For the tyrosine residue the hydroxyl object was added centred on the oxygen atom. Figure 1.1 lists the nature and the position of object for each residue. The SuMo software has weighting value of the different object. All objects have the same weight (1) instead of "carbon alpha" with the value of (0.1). This choice was made in order to avoid to give to the

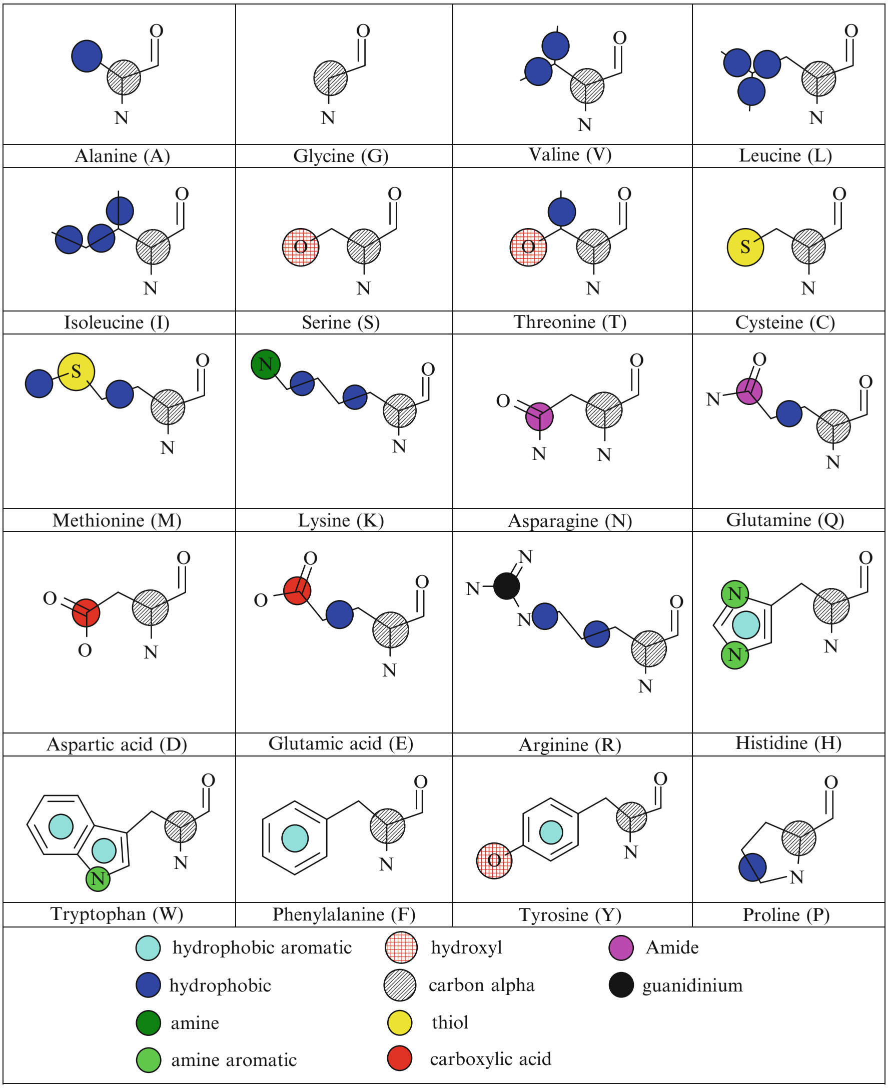


Fig. 1.1 New set of "objects" in SuMo. Correspondence between amino acids and chemical groups as defined in the new set

backbone a too much high importance weight. The new set was tested on the two different protein families previously used for SuMo (Jambon et al. 2003) and a comparative analysis result of the two versions is given hereafter. The definition of the chemical groups that are defined for each amino acid is shown in Fig. 1.1.

1.2.3 Web Interface

From the user point of view, the use of SuMo involves several steps:

1.2.3.1 Choice of a Target Structure

First step, the user has to specify a 3D protein structure as target. This target can be either a file to upload or a PDB id. A selection within this structure can be performed. The user can select either the whole protein or only some chains or some sites bound to ligands. From this selection, SuMo will give the list of retained chemical groups. A non-interactive interface using a simple query language (SuMoQ) is also proposed to advanced users to allow more flexibility in the queries.

1.2.3.2 Choice of SuMo Database

The second step is to select the database to scan. A first database available consists of all the PDB structures in which only redundant chains (sharing 100% sequence identity) have been removed. The second database consists of the ligand binding sites that are found in the PDB database. In order to keep a large panel of conformational variants of the proteins, the variety of structures of proteins of identical or very close sequences is preserved. Selection can contain either the full list of structures in the database or just a subset.

1.2.3.3 Presentation of Results

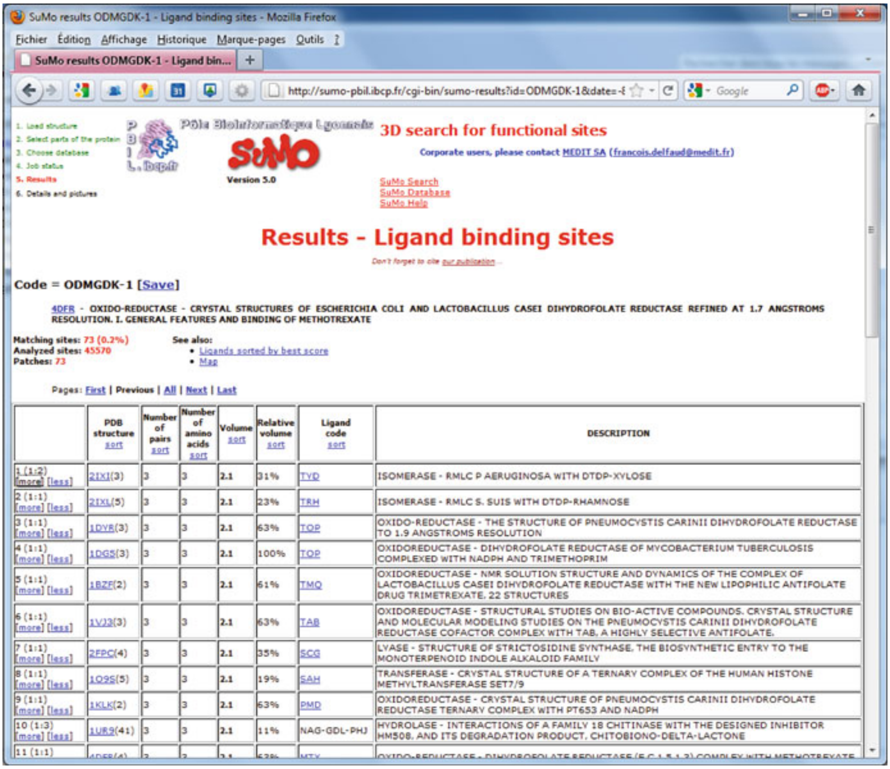
After comparing the selected target against the previously chosen database, results are displayed in a table as a list of potentially interesting similarities (Fig. 1.2).

Results can be sorted according to different criteria, such as the volume of the matched sites, the number of matched pairs of chemical groups, the number of amino acids involved or the SuMo score. The query specification and the results can be saved or exported as text for further analysis.

In the case of scanning the database of ligand binding sites, results are also summarized as a mapping of the potential ligand binding chemical groups in the query structure and as a list of potential ligands sorted by maximal score (volume of the site). Links to the PDBSum web site are also available for complementary informations.

1.2.3.4 Detailed Results

For each pair of matched 3D sites, detailed information is given (Fig. 1.3): description of the chemical groups that matched, parallel view of both sites in the same orientation, direct view in RasMol, links to other resources. Also, various numerical parameters are given as a support for extended analyses of the results.


SuMo results ODMGDK-1 - Ligand binding sites - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
SuMo results ODMGDK-1 - Ligand bin...
http://sumo-pbil.ibcp.fr/cgi-bin/sumo-results?id=ODMGDK-1&date=-8
Google
1. Load structure
2. Select parts of the protein
3. Choose database
4. Job status
5. Results
6. Details and pictures
PBIL.ibcp.fr
Pôle Bioinformatique Lyonnais
SuMo
Version 5.0
SuMo Search
SuMo Database
SuMo Help

3D search for functional sites

Corporate users, please contact MEDIT SA (francois.delfaud@medit.fr)

Results - Ligand binding sites

Don't forget to cite our publication ...

Code = ODMGDK-1 [Save]

4DFR - OXIDO-REDUCTASE - CRYSTAL STRUCTURES OF ESCHERICHIA COLI AND LACTOBACILLUS CASEI DIHYDROFOLATE REDUCTASE REFINED AT 1.7 ANGSTROMS RESOLUTION. I. GENERAL FEATURES AND BINDING OF METHOTREXATE

Matching sites: 73 (0.2%)
Analyzed sites: 45570
Patches: 73

See also:
- Ligands sorted by best score
- Map

Pages: First | Previous | All | Next | Last

	PDB structure sort	Number of pairs sort	Number of amino acids sort	Volume sort	Relative volume sort	Ligand code sort	DESCRIPTION
1 (1:2) [more] [less]	2IXI(3)	3	3	2.1	31%	TYD	ISOMERASE - RMLC P AERUGINOSA WITH DTDP-XYLOSE
2 (1:1) [more] [less]	2IXL(5)	3	3	2.1	23%	TRH	ISOMERASE - RMLC S. SUIS WITH DTDP-RHAMNOSE
3 (1:1) [more] [less]	1DYR(3)	3	3	2.1	63%	TOP	OXIDO-REDUCTASE - THE STRUCTURE OF PNEUMOCYSTIS CARINII DIHYDROFOLATE REDUCTASE TO 1.9 ANGSTROMS RESOLUTION
4 (1:1) [more] [less]	1DG5(3)	3	3	2.1	100%	TOP	OXIDOREDUCTASE - DIHYDROFOLATE REDUCTASE OF MYCOBACTERIUM TUBERCULOSIS COMPLEXED WITH NADPH AND TRIMETHOPRIM
5 (1:1) [more] [less]	1BZF(2)	3	3	2.1	61%	TMQ	OXIDOREDUCTASE - NMR SOLUTION STRUCTURE AND DYNAMICS OF THE COMPLEX OF LACTOBACILLUS CASEI DIHYDROFOLATE REDUCTASE WITH THE NEW LIPOPHILIC ANTIFOLATE DRUG TRIMETREXATE, 22 STRUCTURES
6 (1:1) [more] [less]	1VJ3(3)	3	3	2.1	63%	TAB	OXIDOREDUCTASE - STRUCTURAL STUDIES ON BIO-ACTIVE COMPOUNDS. CRYSTAL STRUCTURE AND MOLECULAR MODELING STUDIES ON THE PNEUMOCYSTIS CARINII DIHYDROFOLATE REDUCTASE COFACTOR COMPLEX WITH TAB, A HIGHLY SELECTIVE ANTIFOLATE.
7 (1:1) [more] [less]	2FPC(4)	3	3	2.1	35%	SCG	LYASE - STRUCTURE OF STRICTOSIDINE SYNTHASE, THE BIOSYNTHETIC ENTRY TO THE MONOTERPENOID INDOLE ALKALOID FAMILY
8 (1:1) [more] [less]	1O9S(5)	3	3	2.1	19%	SAH	TRANSFERASE - CRYSTAL STRUCTURE OF A TERNARY COMPLEX OF THE HUMAN HISTONE METHYLTRANSFERASE SET7/9
9 (1:1) [more] [less]	1KLK(2)	3	3	2.1	63%	PMD	OXIDOREDUCTASE - CRYSTAL STRUCTURE OF PNEUMOCYSTIS CARINII DIHYDROFOLATE REDUCTASE TERNARY COMPLEX WITH PT653 AND NADPH
10 (1:3) [more] [less]	1UR9(41)	3	3	2.1	11%	NAG-GDL-PHJ	HYDROLASE - INTERACTIONS OF A FAMILY 18 CHITINASE WITH THE DESIGNED INHIBITOR HM508, AND ITS DEGRADATION PRODUCT, CHITOBIONO-DELTA-LACTONE
11 (1:1)	4DFR(4)	3	3	2.1	63%	MTX	OXIDO-REDUCTASE - DIHYDROFOLATE REDUCTASE (E.C.1.5.1.3) COMPLEX WITH METHOTREXATE

Fig. 1.2 List of detected ligands binding sites

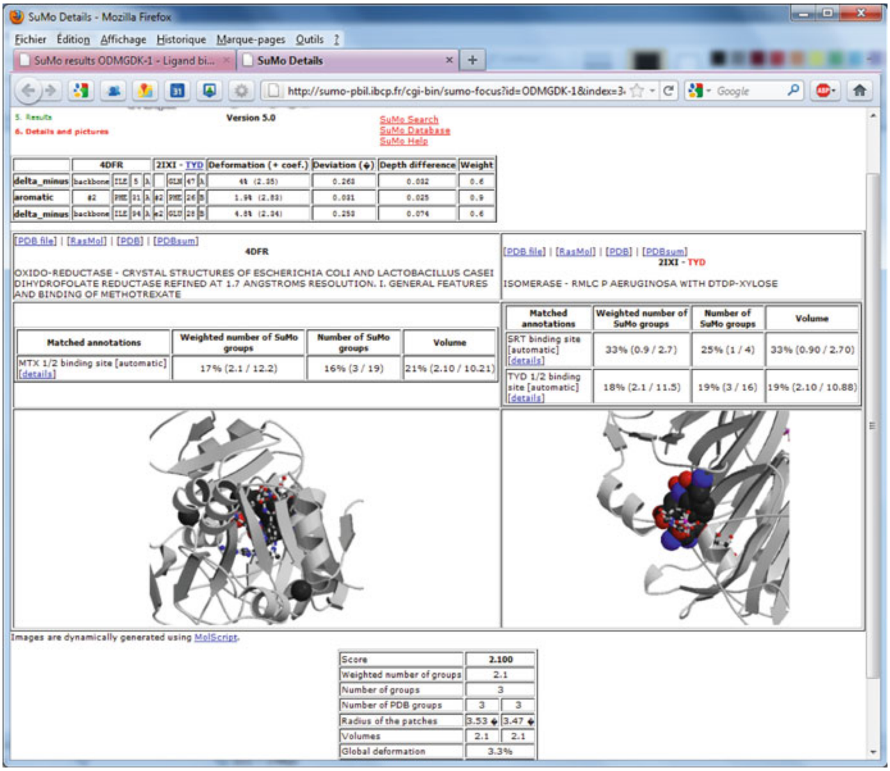
1.3 Results

A comparison of the results obtained with the original definition of objects and the new one described herein is provided on two extreme examples in which the structure-function relationship in these proteins is well-known. First the classical case of convergent evolution of serine proteases illustrates the independence of the SuMo method from fold and sequence similarities.

The second example illustrates over a larger test set the possible discrimination of functional sites from non-functional sites in the legume lectins family, despite high sequence similarity and good overall superimposition as indicated by low RMSD.

1.3.1 Test of New Set on Proteases

To perform this test a special set of models was made. From the PDB, 61 proteases and 970 isomerases were downloaded. For the isomerase (EC 3.1.3) a second set was made to be sure that there is no protease function in these isomerase. The first 970


SuMo Details - Mozilla Firefox
Fichier Édition Affichage Historique Marque-pages Outils ?
SuMo results ODMGDK-1 - Ligand bi...
SuMo Details
http://sumo-pbil.ibcp.fr/cgi-bin/sumo-focus?id=ODMGDK-1&index=3
Google
5. Results
6. Details and pictures
Version 5.0
SuMo Search
SuMo Database
SuMo Help

	4DFR				2IXI - TYD				Deformation (+ coef.)	Deviation (◆)	Depth difference	Weight
delta_minus	backbone	ILE	5	A		GLN	47	A	4% (2.35)	0.263	0.032	0.6
aromatic	#2	PHE	31	A	#2	PHE	26	B	1.9% (2.83)	0.031	0.025	0.9
delta_minus	backbone	ILE	94	A	#2	GLU	28	B	4.8% (2.34)	0.253	0.074	0.6

[PDB file] | [RasMol] | [PDB] | [PDBsum]
4DFR

OXIDO-REDUCTASE - CRYSTAL STRUCTURES OF ESCHERICHIA COLI AND LACTOBACILLUS CASEI DIHYDROFOLATE REDUCTASE REFINED AT 1.7 ANGSTROMS RESOLUTION. I. GENERAL FEATURES AND BINDING OF METHOTREXATE

Matched annotations	Weighted number of SuMo groups	Number of SuMo groups	Volume
MTX 1/2 binding site [automatic] [details]	17% (2.1 / 12.2)	16% (3 / 19)	21% (2.10 / 10.21)

[PDB file] | [RasMol] | [PDB] | [PDBsum]
2IXI - TYD

ISOMERASE - RMLC P AERUGINOSA WITH DTDP-XYLOSE

Matched annotations	Weighted number of SuMo groups	Number of SuMo groups	Volume
SRT binding site [automatic] [details]	33% (0.9 / 2.7)	25% (1 / 4)	33% (0.90 / 2.70)
TYD 1/2 binding site [automatic] [details]	18% (2.1 / 11.5)	19% (3 / 16)	19% (2.10 / 10.88)

Images are dynamically generated using MolScript.

Score	2.100	
Weighted number of groups	2.1	
Number of groups	3	
Number of PDB groups	3	3
Radius of the patches	3.53 ◆	3.47 ◆
Volumes	2.1	2.1
Global deformation	3.3%	

Fig. 1.3 Superimposition of hits in SuMo

entres from the PDB were downloaded and added to the database. For protease, 61 models were chosen and dispatched in several sub families: 16 serine proteases (EC 3.4.21); 8 metallopeptidases (EC 3.4.24); 8 cysteine peptidases (3.4.22); 9 aspartic peptidases (3.4.23); 4 aminopeptidases (EC 3.4.11); 6 metallocarboxypeptidase (EC 3.4.17); 5 serine carboxypeptidases (EC 3.4.16); 1 cysteine carboxypeptidases (EC 3.4.18) and 4 threonine peptidases (EC 3.4.25). The list of serine proteases was: 1A5H, 1BF9, 1BML, 1CEA, 1D3P, 1DDJ, 1EZX, 1FAX, 1FLE, 1FQ3, 1GI7, 1H4W, 1IAU, 1LTO, 1SBC and 1AFQ. The query was defined on the active site of the 1AFQ serine protease model (Kashima et al. 1998). SuMo objects were chosen within three spheres of selection with 10 Å radius and centered on W212, F302 and S218 residues. The selected set of objects to form the query is about 138 different objects located on 58 residues.

The result of the run was analysed, the job takes less than 3 min of running time on a quad cpu Xeon core. The first 13 hits were all serine proteases and the 1SBC entry was found on rank 79, the 2 other last entres were not recovered. The result is presented on Table 1.1. Hits number 34, 37 and 39 were cysteine endo peptidases with PDB code 1 AU0, 1MEM and 1KFU, respectively. 16 proteases among 61 were

Table 1.1 Comparison of new (left panel) and original (right panel) objects on 61 proteases and 970 isomerases

New object data set				Original object data set			
pdb code	Sumo score	Objects		pdb code	Sumo score	Objects	
1AFQ	69.98	233	SP	1H4W	16.76	39	SP
1H4W	52.97	131	SP	1FAX	10.98	25	SP
1A5H	47.16	109	SP	1A5H	8.87	22	SP
1GI7	46.68	106	SP	1GI7	4.9	10	SP
1FAX	45.51	106	SP	1Q6H	4.34	8	iso
1FLE	43.02	96	SP	1XYH	4.12	9	iso
1LTO	39.51	93	SP	3REQ	4.08	7	iso
1BML	38.76	92	SP	1IAU	4.07	8	SP
1D3P	37.35	85	SP	1EZX	4	9	SP
1EZX	36.63	81	SP	1PJH	3.99	8	iso
1FQ3	35.87	94	SP	2GZM	3.95	7	iso
1IAU	35.19	80	SP	1B6C	3.94	7	iso
1DDJ	30.2	65	SP	1N23	3.9	7	iso
1NSS	11.45	33	iso	1GXD	3.9	8	MEP
1JC4	11.05	28	iso	1P5Q	3.9	8	iso
1UPI	10.59	30	iso	2IAM	3.8	6	iso
1ZVC	10.56	24	iso	2OJU	3.78	8	iso
1B6C	10.41	33	iso	1D3P	3.75	9	SP
2VEP	10.38	37	iso	1EQ2	3.75	8	iso
1VGA	10.33	34	iso	1Q6I	3.75	8	iso
2NR0	10.26	23	iso	2CIR	3.69	7	iso
1WDM	10.14	37	iso	2CIS	3.69	7	iso
1Z8K	9.99	22	iso	1TTJ	3.69	7	iso
1R2T	9.89	20	iso	2BI8	3.67	7	iso
2F6Q	9.8	31	iso	1HOT	3.65	6	iso
1RCQ	9.78	35	iso	1FS5	3.65	6	iso
1LZO	9.75	23	iso	1TCO	3.64	7	iso
1O5X	9.75	22	iso	1GR0	3.64	7	iso
1WOB	9.75	22	iso	1ZLI	3.64	7	MCP
1M7O	9.75	22	iso	1JOF	3.64	8	iso
1M7P	9.74	22	iso	1UWY	3.63	6	MCP
1IV8	9.66	24	iso	2P5Y	3.62	7	iso
2BTM	9.61	25	iso	1SEU	3.6	6	iso
1 AU0	9.6	20	CEP	1WLT	3.6	6	iso
1P5G	9.57	26	iso	2IFY	3.6	8	iso
1HG3	9.56	23	iso	1TTJ	3.6	7	iso
1MEM	9.51	18	CEP	1BML	3.6	8	SP
2OK3	9.4	22	iso	2GYI	3.56	7	iso
1KFU	9.37	20	CEP	1PCJ	3.55	7	iso
1IIP	9.37	29	iso	1CY8	3.5	6	iso
2BRJ	9.33	20	iso	2C5E	3.49	7	iso

(continued)

Table 1.1 (continued)

New object data set				Original object data set			
pdb code	Sumo score	Objects		pdb code	Sumo score	Objects	
2HXG	9.28	20	iso	2JFO	3.49	7	iso
1AW2	9.28	19	iso	2H4L	3.49	7	iso
1SUX	9.27	19	iso	1YA7	3.48	7	TEP
1JX1	9.26	22	iso	1YAU	3.48	7	TEP
1R2R	9.25	19	iso	1D6M	3.45	6	iso
2VEN	9.24	19	iso	2VA6	3.45	6	AEP
16/61 Protease : 26,2%				14/61 Protease : 22,9%			
13/16 serine protease: 81,3%				8/16 Serine protease: 50%			

recovered on the first 50 hits which gives a percentage of 26.2%, but 13 on 16 serine protease were recovered with a percentage of 81.3%. The new set gives better results that the previous one with a percentage of recovery about 22.9% of protease and only 8 on 16 serine proteases recovered on the 50 first hits. Difference between the two SuMo object sets is clearly establish on this run because the first 13 lines contain serine proteases for the new set and only 4 entries at the top of the list for the old set. In fact, deeper analyses of the result from the new set of SuMo objects give a better answer. Only 3 serine proteases were not recovered on the first 50 hits: 1SBC (79 hits), 1CEA and 1BF9 are not present. The 1CEA (Mathews et al. 1996) structure is the non covalent complex of the recombinant kringle 1 domain of human plasminogen. This structure has 7 domains, one PAN, 5 kringle and one peptidase domain, the resolved structure have only the end on the PAN and the kringle 1 domain. Even the structure is displayed on the PDB with serine protease EC code; the protease domain is not present in the PDB file. For the second entry "no present" with the PDB code 1BF9, the same conclusion can be made. The structure is the N terminal EGF like domain from human factor VII and can be divided in three domains, Gla EGF1, EGF2 and Serine protease. The PDB structure resolved by NMR gives the EGF1 domain whereas the other domains are absent. Therefore, the new SuMo objects allow to recovered 13 of the 14 true serine proteases thus giving a recovery percentage of 92.9%.

The analysis of the common object or signature is displayed for all serine protease on Table 1.2. The signature was compiled by residue and the number of common objects with the query structure (1AFQ) is given in the table. A value number of 1 means that all objects of a residue were recovered. The serine protease made the peptidase reaction with three catalytic residues: H57, D103 and S195. The H57 and the D102 were found in the 14 "true" serine proteases, only the S195 was not recovered in the 1EZX structure. One of the advantage of the method is the capability to analyse common residues in discontinuous sequence and not only in 2D like classical approach. The 1SBC model has quite nothing in common but the three catalytic residues, it is possible to make a pattern of recognition only based on this three residues to determine if it is a serine protease. In this case, SuMo software can be used

Table 1.2 Result of the run on a set constituted by 61 proteases mixed with 970 isomerases. Serine proteases signatures are displayed with the number of objects recovered by residue. 1 mean all objects of the residue were recovered

	N° aa	query	1AFQ	1H4W	1A5H	1GI7	1FAX	1FLE	1LTO	1BML	1D3P	1EZX	1FQ3	1IAU	1DDJ	1SBC
	V17	1	1	1	1	1	1	1	1	1	1		1	1		
	C42	1	1	1	1	1	1		1	1	1		1	1	1	1
	A55	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	A56	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Catalytic	H57	1	1	1	1	1	1	1/5	1	1	1	3/5	1	1	1/5	2/5
	C58	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	G59	1	1	1	1	1	1		1	1		1			1	
	Y94	1	1	1	1/2	1	1/2				1/2	1/2	1	1		
	N95	1	1	1	1	1	1					1	1	1		
	S96	1	1	1/2	1/2		1/2					1/2				
	L97	1	1	1/3	1/3		1/3					1/3	1/3	1/3		
	T98	1	1	1	1		2/3				1/3	1	1/3	1/3		
	I99	1	1	1	1/4	1/4	1/4		1/4		2/4	2/4	1/4	1/4		1/4
	N100	1	1	1/2	1/2	1/2	1/2	1/2	1/2		1/2	1	1/2	1/2		
	N101	1	1	1	1	1	1	1	1		1	1	1	1		
Catalytic	D102	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1/2
	L143	1	1		1	1	1	2/3		1			1	1		
	Y146	1	1													
	N165	1	1	1					1	1					1	
	C168	1	1	1	1/2	1	1	1	1	1		1	1/2	1/2	1	
	K169	1	1	1		1	1	1/2		1		1			1/2	
	Y171	1	1	1			1	1				1		1		
	W172	1	1	5/6		2/6		4/6				3/6		1/6	1/6	
	G173	1	1			1		1						1	1	
	T174	1	1	1		1/2		1/2				1/2		1/2	1/2	

(continued)

Table 1.2 (continued)

	N° aa	query	1AFQ	1H4W	1A5H	1GI7	1FAX	1FLE	1LTO	1BML	1D3P	1EZX	1FQ3	1IAU	1DDJ	1SBC
	K175	1	1	1	3/4	2/4	1/4	2/4			2/4	2/4		1/4	2/4	
	I176	1	1	1	3/4	1	1	1	2/4	1	3/4	1		3/4	1	
	K177	1	1	2/4	2/4	2/4	2/4	1	2/4	1/4	1/4	2/4		1/4	1/4	
	A179	1	1	1/2	1/2			1/2	1/2	1/2		1/2	1/2	1/2	1/2	
	M180	1	1	1	1	1	1	1	1	2/4	1	1	2/4	2/4	2/4	
	I181	1	1	1/2	1	1	1/2	1	1	1	1/2	1/2	1	1	1	
	C182	1	1	1	1	1	1	1	1	1	1/2	1	1	1/2	1	
	A183	1	1	1	1	1	1	1	1	1	1	1		1/2	1	
	G184	1	1	1	1	1	1	1	1	1		1		1	1	
	S189	1	1	1/2	1/2	1/2	1/2	1	1/2	1/2	1/2		1	1		
	S190	1	1	1	1/2	1/2	1/2	1/2	1/2	1	1/2		1/2	1/2		
	C191	1	1	1	1	1	1	1	1	1	1		1/2	1/2		
	M192	1	1	1/2	1/2	1/4	1/4	1/4	1/4	1/4	1/4		1/4	1/4		
	G193	1	1	1	1	1	1	1	1	1	1		1	1		1
	D194	1	1	1	1/2	1	1	1	1	1	1		1	1	1/2	1/2
Catalytic	S195	1	1	1	1	1	1	1/2	1	1/2	1		1	1	1/2	1/2
	G196	1	1	1	1	1	1	1	1	1	1		1	1	1	
	I212	1	1	1	1	1	1	1	1	1	1	1	1	1	2/3	
	V213	1	1	1	1	1	1	1	1	1	1	1	1	1	1	2/3
	S214	1	1	1	1	1	1	1	1	1	1	1/2	1	1	1	1/2
	W215	1	1	1	1	5/6	1	2/6	1/6	5/6	1	2/6	4/6	2/6	1/6	1/6
	2 G16	1	1	1	1	1	1			1	1		1	1	1	
	S217	1	1													
	S218	1	1													
	T219	1	1					1/3								

	C220	1	1	1	1	1	1	1	1/2	1	1					
	S221	1	1				1/2	1/2			1/2					
	T224	1	1		2/3	1/3		1/3							1/3	
	P225	1	1	1	1	1		1	1	1		1	1	1	1	
	G226	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
	V227	1	1	1	1	1	1	1	1	1	1/3	1	1	1	1	1/3
	Y228	1	1	1	1	1	3/4	3/4	1	1	1	1	1/4	1/4	1	
	A229	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1/2

to detect secondary site or "moon lighting" sites of proteins. The global analysis the 14 serine proteases shows conserved area and empty band showing that in there area there are no common objects with 1AFQ (Kashima et al. 1998).

1.3.2 Test of New Set on Lectins

Lectin proteins have a specific binding capability for oligosaccharide and particularly glycoprotein. To make a comparison with the previous SuMo work, the same lectin set was chosen. We choose a set of 90 3D models of lectin extracted from the Protein Data Bank (PDB). In this set 18 lectins are not able to bind oligosaccharide (designed hereafter as false lectin). The 2PEL (Banerjee et al. 1996) structure is the peanut lectin whose structure was solved by X-ray crystallography at 2.25 Å resolution. The LAT ligand is a lactose molecule and the structure has been co-crystallized with two lactose molecules bound in two different sites LAT and LBT. The reference set of objet was defined as all objects around 6.5 Å of the LAT ligand site. This distance selects 16 objects on 2PEL model, which are distributed on the 12 residues: D80, A82, D83, G103, G104, Y125, N127, E129, S211, L212, G213 and G214.

The comparison was performed on a double cpu Xeon type dual core and the run took less than 1 min of CPU time. The 6 best first hits were peanut lectins with different ligands under reference: 1QF3, 1CR7, 2PEL, 2TEP, 1BZW, 1CIW. The number of matching objects ranges from 16 to 12, the L212 and the carbon alpha of the S211 were not recovered for 1BZW and 1CIW structures. The next hits give concanavilin structure (PDB code: 1NLS) with 9 objects. Most hits have 5 or 6 objects in common with the query and belong to the concanavilin family. The program classified 93 hits with 16 to 3 objects. The analysis reveals that only 2UU8 with 7 common objects is present and the rest of the 17 false lectins are not present in the listing. This result compared favorably with the same query on the same server from the previous set of object reveal better hits. Indeed, by using the previous set 14 of the 18 false lectins were present in hits, 7 has more than 5 common objects and 7 have less 4 common, only 4 models are not in the list (Table 1.3). The false lectins not recovered are: 1IOA, 1DQ2, 1ENQ and 1QFD. It is interesting to notice than 2UU8 have 5 objects on the previous set and 7 with the new one. 2UU8 have common object dispatched on 6 residues. Here is the list of matching pairs of residue on 2PEL vs. 2UU8: A82/A207, D83/D208, G103/G227, G104/R228 (carbon alpha), Y125/Y12 and N127/N14. It is interesting to notice than the carbon alpha of the G104 of the 2PEL lectin is matching with the R228 of the false lectin 2UU8. The weight of "carbon alpha" object is minimal in order to not favor proteins with the same fold, so the score 2UU8 is about 4.44 versus 8.05 for 2PEL and are hits number 20 / 93. Among the 12 residues, 6 has been paired explaining the "good" hit, but the mutation on glycine 104 in arginine on the binding has probably disabled the capability of this protein to bind with an oligosaccharide.

Table 1.3 Result obtained with lectin 2PEL as query. On the table only false lectins are displayed. If number of SuMo objects is higher than 5, the entry is in dark gray. Otherwise, it is in light gray

PDB code	New object set		Previous object set	
	SuMo score	Object Number	SuMo score	Object Number
1IOA	Non present	Non present	Non present	Non present
1HSS	Non present	Non present	3.49	7
3AIT	Non present	Non present	2.45	4
1AVB	Non present	Non present	3	6
4AIT	Non present	Non present	3.65	6
1CLV	Non present	Non present	2.45	4
1HOE	Non present	Non present	2.45	4
1HTX	Non present	Non present	2.15	3
1CN1	Non present	Non present	2.44	5
1APN	Non present	Non present	2.44	5
1DQ2	Non present	Non present	Non present	Non present
1ENQ	Non present	Non present	Non present	Non present
1OK0	Non present	Non present	2.45	4
1QFD	Non present	Non present	Non present	Non present
2AIT	Non present	Non present	2.45	4
1CES	Non present	Non present	2.44	5
2UU8	4.44	7	2.29	5
1ENS	Non present	Non present	2.4	4

1.3.3 *New Biological Application on the Search for a Common Site in Betalactam Molecule*

A betalactam class molecule is the most prescribed antibiotic drug (Cars et al. 2001; Coenen et al. 2006; Ferech et al. 2006). Betalactam makes a covalent bond on a specific site on the Penicillin Binding Protein (PBP) between one lysine and the betalactam cycle. Several complexes of the PBP and betalactam molecule were solved and deposited into the PDB. For this study, we took a PBP-2X protein solved with a cefuroxime molecule (Gordon et al. 2000). The structure is deposited under the reference 1QMF and its structure was solved by X ray diffraction at a 2.8 Å resolution. The whole protein was converted to a list of objects thanks to our dictionary of new objects and we used the CES site with a selection of SuMo objects around the molecule with a range of 6 Å. This selection involves 36 objects and 255 triplets. The run was carried out with the complete surface of all proteins of the Complete Protein Data Bank. The calculation was made on a quad core Linux computer and took roughly 15 h. The first 45 entries were listed in Table 1.4.

The first 17 entries correspond to different PBP (PDB-2B and PDB-1A) stored in the databank. The sequence identities computed after alignment by using ClustalW is 89/705(12.63%) for PBP-2X and PBB-2B and 81/704 (11.51%) for PBP-1A and PBP-2X. After the reference site, SuMo identified 8 PBP-2X which have very similar sequences. A second set of entries after PBP-2X is composed first by 3 PBP-2B and 3 PBP-1A and at least 3 other PBP's. SuMo program computed its search based on

Table 1.4 Result of SuMo run using the CES binding site of the Penicillin Binding Protein (1QMF) as reference against all the proteins of the Protein Data Bank (*PDB*). Only the first 45 entries are displayed. The white to light grays entries are all penicillin binding proteins (*PBP*). The dark grays entries are beta-lactamase proteins

Hit Id	PDB Id	SCF Count	Sumo Score	Header line of the PDB structure
1	1QMF	67	33.02	Penicillin-binding protein 2X (PBP-2X)
2	2ZC3	51	28.15	Penicillin-binding protein 2X (PBP 2X)
3	2ZC4	50	27.16	Penicillin-binding protein 2X (PBP 2X)
4	1QME	50	27.01	Penicillin-binding protein 2X (PBP-2X)
5	2Z2L	48	26.95	Penicillin-binding protein 2X (PBP2X)
6	2Z2M	43	24.96	Cefditoren-acylated penicillin-binding protein 2X
7	1RP5	48	23.615	PBP2X from *Streptococcus pneumoniae* strain 5259 WITH REDUCED susceptibility to beta-lactam antibiotics
8	1PYY	43	22.89	Double mutant PBP2X T338A/M339F from *Streptococcus pneumoniae* strain R6 AT 2.4 A resolution
9	2WAF	34	19.23	Penicillin-binding protein 2B (PBP-2B)
10	2WAD	30	16.61	Penicillin-binding protein 2B (PBP-2B)
11	2WAE	31	16.57	Penicillin-binding protein 2B (PBP-2B)
12	2ZC5	27	14.09	Penicillin-binding protein 1A (PBP 1A) acyl-enzyme complex (Biapenem)
13	2C5W	25	13.91	Penicillin-binding protein 1A (PBP-1A) acyl-enzyme complex (Cefotaxime)
14	2ZC6	25	13.30	Penicillin-binding protein 1A (PBP 1A) acyl-enzyme complex (Tebipenem)
15	2JE5	24	13.25	Structural and mechanistic basis of penicillin binding protein inhibition by lactivicins
16	2JCH	22	12.71	Structural and mechanistic basis of penicillin binding protein inhibition by lactivicins

(continued)

Table 1.4 (continued)

Hit Id	PDB Id	SCF Count	Sumo Score	Header line of the PDB structure
17	3EQU	21	12.66	Crystal structure of penicillin-binding protein 2 from *Neisseria gonorrhoeae*
18	1YLZ	24	12.49	X-ray crystallographic structure of CTX-M14 beta-lactamase complexed with ceftazidime-like boronic acid
19	1MWR	21	12.18	Structure of semet penicillin binding protein 2a from methicillin resistant *Staphylococcus aureus*
20	2V2F	21	12.15	Crystal structure of PBP1A from drug-resistant strain 5204 FROM *Streptococcus pneumoniae*
21	1IYP	21	11.86	TOHO-1 beta-lactamase in complex with cephalothin
22	1IYO	21	11.75	TOHO-1 beta-lactamase in complex with cefotaxime
23	2JCI	20	11.40	Structural insights into the catalytic mechanism and the role of *Streptococcus pneumoniae* PBP1B
24	1YLW	21	11.37	X-ray structure of CTX-M-16 beta-lactamase
25	2ZQA	19	11.33	Cefotaxime acyl-intermediate structure of class a beta-lacta TOHO-1 E166A/R274N/R276N triple mutant
26	1YLY	22	11.31	X-RAY CRYSTALLOGRAPHIC STRUCTURE OF CTX-M-9 BETA-LACTAMASE COMPLEXED WITH CEFTAZIDIME-LIKE BORONIC ACID
27	2ZQC	18	10.94	Aztreonam acyl-intermediate structure of class a beta-lactam TOHO-1 E166A/R274N/R276N triple mutant
28	2ZQ9	19	10.83	Cephalothin acyl-intermediate structure of class a beta-lactamase TOHO-1 triple mutant
29	2FFF	19	10.80	Open form of a class a transpeptidase domain (PBP)

(continued)

Table 1.4 (continued)

Hit Id	PDB Id	SCF Count	Sumo Score	Header line of the PDB structure
30	2Q9M	19	10.76	4-substituted trinems as broad spectrum-lactamase inhibitors: structure-based design, synthesis and biological activity
31	1TEM	19	10.52	6 alpha hydroxymethyl penicilloic acid acylated on the TEM- 1 BETA-lactamase from *Escherichia coli*
32	2A49	18	10.45	Crystal structure of clavulanic acid bound TO E166A VARIANT OF SHV-1 beta-lactamase
33	2C6W	19	10.32	Penicillin-binding protein 1a (PBP-1A) from *Streptococcus pneumoniae*
34	3DWK	19	10.27	Identification of dynamic structural motifs involved in peptidoglycan glycosyltransfer
35	1FQG	18	10.15	Molecular structure of the acyl-enzyme intermediate in tem- 1 beta-lactamase
36	3BFC	21	9.98	Class a beta-lactamase sed-G238C complexed with imipenem
37	1YMS	16	9.93	X-ray crystallographic structure of CTX-M-9 beta-lactamase complexed with nafcinin-like boronic acid inhibitor
38	1XKZ	17	9.87	Crystal structure of the acylated beta-lactam sensor domain of BLAR1 from *S. aureus*
39	1BLS	17	9.86	Crystallographic structure of a phosphonate derivative of the enterobacter cloacae p99 cephalosporinase: mechanistic interpretation of a beta-lactamase transition state analog
40	1VQQ	17	9.80	Structure of penicillin binding protein 2a from methicillin resistant *Staphylococcus aureus.*
41	2ZQD	16	9.80	Ceftazidime acyl-intermediate structure of class a beta-lact toho-1 E166A/R274N/R276N TRIPLE MUTANT

(continued)

Table 1.4 (continued)

Hit Id	PDB Id	SCF Count	Sumo Score	Header line of the PDB structure
42	1IYQ	17	9.79	TOHO-1 beta-lactamase in complex with benzylpenicillin
43	2ZD8	20	9.67	SHV-1 class a beta-lactamase complexed with meropenem
44	1PIO	16	9.42	An engineered *Staphylococcus aureus* PC1 beta-lactamase that hydrolyses third generation cephalosporins
45	1CK3	15	9.41	N276D mutant of *Escherichia coli* TEM-1 beta-lactamase

triplet of objects, protein with same fold and very similar site as PBP protein, are highly ranked even if the percentage of identity of sequence is very low. After PBP, the analysis of hits reveals another set composed by betalactamase. Betalactam molecule binds on this protein and its cycle is opened by the enzyme. When the betalactam cycle is open the drug loses its antibiotic effect. The protein is involved in the mechanism of antibiotic resistance in bacteria. After line 17, entries 21 (among the last 28) are betalactamase proteins and 7 other PBP different fromPBP-2X. The sequence identity between the first betalactamase (code 1YLZ) (Chen et al. 2005) and reference PBP-2X (1QMF) is only 50/702: 7.19%. The analysis of the first beta lactamase revealed that this protein is involved in the resistance because this enzyme of this strain binds on penicillin drug and also third generation cephalosporin drug. This PBP was crystallized with ceftazidime which is third generation cephalosporin. A BLAST run was used to find proteins with similar sequence of the PBP-2X on the sequence database of PDB model, we found all PBP-2X but no other PBP and no betalactamase structure. The two PDB structures were superimposed by using the Sybyl X software (Tripos 2010), the RMSD found was as high as 20 Å and the two binding sites are located at the opposite directions in the fitted molecules. On the contrary, the analysis of the superposition provided by SuMo software reveals that mostly all objects of the sites are commonly shared by the two structures. The three residues which could play a catalytic role have a equivalent object on the two structures, the K340 of the PBP-2X is replaced by the K73 on the betalactamase, the S337 by S70 and S395 by S130. The superposition of the both sites resulting from the SuMo analysis is displayed on the Fig. 1.4.

1.4 Discussion

Since the development of the original SuMO method in 2003 (Jambon et al. 2003), several improvements have been performed either in our team (Jambon et al. 2005) or by others (Sperandio et al. 2007), including protein-ligand interactions exposed

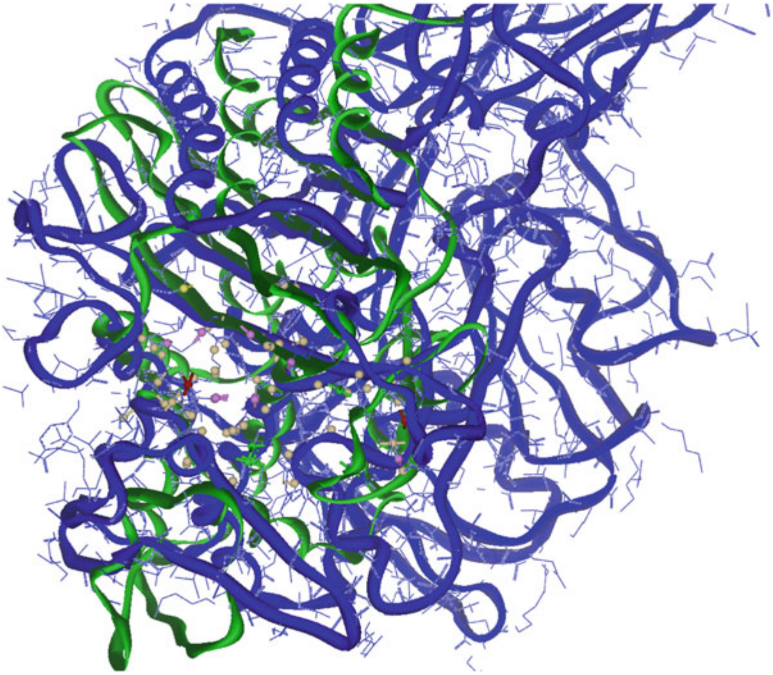


Fig. 1.4 Superimposition, obtained by SuMo software, of the Penicillin Binding Protein (1QMF) in blue with the beta – lactamase (1YLZ) protein in green. Sites are well superimposed (RMSD = 2.17 Å)

at the surface of a protein (Doppelt-Azeroual et al. 2010) or Fragments-Based Drug Design (Moriaud et al. 2009). A similar approach using clouds of atoms has been recently described (Hoffmann et al. 2010) and today a database of protein complexes suitable for a critical assessment of predicted interaction is available is available to check blind predictions (Janin et al. 2003; Janin 2010). In this chapter, we describe a revised strategy to define standardized “SuMo objects” in order to improve the quality of the results and to decrease the CPU time. Since the number of objects has been decreased in this new set, this contributes to a simpler description of the molecules to be compared. This lead to better response time even if the size of the PDB has dramatically increased (21339 protein structures in 2003 versus 68353 in June 2011). The results obtained by SuMo by using the new set of object favorably compare with those obtained with the classical set. This is confirmed in the comparative study that used the 2 standards protein test sets of the original study. For example, 2 additional proteases were identified and 13 instead of 8 among 16 have been correctly assigned as serine proteases. Even more spectacular are the results obtained with the Penicillin Binding Protein (PBP) and beta lactamases. The two proteins families are well separated at the function level giving rise to two well separated blocks in the list. However, starting with a particular protein of a given family (1QMF), the SuMo program was able to catch proteins that all share the capability to bind betalactam molecules. The SuMo approach is a tentative towards automatic functional annotations of protein of unkown functions based solely on their 3D structures. In the future, efforts will be made to go towards protein-protein interactions capabilities as SuMo can be seen as the foundation of the vocabulary used in a grammar still to be discovered about the rules that govern the molecular language of interaction.

Acknowledgements Thanks are due to Martin Jambon as the main author of the original SuMo program written in OCAML.

References

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

Ballester PJ, Richards WG (2007) Ultrafast shape recognition to search compound databases for similar molecular shapes. J Comput Chem 28:1711–1723

Banerjee RDK, Ravishankar R, Suguna K, Surolia A, Vijayan M (1996) Conformation, protein-carbohydrate interactions and a novel subunit association in the refined structure of peanut lectin-lactose complex. J Mol Biol 259:281–296

Bertolazzi P, Guerra C, Liuzzi G (2010) A global optimization algorithm for protein surface alignment. BMC Bioinformatics 11:488

Capra JA, Singh M (2007) Predicting functionally important residues from sequence conservation. Bioinformatics 23:1875–1882

Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5:e1000585

Cars O, Molstad S, Melander A (2001) Variation in antibiotic use in the European Union. Lancet 357:1851–1853

Chen Y, Shoichet B, Bonnet R (2005) Structure, function, and inhibition along the reaction coordinate of CTX-M beta-lactamases. J Am Chem Soc 127:5423–5434

Coenen S, Ferech M, Dvorakova K, Hendrickx E, Suetens C, Goossens H (2006) European surveillance of antimicrobial consumption (ESAC): outpatient cephalosporin use in Europe. J Antimicrob Chemother 58:413–417

Doppelt-Azeroual O, Delfaud F, Moriaud F, de Brevern AG (2010) Fast and automated functional classification with MED-SuMo: an application on purine-binding proteins. Protein Sci 19:847–867

Erdin S, Ward RM, Venner E, Lichtarge O (2010) Evolutionary trace annotation of protein function in the structural proteome. J Mol Biol 396:1451–1473

Ferech M, Coenen S, Dvorakova K, Hendrickx E, Suetens C, Goossens H (2006) European surveillance of antimicrobial consumption (ESAC): outpatient penicillin use in Europe. J Antimicrob Chemother 58:408–412

Gordon EJ, Mouz N, Duee E, Dideberg O (2000) The crystal structure of the penicillin-binding protein 2x from *Streptococcus pneumoniae* and its acyl-enzyme form: implication in drug resistance. J Mol Biol 299(2):477–485

Hoffmann B, Zaslavskiy M, Vert JP, Stoven V (2010) A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. BMC Bioinformatics 11(1):99

Holm L, Sander C (1997) Dali/FSSP classification of three-dimensional protein folds. Nucleic Acids Res 25:231–234

Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJA (2008) The 20 years of PROSITE. Nucleic Acids Res 36:D245–D249

Jambon M, Imberty A, Deleage G, Geourjon C (2003) A new bioinformatic approach to detect common 3D sites in protein structures. Protein-Struct Funct Genet 52:137–145

Jambon M, Andrieu O, Combet C, Deleage G, Delfaud F, Geourjon C (2005) The SuMo server: 3D search for protein functional sites. Bioinformatics 21:3929–3930

Janin J (2010) Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6:2351–2362

Janin J, Henrick K, Moult J, Ten Eyck L, Sternberg MJE, Vajda S, Vasker I, Wodak SJ (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Protein-Struct Funct Bioinform 52:2–9

Kashima A, Inoue Y, Sugio S, Maeda I, Nose T, Shimohigashi Y (1998) X-ray crystal structure of a dipeptide-chymotrypsin complex in an inhibitory interaction. Eur J Biochem 255:12–23

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 Bioinformatics 9(1):17

Mathews II, Vanderhoff-Hanaver P, Castellino FJ, Tulinsky A (1996) Crystal structures of the recombinant kringle 1 domain of human plasminogen in complexes with the ligands epsilon-aminocaproic acid and trans-4-(aminomethyl)cyclohexane-1-carboxylic acid. Biochemistry 35:2567–2576

Moriaud F, Doppelt-Azeroual O, Martin L, Oguievetskaia K, Koch K, Vorotyntsev A, Adcock SA, Delfaud F (2009) Computational fragment-based approach at PDB scale by protein local similarity. J Chem Inf Model 49:280–294

Pearson WR (1991) Searching protein-sequence libraries – comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. Genomics 11:635–650

Reisen F, Weisel M, Kriegl JM, Schneider G (2010) Self-organizing fuzzy graphs for structure-based comparison of protein pockets. J Proteome Res 9:6498–6510

Sael L, La D, Li B, Rustamov R, Kihara D (2008) Rapid comparison of properties on protein surface. Protein-Struct Funct Bioinform 73:1–10

Schalon C, Surgand JS, Kellenberger E, Rognan D (2008) A simple and fuzzy method to align and compare druggable ligand-binding sites. Protein-Struct Funct Bioinform 71:1755–1778

Shulman-Peleg A, Mintz S, Nussinov R, Wolfson HJ (2004) Protein-protein interfaces: recognition of similar spatial and chemical organizations. Algorithm Bioinform Proc 3240:194–205

Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38:D161–D166

Sonavane S, Chakrabarti P (2010) Prediction of active site cleft using support vector machines. J Chem Inf Model 50:2266–2273

Sperandio O, Andrieu O, Miteva MA, Vo MQ, Souaille M, Delfaud F, Villoutreix BO (2007) MED-SuMoLig: a new ligand-based screening tool for efficient scaffold hopping. J Chem Inf Model 47:1097–1110

Tripos (2010) Sybyl X. St. Louis, MO 63144–2319 USA, Tripos Inc

Venkatraman V, Sael L, Kihara D (2009) Potential for protein surface shape analysis using spherical harmonics and 3D Zernike descriptors. Cell Biochem Biophys 54:23–32

Via A, Ferre F, Brannetti B, Helmer-Citterich M (2000) Protein surface similarities: a survey of methods to describe and compare protein surfaces. Cell Mol Life Sci 57:1970–1977

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

Ward RM, Venner E, Daines B, Murray S, Erdin S, Kristensen DM, Lichtarge O (2009) Evolutionary trace annotation server: automated enzyme function prediction in protein structures using 3D templates. Bioinformatics 25:1426–1427

Weskamp N, Kuhn D, Hullermeier E, Klebe G (2004) Efficient similarity search in protein structure databases by k-clique hashing. Bioinformatics 20:1522–1526

Chapter 2
Identification of Pockets on Protein Surface to Predict Protein–Ligand Binding Sites

Bingding Huang

Keywords LIGSITE^{csc} • MetaPocket • Pocket identification • Protein-ligand binding sites • 3D grid • Residues conservation • Cavity • Pocket • Mathematical morphology • Q-SiteFinder • Inside protein • Near surface • In solvent • Cubic diagonals

2.1 Introduction

Proteins perform their biological functions in different cell processes mainly by interacting with other molecules such as other proteins, ligands, DNAs and RNAs etc. Not all but only parts of residues in proteins are involved in such interactions. Therefore, identification of these interacting residues on a protein is of great importance to understanding of protein functions. In the variety of interactions, the interactions between proteins and ligands have been widely studied in protein-ligand docking, in virtual screening and structure-based drug design etc. There exist a number of cavities or pocket sites on protein surface where small molecules might bind. Therefore, identification of such pocket sites is often the first step in protein ligand-binding site prediction. Many computational algorithms and tools have been developed in recent decades to predict protein-ligand binding site from identification of pockets on protein structures, such as POCKET (Levitt and Banaszak 1992), LIGSITE (Hendlich et al. 1997), CAST (Dundas et al. 2006; Binkowski et al. 2003), LIGSITE$^{CS/C}$ (Huang and Schroeder 2006), PASS (Brady and Stouten 2000),

B. Huang (✉)
Systems Biology Division, Zhejiang-California International NanoSystems Institute,
Zhejiang University, Kaixuan Road 268, 310029 Hangzhou, China

Bioinformatics Group, Biotechnology Center, Technical University of Dresden,
Tatzberg 47, 01307 Dresden, Germany
e-mail: bhuang@biotec.tu-dresden.de

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_2,

Q-SiteFinder (Laurie and Jackson 2005), SURFNET (Laskowski 1995), Fpocket (Le Guilloux et al. 2009), GHECOM (Kawabata 2010), ConCavity (Capra et al. 2009), POCASA (Yu et al. 2010), PocketPicker (Weisel et al. 2007), SiteHound (Ghersi and Sanchez 2009; Hernandez et al. 2009) and so on. Some of these methods have been described in details in other chapters. Most of the existing methods for protein-ligand binding site prediction can be classified into two types: geometry-based and energy-based. The geometry-based methods can be further classified into grid-based, sphere-based and α-shape-based (Kawabata 2010; Yu et al. 2010). In the grid based methods, the protein structure is projected into a 3D grid and the grid points are categorized into different types such as “outside protein”, “inside protein” and “near protein surface” according to their positions related to the protein. Then those grid points not inside protein are clustered using some geometry attributes and those grids points at the pocket sites can be recognized in the end. LIGSITE^{CS}, GHECOM, PocketPicker and ConCavity are the representatives of such type. In LIGSITE^{cs}, the grid points are categorized into three types: inside protein, near surface and in the solvent. For all the solvent points, a seven-direction scanning is applied. Every grid point will be evaluated by the number of SSS (surface-solvent-surface) event it has, and if the grid point has more or equal than five such events, it normally locates at a pocket site point. LIGSITE^{cs} will be explained in details in the next section. GHECOM also firstly projects the protein into a 3D grid, and the geometry attribute used in this method is mathematical morphology. It uses the theory of mathematical morphology to define the pocket region on protein surface. In mathematical morphology (Masuya and Doi 1995), there are four basic operations of dilation, erosion, opening and closing for a probe to define a pocket site. In ConCavity, a 3D grid is constructed to include the protein as well. Each grid point is evaluated and scored by the structural information and the evolutional information. In the end, the regions with many high-scoring grid points are considered to be pocket sites. In the sphere-based approaches, the common strategy is to fulfill the spheres on protein surface layer by layer and a cutting method is applied when fulfilling. The final pocket sites are that those regions which are in rich of such spheres. This kind of methods include SURFNET, PASS, PHECOM (Kawabata and Go 2007) and POCASA (Yu et al. 2010). Approaches based on α-shape include CAST and Fpocket. CAST computes the triangulations of the protein’s surface atoms and these triangulations are grouped by letting small sized ones flow towards the neighboring larger one. The pocket sites are the collection of empty triangles. Different from CAST, Fpocket uses the idea of α- sphere which is a sphere contacting four atoms on its boundary and containing no inside atom. The next step is to identify clusters of spheres close together and those clusters are potential pocket sites. In contrast to geometry-based methods, Q-SiteFinder (Laurie and Jackson 2005) aims to find pocket sites by computing the interaction energy between protein atoms and a small molecule probe. In Q-SiteFinder, layers of methyl (—CH3) probes are initialized on protein surface to calculate the van der Waals interaction energy between the protein atoms and the probes. Then the probes are clustered into many groups and are ranked by the total energy of probes. Those clusters with high energy will be the potential ligand binding sites. SiteHound (Ghersi and Sanchez 2009; Hernandez et al. 2009) is similar to Q-SiteFinder but it includes Lennard-Jones and

Table 2.1 Summary of existing methods for protein-ligands binding site prediction

	Strategy			
	Geometry			
Name	Grid	Sphere	α-shape	Energy
LIGSITE^{cs}	√			
SURFNET		√		
PASS		√		
Q-SiteFinder				√
Fpocket			√	
POCASA	√	√		
GHECOM	√			
ConCavity	√			
PocketPicker	√			
CAST			√	
SiteHund				√

The first eight methods are included in metaPocket

electrostatics energy terms and uses different types of probes to calculate interaction energy. Table 2.1 briefly summarizes the category of these existing computational methods.

In this chapter, we will focus on the grid-based method LIGSITE^{csc} and a consensus method metaPocket (Huang 2009; Zhang et al. 2011), which were developed in our group. In the next sections, we will explain the detailed algorithm of LIGSITE^{csc} and metaPocket, then the performance of these methods with other methods will be compared on the same test data-sets using the same evaluation criteria.

2.2 LIGSITE^{csc} Approach

In our LIGSITE^{csc} approach, we introduced two extensions based on LIGSITE: First, instead of capturing protein-solvent-protein events, we capture the more accurate surface-solvent-surface events using the protein's Connolly surface (Connolly 1983), and not the protein's atoms. We call this extension LIGSITE^{cs} (cs = Connolly surface). Second, we re-rank the pockets identified by the surface-solvent-surface events by the degree of conservation of the involved surface residues. We call this extension LIGSITE^{csc} (csc = Connolly surface and conservation).

The LIGSITE^{csc} algorithm proceeds as follows. First, the protein is projected onto a 3D grid (Fig. 2.1). In order to minimize the necessary grid size, we apply principal component analysis so that the principal axis of the protein aligns with the x-axis, the second principal axis with the y-axis, and the third with the z-axis. Such rotation does not affect the quality of the results and it only minimizes the necessary grid size. For each grid, we use a step size of 1.0 Å (grid space). Different grid spaces have been tested as well. Second, grid points are classified into three categories: "inside protein", "near surface", or "in solvent" using the following rules: a grid

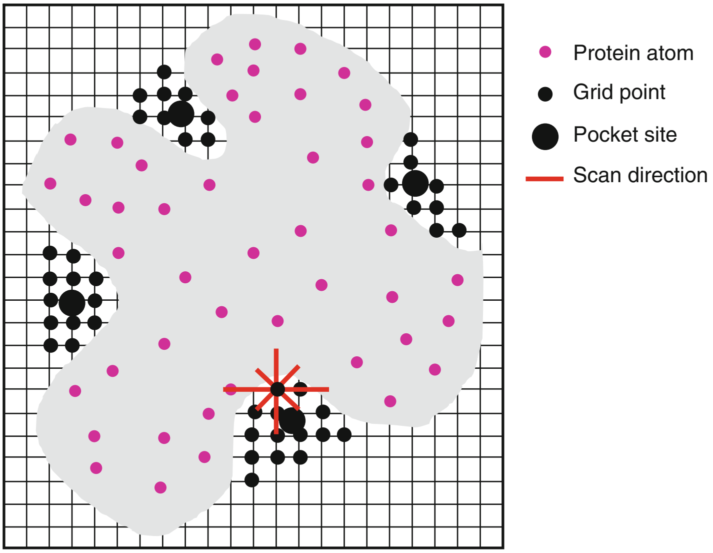


Fig. 2.1 The detailed algorithm of LIGSITE^{csc}. The protein is projected into a 3D grid (here 2D) and the grid points located at the pocket sites are identified by scanning seven directions (four directions for 2D) for solvent-surface-solvent (SSS) events

point is marked as "inside protein" if there is at least one protein atom within 1.6 Å. Next, the solvent excluded surface is calculated using the Connolly algorithm (Connolly 1983) and the surface vertices' coordinates are stored. In the Connolly algorithm, a hypothetical probe sphere (usual radius 1.4 Å) rolls over the protein. The Connolly surface is a combination of the van der Waals surface of the protein and the probe spheres surface, if the probe is in contact with more than one atom. A grid point is marked as "near surface" if a surface vertex is within 1.0 Å. All the other grid points are marked as "in solvent". A sequence of grid points, which starts and ends with "near surface" grid points and which has "in solvent" grid points in between, is called a surface-solvent-surface (SSS) event. LIGSITE^{csc} scans seven directions, the x, y, z directions and four cubic diagonals, for such SSS events. If the number of surface-solvent-surface events of a solvent grid exceeds a minimal threshold (MINSSS, 5 in this work), then this grid is marked as pocket. Finally, all pocket grid points are clustered according to their spatial proximity. I.e. if a pocket grid point is within 3 Å to a pocket grid point cluster, it is added to this cluster. Otherwise, it becomes a new cluster. Next, the clusters are ranked by the number of grid points in the cluster. The top three clusters are retained and their centres of mass are used to represent the predicted pocket sites. This first extension to the basic LIGSITE algorithm is called LIGSITE^{cs}. For LIGSITE^{csc}, the top three pocket sites are re-ranked according to the degree of conservation of the involved surface residues around the pocket sites. To be precise, the conservation score is the average conservation of all residues within a sphere of certain radius (8 Å here) of the centre of mass of the cluster. The conservation score for each residue in a given PDB ID is obtained from the ConSurf-HSSP database (Glaser et al. 2005), ranging from 1 (less conserved)

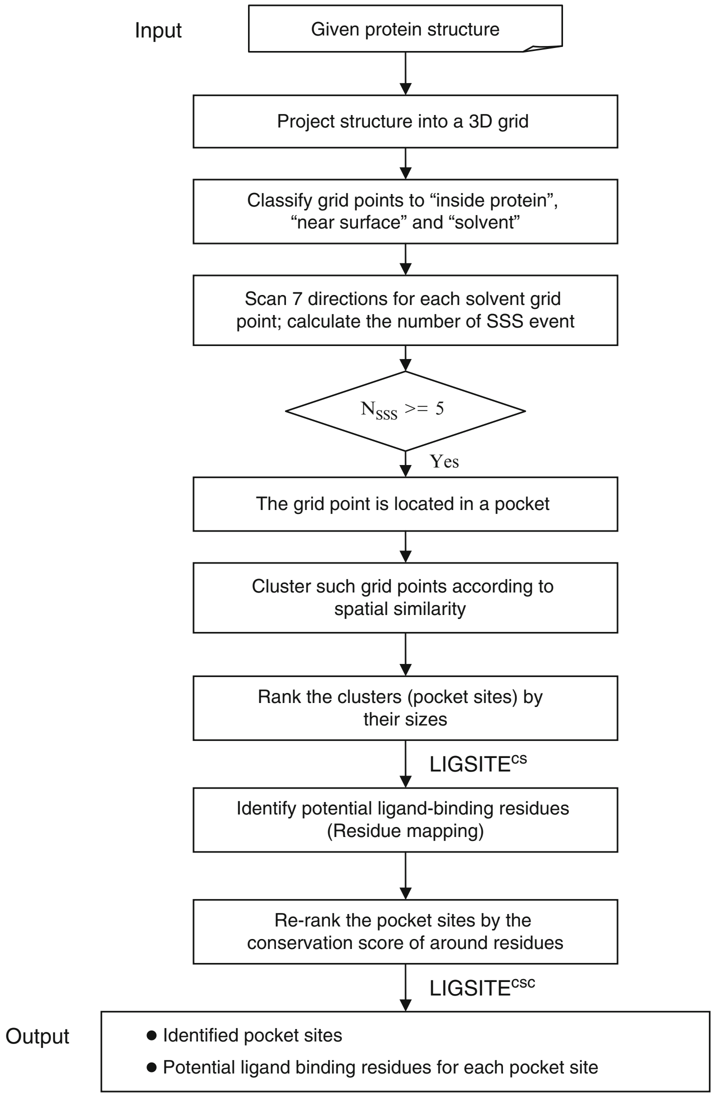

Input
Given protein structure
Project structure into a 3D grid
Classify grid points to “inside protein”, “near surface” and “solvent”
Scan 7 directions for each solvent grid point; calculate the number of SSS event
$N_{SSS} >= 5$
Yes
The grid point is located in a pocket
Cluster such grid points according to spatial similarity
Rank the clusters (pocket sites) by their sizes
LIGSITE^{cs}
Identify potential ligand-binding residues (Residue mapping)
Re-rank the pocket sites by the conservation score of around residues
LIGSITE^{csc}
Output
● Identified pocket sites
● Potential ligand binding residues for each pocket site

Fig. 2.2 The detailed workflow of LIGSITE$^{cs/c}$. The identified pocket sites in LIGSITE^{cs} are ranked by the pocket size and then re-ranked by residue conservation scores in LIGSITE^{csc}

to 9 (more conserved). This ConSurf-HSSP database pre-calculates the conservation score for all the PDB files in the PDB. However, if the users submit a new protein structure without any PDB ID, it is impossible to retrieve the conservation score from ConSurf-HSSP and thus the last re-ranking step LIGSITE^{csc} could not be applied. In such cases, geometric ranking LIGSITE^{cs} will be applied and the pocket sites are thus ranked by the pocket sizes rather than conservation score in the end. The whole process of LIGSITE^{csc} is illustrated in Fig. 2.2 in details.

2.3 MetaPocket Approach

There are two versions of metaPocket approach, MetaPocket1.0 and MetaPocket 2.0. MetPocket1.0 was developed in 2009 and it only contained four methods and the web-server is at http://metapocket.eml.org (Huang 2009). MetaPocket2.0 is an extension of metaPocket1.0 and contained four more methods developed between 2009 and 2010, recently published in the Bioinformatics journal (Zhang et al. 2011). Here we only mainly describe it as metaPocket since there is no big difference between version 1.0 and 2.0, except that four more methods are included in version 2.0.

In this section we will describe the algorithm and workflow for MetaPocket in details. In a word, MetaPocket is a comprehensive method in which the predicted sites from eight methods: LIGSITE^{cs}, PASS, Q-SiteFinder, SURFNET, Fpocket, GHECOM, ConCavity and POCASA are combined together to improve the protein-ligand binding prediction success rate. These eight methods are chosen because their developers provide source codes or executable binary and web-server available freely. MetaPocket proceeds in three steps to work: calling all single methods, meta-pockets generation and potential ligand-binding residue mapping. MetaPocket takes a standard PDB file as input, and outputs the prediction pockets and also the prediction pockets of all the successfully running single methods, and the potential ligand-binding residues around each meta-pocket. The whole workflow of metaPocket is illustrated in Fig. 2.3 and each step is explained in details as below.

Calling all single methods. In this step, the input protein structure is sent to all the single methods in parallel and separately to save total running time. For LIGSITEcs, PASS, SURFNET, GHECOM, Fpocket and ConCavity, their executable binary programs are run locally to do the prediction. For Q-SiteFinder and POCASA, python scripts are implemented to submit the protein structure to their web servers and the results are retrieved from the remote servers automatically. Thus these two methods depend on internet connection or the status of their web-servers and could fail sometimes due to bad connection and showdown of web-servers. As results, LIGSITE^{cs}, PASS and SURFNET output different clusters of grid points and the mass center of these clusters is used to represent the pocket site. For the other five methods, pocket sites are indicated by clustered probes. Thus, the mass center of each cluster is calculated and then is used as the representative point of the identified pocket sites. As we note that, each identified pocket site from every method is ranked by different scoring functions, either by the number of grid points or by the size of cluster. Thus, we can not directly compare the rankings among each pocket from different methods. To make them comparable, the z-score is calculated separately for each site in different methods according to Formula 2.1. This z-score will be used later as final scoring function in metaPocket method.

$$Z - score_i = \frac{X_i - \overline{X}}{\sigma} \tag{2.1}$$

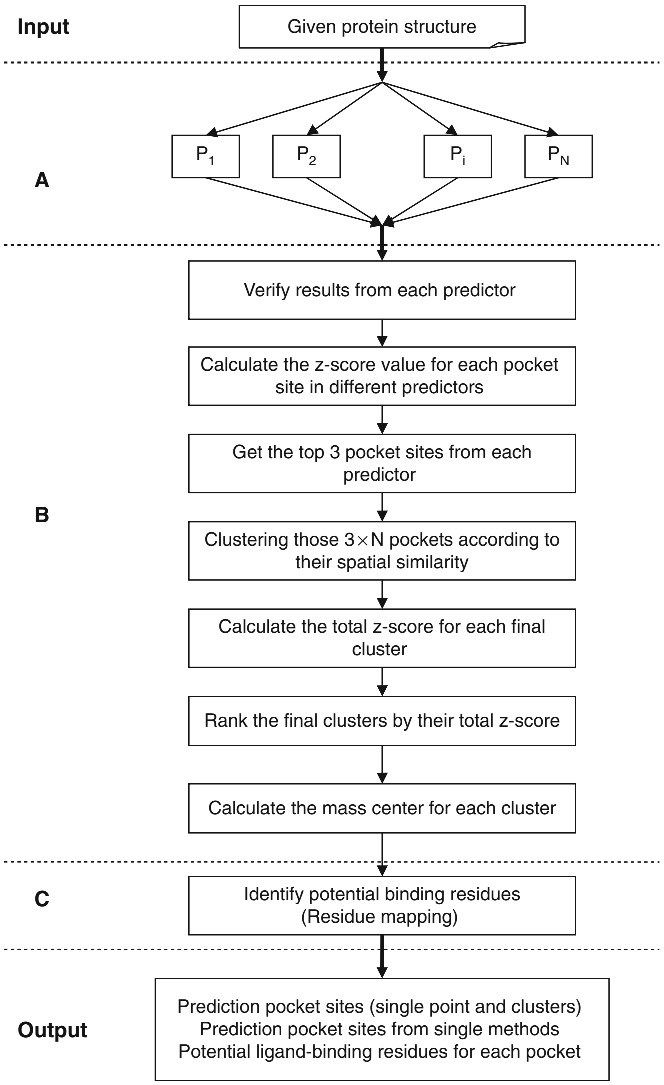

Input
Given protein structure
A
P_1
P_2
P_i
P_N
B
Verify results from each predictor
Calculate the z-score value for each pocket site in different predictors
Get the top 3 pocket sites from each predictor
Clustering those 3×N pockets according to their spatial similarity
Calculate the total z-score for each final cluster
Rank the final clusters by their total z-score
Calculate the mass center for each cluster
C
Identify potential binding residues (Residue mapping)
Output
Prediction pocket sites (single point and clusters)
Prediction pocket sites from single methods
Potential ligand-binding residues for each pocket

Fig. 2.3 The illustration of the workflow in metaPocket. Step A: calling each single method. Step B: generating meta-pocket sites. Step C: mapping potential ligand-binding residues

Generating meta-pocket sites. After calling each method, metaPocket only takes the first three pocket sites from each method into account. Thus, totally there are 24 pocket sites and these pocket sites are somehow overlapped spatially. To identify those overlapped pocket sites (we call them as “meta-pocket” sites), we use

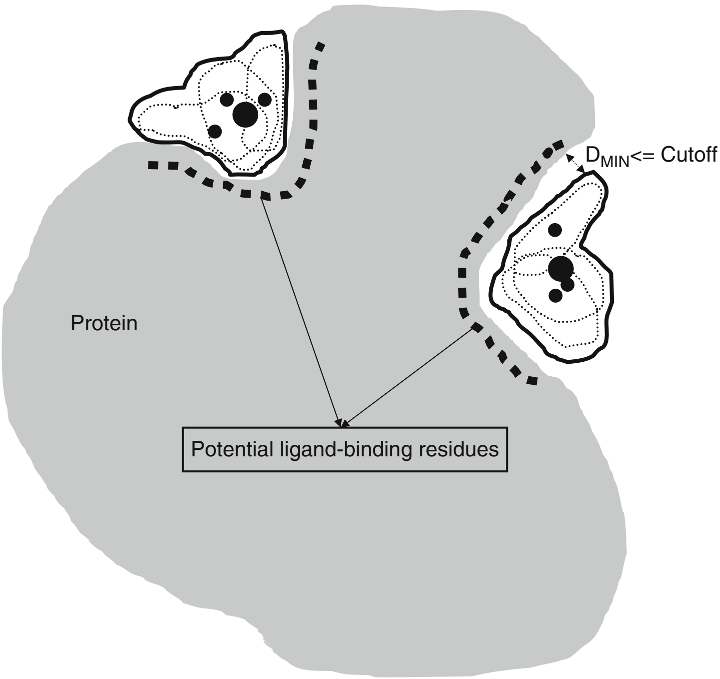


Fig. 2.4 The illustration of residue mapping step in metaPocket. The smaller spheres indicate the pocket sites from single methods and the bigger sphere indicates the meta-pocket site of metaPocket. The regions surrounded by thin dotted lines out of protein are the original clusters of each single method. The region surrounded by the thicker solid line is the cluster of the meta-pocket generated by metaPocket after merging all the clusters of single methods. The dotted line in the protein indicates the potential ligand-binding residues within a certain distance (threshold: D_{MIN}) to the cluster of meta-pocket sites

hierarchical clustering approach (single-linkage clustering) to cluster these 24 single pocket sites according to their spatial similarity. The distance cut-off threshold is set to 8 Å here. That is, two single pocket sites will be clustered into one meta-pocket site if they are within 8 Å. After clustering, the total z-score for each cluster is calculated and serve as the final scoring function to re-rank the final meta-pocket sites. In the end, the mass center for each final cluster is calculated and is represented as the final meta-pocket site in the output of MetaPocket.

Mapping ligand-binding residues around the pocket site. The purpose of this step is to identify the functional residues around the identified meta-pocket site which could be the potential ligand binding sites on protein surface. As illustrated in Fig. 2.4, metaPocket uses a synthetically way to identify those residues which might contribute to protein-ligand interaction. As we mentioned above, each method outputs a cluster of probe points for each pocket site. In this step, the probe points from each single method are merged in the same meta-pocket site. Then a big cluster of probe points is obtained for each meta-pocket site. Those surface residues, whose any atoms are within a certain distance (5 Å used here) to the probe points in the

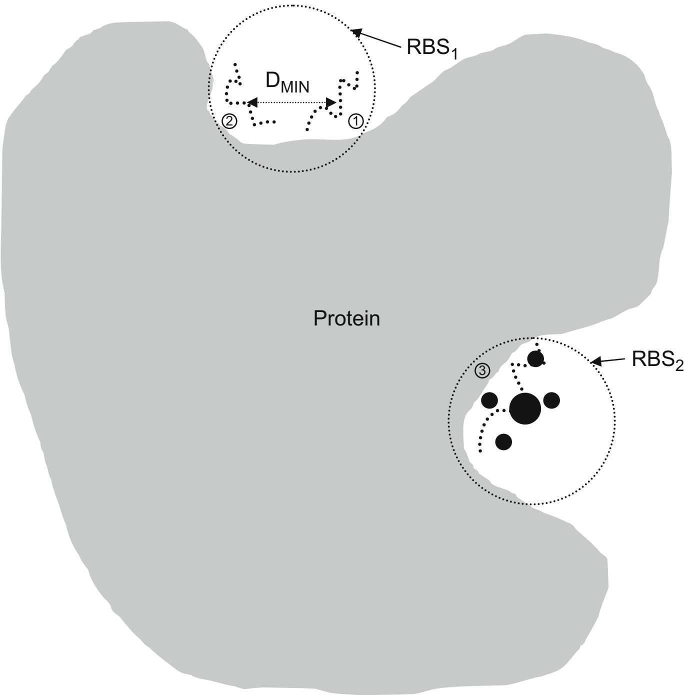


Fig. 2.5 The success definition of metaPocket. Ligands are illustrated as dotted lines and marked by 1, 2 and 3. First all the ligands bound on protein surface are clustered by their spatial distance using cutoff value DMIN. Here two real ligand binding sites are shown in circle and marked as RBS1 and RBS2 with three ligands. In RBS2, the solid spheres in different size are shown. The smaller spheres indicate the pocket sites from single methods and the bigger sphere indicates the meta-pocket site of metaPocket. If these sites are within 4 Å to any atom of the ligand, then this real ligand-binding site is successfully detected

cluster, are the potential ligand binding residues. The surface residues are defined using the NACCESS program whose relative solvent accessible surface area is more than 20%.

2.4 Evaluations of LIGSITE^{csc}, MetaPocket and Other Approaches

To evaluate and compare metaPocket with other single methods fairly, the same performance measurement and data-set should be used. It is noted that for some proteins in the data-sets we used here, more than one ligand is bound. These ligands might be separated in different pocket sites but sometimes occupy the same region on protein surface, for example, those co-factors and substrates. As illustrated in Fig. 2.5, first, we define the real ligand binding sites (RBS), which are those regions on protein surface where one or more ligands are bound. If two ligands are closed to each other (distance threshold 5 Å), they are defined to share the same RBS. Here we define that one RBS is predicted correctly if it is located at the identified pocket sites, i.e. any atom of the ligand is within 4 Å to the mass center of this pocket. We also define that a prediction is a hit if at least one RBS in the given protein is detected correctly in a certain number of top predictions. The top 1 to top 3 identified pocket sites from metaPocket and other methods are evaluated separately in this work. Thus, to compare the performance of different approaches quantitatively, the Success Rate (SR) is calculated according to the following formula:

$$Success\_Rate = \frac{N_{HIT}}{N_P} \tag{2.2}$$

Where N_P is the total number of proteins in the dataset; N_{HIT} is the total number of hit prediction. The success rate is calculated for all the methods for the top 1, top 2 and top 3 predictions, respectively.

2.4.1 Test Datasets

In the evaluation step, different datasets are being used, including 48 bound/unbound protein-ligand complexes, 210 bound protein-ligand complexes and 198 drug-target complexes. These datasets are described in details in the relevant publications (Huang 2009; Zhang et al. 2011; Huang and Schroeder 2006). For the bound protein-ligand complexes, first the ligands are removed and only the protein structures are input for pocket identification. Then the ligands are put back for success rate calculation. For the unbound prediction, the unbound protein structures are input for pocket identification and then aligned to bound protein structures. The ligands bound in the bound proteins are then used to calculate the success rate. The detailed description and the PDB structures of these three data-sets can be freely downloaded from metaPocket web-server http://projects.biotec.tu-dresden.de/metapocket.

2.5 Results

In this section, we will describe the prediction results of metaPocket, as well as the results for those eight single individual methods.

Table 2.2 shows the success rates for metaPocket and the eight single methods for these three datasets. Overall, metaPocket archived better result than each of the eight single methods. In the top1 and top2 prediction for drug-target data-set, LIGSITE^{cs} performed best among the eight single methods and metaPocket increased the success rate by 13%. In the top3 predictions, Q-SiteFinder is the best method and metaPocket also has 12% improvements. The reason why metaPocket improves the success rate is that it takes the overlapping prediction results from different

Table 2.2 The success rates (%) of the top 3 predictions by metaPocket and 8 different methods on the 48 bound/unbound dataset, 210 bound dataset and 198 drug-target dataset

	48 bound/unbound dataset			210 bound dataset			198 drug-target dataset		
Method	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
MetaPocket	**85/79**	**92/90**	**96/94**	**81**	**91**	**95**	**61**	**70**	**74**
LIGSITE^{CS}	**81/**71	**90/**79	**92/85**	**70**	**81**	88	**48**	**57**	61
PASS	58/58	79/65	83/77	51	71	79	35	50	56
Q-SiteFinder	75/52	83/60	90/75	72	85	**90**	40	54	**62**
SURFNET	42/42	56/58	60/62	42	53	57	24	30	34
GHECOM	67/67	81/77	83/81	63	73	79	39	51	56
ConCavity	75/**73**	81/**81**	83/83	73	81	84	47	53	56
Fpocket	65/56	79/77	81/85	61	73	81	31	48	57
POCASA	65/55	79/60	81/67	75	75	80	43	54	56

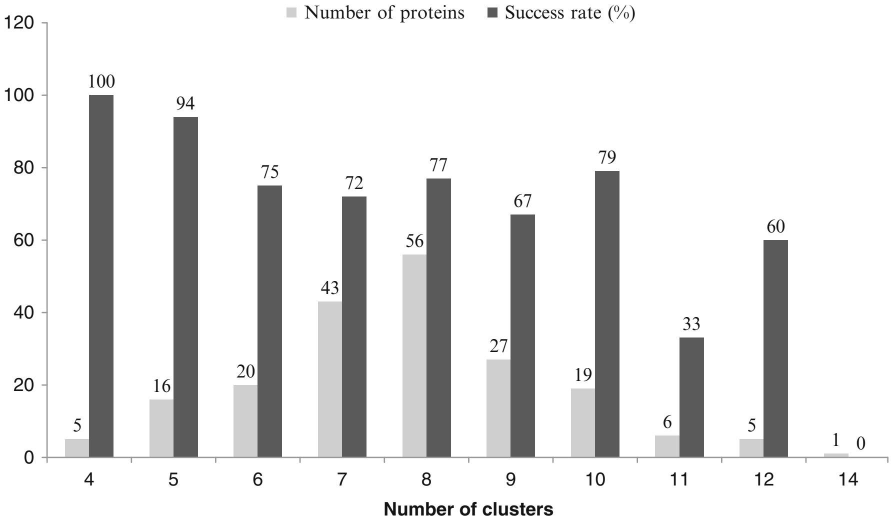


Fig. 2.6 The metaPocket prediction success rates on the drug-target data-set at the top 3 versus the number of clusters (meta-pocket sites). The number of proteins is also indicated

approaches. In general, one pocket site has higher probability to be a real ligand binding site when it is picked out by multiple methods as top predictions. This is not surprising because different pocket detection methods use different scoring functions to rank these cavities and metaPocket clusters all the identified pocket sites according to their spatial distance and re-ranks them by summing up the z-scores of different methods.

In the combining procedure of metaPocket, only the top 3 pocket sites from each of 8 single methods are taken into account and these 24 pocket sites are clustered into different clusters (so called meta-pocket site) according to their spatial similarity. In the evaluation of metaPocket on the drug-target dataset, the number of final clusters for each protein and the prediction success rates of metaPocket on those proteins are quite diverse. Figure 2.6 shows the distribution of the number of proteins with different number of clusters on the drug-target dataset, and the success rates for those proteins having the same number of clusters. Overall, the number of clusters ranges from 4 to 14, which means there are 4–14 cavities (meta-pocket sites) on protein surfaces generally. It is note that there are 5 cases, in which those 24 pockets are clustered into 4 clusters, meaning that those 5 proteins only have 4 big cavities on their surfaces and all the 8 methods correctly picked them up at their top 3 predictions. In these five cases, metaPocket all predicted the ligand-binding sites correctly. In contrast, there is only one case that the number of final clusters is 14, which indicates that this protein has 14 cavities on its surface and each of 8 methods picked up different pockets at their top 3 predictions. The real ligand binds to one of those 14 cavities and metaPocket failed to recognize it correctly at the top 3 predictions. As shown in Fig. 2.6, most of the proteins have 7 (43 cases) or 8 (56 cases) cavities

Table 2.3 Number of hit proteins in each pocket prediction class on the drug-target dataset for each method

Method	1st pocket	2nd pocket	3rd pocket	None
MetaPocket	121	17	9	51
LIGSITE^{CS}	95	18	7	78
PASS	69	30	11	88
Q-SiteFinder	79	28	16	75
SURFNET	46	11	8	133
GHECOM	78	22	10	88
ConCavity	93	12	6	87
Fpocket	61	34	17	86
POCASA	83	23	4	88

on surface generally and obviously there is no strong correlation between the number of cavities and the prediction success rate of metaPocket.

It is believed that ligands trend to binds to the large pocket site on protein surface. In order to check whether ligands bind to large pockets on protein surface, we conducted a statistical analysis to assess the possibility that a real ligand-binding site locates at the top 3 identified pockets. Here the identified pocket sites are classified into four different classes: the actual ligand binding site locates at the first, the second, the third pocket, or at none of these top 3 pockets (Table 2.3). In the top 3 predictions of metaPocket, there were 121 (61%) cases that the top-1 predicted pocket is the real ligand-binding site. There were 17 and 9 cases that the second, the third prediction pocket is the real ligand-binding site, respectively. However, there were 51 cases for which the metaPocket failed to detect the real ligand-binding site (RBS) among the top 3 predictions. Among the 121 cases that ligands were predicted to bind to the first pocket site in metaPocket, in 94 (78%) cases, the predictions overlap with one of the top 3 identified pockets identified by all of the 8 single methods and in 17 (14%) cases the predictions overlap with one of the top 3 identified pockets identified by 7 out of the 8 single methods. Only in 12 of the 121 cases, the real-ligand binding sites were predicted by all 8 single methods at the top-1 prediction.

2.6 Conclusion

To make LIGSITE^{csc} and metaPocket available to the community, we built easy-to-use web-servers and make them online at http://projects.biotec.tu-dresden.de/pocket/ and http://projects.biotec.tu-dresden.de/metapocket. Generally it only takes several seconds for LIGSITE^{csc} to finish pocket identification, depending on the size of protein. In metaPocket, eight single methods are called in parallel to reduce computational time. Each of eight single methods is treated as a plug-in and thus it is very easy to add other new methods into metaPocket, to further improve ligand binding site prediction success rate. Please note that some of these 8 methods might fail to return any prediction results due to some reasons. This plug-in pattern makes metaPocket automatically detect the failed methods and the metaPocket algorithm is only applied to those results from successful methods. The users can provide a PDB ID and a chain ID or upload their own structures. The metaPocket server will output the prediction results from eight single methods and the meta-pocket sites of metaPocket based on those results. The predicted pocket sites and those surrounding residues can be downloaded as standard PDB files to be investigated locally by the users in PyMOL (Delano 2002) or directly be visualized on the server based on JMOL (http://www.jmol.org) plug-in. Normally it only takes about 10 seconds to a few minutes for metaPocket to finish pocket identification depending on the size of protein. We envisage that our metaPocket web-server will become an all-in-one tool for protein ligand binding site prediction to the community and provide useful guide to structure-based functional annotation, site-directed mutagenesis experiments, protein-ligand docking and large scale virtual screening.

With more and more efforts being made in this field, many free computational software and web-servers are available for pocket identification and protein-ligand binding site prediction. The goal of our metaPocket approach is to combine all these free tools together and improve the ligand-binding site prediction success rate. We believe that our web-server will provide the users a comprehensive web tool in protein-ligand binding site prediction. In the future, we will continue working on it and hope to include more and more algorithms and tools into our metaPocket server.

References

Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 31(13):3352–3355

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

Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. PLoS Comput Biol 5(12):e1000585

Connolly M (1983) Analytical molecular surface calculation. J Appl Cryst 16:548–558

Delano W (2002) The PyMOL Molecular Graphics System. http://pymol.sourceforge.net/

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (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

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

Glaser F, Rosenberg Y, Kessel A, Pupko T, Ben-Tal N (2005) The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. Proteins 58:610–617

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

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 (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. OMICS 13(4):325–330

Huang B, Schroeder M (2006) LIGSITE^{csc}: predicting ligand binding sites using the Connolly surface and degree of conservation. BMC Struct Biol 6(1):19

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

Laskowski RA (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. J Mol Graph 13(5):323–330, 307–308

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

Levitt D, Banaszak L (1992) POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. J Mol Graph 10:229–234

Masuya M, Doi J (1995) Detection and grometric modeling of molecular surfaces and cavities using digital mathematical morphological operations. J Mol Graph 13(6):331–336

Weisel M, Proschak E, Schneider G (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. Chem Cent J 1:7

Yu J, Zhou Y, Tanaka I, Yao M (2010) Roll: a new algorithm for the detection of protein pockets and cavities with a rolling probe sphere. Bioinformatics 26(1):46–52

Zhang Z, Li Y, Lin B, Schroeder M, Huang B (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. Bioinformatics 27(15):2083–2088

Chapter 3
Can the Structure of the Hydrophobic Core Determine the Complexation Site?

Mateusz Banach, Leszek Konieczny, and Irena Roterman-Konieczna

Keywords Hydrophobic core • Oil drop • Tertiary structure stabilization • Fuzzy oil drop • Effective atom • Gauss function • Pair-wise interaction • Theoretical hydrophobicity distribution • Observed hydrophobicity distribution • Idealized hydrophobicity distribution • Empirical hydrophobicity distribution • Hydrophobicity deficiency • Hydrophobicity excess • Kullback–Leibler entropy • Divergence entropy • Random distribution • Structural discordance • Downhill proteins • Antifreeze proteins • Fast-folding proteins

3.1 Introduction

Stabilization of the tertiary protein structure is most often attributed to hydrophobic interactions, although this type of interaction is not specifically reflected in protein force fields. Initial attempts to extend the analysis of traditional nonbinding interactions with factors representing hydrophobic interactions (Levitt 1976) were not particularly successful, even though the influence of the aqueous environment on molecular dynamics cannot be underestimated in respect to experimental observations.

M. Banach • L. Konieczny • I. Roterman-Konieczna (✉)
Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College,
Lazarza 16, 31-530 Cracow, Poland
e-mail: myroterm@cyf-kr.edu.pl

Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University,
Reymonta 4, 30-059 Cracow, Poland

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_3,

Simulating dynamic processes in an environment where the presence of water is modeled as a large quantity of individual water molecules (each of which comprises either a single effective atom or three separate atoms arranged into a geometric shape and associated with a specific charge) is computationally difficult due to the large number of atoms involved. More specifically, the number of interacting atoms (in the protein's neighborhood as well as in the protein itself) may reach several dozen or even several hundred (Zobnina and Roterman 2009). Compounding this problem is the fact that interactions between individual water molecules and specific atoms belonging to the protein body are highly local – thus, modeling them individually does not reflect the holistic influence of water on the protein (which, as discussed in (Zobnina and Roterman 2009), drives the structural arrangement and optimization of the entire molecule rather than its constituent parts).

Stabilization mediated by hydrophobic interactions (expressing the influence of water on the protein as a whole) can be explained on the basis of the "oil drop" model (Kauzmann 1959). The model introduces the notion of a "hydrophobic core". It claims that the eventual distribution of hydrophobic and hydrophilic residues in the protein body is determined by the aqueous environment. Hydrophilic residues tend to migrate to the surface of the protein while hydrophobic residues are internalized (Kauzmann 1959).

Structural stabilization of the protein is also associated with optimization of nonbinding interactions (electrostatic and van der Waals potentials), although the optimization processes involved differ from those covered by the "fuzzy oil drop" model (discussed in this chapter). The distribution of hydrophobic interactions may indicate active sites, responsible for binding ligands and protein complexation.

3.2 Description of the Model

3.2.1 Theoretical (Idealized) Hydrophobicity Distribution

Our model assumes the existence of an idealized hydrophobicity distribution which is treated as the "target". This distribution involves a hydrophobic core (where the concentration of hydrophobicity reaches its highest value) located at the geometric center of the molecule. As we move away from the center, hydrophobicity decreases stochastically, reaching a value close to 0 on the molecule surface.

In order to accurately model such a structure, a 3D Gauss function can be applied (Konieczny et al. 2006). Traditionally, Gauss functions are used to model stochastic distribution, whereas in our model they reflect the distribution of hydrophobicity. For the sake of interpretational consistency, we can state that the values of this function correspond to the probability that hydrophobic conditions will be encountered at specific locations within the protein body.

The corresponding Gauss function is given by the following equation:

$$\tilde{H}_i^t = \frac{1}{\tilde{H}_{sum}^t}\exp\left(\frac{-(x_i-\bar{x})^2}{2s_x^2}\right)\exp\left(\frac{-(y_i-\bar{y})^2}{2\sigma_y^2}\right)\exp\left(\frac{-(z_i-\bar{z})^2}{2\sigma_z^2}\right)$$

where $\tilde{H}_i^t$ denotes the hydrophobicity density at coordinates x_i, y_i, z_i (the position of the effective atom for the *i-th* residue), while parameters $\bar{x}, \bar{y}, \bar{z}$ denote the coordinates of the central point of the ellipsoid (treated as the geometric center of the entire molecule) whose size is determined by values $\sigma_x, \sigma_y, \sigma_z$ (calculated as 1/3 of the greatest distance along each axis). The calculation of $\sigma_x, \sigma_y, \sigma_z$ follows a predetermined orientation of the molecule, with its geometric center located at the origin of the coordinate system. Inter-atomic pairwise distances are calculated for the whole molecule. The greatest distance determines the orientation of the molecule along the X axis, while the greatest distance between two projections of atom positions on the YZ plane determines its orientation along the Y axis. Values $\sigma_x, \sigma_y, \sigma_z$ can be calculated for the orientation described above. Their sum is used as a normalizing factor for the distribution.

The value of the presented function at coordinates *x, y, z* is interpreted as the corresponding hydrophobicity density.

In order to ensure uniformity of the presented model we need to determine the preferred spatial orientation of the protein molecule. The corresponding procedure is as follows:

1. Calculate the coordinates of the geometric center of the molecule;
2. Shift the molecule in such way that its geometric center coincides with the origin of the coordinate system;
3. Following the transformation applied in step 2, calculate the greatest pairwise distance between two atoms belonging to the molecule;
4. The atoms identified in step 3 determine the spatial orientation of the molecule – the line which connects them should be parallel to (or coincide with) the X axis of the coordinate system;
5. Given the new orientation of the protein molecule, project the positions of its atoms onto the YZ plane and locate two atoms for which the pair-wise distance between the corresponding projections is greatest;
6. The line connecting the two atoms identified in step 5 should run parallel to (or coincide with) the Y axis of the coordinate system. In order to achieve this, rotate the molecule about the X axis, as required.
7. Given the new orientation of the molecule, locate atoms which are separated by the greatest distance from the center of the coordinate system along each axis (two atoms for each axis);
8. Increase the distances determined in step 7 by 9 Å (the cutoff distance for hydrophobic interactions) in each direction;
9. Divide the distances calculated in step 8 by 6, thereby deriving values for $\sigma_x, \sigma_y, \sigma_z$ (in accordance with the three-sigma rule).

The H_{sum}^{t} coefficient (aggregate Gauss function value for all N points) is applied to normalize the distribution.

3.2.2 *Observed (Empirical) Hydrophobicity Distribution*

The probabilistic distribution of hydrophobicity can be understood as a reference structure, determining the "idealized" shape of the protein's hydrophobic core. It should be noted that the actual structure of the core may not fully correspond to this idealized model. In an actual protein the observed hydrophobicity distribution can be established on the basis of the locations of hydrophobic and hydrophilic residues, according to the function proposed by Levitt (1976). Levitt's function enables us to calculate the potency of hydrophobic interactions between specific residues relative to their mutual distance and their own hydrophobicity. It is given as:

$$\tilde{H}_i^e = \frac{1}{\tilde{H}_{sum}^e} \sum_j \begin{cases} \left(H_i^r + H_j^r\right)\left(1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right), & \text{for}\, r_{ij} \le c \\ 0, \text{for}\, r_{ij} > c \end{cases}$$

where $\tilde{H}_i^e$ denotes the experimentally observed hydrophobicity density at a certain point (specifically, at the position of the effective atom of the *i-th* residue). Hydrophobic interactions can be aggregated in a distance-dependent form (as given in the formula) with cutoff distance *c* equal to 9 Å (as proposed by Levitt). This aggregate value acts as a normalizing factor for the distribution. Values H_i^r, H_j^r express the hydrophobicity of the *i-th* and *j-th* residues respectively (following the scale presented in Brylinski et al. (2007a)).

Good agreement between both distributions is observed in globular proteins which possess a highly regular hydrophobic core localized centrally in the protein body with hydrophilic residues exposed on the surface. Hydrophobicity density decreases in accordance with the Gauss function, reaching values close to zero on the protein surface, which is why the molecule remains soluble.

H^r coefficients are used to express the hydrophobicity of each amino acid (any scale can be applied here). The *r* values determine distances between pairs of interacting residues (specifically, between their effective atoms whose positions are derived by averaging out the locations of all atoms belonging to side chains). *c* is the cutoff distance for hydrophobic interactions, which – following (Levitt 1976) – was taken as 9 Å. Introducing this value into the idealized hydrophobicity distribution broadens the area in which the molecule "perceives" hydrophobic interactions.

The normalization coefficient enables us to interpret values of the presented function as the likelihood that hydrophobic conditions will be encountered at specific points in the protein body. Following normalization both distributions may be used to calculate differences between the theoretical and observed hydrophobicity (or its likelihood) at any coordinates.

Since both distributions are normalized (via a coefficient which appears in both equations) the irregularity of the hydrophobicity distribution in actual proteins may be measured by comparing idealized and observed values, applying the following expression:

$$\Delta\tilde{H}_i = \tilde{H}_i^t - \tilde{H}_i^e$$

Local hydrophobicity deficiencies (expressed by large positive values of $\Delta\tilde{H}_i$) are thought to correspond to cavities capable of binding ligands. On the other hand, it is assumed that excess hydrophobicity (expressed by low negative values of $\Delta\tilde{H}_i$), particularly when observed on the protein surface, may trigger protein-protein complexation.

Thus, the positions of local minima and maxima in the $\Delta\tilde{H}$ profile may indicate residues involved in protein-protein interactions, ligand complexation or other types of interaction.

The validity of the presented model may be verified by analyzing actual proteins, both accordant with and divergent from theoretical assumptions. Identifying a protein as structurally accordant can be treated as an argument in support of the model, reflecting the influence of the aqueous environment on the protein body. On the other hand, when serious discrepancies between actual and predicted structures are observed, a thorough analysis of their underlying causes may lead to useful conclusions. Determining the reasons behind irregularities in the structure of the hydrophobic core may yield fresh insight into the mechanisms of protein folding.

A sample differential profile (highlighting the discrepancies between the expected and observed hydrophobicity distributions) is shown in Figs. 3.1 and 3.2. Figure 3.1 depicts a protein whose structure is highly accordant with theoretical predictions (1BDD) (Gouda et al. 1992), while Fig. 3.2 represents a case of poor agreement between the model and observed properties (1 G58) (Ramoni et al. 2001).

1BDD (60 aa) is a recombinant B domain (FB) of the staphylococcal protein A, which specifically binds to the Fc portion of immunoglobulin G. Its $\Delta\tilde{H}_i$ profile is shown in Fig. 3.1a 1 G58 (159 aa) is an odorant-binding protein in form of homodimer which complexes its ligand (1-octen-3-ol). Figure 3.2a presents the $\Delta\tilde{H}_i$ profile for this protein. Residues involved in ligand complexation and monomeric unit binding have been highlighted.

Figure 3.1b depicts the theoretical (T) and observed (O) distributions of hydrophobicity for the 1BDD protein while Fig. 3.2b presents the corresponding distributions for 1 G85. While the theoretical and observed distributions are in good agreement for 1BDD (as can be seen in Fig. 3.1b) they remain substantially divergent in the case of 1 G85 (Fig. 3.2b)

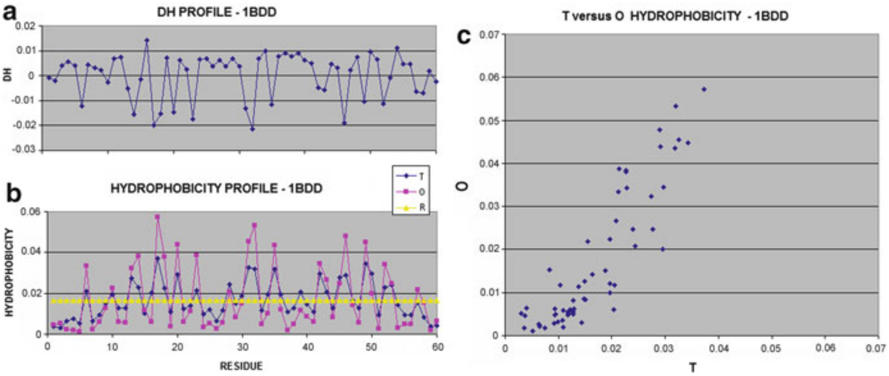


Fig. 3.1 Hydrophobicity distribution profiles for protein 1BDD whose hydrophobic core is structurally accordant with the assumed model: (**a**) differences between expected and observed hydrophobicity; (**b**) theoretical (*T*) and observed (*O*) and random (*R*) hydrophobicity distributions; (**c**) relation between theoretical (*T*) and observed (*O*) hydrophobicity values. *Pink squares* represent residues involved in protein complexation; *yellow triangles* represent residues involved in ligand binding

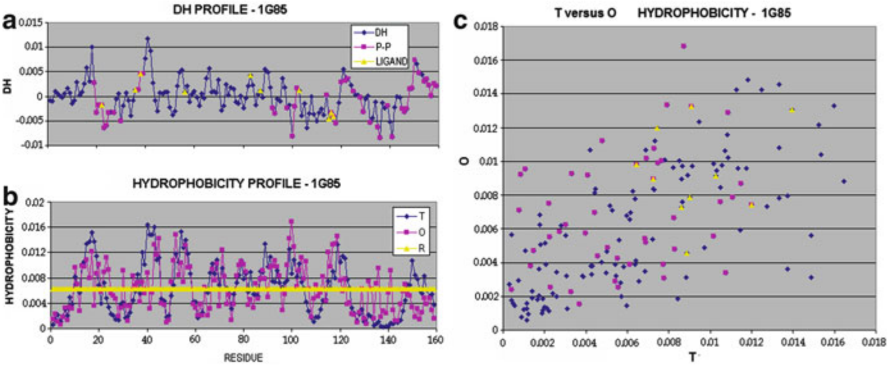


Fig. 3.2 Hydrophobicity distribution profiles for protein 1 G85 whose hydrophobic core is structurally discordant with the assumed model: (**a**) differences between expected and observed hydrophobicity; (**b**) theoretical (*T*), observed (*O*) and random (*R*) hydrophobicity distributions; (**c**) relation between theoretical (*T*) and observed (*O*) hydrophobicity values; *Pink squares* represent residues involved in protein complexation; *yellow triangles* represent residues involved in ligand binding

Of note is the arrangement of points representing the relation between expected (T) and observed (O) hydrophobicity distributions: for 1BDD they follow a linear pattern and exhibit little variance, while for 1 G85 their relative spread is much greater. This phenomenon is most likely caused by residues responsible for molecular interactions, present in 1 G85.

3.2.3 Theoretical Versus Experimental Hydrophobicity

Standardization also enables quantitative comparison of both distributions. This analysis can be based on the following definition of distance entropy proposed by Kulback and Leibler (Nalewajski 2006):

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

where *P(i)* denotes the observed probability (hydrophobicity density) as applied to the *i*-th residue, while *Q(i)* denotes the expected (target distribution) hydrophobicity for the same residue.

Two target (reference) distributions are considered when interpreting D_{KL} values. It should be noted that since D_{KL} expresses entropy, only a relative comparison between pairs of distributions can be meaningful in this scope. Hence, one D_{KL} value reflects the distance between observed (O) and theoretical (T) distributions, while another value is derived as the distance between observed (O) and random (R) distributions. Random distribution can be obtained by assigning equal hydrophobicity to each residue (thus, $R_i = 1/N$, where N is the number of residues in the polypeptide chain).

To simplify the notation, the following comparison between observed and theoretically expected distributions is introduced (Banach and Roterman 2009):

$$O / T = \sum_{i=1}^{N} O_i \log_2 \frac{O_i}{T_i}$$

Consequently, the distance between observed and random distributions is given as:

$$O / R = \sum_{i=1}^{N} O_i \log_2 \frac{O_i}{R_i}$$

The relation between O/T and O/R is taken as a classification criterion when determining whether a given protein is structurally accordant with the theoretical model. If O/T is greater than O/R, the protein does not conform to the model. We are currently developing computational tools which will enable us to objectively quantify (cluster analysis) the degree of accordance based on cluster analysis.

1 G85 exhibits far greater discordance between O and T than the protein presented in Fig. 3.1c. This discordance appears to result from distortions caused by the presence of an external molecule (Table 3.1).

An important question arises: why do some proteins exhibit significant structural discordance between theoretical expectations and observed properties of their hydrophobic cores?

Table 3.1 O/T and O/R values for 1BDD, 1 G85 and 3DRC respectively, presenting the degree of accordance between the observed hydrophobic core and the assumed "fuzzy oil drop" model

				Divergence entropy [bit]	
Protein	Protein Characteristics	Structural Form	Residues involved	O/T	O/R
1BDD	Immunoglobulin binding protein	Complete		**0.186**	**0.585**
1 G85	Odorant binding protein	Complete		0.260	0.259
	1-Octen-3-ol	No ligand	22, 36, 38, 56, 83, 87, 103, 115–118	**0.213**	**0.303**
	Homodimer	No P-P	19–25, 27, 30, 37–39, 92, 93, 98, 100–102, 114, 116–118, 120, 122, 123, 125, 128, 129, 135, 136, 139, 141, 145–150, 152, 154–159	**0.202**	**0.294**
3DRC	Dihydrofolate reductase	Complete		**0.154**	**0.184**
	Methotrexate	No ligand	93, 98, 100	**0.149**	**0.193**
	Cl^-	No ion	142, 144, 145	**0.155**	**0.185**
	EC#:1.5.1.3	No catalytic	11, 20	**0.154**	**0.188**

Instances where O/T < O/R have been highlighted

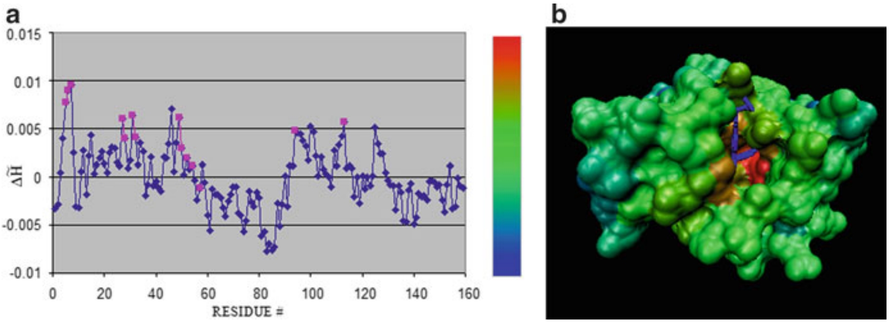


Fig. 3.3 $\Delta\tilde{H}_i$ profile and volumetric hydrophobicity distribution in 3DRC. (**a**) $\Delta\tilde{H}_i$ profile, with ligand-binding residues tagged in *pink*; (**b**) 3D representation of 3DRC with attached ligand (*dark blue*). *Red* areas indicate residues with high $\Delta\tilde{H}_i$ values – their placement in close proximity to the ligand suggests that such residues are involved in generating binding pockets

Analysis of residues for which $\Delta\tilde{H}_i$ reaches local maxima suggests that these residues are associated with localized hydrophobicity deficiencies. By the same token, residues for which $\Delta\tilde{H}_i$ reaches local minima point to areas of excess hydrophobicity.

Hydrophobicity deficiencies may be caused by the proximity of ligand-binding pockets while residues with excess hydrophobicity (local minima in the $\Delta\tilde{H}_i$ profile) permit protein-protein interactions and can therefore be responsible for protein complexation (if they are located on the surface of the protein).

Another question may be asked at this point: how can the locations of such anomalous residues in the protein's volumetric structure be determined? In order to better illustrate this issue we can depict the distribution of hydrophobicity in a folded protein using a color gradient (Fig. 3.3). Red areas indicate high $\Delta\tilde{H}_i$ values while blue areas correspond to $\Delta\tilde{H}_i$ minima. Green residues are consistent with theoretical predictions with respect to hydrophobicity.

Figure 3.3 also suggests that at least in some proteins the structure of the hydrophobic core is accordant with the theoretical model and that the factor which likely triggers distortions in the core structure is the presence of the ligand.

The following questions should now be posed:

1. Has the ligand attached to the protein molecule by compensating for its hydrophobicity deficiencies?
2. Can the ligand be responsible for irregularities which emerge in the protein structure during folding and which ensure high specificity of the resulting binding pocket?

If the former assumption holds then $\Delta\tilde{H}_i$ maxima in the hydrophobicity profile should point to the binding sites of hydrophobic ligands. Moreover, by binding to the protein molecule the ligand should compensate for its hydrophobicity deficiencies, resulting in a perfect "oil drop" structure (as predicted by the 3D Gauss function).

If, in turn, the latter assumption is true then it follows that the ligand affects the polypeptide chain during the folding process. As the ligand tries to find an optimal place to adhere to the emerging protein, the polypeptide chain "acknowledges" its presence and folds in a manner consistent with the presence of the ligand. This phenomenon can explain the highly selective nature of certain proteins (in terms of binding ligands) and enables us to search for potential binding sites based on the distribution of local maxima in the $\Delta\tilde{H}_i$ profile.

The 3DRC protein is an enzyme (dihydrofolate reductase - EC#:1.5.1.3) (3 C Warren et al. 1991) which forms a complex with methotrexate. The structure of its hydrophobic core roughly corresponds to the theoretical model, although eliminating residues responsible for ligand complexation significantly improves this alignment. Elimination of residues involved in ion binding (Cl^-) does not affect the remainder of the protein as far as the structure of its hydrophobic core is concerned. Similarly, cleaving catalytic residues has negligible impact on the core. There are, however, molecules in which the act of eliminating residues responsible for enzymatic catalysis (based on O/T and O/R values) greatly improves the agreement between the theoretical and observed hydrophobicity distributions in the remainder of the protein body.

Calculation of O/T and O/R for molecules stripped of interacting residues requires repeated normalization of O and T values (for polypeptide chains in which the corresponding residues were eliminated). The data given in Table 3.1 presents the protein 1BDD – the protein representing the hydrophobicity core structure accordant with the idealized form expressed by 3-D Gauss function.

Protein 1 G85 in its complete form represents the hydrophobic core not accordant with idealized structure. The elimination of residues (# residues engaged in ligand complexation) engaged in ligand binding (NO LIGAND) reveals the structure of the rest of protein molecule as representing hydrophobicity core structure accordant with the idealized form. It means that the deformation of hydrophobic core structure is introduced by residues engaged in ligand binding – and in consequence – presence of ligand influences the core structure. The elimination of residues engaged in protein-protein interaction (NO P-P) reveals the structure of the hydrophobic core accordant with assumed model.

Protein 3DRC represents the structure of hydrophobic core as accordant with idealized structure. Elimination of residues engaged in ligand binding (NO LIGAND) either ion complexation (NO ION) or residues engaged in catalytic activity (NO CATALYTIC) does not change the status of the hydrophobic core status which is accordant with the assumed one in all these cases.

Before we analyze the role of the ligand in protein folding we should first determine to what extent the presented "idealized" hydrophobic core model is realized in actual proteins. To achieve this goal a study has been conducted, where a variety of small proteins (ca. 60 amino acids) was checked for structural accordance with theoretical predictions. The analyzed proteins exhibited a broad range of properties and biological activity profiles. They included enzymes, chaperonins, metal-complexing proteins, scaffold proteins, protein-ligand complexes etc. (Prymula et al. 2009, 2010; Prymula and Roterman 2009; Minervini et al. 2008). Our analysis points to two groups of proteins whose structure is highly accordant with the presented model: "antifreeze" proteins (DeVries and Wohlschlag 1969; Jarov et al. 2004) and "downhill" (fast-folding) proteins (Fisher and DeLisa 2008; Dyer 2007; Ozkan et al. 2002; Zhu et al. 2003). Both groups confirm the predictions of the model (Banach et al. 2012; Roterman et al. 2011).

Fast-folding (or "downhill") proteins have been experimentally proven to possess the ability to undergo rapid and reversible folding *in vitro*. This property suggests high spontaneity of the folding process, with little reliance on external conditions. Conformance with theoretical predictions was assessed on the basis of distance entropy values (O/T and O/R) (Banach et al. 2011).

Analyzing examples of structurally accordant proteins validates our model by confirming that such proteins do indeed exist. The physiochemical properties of these proteins ("fast-folding" group) suggest that their structure is only affected by the aqueous environment. Thus, a model acknowledging the relationship between the polypeptide chain and the aqueous environment seems sufficient to determine the structural ordering of protein molecules.

When discussing antifreeze proteins, the influence of mutations should be taken into account (Banach et al. 2012). The PDB database usually lists several mutations per protein. Analysis of the hydrophobic core of 1MSI (Jia et al. 1996) indicates that while most mutations do not significantly alter the structure of this protein, specific mutations at position 16 (A16M, A16T, A16M, A16C, A16R, A16Y) result in reorganization of the hydrophobic core in a way which breaks conformance with our model. This can have far-reaching implications for the shape of the entire protein molecule and for its biological properties (Banach et al. 2011).

Having presented a selection of proteins whose structure follows the presented model we should devote our attention to the observed discrepancies. Among enzymes with well-defined active sites hydrolases appear to exhibit particularly good agreement with the "fuzzy oil drop" model. Studying their $\Delta\tilde{H}_i$ profiles points to specific areas where selected amino acids diverge from the model in terms of hydrophobicity. These amino acids correspond to sites of enzymatic activity, i.e. the binding pockets (which translate into $\Delta\tilde{H}_i$ profile maxima).

From among many classes of enzymes, our model is particularly efficient in predicting the active sites of hydrolases (Brylinski et al. 2007b, c) – this is why hydrolases have been singled out for in-depth analysis, which is presented in (Prymula et al. 2011).

3.3 Summary

In summarizing the presented work we should state that irregularities in the structure of the protein's hydrophobic core, triggered e.g. by the presence of a ligand, provide a good starting point for identification of active sites (ligand-binding and protein complexation areas). These irregularities correspond to minima or maxima of the $\Delta\tilde{H}_i$ hydrophobicity profile, which, in turn, indicates which residues are in direct contact with a ligand or with another protein molecule. Measuring deformations of the hydrophobic core yields valuable insight into the organization and structuring of the molecule as a whole. The distribution of electrostatic charges appears close to random (Marchewka et al. 2011), which suggests local optimization of electrostatic fields. Contrary to electrostatic forces, hydrophobic interactions cannot be meaningfully optimized in a pair-wise fashion – instead, hydrophobicity optimization should take into account the protein's $\Delta\tilde{H}_i$ profile. Low $\Delta\tilde{H}_i$ values (close to 0) indicate that the structure of the protein's hydrophobic core approximates the theoretical ideal. The key aspect of our work is relating deviations from theoretical predictions to the presence of ligands or other factors which may affect the distribution of hydrophobicity in the protein molecule. The "fuzzy oil drop" model was applied to simulate the environment for folding process (Brylinski et al. 2006a, b). Particularly the presence of external force field of hydrophobic character (fuzzy oil drop model) accompanying the protein folding process in the presence of ligands revealed the role of ligand directing the folding process toward the specific cavity binding the specific ligand as it was done for ribonuclease (Brylinski et al. 2006c) and hemoglobin (Brylinski et al. 2007a).

The reports supporting our assumption about the necessary participation of ligand in folding process can be found in Choi et al. (2008), Wittung-Stafshede (2002), Kopecká et al. (2011), Bushmarina et al. (2006), Kayatekin et al. (2008), Curnow and Booth (2010) although the opposite interpretation is also reported (Sakamoto et al. 2011; Bushmarina et al. 2006)

References

Banach M, Roterman I (2009) Recognition of protein complexation based on hydrophobicity distribution. Bioinformation 4(3):98–100

Banach M, Prymula K, Konieczny L, Roterman I (2011) "Fuzzy oil drop" model verified positively. Bioinformation 5(9):375–377

Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I (2012) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J Mol Model 18(1):229–237

Brylinski M, Konieczny L, Roterman I (2006a) Hydrophobic collapse in late-stage folding (in silico) of bovine pancreatic trypsin inhibitor. Biochimie 88(9):1229–1239

Brylinski M, Konieczny L, Roterman I (2006b) Fuzzy-oil-drop hydrophobic force field–a model to represent late-stage folding (in silico) of lysozyme. J Biomol Struct Dyn 23(5):519–528

Brylinski M, Konieczny L, Roterman I (2006c) Hydrophobic collapse in (in silico) protein folding. Comput Biol Chem 30(4):255–267

Brylinski M, Konieczny L, Roterman I (2007a) Is the protein folding an aim-oriented process? Human haemoglobin as example. Int J Bioinform Res Appl 3(2):234–260

Brylinski M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, Konieczny L, Roterman I (2007b) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3(5):e94, Epub

Brylinski M, Kochanczyk M, Broniatowska E, Roterman I (2007c) Localization of ligand binding site in proteins identified in silico. J Mol Model 13(6–7):665–675

Bushmarina NA, Blanchet CE, Vernier G, Forge V (2006) Cofactor effects on the protein folding reaction: acceleration of a-lactalbumin refolding by metal ions. Protein Sci 15:659–671

Choi SI, Han KS, Kim CW, Ryu K-S, Kim BH et al (2008) Protein solubility and folding enhancement by interaction with RNA. PLoS One 3(7):e2677

Curnow P, Booth PJ (2010) The contribution of a covalently bound cofactor to the folding and thermodynamic stability of an integral membrane protein. J Mol Biol 403:630–642

DeVries AL, Wohlschlag DE (1969) Freezing resistance in some Antarctic fishes. Science 163(3871):1073–1075

Dyer RB (2007) Ultrafast and downhill protein folding. Curr Opin Struct Biol 17:38–47

Fisher AC, DeLisa MP (2008) Laboratory evolution of fast-folding green fluorescent protein using secretory pathway quality control. PLoS One 3(6):e2351

Gouda H, Torigoe H, Saito A, Sato M, Arata Y, Shimada I (1992) Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. Biochemistry 31:9665–9672

Jia Z, DeLuca CI, Chao H, Davies PL (1996) Structural basis for the binding of a globular antifreeze protein to ice. Nature 384:285–288

Jorov A, Zhorov BS, Yang DS (2004) Theoretical study of interaction of winter flounder antifreeze protein with ice. Protein Sci 13:1524–1537

Kayatekin C, Zitzewitz JA, Matthews CR (2008) Zinc binding modulates the entire folding free energy surface of human Cu, Zn superoxide dismutase. J Mol Biol 384(2):540–555

Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14:1–63

Konieczny L, Brylinski M, Roterman I (2006) Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol 6(1–2):15–22

Kopecká J, Krijt J, Raková K, Kožich V (2011) Restoring assembly and activity of cystathionine β-synthase mutants by ligands and chemical chaperones. J Inherit Metab Dis 34:39–48

Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107

Marchewka D, Banach M, Roterman I (2011) Internal force field in proteins seen by divergence entropy. Bioinformation 6(8):300–302

Minervini G, Evangelista G, Polticelli F, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Wiśniowski Z, Matczyńska E, Prymula K, Roterman I (2008) Never born proteins as a test case for ab initio protein structures prediction. Bioinformation 3(4):177–179

Nalewajski RF (2006) Information theory of molecular systems. Elsevier, Amsterdam. ISBN 978-0-444-51966-5

Ozkan SB, Dill K, Bahar I (2002) Fast-folding protein kinetics, hidden intermediates and the sequential stabilization model. Protein Sci 11:1958–1970

Prymula K, Roterman I (2009) Functional characteristics of small proteins (70 amino acid residues) forming protein-nucleic acid complexes. J Biomol Struct Dyn 26(6):663–677

Prymula K, Piwowar M, Kochanczyk M, Flis L, Malawski M, Szepieniec T, Evangelista G, Minervini G, Polticelli F, Wiśniowski Z, Sałapa K, Matczyńska E, Roterman I (2009) In silico structural study of random amino acid sequence proteins not present in nature. Chem Biodivers 6(12):2311–2336

Prymula K, Sałapa K, Roterman I (2010) "Fuzzy oil drop" model applied to individual small proteins built of 70 amino acids. J Mol Model 16(7):1269–1282

Prymula K, Jadczyk T, Roterman I (2011) Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction. J Comput Aided Mol Des 25(2):117–133

Ramoni R, Vincent F, Grolli S, Conti V, Malosse C, Boyer FD, Nagnan-Le Meillour P, Spinelli S, Cambillau C, Tegoni M (2001) The insect attractant 1-octen-3-ol is the natural ligand of bovine odorant-binding protein. J Biol Chem 276:7150–7155

Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M (2011) Two-intermediate model to characterize the structure of fast-folding proteins. J Theor Biol 283(1):60–70

Sakamoto K, Bu G, Chen S, Takei Y, Hibi K, Kodera Y, McCormick LM, NakaoA NM, Muramatsu T, Kadomatsu K (2011) Premature ligand-receptor interaction during biosynthesis limits the production of growth factor midkine and its receptor LDL receptor-related protein 1. J Biol Chem 286(10):8405–8413

Warren MS, Brown KA, Farnum MF, Howell EE, Kraut J (1991) Investigation of the functional role of tryptophan-22 in Escherichia coli dihydrofolate reductase by site-directed mutagenesis. Biochemistry 30:11092–11103

Wittung-Stafshede P (2002) Role of cofactors in protein folding. Acc Chem Res 35(4):201–208

Zobnina V, Roterman I (2009) Application of the fuzzy-oil-drop model to membrane protein simulation. Proteins 77(2):378–394

Zhu Y, Alonso DO, Maki K, Huang CY, Lahr SJ, Daggett V, Roder H, DeGrado WF, Gai F (2003) Ultrafast folding of alpha3d: a de novo designed three-helix bundle protein. Proc Natl Acad Sci USA 100:15486–15491

Chapter 4
Comparative Analysis of Techniques Oriented on the Recognition of Ligand Binding Area in Proteins

Paweł Alejster, Mateusz Banach, Wiktor Jurkowski, Damian Marchewka, and Irena Roterman-Konieczna

Keywords Geometric analysis • Knowledge mining • F-measure • MCC • ROC curve • Precision • Recall • True positive • False positive • True negative • False negative • Sensitivity • Comparative analysis • Receiver operating characteristic • False positive rate • True positive rate • CASTp • Pocket-finder • QSite-finder • SuMo • ConSurf • Computed atlas of surface topography of proteins • Conservative residues • SuMo – surfing the molecules • Target protein • Easy proteins • Hard proteins • Fuzzy oil drop

4.1 Introduction

This chapter presents an analysis of the various models implemented by software packages which enable computerized identification of ligand binding sites.

In general, two distinct classes of models can be defined: those which rely on geometric analysis of binding pockets (CASTp, Q-Site-Finder, Pocket-Finder) and those based on knowledge mining (SuMo, ConSurf and the "fuzzy oil drop" model).

P. Alejster • M. Banach • D. Marchewka • I. Roterman-Konieczna (✉)
Department of Bioinformatics and Telemedicine, Jagiellonian University – Medical College,
Lazarza 16, 31-530 Cracow, Poland
e-mail: myroterm@cyf-kr.edu.pl

Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University,
Reymonta 4, 30-059 Cracow, Poland

W. Jurkowski (✉)
Computational Biology Group, Luxembourg Centre for Systems Biomedicine, University of
Luxembourg, Campus Belval 7, avenue des Hauts-Fourneaux, Esch-Belval, L 4362, Luxembourg

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_4,

The authors of the ConSurf package assumed that biologically active residues (including residues responsible for ligand binding) are conservative in character – thus, their tool searches for such conservative residues and analyzes their ability to bind ligands.

The FOD method ("fuzzy oil drop") as described in preceding chapter exploits observable differences in hydrophobicity as a useful criterion for identification of binding pockets.

The SuMo package starts with a protein-ligand complex and derives similarity metrics to determine which site is most likely responsible for binding the specific ligand.

Although each of these packages applies different methods, results are usually presented in the form of a ranked list, suggesting the most probable solutions to each problem. In our analysis we will always focus on the topmost (i.e. highest ranked) and bottommost (i. e. lowest ranked) solution from each list. The technique based on "fuzzy oil drop" model identifies only one binding area due to the form of the criterion used for recognition.

4.2 The Object of Comparative Analysis

4.2.1 Ligands – The Recognition of Their Binding Cavity in Proteins

For our comparative study we have selected proteins which form complexes with NAD^+ and FMN. When choosing ligands we considered their size (preferring large molecules) as well as the relative differences in stability of protein-ligand complexes. Both selected ligands are classified as nucleotide-like. FMN yields a stable complex, while NAD^+ is only transiently associated with the given protein (enzyme) as its complexation is rather weak (Kamburov et al. 2011; Tsai et al. 2009).

Analysis of selected proteins (enzymes) results in a comparative assessment of the accuracy of various software packages. Moreover, it also enables us to conclude of the relation between the stability of each protein-ligand complex and its biological role, as well as the properties of the binding pocket itself.

Flavin mononucleotide (FMN), or **riboflavin-5′-phosphate** (produced from riboflavin (vitamin B2) by the enzyme riboflavin kinase) acts as prosthetic group of various oxidoreductases (including NADH dehydrogenase). In NADH dehydrogenase FMN plays the role of electron carrier by being alternately oxidized (FMN) and reduced (FMNH2). FMN is a stronger oxidizing agent than NAD^+ due to its participation in both one- and two-electron transfers. It also acts as a cofactor in optical receptors sensitive to blue light (Joosten and van Berkel 2007).

NAD^+ is a dinucleotide consisting of adenosine monophosphate and nicotinamide linked by an anhydrous bridge. NAD^+ binds one proton and two electrons which act upon the amide moiety of nicotinamide (Pollak et al. 2007). A second proton is expelled into the reaction environment. Following reduction, NAD^+ (NADH) is oxidized by complex I of the respiratory chain. As a result of electron transfer (triggered by later stages of the respiratory chain), an electrochemical gradient emerges. ATP synthase can then exploit this gradient to synthesize universal energy carriers (ATP molecules) (Pollak et al. 2007).

4.2.2 Protein Data Set

The list of proteins taken for analysis was determined from PDB based on a keyword search (with FMN and NAD$^+$ as the keywords). Proteins complexed to more than the selected ligands were excluded from analysis (Rose et al. 2011).

In order to limit the redundancy of the protein test set we selected targets with pairwise sequence identity below 30 %, as reported by ClustalW (Chenna et al. 2003).

4.3 Comparison Methodology

Comparative analysis focused on the following criteria of correctness:

1. F-measure
2. MCC
3. ROC curves

These parameters were calculated to validate the end results from each of the presented programs. All of them are based on the following metrics: TP (*true positive*) – the number of residues correctly identified as involved in binding ligands; FP (*false positive*) – the number of residues falsely suspected of involvement in binding ligands (contrary to experimental data); TN (*true negative*) – the number of residues correctly identified as not involved in binding ligands, and FN (*false negative*) – the number of residues which are experimentally known to be involved in binding ligands but were not identified as such by the tested software.

TP and TN calculations were based upon the "gold standard" provided by the PDBSum database (Laskowski 2009).

4.3.1 F-Measure

F-measure specifies the so-called precision and recall properties, which are often referred to when determining the correctness of pattern recognition algorithms. They give a formal meaning to the notion of accuracy, expressing the number of cases (instances) in which a correct solution has been reached (Olson and Delen 2008; van Rijsbergen 1979).

The measure of exactness (or fidelity) is assumed to express precision, while completeness is measured by recall. Recall is computed as the fraction of correct cases among all cases that *actually* belong to the relevant subset, while precision is the fraction of correct cases among those that the algorithm *believes* to belong to the relevant subset.

High recall is understood as not missing anything (however, it may involve returning a lot of useless results, i.e. low precision). In contrast, high precision describes a situation where all of the returned results are relevant, although not all relevant results may have been returned (low recall).

F-measure is given by the following formula:

$$F - measure = \frac{TP}{TP + FP + TP + FN}$$

It effectively integrates both contributory measures (precision and recall), acknowledging the effect of TP, FP and FN:

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Higher F-measure values indicate better (more accurate) solutions.

4.3.2 MCC

Another correctness measure sometimes applied in research is called MCC – the Matthews Correlation Coefficient (Altman and Bland 1994; Baldi et al. 2000; Matthews 1975; Carugo 2007). It is derived directly from the so-called confusion matrix and given by the following formula:

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Sensitivity (also called recall rate) measures the proportion of actual positives which are correctly identified as such (number of residues correctly recognized as involved in ligand binding). In contrast, specificity is the fraction of correctly identified negatives (number of residues correctly recognized as not involved in ligand binding). It is worth noting that these coefficients closely correspond to the concept of type I and type II errors.

4.3.3 ROC Curve – Receiver Operating Characteristic

Our comparative study is further augmented by ROC curve analysis (Fawcett 2006).

The receiver operating characteristic (or ROC curve for short) is a graphical representation of the relation between sensitivity (expressed as TPR, i.e. true positive rate) and specificity (determined by FPR – false positive rate) for a binary classification which bases upon some measurable quantity.

ROC constitutes the graphical equivalent of the so-called contingency table (or confusion matrix):

	Actual value	
Prediction	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

TPR (equivalent to sensitivity) is expressed as:

$$TPR = \frac{TP}{TP + FN} \qquad \text{Y-axis}$$

FPR (equivalent to 1-specificity) is expressed as:

$$FPR = \frac{FP}{TN + FP} \qquad \text{X-axis}$$

As the ROC curve is meant to visualize the dependence of TPR on FPR, FPR values are typically plotted along the X axis, while the Y axis represents TPR values.

The “fuzzy oil drop” model focuses on $\Delta\tilde{H}$ – the difference between the expected and observed potency of hydrophobic interactions at specific points in the protein molecule. We assume that significant discrepancies between these two values point to the presence of a ligand which distorts the protein’s own structural form. In this sense, the binary classification mentioned above determines which amino acids, representing local maxima (or minima) on the $\Delta\tilde{H}$ scale are actually involved in binding ligands. This comparative method of determining the accuracy of theoretical predictions is only applicable to the “fuzzy oil drop” model due to variations in the $\Delta\tilde{H}$ cutoff values for TPR and FPR parameters respectively. Since classification can focus either on local maxima (hydrophobicity deficiency) or minima (hydrophobicity excess), ROC curves are plotted for each criterion separately. The goal is to determine whether the ligand binds to a cavity representing a local hydrophobicity deficiency, or is attracted to residues characterized by excess hydrophobicity where its presence can shield such areas from direct contact with water.

The ROC analysis is applicable solely for “fuzzy oil drop” model since this model only bases on the quantitative measurements of the identification criterion which is the value (cutoff level) for $\Delta\tilde{H}$ values. The other models deliver only binary solutions expressed as YES – residue engaged in complexation and NO – residue not engaged in complexation. This is why the ROC curve analysis is presented only to interpret the results based on “fuzzy oil drop” model.

4.4 Tools Under Consideration

4.4.1 Geometry-Based Techniques

4.4.1.1 CASTp – Computed Atlas of Surface Topography of Proteins (http://sts.bioengr.uic.edu/castp)

CASTp identifies ligand binding sites by studying the geometric properties of protein pockets under the assumption that ligands are naturally attracted to depressions in the protein body. Thus, its core algorithm searches the 3D representation of the protein for pockets capable of housing a solvent molecule with a diameter of 1.4 Å. The authors refer to such pockets as "mouths". In contrast, a cavity is a depression which remains inaccessible to the solvent molecule and therefore does not have the properties of a "mouth". Identification of binding sites bases on a computational geometry model capable of locating "pockets" and "cavities". Cavity determination parameters are not dependent on the rotation of the molecule. Moreover, CASTp does not employ grid coordinate analysis (Liang et al. 1998a). Cavities are identified by way of weighted Delaunay triangulation and applying the alpha complex for shape measurements (Edelsbrunner and Mucke 1994; Edelsbrunner 1995; Facello 1995; Edelsbrunner and Shah 1996; Edelsbrunner et al. 1998, 1995). These methods return the surface of the accessible pockets as well as of internal (inaccessible) cavities. For each cavity the program calculates its area, volume and solvent-accessible surface (with respect to molecular surface).

In our analysis from among the files listing atoms representing all detected pockets and cavities, the extracted amino acids corresponding to a single, specific protein chain were selected and then compared with the reference database (PDBSum – used as the golden standard).

A detailed discussion of CASTp algorithms can be found in Liang et al. (1998a, b); Binkowski et al. (2003) and Dundas et al. (2006).

4.4.1.2 Pocket-Finder (http://www.modelling.leeds.ac.uk/pocketfinder/)

Pocket-Finder is an extension of an existing software package called Ligsite, developed by Hendlich et al. (1997). It uses a grid system with a resolution of 0.9 Å, centered upon the target protein. A scanning probe with a radius of 1.6 Å along each axis is used. This probe also enables testing cubic diagonals. Identifying a pocket requires locating an area where a grid point which belongs to the protein molecule is adjacent to a grid point which represents empty space, which is itself adjacent to another protein-bound point. Identification of the status of each grid point is performed in seven directions, resulting in seven separate results for each point. If five of these results are positive, the empty space is treated as a cavity.

PDB-derived proteins are scanned for ligands. If the contact between molecule and protein is possible, this molecule is treated as possible ligand.

The Pocket-Finder tool applies geometric criteria to locate potential binding sites and the residues which attach to them. These areas are then listed in the order of diminishing probability that a given pocket actually represents a binding site. We have conducted calculations for a set of PDB proteins by applying default criteria, i.e. stripping ligand atoms prior to scanning the structure of the protein for potential binding pockets.

From among the listed atoms identified as belonging to the first detected pocket we have extracted amino acids corresponding to a single, specific protein chain and then compared these results with the reference database (PDBSum – treated as golden standard).

If no atoms could be identified/displayed for the first detected site (“Residues” box) we selected the first nonempty site.

4.4.1.3 Q-Site-Finder (http://bmbpcu36.leeds.ac.uk/qsitefinder/help.html)

Q-Site-Finder probes the surface of the protein by using the hydrophobic -CH3 moiety as a tester which attempts to bind to the protein body. The tool analyzes the energy of interactions and optimizes the resulting complex. Energy values are then subjected to clustering and a ranking list is produced, presenting the most optimal arrangements. Such optimized structures are treated as potential complexation sites (Laurie and Jackson 2005).

4.4.2 Knowledge-Based Tools

4.4.2.1 ConSurf – Conservative Residues (http://consurf.tau.ac.il/)

The ConSurf-DB program attempts to find evolutionary conservative profiles for proteins stored in the PDB database (Landau et al. 2005). Amino acid sequences similar to the one being analyzed are aggregated and subjected to multiple alignment passes using PSI-BLAST and MUSCLE tools. The Rate4Site algorithm based on empirical Bayesian inference is implemented in ConSurf to measure the degree of evolutionary conservativeness for each amino acid in the polypeptide chain. Phylogenetic relations are taken into account when identifying relations between the aligned proteins and the stochastic nature of the evolutionary process. A particularly helpful feature is the ability to assess the pattern visually, using a 3D representation of the protein body to determine which residues are important from the point of view of the protein’s biological properties (Goldenberg et al. 2009).

Input data is given as a .pdb file containing the structure of a protein. Upon parsing the file, the program automatically performs a search for homologous proteins with a well-known 3D structure, using the PSI-BLAST tool (Altschul et al. 1997). Default sequence alignment is obtained by using the MUSCLE algorithm (Edgar 2004), although this can be replaced with CLUSTALW (Thompson et al. 1994).

In our calculations we have applied MUSCLE, as the authors claim that it is both more efficient and more accurate than the alternative solution. Subsequently, the tool constructs a phylogenetic tree by applying the neighbour joining (NJ) algorithm (Pupko et al. 2002). The resulting tree consists of homologues singled out in the previous stage. Finally, a sequence of "conservation scores" is calculated using empirical Bayesian (Mayrose et al. 2004) or Maximum Likelihood (Pupko et al. 2002) algorithms. In our research we applied the default "Evolutionary Substitution" settings.

The end result is a three-dimensional representation of the protein structure, which can be visualized using FirstGlance (Ashkenazy et al. 2010). The surface of the protein is tagged with "conservativeness scores" using various colors. ConSurf also generates output pdb files, containing the structure of the protein along with the identified amino acids for which "conservativeness scores" have been determined.

4.4.2.2 Fuzzy Oil Drop Model

This model relies on identifying irregularities in the distribution of hydrophobicity within the protein molecule. These irregularities are then compared with an idealized, theoretical distribution obtained by using a 3D Gauss function (Konieczny et al. 2006; Banach et al. 2012) as representing the highest hydropgobicity at the central part of ellipsoid (or sphere if the size of drop is equal along each direction) with hydrophobicity decrease according to the increase of distance versus the center of the protein molecule reaching the level close to zero at the surface of protein body. Such idealized hydrophobicity distribution is expected to be identified in a special group of proteins (like downhill proteins). It is assumed that the irregularity of hydrophobicity distribution in protein body represents the intentional character what means it is function related. Thus comparison of expected and theoretical $\tilde{H}$ profiles is performed. Plotting the distribution of $\Delta\tilde{H}$ (differences between expected and observed hydrophobicity) along the polypeptide chain reveals residues for which $\Delta\tilde{H}$ reaches high values on the positive or negative scale. According to the theoretical model, the former are suspected of involvement in binding ligands (usually of the hydrophobic or emphiphilic variety) while the latter – if exposed on the surface of the protein – may participate in protein complexation, resulting in multi-protein aggregates.

Ligand binding sites are thus determined by searching for residues with either very high or very low $\Delta\tilde{H}$ values (local minima or maxima). Proper identification relies on selecting a cutoff threshold, starting with the $\Delta\tilde{H}$ profile maximum. Performing calculations for consecutive cutoff thresholds yields a ROC curve which represents the given protein. Accuracy of predictions may then be determined by calculating the surface area bounded by two curves – a diagonal and the TPR-vs-FPR curve. The greater the area, the more accurate the results.

4.4.2.3 SuMo (Surfing the Molecules)

(http://sumo-pbil.ibcp.fr/cgi-bin/sumo-help?topic=prediction)

The model implemented by this program is based on the analysis of 20,000 proteins with known 3D structures. 11,000 categories of ligand binding sites have been recognized by the authors (Jambon et al. 2003, 2005). The protein under consideration is compared against all categories. The geometry of the binding site is represented as a simplified graph, based on chemical group triplets. The positions of groups with respect to one another (edge lengths) determine each functional triangle. Graph similarity is taken as an identification criterion, where the resulting graph is compared with the 11,000 previously mentioned categories. Potential protein-ligand interaction sites (as well as the recognized ligand molecules) are produced as output, along with a ranking list of all possible solutions.

The analysis described in this work bases on the following user-defined options:

- single selected chain;
- single, specific ligand (FMN or NAD$^+$);
- amino acid numbers extracted from a text-based output file written by the software tool;
- TP, FP, FN and TN values calculated by relying on PDBSum data (Laskowski 2009).

4.5 Ligand Binding Site Recognition – Comparable Analysis

Results obtained with CASTp, Pocket-Finder, QSite-Finder, ConSurf, Sumo and the Fuzzy Oil Drop (FOD) model were subjected to validation, as described above. Rankings obtained by applying the MCC criterion diverge from corresponding F-measure rankings – in particular, the top and bottom parts of each list tend to contain the same proteins, but in a different order. Proteins for which MCC and F-measure values could not be calculated, or for which the programs did not arrive at any solutions, were omitted. We also skipped proteins which could not be correctly processed by a given program due to size restrictions – thus, a different number of solution is listed for each software package.

Results are presented separately for NAD$^+$ and FMN ligands and subdivided into geometry- and knowledge-based models. Within each group software packages are discussed alphabetically.

4.5.1 NAD$^+$ Complexing Proteins

4.5.1.1 Geometry-Based Packages

Figure 4.1 presents the results of geometry-based analysis of NAD$^+$ binding pockets, as performed by the CASTp tool. The MCC measure exhibited the lowest variance of binding site identification validity (compared to other software packages),

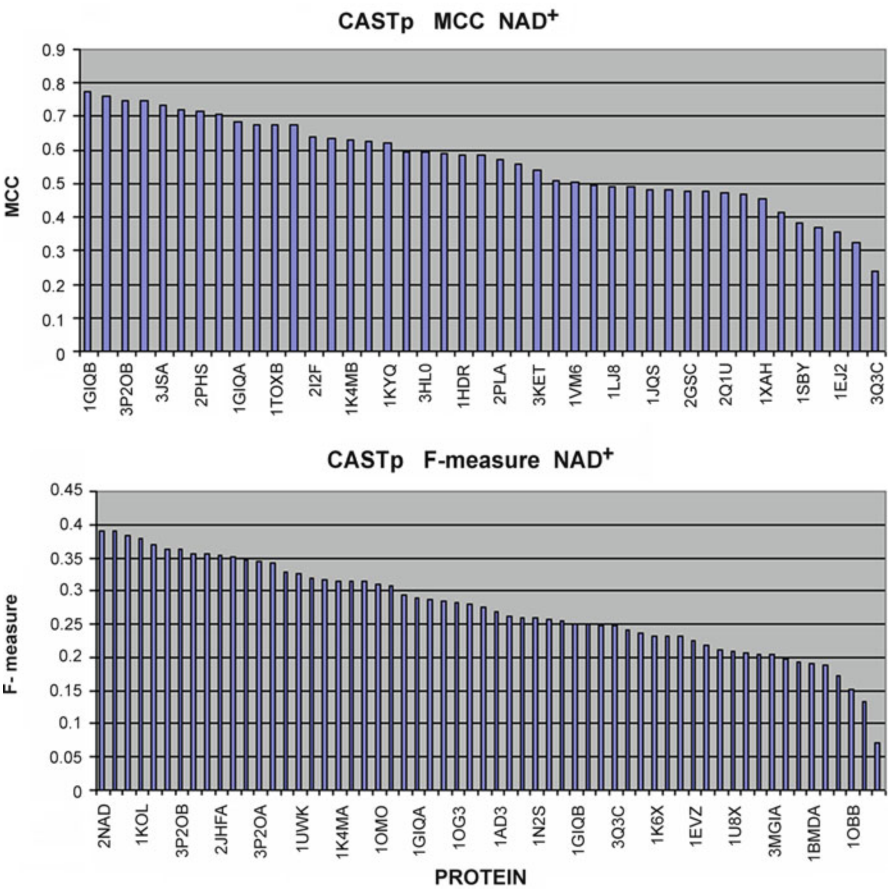


Fig. 4.1 Comparison of NAD$^+$-complexing proteins as reported by CASTp. *Top*: MCC; *bottom*: F-measure

ranging from approximately 0.8 to a little over 0.2. Both measures (MCC and F-measure) remained in fairly good agreement with each other – thus, we can conclude that the results of NAD$^+$ binding site identification as returned by CASTp are relatively trustworthy.

Of note are the quantitative differences between MCC and F-measure rankings, which affect the ordering of each list (although they are usually restricted to close neighbors).

MCC values reported by PocketFinder indicate relative uniformity of results, with approximately half of the tested proteins accorded values between 0.4 and 0.5 (Fig. 4.2). Only two proteins scored negative values on the MCC scale. In contrast, F-measure values are significantly more varied (a characteristic shared with other software packages), with most results falling in the 0.1–0.3 range.

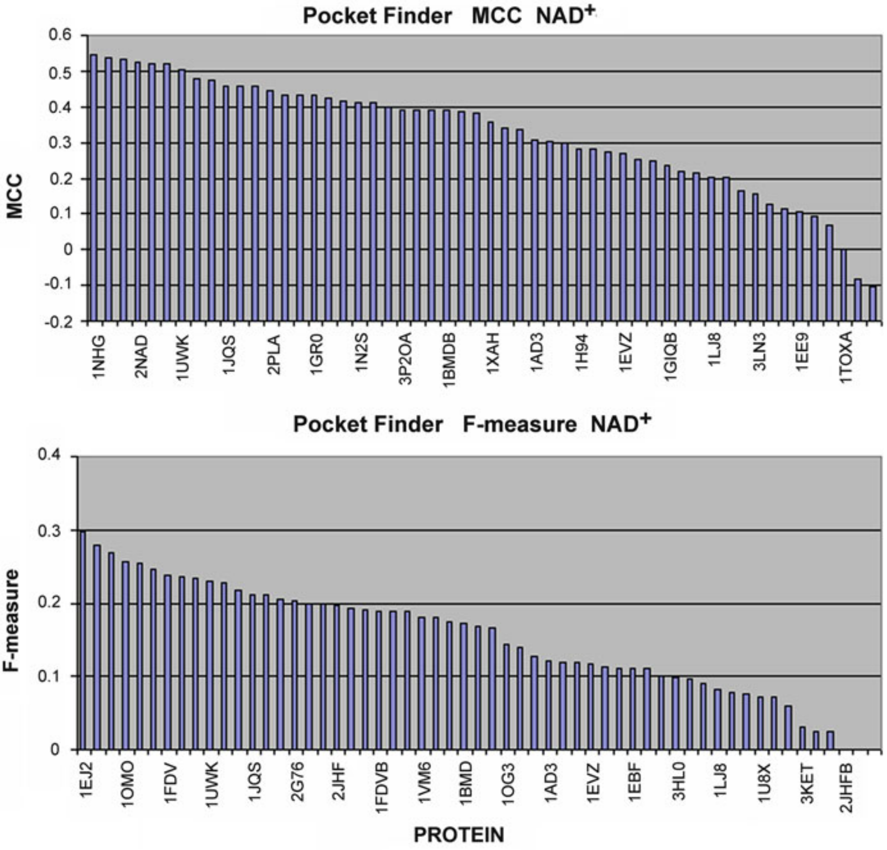


Fig. 4.2 Comparison of MCC and F-measure results for NAD$^+$-complexing proteins as reported by PocketFinder

High MCC values reported by Q-SiteFinder indicate accurate solutions, although results were more varied than those obtained by CASTp. For several proteins MCC values turned out negative, which suggests the majority of FN and FP (Fig. 4.3).

4.5.1.2 Knowledge-Based Tools

The validity of results derived by tracking evolutionary relationships falls in the 0.15–0.55 range, with only a single negative MCC result. Comparing MCC and F-measure values indicates that both methods yield relatively similar results for a broad range of proteins. We can conclude that the evolutionary approach is a useful method for studying the properties of various proteins. On a more general level, it seems that the generation of binding pockets in NAD$^+$-complexing proteins is significantly determined by evolutionary factors (Fig. 4.4).

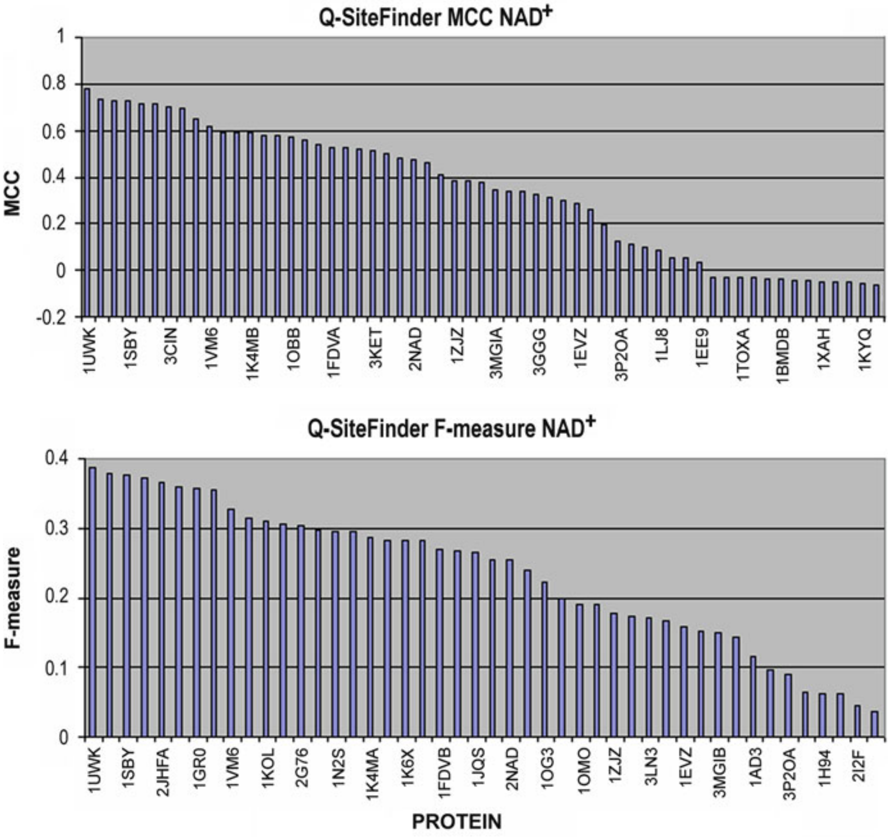


Fig. 4.3 Comparison of MCC and F-measure results for NAD$^+$-complexing proteins as reported by Q-SiteFinder

The assessment of F-measure values produced by the FOD model points to several proteins where such values are particularly high. In each case we find that the ligand is anchored in a very deep pocket – it seems that the presence of such a deep void distorts the structure of the protein's hydrophobic core (Fig. 4.5).

The FOD method is the only one of the presented techniques where results depend on an assumed cutoff threshold, establishing a discrete transformation over the $\Delta\tilde{H}$ function (i.e. the difference between the assumed and observed hydrophobicity distribution). All results discussed in this chapter are based on an 80% threshold, where only those residues for which the $\Delta\tilde{H}$ function value is above 80% of its peak are suspected of involvement in binding ligands.

MCC values fell in the 0.1–0.5 range for most proteins, with approximately 10 cases of negative values being reported.

The SuMo package returns fairly consistent results for MCC and F-measure, singling out several proteins where the values are particularly high and producing

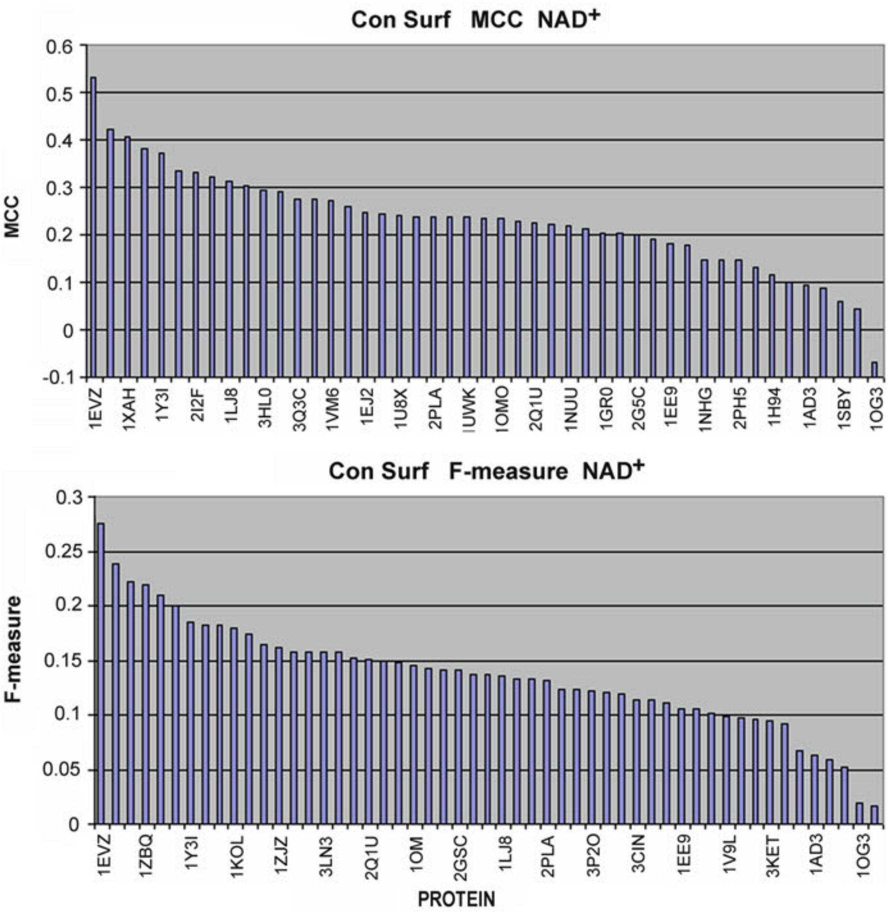


Fig. 4.4 Comparison of MCC and F-measure results for NAD$^+$-complexing proteins as reported by ConSurf

comparable results for most other proteins. In approximately 10 cases MCC values were below 0 (Fig. 4.6).

4.5.2 *FMN Binding Site Identification*

4.5.2.1 Geometry-Based Packages

The results produced by the CASTp tool for FMN-binding proteins are presented in Fig. 4.7. The figure omits cases in which the returned values were below 0. Three proteins approach 1 on the MCC scale, while F-measure values indicate high reliability of most of the obtained results (values between 0.2 and 0.5).

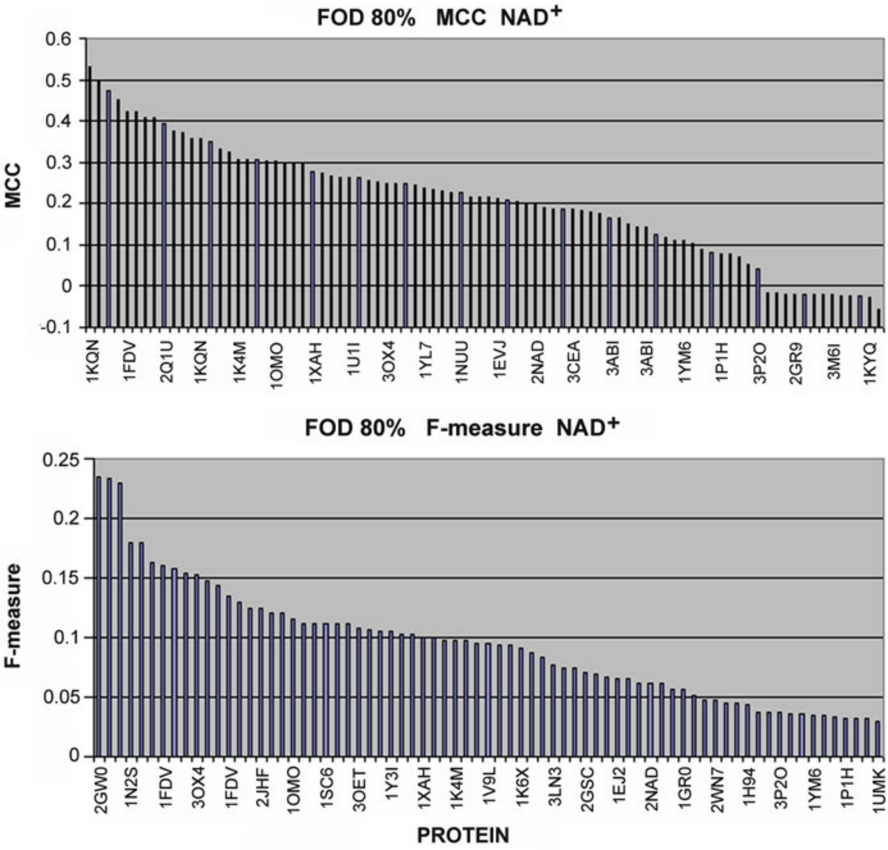


Fig. 4.5 Comparison of MCC and F-measure results for NAD+-complexing proteins as reported by the FOD model. The cutoff threshold was established at 80 %, meaning that only values beyond 80 % of the maximum were considered valid

The PocketFinder tool also produced high MCC values for most of the tested proteins, with only two molecules ranked below 0 (Fig. 4.8). As with most other software packages, F-measure scores are somewhat more diverse than MCC results, although they remain comparable (falling between 0.2 and 0.3 in most cases).

The QSiteFinder tool identified a relatively numerous group of proteins with high MCC and F-measure scores – notably a set of molecules for which MCC values are in the 0.4–1.0 range (Fig. 4.9). These results are supported by the corresponding F-measure scores.

4.5.2.2 Knowledge-Based Packages

Consistent F-measure results were obtained by applying the ConSurf package. Good agreement between MCC and F-measure scores indicates high reliability of evolutionary methods in the context of FMN-complexing proteins (Fig. 4.10).

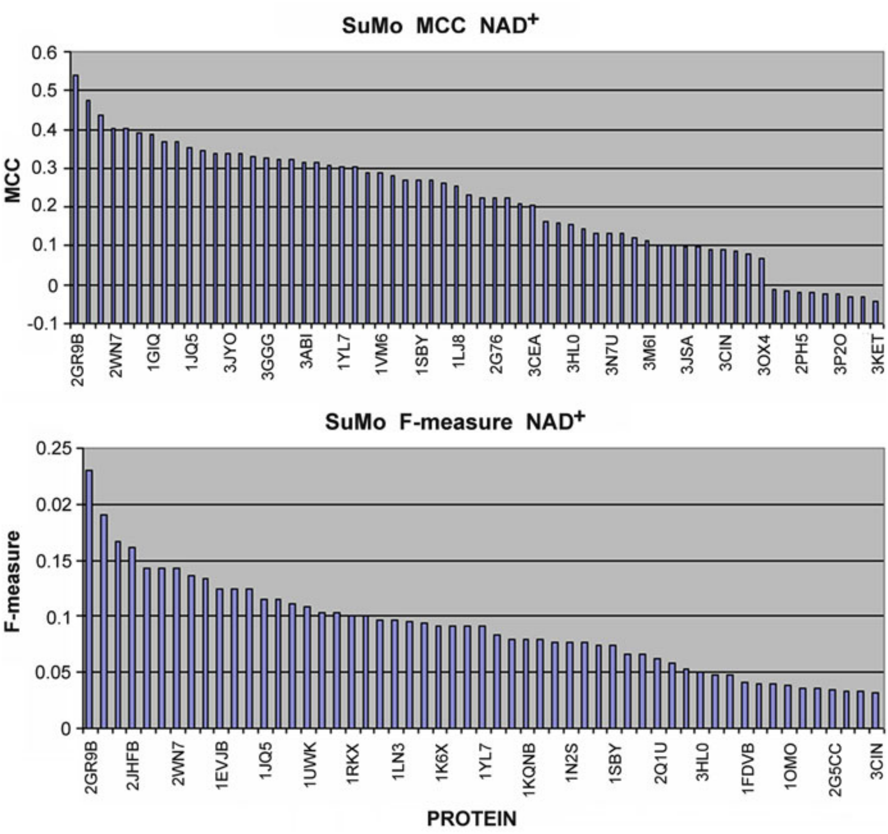


Fig. 4.6 Comparison of MCC and F-measure results for NAD$^+$-complexing proteins as reported by Sumo

We can conclude that the pocket to which this ligand binds (and hence the biological function performed by the protein) remain fairly conservative from an evolutionary viewpoint.

The FOD model yielded significantly lower scores for FMN binding sites compared with NAD$^+$ (Fig. 4.11). Of particular note are the low values of the F-measure metric. This suggests that the mechanism responsible for FMN complexation affects the structure of the protein's hydrophobic core to a far lesser degree than in the case of NAD$^+$. Indeed, FMN complexation is relatively static, i.e. the ligand remains securely lodged in a specific binding pocket, whereas NAD$^+$ binding is more dynamic and the complex exists only for a brief while, limiting the likelihood of errors. This is why it may be easier to distinguish NAD$^+$ complexation sites.

Similarly to other tools (except the FOD model), SuMo reports higher values of MCC and F-measure metrics for FMN as compared to NAD$^+$ (Fig. 4.12). It appears that FMN binding sites are more accurately determined by the geometry of the binding pocket than NAD$^+$ binding sites. MCC scores seem fairly consistent, while F-measure values are more varied.

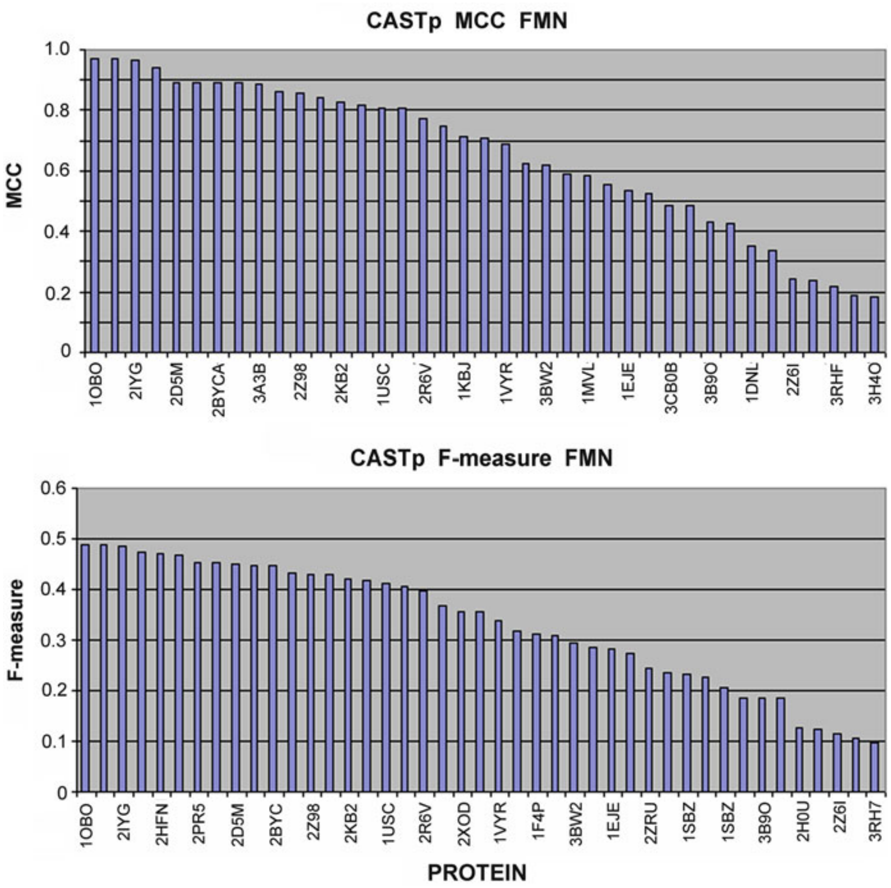


Fig. 4.7 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by CASTp

4.5.3 Properties of Target Proteins

Further analysis of the presented results suggests that proteins can be divided into “easy” and “difficult” from the point of view of identifying binding pockets. In order to visualize the aggregate scores, we have compared 10 proteins for which each program reported (respectively) best- and worst-case results. This comparison is presented in Table 4.1 and in Figs 4.13 and 4.14. Unique matches, i.e. proteins for which binding pockets were correctly identified only by a single program (corresponding to placement in the top 10 results on a given list), have been highlighted.

Comparing Figs. 4.13 and 4.14 reveals differences with respect to the protein being considered as well as the applied computational model. Even though all of the proteins on each list belong to the “top 10” group, their corresponding MCC and F-measure scores are quite varied.

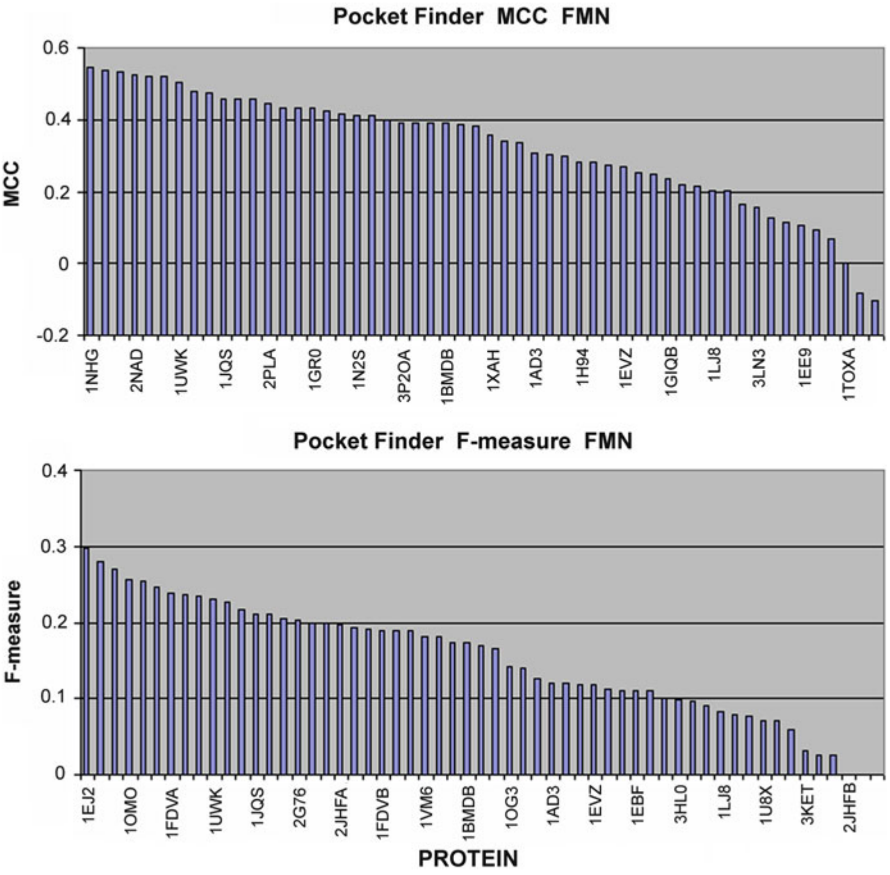


Fig. 4.8 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by PocketFinder

Differences in the validity of binding site identification for NAD$^+$- and FMN-complexing proteins are depicted in Figs. 4.15 and 4.16 respectively. We have selected proteins which obtained a score of 0.8–1.0 on the MCC scale and a score of 0.4–0.5 on the F-measure scale for FMN, as well as those with MCC values of approximately 0.8 and F-measure values of approximately 0.4 for NAD$^+$. The presented tables compare the scores obtained for each of those proteins using various theoretical models. As can be seen, different tools exhibit varying degrees of accuracy in identifying specific binding pockets.

An interesting conclusion arises with respect to FMN complexation sites: in this scope all of the analyzed software packages seem to have arrived at an identical set of "difficult" proteins, while also producing consistent results for "easy" ones. Result sets obtained from ConSurf and CASTp contained no unique matches. Thus, both packages can be described as relatively trustworthy but also largely incapable of handling unusual situations.

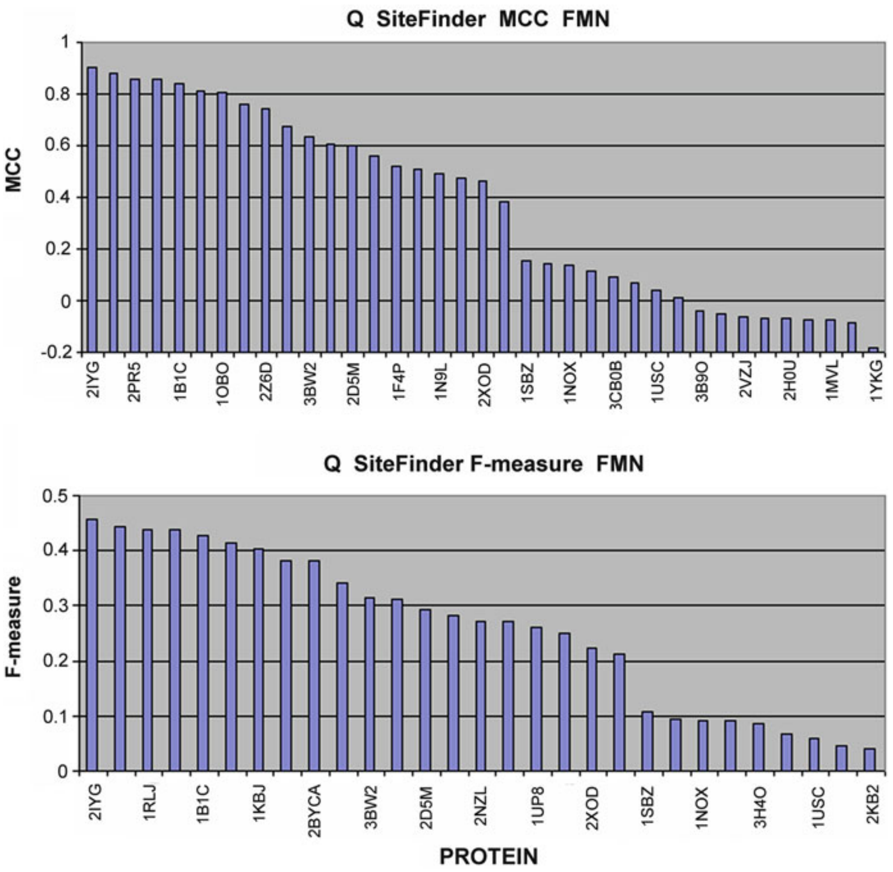


Fig. 4.9 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by QSiteFinder

NAD+ identification results were far more varied, with some programs (particularly CASTp and SuMo) producing correct results for proteins which could not be correctly processed by most other packages. We should also note that the phenomenon of polypeptide chain complexation (in the form of dimers, tetramers and – occassionally – decamers) presents significant problems for the FOD model as it distorts each chain's hydrophobic core, producing many false positives.

Top row (left to right): 2BYC, 1B1C, 2IYG (binding site easy to identify for all programs except FOD model)

Bottom row (left to right): 3H4O, 3CB0, 3A3B.

In most cases, accurate identification of binding sites was possible for the following proteins:

The structural and biological characteristics of the "easy" proteins does not allow to define the common criteria for the successful prediction. Neither monomeric/

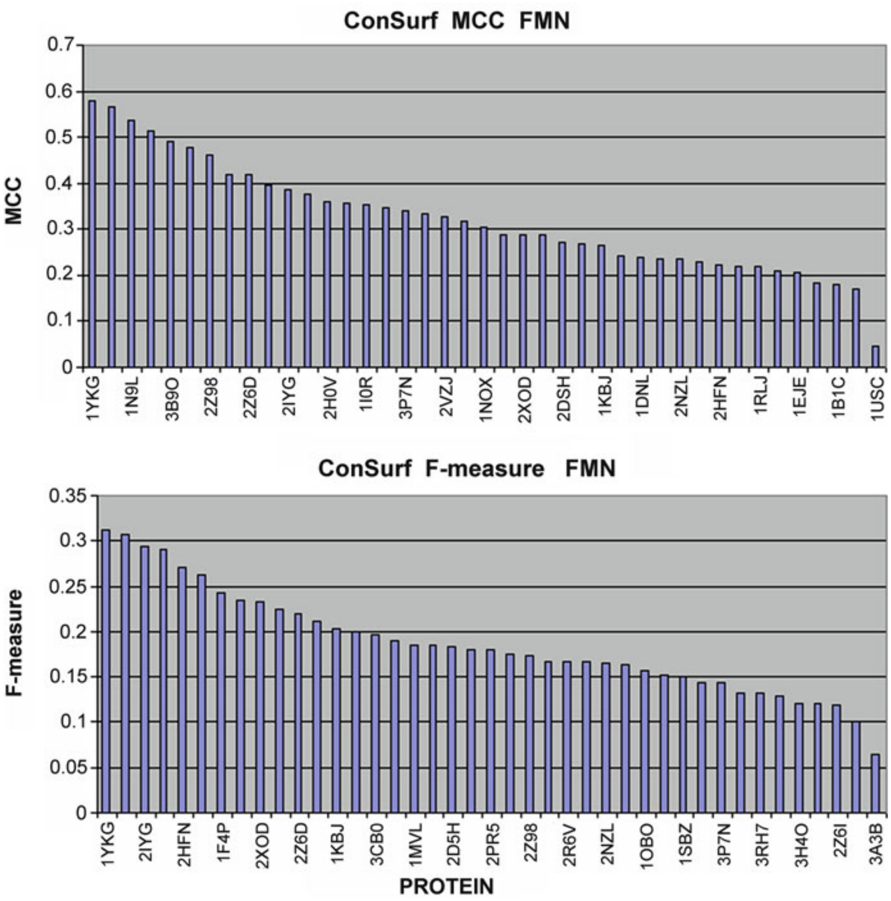


Fig. 4.10 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by ConSurf

complex form nor biological activity has been recognized as common for this group of proteins (Table 4.2). Only the visual analysis of the 3-D structure presentation suggest the well defined cavity in "easy" proteins.

The same must be concluded in respect to "hard" proteins. No criteria for this group of proteins can be defined to characterize their specificity in respect to their predictability (Table 4.3).

FMN binding site identification proved difficult for the following proteins shown in Table 4.3.

Top row (left to right): 2JHF, 1SBY, 3ABI (binding site easy to identify for most programs)

Bottom row (left to right): 3P2O, 1EE9, 1 AD3 (binding site difficult to identify for most programs)

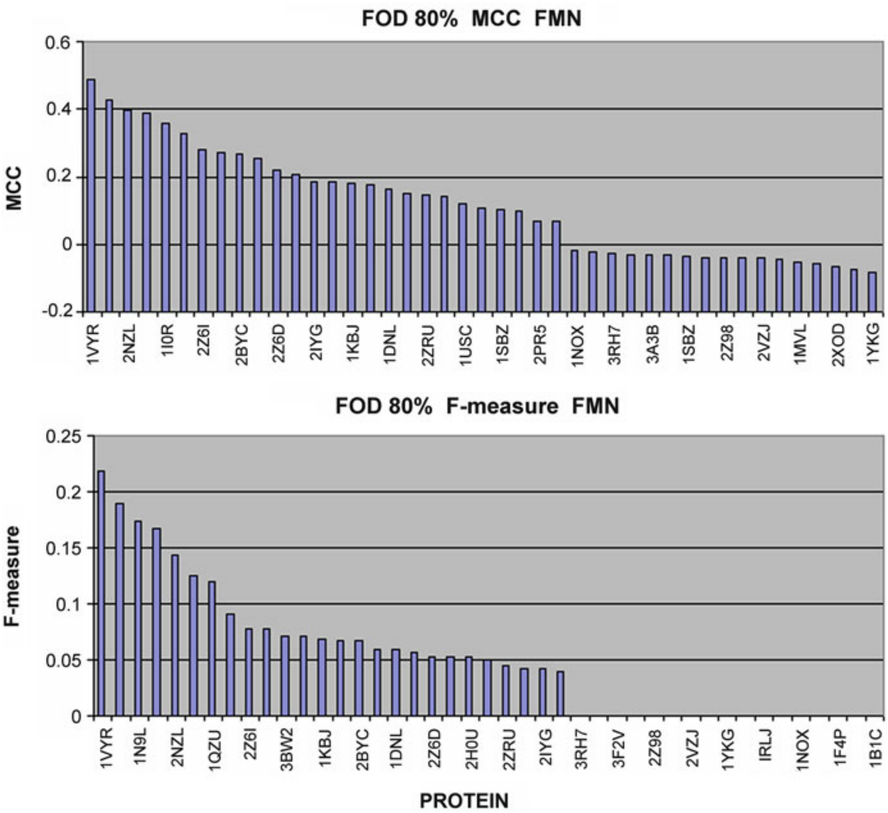


Fig. 4.11 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by the FOD model

The group of “easy” proteins complexing NAD$^+$ (shown in Table 4.4) reveals more common characteristics suggesting that probably the dimeric proteins of the class EC.1.1.1.1 (oxydoreductase) represent the structures with well defined ligand binding cavity.

The “hard” proteins (in respect to NAD+binding side identification) appeared to be all enzymes of bacterial origin (except one protein which is mammalian). No common characteristics can be given in this point to make predictable the success or failure in identification (Table 4.5). The characteristics of cavities for “easy” and “hard” proteins is given in Table 4.6.

Table 4.6 reveals the general difference in NAD+and FMN binding cavity which is much larger for NAD+although the other parameters seem to be quite comparable for these two ligands. The comparison between “easy” and “hard” reveals differences of vertices, buried vertices and average depth which are lower for FMN binding cavities in proteins recognized as “hard”.

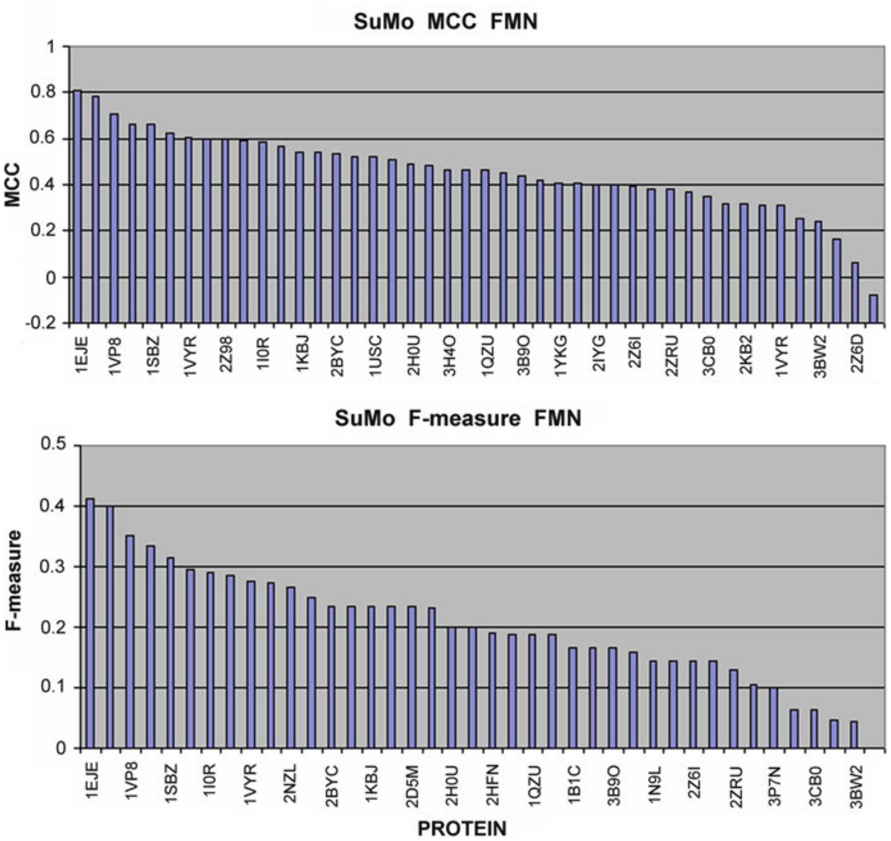


Fig. 4.12 Comparison of MCC and F-measure results for FMN-complexing proteins as reported by SuMo

4.5.4 Analysis of Individual Cases

Table 4.1 lists proteins for which identification proved easy, as well as those for which the programs produced poor results. Most proteins were correctly processed by various tools although there are also cases where only a single program was able to produce correct results. This section presents such proteins in more detail.

Correct NAD$^+$ binding site identification was most frequently obtained for 2JHF (alcohol dehydrogenase EC 1.1.1.1.) Of all the tested tools only ConSurf was unable to correctly process this protein, which suggests that the relevant binding site is specific and non-conservative in character.

The numbers of unique NAD$^+$ binding pocket matches returned by each program are as follows: CASTp – 7, PocketFinder – 3, QSiteFinder – 5, FOD – 6, SuMo – 8. Corresponding values for FMN are: QSiteFinder – 2, PocketFinder – 3, FOD – 4, SuMo – 5. All these results refer to the top 10 places on each ranking list.

Table 4.1 Comparison of worst- (LOW) and best-case (HIGH) results for each ligand obtained by various software packages

	NAD+				FMN			
	Low		High		Low		High	
Ligand	MCC	F-m	MCC	F-m	MCC	F-m	MCC	F-m
Geometry based models								
CASTp	3CIN	1BMDA	1GIQB	2NAD	2VZJ	2VZJ	1YKG	1OBO
	3MGI	2Q1U	3ABI	1SBY	3B9O	1SBZ	2BYC	1B1C
	2PH5	1EE9	**3P2OB**	1EJ2	1SBZ	2WQF	1B1C	2IYG
	1LJ8	1XAH	**1EBF**	1KOL	1DNL	3B9O	2IYG	1RLJ
	1BMDA	1NHG	**3JSA**	1NHG	2WQF	1DNL	2BYC	2D5M
	1BMDB	1SBY	**3LN3**	1Y3I	2Z6I	2H0U	2HFN	2PR5
	1EE9	1 AD3	**2PH5**	**3P2OB**	2H0U	3H4O	2 KB2	2BYCA
	1OBB	1EJ2	**1 H94**	1ZJZ	3RHF	2Z6I	1F4P	1YKG
	3LN3	1Y3I	1GIQA	1VM6	3F2V	3F2V	3BW2	3A3B
	1TOX	3Q3C	1UWK	2JHFA	3H4O	3RH7	2XOD	2NZL
PocketFinder	1GIQA	1EE9	1EJ2	1NHG	3A3B	3F2V	2Z6D	2Z6D
	3LN3	1U8X	1ZJZ	1EJ2	3B9O	3A3B	1OBO	1OBO
	2WN7	1Y3I	1NHG	**1OMO**	1SBZ	2Z6I	2PR5	1RLJ
	1DHS	2WN7	**1OMO**	2NAD	2Z6I	2WQF	1RLJ	2PR5
	1EE9	3KET	1SBY	1ZJZ	2WQF	1SBZ	**3CB0B**	1USC
	1U8X	1KYQ	2NAD	**1SBY**	1MVL	1NOX	**1USC**	1B1C
	1Y3I	1TOXA	**1FDVA**	1UWK	1NOX	1MVL	1B1C	2R6V
	1TOXA	2JHFB	3ABI	**1FDVA**	3H4O	2H0U	1I0R	2IYG
	3KET	3P2OB	1HDR	3ABI	2H0U	3H4O	1EJE	2D5M
	1KYQ	1EVJ	1UWK	1JQ5	1QZU	1QZU	**2R6V**	3CB0B
Q SiteFinder	3HL0	3MGIB	1UWK	1UWK	1I0R	1VYR	2IYG	2IYG
	2GWL	1EJ2	**2Q1U**	**2Q1U**	1SBZ	3B9O	3A3B	3A3B
	1BMDB	1 AD3	2JHFB	1SBY	2WQF	3F2V	2PR5	1RLJ
	1BMDA	2G5C	1SBY	2JHFB	1NOX	2VZJ	1RLJ	2PR5
	1U8X	3P2OA	2JHFA	2JHFA	3CB0B	2 KB2	1B1C	1B1C
	1XAH	3P2OB	3ABI	**3CIN**	3H4O	2H0U	**1KBJ**	1OBO
	3JSA	1 H94	**3CIN**	**1GR0**	2Z6I	1DNL	1OBO	**1KBJ**
	1DHS	1LJ8	**1GR0**	3ABI	1USC	1MVL	2BYCA	2Z6D
	1KYQ	2I2F	2PLA	1VM6	1VYR	1QZU	2Z6D	2BYCA
	1EBF	**1EE9**	1VM6	2PLA	2 KB2	1YKG	**3P7NB**	**3P7NB**
Knowledge-based models								
ConSurf	1V9L	1V9L	**1EVZ**	**1EVZ**	1V9L	2PH5	1YKG	1YKG
	1K6X	1K6X	1HDR	1HDR	1K6X	1SBY	2BYC	2BYC
	2JHFA	2JHFA	1K4MA	**1XAH**	2JHFA	2JHFA	1B1C	1B1C
	3KET	3KET	**1ZBQ**	**1ZBQ**	3KET	1 H94	2IYG	2IYG
	2PH5	2PH5	**1XAH**	1EVJ	2PH5	1FDVA	2BYC	2BYC
	1 H94	1 H94	1EVJ	1K4MA	1 H94	1 AD3	2HFN	2HFN
	1 AD3	1 AD3	1Y3I	1Y3I	1 AD3	3KET	2 KB2	2 KB2
	1KYQ	1KYQ	1EJ2	2 G76	1KYQ	1KYQ	1F4P	1F4P
	1OG3	1OG3	2 G76	2I2F	1OG3	2GR9	3BW2	3BW2
	2GR9	2GR9	1KOL	**1P1H**	2GR9	1OG3	2XOD	2XOD
FOD	2GR9	1UWK	1KQN	2GWL	1N9L	2D5M	1EJE	1VYR

(continued)

Table 4.1 (continued)

	NAD+				FMN			
	Low		High		Low		High	
Ligand	MCC	F-m	MCC	F-m	MCC	F-m	MCC	F-m
Cutoff 80 %	1KYQ	1GIQ	2GWL	1KQN	2NZL	1YKG	2Z98	3CB0
	1TOX	1KYQ	1HDR	1HDR	3CB0	1SBZ	1OBO	1N9L
	3JYO	1P1H	1N2S	1N2S	1I0R	1RLJ	2VZJ	3B9O
	3M6I	1TOX	1FDV	1OBB	3B9O	1OBO	2D5M	2NZL
	1TOX	2GR9	2I2F	2Q1U	2Z6I	1NOX	1MVL	1I0R
	3JSA	3JSA	1KQN	1FDV	1VP8	1MVL	1B1C	1QZU
	3M6I	3JYO	1OBB	2I2F	2BYC	1F4P	2XOD	2R6V
	1KYQ	3KET	2Q1U	1KQN	2BYC	1EJE	1RLJ	2Z6I
	3KET	3M6I	1FDV	3OX4	2Z6D	1B1C	1YKG	1VP8
SuMo	3CIN	3OX4	**2GR9**	**2GR9**	2Z6I	3CB0	1EJE	1EJE
	1DHS	1TOX	**1BMD**	**1BMD**	3RH7	2 KB2	1F4P	1F4P
	1GR0	1V9L	**1OG3**	**1OG3**	2ZRU	2 KB2	**1VP8**	**1VP8**
	1SC6	2PH5	2JHF	**2WN7**	2 KB2	1NOX	**1MVL**	**1MVL**
	1TOX	1SC6	**1 AD3**	**1 AD3**	3P7N	1VYR	**1SBZ**	**1SBZ**
	1V9L	1DHS	**1KYQ**	2JHF	3A3B	3P7N	1DNL	2WQF
	1ZBQ	3P2O	**2WN7**	1GIQ	3CB0	3BW2	1I0R	1VYR
	2PH5	1GR0	**3JYO**	2PLA	2PR5	3A3B	2WQF	1DNL
	3KET	1ZBQ	1GIQ	**1EVJ**	3BW2	2Z6D	1VYR	**2Z98**
	3P2O	3KET	1EVJ	1JQ5	2Z6D	2PR5	**2Z98**	2NZL

PDB codes printed in boldface indicate unique matches (i.e. matches found only by a single software package). Underlined codes indicate proteins with the highest frequency of matches for a given classification

For FMN, the most correctly identified binding site appeared to be in 2BYC protein (CASTp, QsiteFinder, ConSurf).

The biological properties of the analyzed proteins are very diverse and do not seem to correspond to binding pocket prediction accuracy. The study group includes enzymes and transport proteins, monomers as well as complexes consisting of individual subunits. It would be difficult to attribute identification accuracy to any common property putatively shared by all of the presented proteins.

The 2Z6I protein was very accurately analyzed by the FOD model but posed significant problems for QSiteFinder, CASTp, SuMo and PocketFinder. FOD also yielded accurate results for 1N9L, contrary to SuMo and PocketFinder. This suggests that the geometry of the binding pocket in these proteins is rather generic and that the interaction between the ligand and the hydrophobic core plays a decisive role.

An entirely different situation occurs in 2BYC which was accurately processed by all models except FOD. It seems that in 2BYC the presence of a ligand does not distort the protein's own hydrophobic core to an appreciable degree.

Determining the factors responsible for binding pocket prediction accuracy is a complicated problem. Proteins which could be easily identified by geometry-based

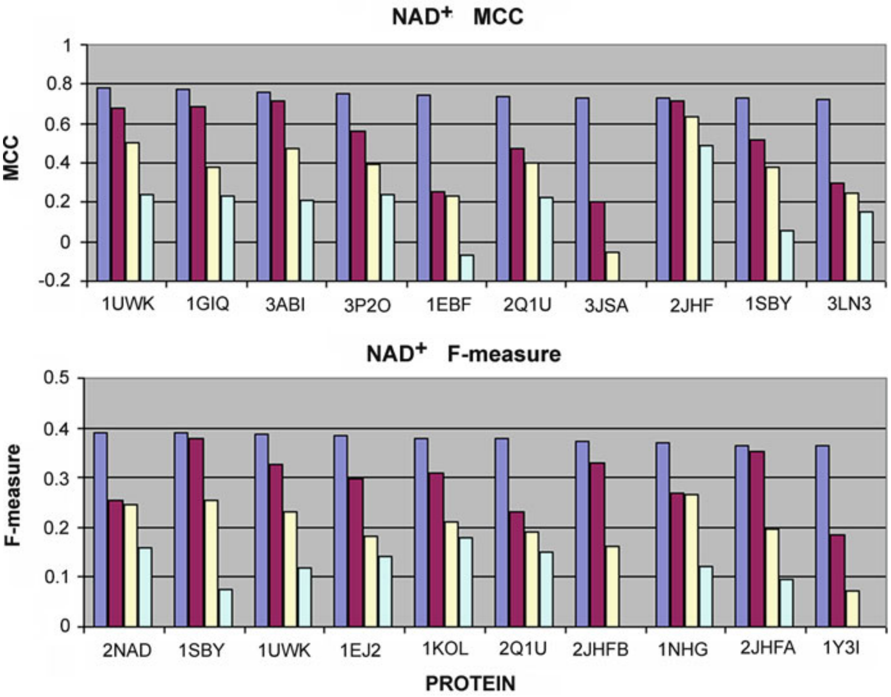


Fig. 4.13 MCC and F-measure scores for NAD$^+$-complexing proteins correctly processed by most of the analyzed tools. Only the top four scores have been taken into account for each program

tools are likely to contain well-ordered pockets whose geometry closely corresponds to the requirements of a specific ligand. This hypothesis is supported by the high consistency of results returned by various geometry-based packages. A selection of best- and worst-case identification scenarios is presented in Figs. 4.15 and 4.16.

An interesting case is the 2Z6I protein (oxidoreductase E.C. 1.3.1.9 in complex with the FMN ligand). It was accurately recognized by the FOD model, as well as by QSiteFinder, CASTp, SuMo and PocketFinder. 2Z6I is a globular molecule with a deeply embedded ligand which affects the structure of the hydrophobic core and suggests significant distortion (hydrophobicitiy deficiency) caused by the binding pocket depth (Fig. 4.17).

Similarly, accurate FOD results were obtained for the 1N9L protein (electron transport – putative blue light receptor). These results are presented in Fig. 4.17. Again, we are dealing with a globular molecule with a deeply embedded FMN ligand. Plotting ROC curves for both proteins (and particularly for 2Z6I) reveals that the area bounded by each curve and the corresponding diagonal is quite large. Residues responsible for binding ligands are highlighted on $\Delta\tilde{H}$ profile plots. Results for a cutoff threshold of 0.004 are presented in Table 4.6.

Table 4.6 compares proteins which yielded correct and incorrect results when applying the ROC curve method (based on the FOD model). The criterion of correctness is the area bounded by the ROC curve and the corresponding diagonal.

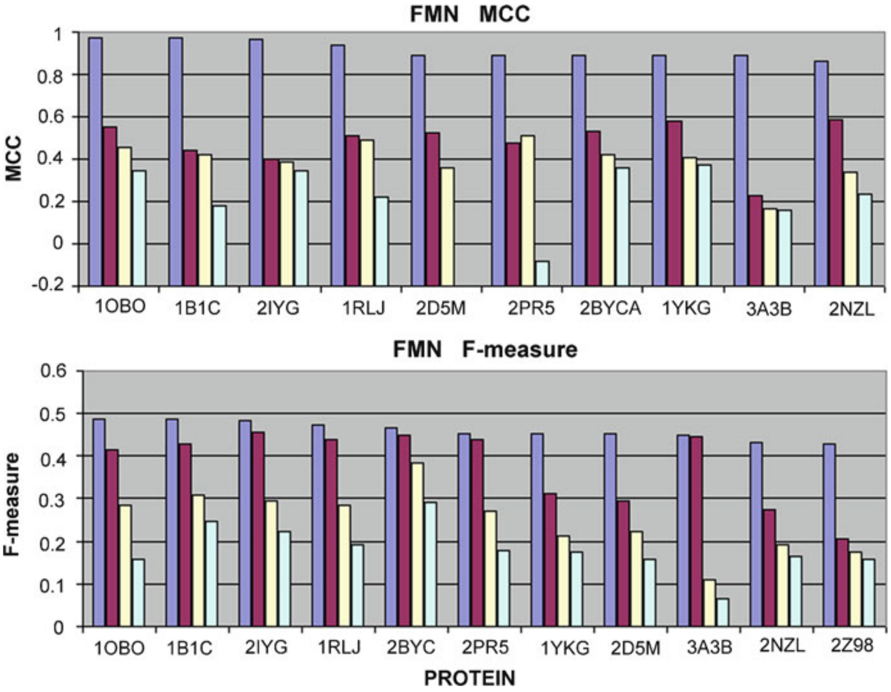


Fig. 4.14 MCC and F-measure scores for FMN-complexing proteins correctly processed by most of the analyzed tools. Only the top four scores have been taken into account for each program

Top row – FMN-binding proteins

Left column – Example protein (3F2V) for which FOD incorrectly identifies ligand binding residues (Table 4.6) The catalytic residue represents the global minimum of the $\Delta\tilde{H}$ profile, i.e. excess hydrophobicity on the protein surface. This residue has been correctly identified as involved in catalytic activity.

Right column – Example protein (2NZL) for which FMN binding residues have been correctly identified. The distribution of catalytic residues suggests correct identification of the enzymatic active site, consisting of amino acids to which the FOD model attributes hydrophobicity deficiencies.

Bottom row – NAD$^+$-binding proteins

Left column – Example protein (1BMD) for which FOD incorrectly identifies ligand binding residues. Enzymatically-active residues represent local maxima of the $\Delta\tilde{H}$ profile, which could be useful in identifying the corresponding catalytic active site.

Right column – Example protein (1 AD3) for which FOD incorrectly identifies ligand binding residues. The distribution of catalytic residues suggests correct identification of the enzymatic active site, consisting of amino acids with peak $\Delta\tilde{H}$ values.

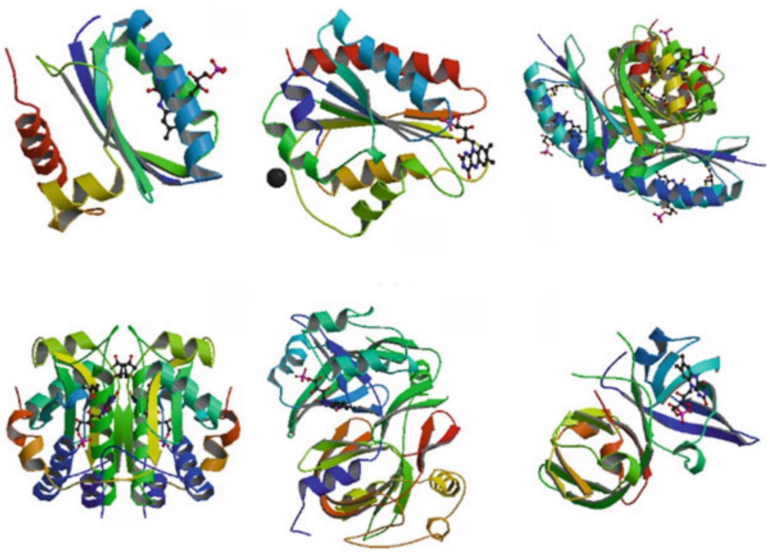


Fig. 4.15 "Easy" (*top row*) and "difficult" (*bottom row*) proteins, with FMN binding site identification accuracy taken as the criterion of difficulty

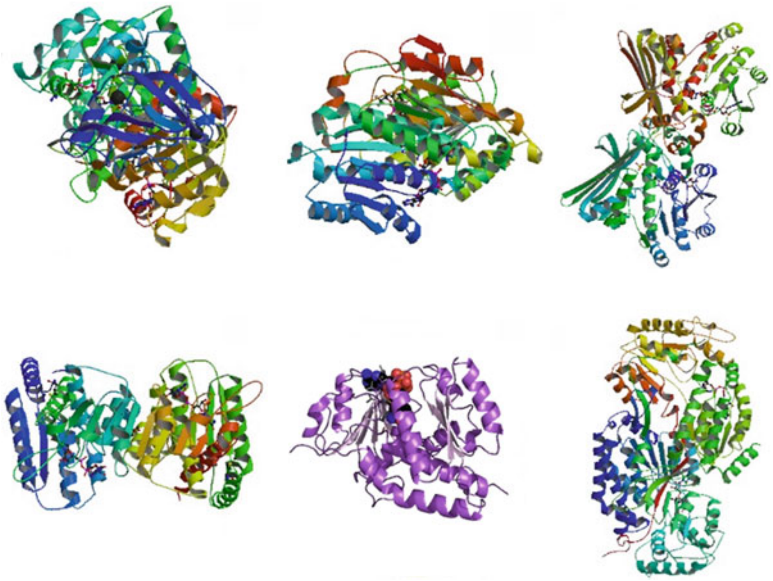


Fig. 4.16 "Easy" (*top row*) and "difficult" (*bottom row*) proteins, with NAD^+ binding site identification accuracy taken as the criterion of difficulty

Table 4.2 The short presentation of "easy" proteins to visualize their different characteristics

PDB-ID	Enzyme	Biological activity	Name	Complex	Source organism
2BYC		Blue light receptor from the BLUF family	Signaling protein	Dimer	*Rhodobacter sphaeroides*
1B1C	E.C.1.6.2.4	FMN-binding domain of human cytochrome p450 reductase	Redox	Monomer	HS monomer
2IYG		Dark-state structure of the bluf domain of the rhodobacterial protein APPA	Signal transduction	Dimer	*Rhodobacter sphaeroides*

Table 4.3 The short presentation of "hard" proteins to visualize their different characteristics

PDB-ID	Enzyme	Biological activity	Name	Complex	Source organism
3H4O		Nitroreductase family protein	Oxydoreductase family	Monomer	*Clostridium difficile* 630
3CB0	E.C.1.14.13.3	4-Hydroxyphe-nylacetate 3-monooxygenase	Oxidoreductase	Tetramer	*Brucella melitensis*
3A3B		Lumazine protein	Luminescent protein	Dimer	*Photobacterium kishitanii*

Table 4.4 Proteins (and their short characteristics) for which NAD+ binding sites could be accurately identified

PDB-ID	Enzyme	Biological activity	Name	Complex	Source organism
2JHF	E.C.1.1.1.1	Alcohol dehydrogenase E chain	Oxydoreductase	Dimer	*Equus Caballus* horse liver
1SBY	E.C.1.1.1.1	Alcohol dehydrogenase from *Drosophila Lebanonensis*	Oxydoreductase	Dimer	*Scaptodrosophila Lebanonensis*
3ABI		Putative uncharacterized protein ph1688	Unknown	Dimer	*Pyrococcus Horikoshi*

Studying the graphical representation of FOD results (Fig. 4.18) reveals the reasons behind the incorrect recognition of residues involved in ligand binding. Poor results for FMN-binding residues seem to be associated with the fact that this ligand is bound on the surface of the protein, without the need for a deep pocket. Hydrogen bonds between the phosphate moiety and polar residues exposed on the protein surface (responsible for stabilization of the resulting complex) do not significantly distort the shape of the protein's hydrophobic core.

Table 4.5 Characteristics of easy and hard predictable binding cavity. The values given are as follows: volume, accessible vertices, buried vertices, average depth for cavities binding particular ligand. Values given are calculated according to PDBSum data

	FMN	NAD$^+$
Easy	1055.27	3376.99
	72.36	74.21
	11.48	12.90
	10.15	12.89
Hard	1141.74	3423.98
	41.35	69.78
	6.43	13.00
	8.10	13.00

Table 4.6 Best- and worst-case results for the FOD model with the $\Delta\tilde{H}$ cutoff threshold set at 80%–this means that calculated $\Delta\tilde{H}$ value (i.e. the difference between expected and observed hydrophobicity) for particular residue is in excess of 80% of the peak value computed for the entire protein. Such residue is suspected of involvement in binding pocket generation

	NAD$^+$		FMN	
	Best	Worst	Best	Worst
FOD	3HL0	2PH5	3P7N	**3BW2**
	2GWL	2WN7	3A3B	3CB0
Cutoff 80%	**1BMD**	1 H94	1QZU	**1VYR**
	1U8X	**2GR9**	1F4P	**2WQF**
	1XAH	**1 AD3**	1OBO	2Z6I
	3JSA	**1GIQ**	2Z98	1N9L
	1DHS	**1TOX**	2VZJ	**2Z6D**
	1KYQ	**1KYQ**	3F2V	3B9O
	1EBF	**3KET**	1MVL	**1DNL**

In the case of proteins depicted in Fig. 4.18, NAD$^+$ binding occurs in close proximity to catalytic residues (marked by red circles). The presented examples indicate that local maxima (or the global minimum, in the case of 3F2V) point to residues with catalytic properties. This is likely why the identification of catalytic residues, as presented in Prymula et al. (2011), appears to be significantly more accurate than identification of NAD$^+$ and FMN binding sites, which is the subject of this discussion.

4.5.5 Summary

It should be noted that the programs presented in this chapter apply diverse techniques to identify ligand binding sites. Low consistency of results (with the exception of the universally poor identification of FMN sites which appears to be restricted

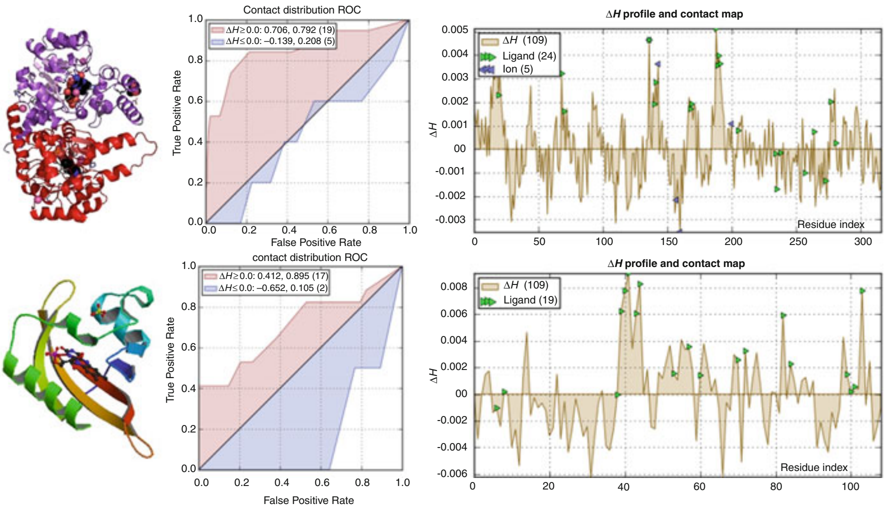


Fig. 4.17 Examples of proteins for which the FOD model correctly identified binding sites. The top row presents 2Z6I – its 3D structure (with a clearly visible deep binding pocket) as well as the ROC curve indicating the relation between TPR and FPR. The area marked in red represents positive $\Delta\tilde{H}$ results while the area marked in blue indicates TPR and FPR coefficients for negative $\Delta\tilde{H}$ values. The attached plots represent $\Delta\tilde{H}$ profiles for both proteins (*top* – 2Z6I; *bottom* – 1N9L), with highlighted ligand binding sites, as well as ion binding sites in the case of 1N9L (which, however, is out of scope of this discussion). 3D images have been derived from PDBSum (for 2Z6I) and PDB (1N9L)

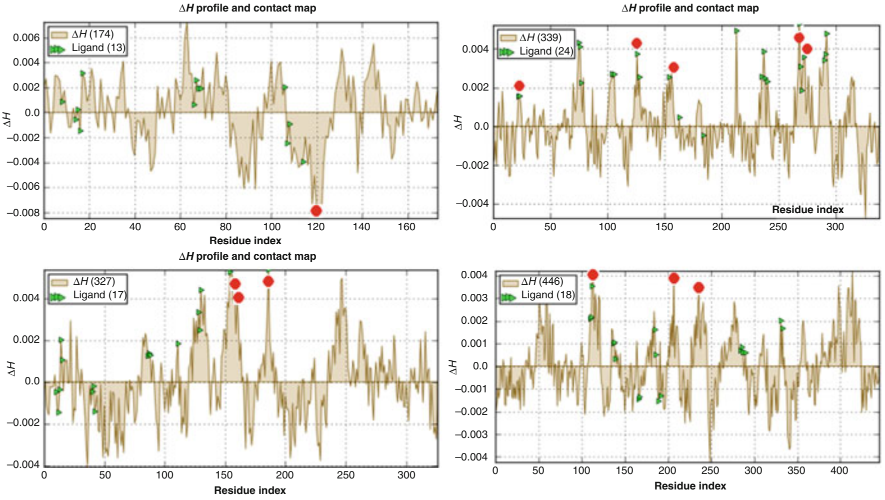


Fig. 4.18 $\Delta\tilde{H}$ profiles, with ligand binding residues marked by *green* triangles and catalytic residues marked by *red* circles

to a specific group of proteins) indicates that binding pockets emerge through many different mechanisms. This phenomenon could be related to the diverse biological activity attributed to proteins with FMN and NAD$^+$ cofactors.

The structural similarity of binding pockets – as indicated by geometry-bases tools – is clearly discernible, although limited in scope.

In the FOD method, the most important factor affecting the identification of binding pockets (by searching for residues which represent local distortions in the structure of the protein's hydrophobic core) appears to be the presence of additional polypeptide chains determining the protein's quarternary structure. Of nearly equal importance is the size of polypeptide chains: large single-core proteins can be subdivided into domains where the hydrophobic core structure exhibits better ordering than in the case of a multi-domain multi-core molecule. The phenomenon of ligands being bound in the inter-chain space (between adjacent subunits) also affects the accuracy of predictions returned by the FOD model.

Given the presented results, it seems valid to conclude that binding pockets are generated through many different mechanisms and that the applicability of each theoretical model is typically limited to a specific subset of proteins, yielding poor results for proteins which do not share the preferred structural properties – even if their general structure (or purpose) is similar.

The set of programs discussed in this chapter was also tested in the context of active site identification in hydrolases (Prymula et al. 2011). In that study, knowledge-based tools proved more reliable than geometry-based packages. In contrast, geometry-based software appears to return better results for FMN and NAD$^+$ binding site identification, as described above. We can therefore conclude that the binding geometry of these ligands is highly deterministic and specific, whereas evolutionary factors play a more pronounced role in shaping enzymatic active sites. Distortions of the protein's hydrophobic core are more closely related to ligand binding sites (Fig. 4.18). It appears that the structure of FMN and NAD$^+$ pockets is local in character and does not affect the shape of molecule as a whole – contrary to active centers in hydrolases.

The comparison presented in Fig. 4.18 explains some of the difficulties involved in identifying ligand binding residues by way of the FOD model. The $\Delta\tilde{H}$ profile indicates that the hydrophobicity attributed to such residues remains in agreement with statistical predictions and therefore does not trigger distortions in the protein's core. This is especially evident in the case of 3F2V (Fig. 4.18), where the placement of the active site (and thus of the substrate) is tied to a specific deformation and can therefore be accurately identified.

The presented examples also point to the role of the environment surrounding the ligand binding and catalytic residues. Enzymatic active sites need to be shielded from water and are usually located in deep pockets, where their hydrophobicity deficiency can be clearly discerned on the $\Delta\tilde{H}$ plot (see protein 2Z6I in Fig. 4.17). Binding ligands involves compensating the protein's own hydropbhobicity deficiencies with the ligand's own excess hydrophobicity, resulting in a droplike core structure. In some cases, however (as depicted in Fig. 4.18, protein 1BMD), the ligand may bind at locations with variable hydrophobicity conditions. Since certain ligands play a part in proton or electron transport they need to be present at various points along the transport pathways. This implies that they must be able to bind at many different locations: from the surface of the molecule (where hydrophobic bonds may form between the protein's surface residues and the ligand's phosphate moiety, particularly in the case of FMN) all the way to deep within binding pockets (where they can be traced by the FOD model, as is the case with all of the presented proteins to which this model has been successfully applied).

An interesting observation can be made with respect to the ligand binding site/enzymatic active site relationship. The FOD model seems to suggest the need for close proximity between the transport unit (responsible for moving electrons or protons) and the catalytic site, while still retaining functional separation (different residues responsible for each function, with different deviations from the idealized hydrophobicity model as indicated by corresponding $\Delta\tilde{H}$ profile values, suggesting different placement in the protein body and dedicated operating environments for transport ligands and catalytic active sites).

References

Altman DG, Bland JM (1994) Diagnostic tests 1: sensitivity and specificity. BMJ 308(6943):1552

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402

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. doi:10.1093/nar/gkq399, PMID: 20478830

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424

Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I (2012) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J Mol Model 18(1):229–237

Binkowski TA, Naghibzadeh S, Liang J (2003) CASTp: computed atlas of surface topography of proteins. Nucleic Acids Res 31:3352–3355

Carugo O (2007) Detailed estimation of bioinformatics prediction reliability through the fragmented prediction performance plots. BMC Bioinformatics 8:380, PMID:17931407

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the clustal series of programs. Nucleic Acids Res 31(13):3497–3500

Dundas J, Ouyang Z, Tseng J, Binkowski A, Turpaz Y, Liang J (2006) CASTp: computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. Nucleic Acids Res 34:W116–W118

Edelsbrunner H (1995) The union of balls and its dual shape. Disc Comput Geom 13:415–440

Edelsbrunner H, Mucke EP (1994) Three-dimensional alpha shapes. ACM Trans Graphics 13:43–72

Edelsbrunner H, Shah NR (1996) Incremental topological flipping works for regular triangulations. Algorithmica 15:223–241

Edelsbrunner H, Facello M, Fu P, Liang J (1995) Measuring proteins and voids in proteins. In: Proceedings of the 28th annual Hawaii international conference on system sciences. IEEE Computer Society Press, Los Alamitos, pp 256–264

Edelsbrunner H, Facello M, Liang J (1998) On the definition and the construction of pockets in macromolecules. Disc Appl Math 88:83–102

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32(5):1792–1797

Facello MA (1995) Implementation of a randomized algorithm for Delaunay and regular triangulations in three dimensions. Comput Aided Geom Des 12:349–370

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874

Goldenberg O, Erez E, Nimrod G, Ben-Tal N (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. Nucleic Acids Res 37, Database issue D323–D327

Hendlich M, Rippmann F, Barnickel G (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. J Mol Graph Model 15:359–363

Jambon M, Imberty A, Deléage G, Geourjon G (2003) A new bioinformatics approach to detect common 3D sites in protein structures. Proteins 52:137–145

Jambon M, Andrieu O, Combet C, Deléage G, Delfaud F, Geourjon C (2005) The SuMo server: 3D search for protein functional sites. Bioinformatics 21:3929–3930

Joosten V, van Berkel WJ (2007) Flavoenzymes. Curr Opin Chem Biol 11(2):195–202

Kamburov A, Pentchev K, Galicka H, Wierling C, Lehrach H, Herwig R (2011) ConsensusPathDB: toward a more complete picture of cell biology. Nucleic Acid res, 39 (Database issue):D712–D717

Konieczny L, Brylinski M, Roterman I (2006) Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol 6:15–22

Landau M, Mayrose I, Rosenberg Y, Glaser F, Martz E, Pupko T, Ben-Tal N (2005) ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. Nucleic Acids Res 33:W299–W302

Laskowski RA (2009) PDBsum new things. Nucleic Acids Res 37:D355–D359

Laurie ATR, Jackson RM (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. Bioinformatics 21:1908–1916

Liang J, Edelsbrunner H, Woodward C (1998a) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci 7:1884–1897

Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998b) Analytical shape computation of macromolecules. II. Identification and computation of inaccessible cavities in proteins. Proteins Struct Funct Genet 33:18–29

Matthews BW (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 405:442–451

Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior. Mol Biol Evol 21:1781–1791

Olson DL, Delen D (2008) Advanced data mining techniques. Springer-Verlag Berlin, Heidelberg, p 138. ISBN 3540769161

Pollak N, Dölle C, Ziegler M (2007) The power to reduce: pyridine nucleotides – small molecules with a multitude of functions. Biochem J 402:205–218

Prymula K, Jadczyk T, Roterman I (2011) Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction. J Comput Aided Mol Des 25(2):117–133

Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. Bioinformatics 18:71–77

Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res. 39 (Database issue):D392–D401

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 22:4673–4680

Tsai C-J, Ma B, Nussinov R (2009) Protein-protein interaction networks: how can a hub protein bind so many different partners? Trends Biochem Sci 34(12):594–600

van Rijsbergen CV (1979) Information retrieval, 2nd edn. Butterworth, London; Boston. ISBN 0-408-70929-4

Chapter 5
Docking Predictions of Protein-Protein Interactions and Their Assessment: The CAPRI Experiment

Joël Janin

Abstract Protein-protein docking was born in the 1970s as a tool to analyze macromolecular recognition. It developed afterwards into a method of prediction of the mode of association between proteins of known structure. Since 2001, the performance of docking procedures has been assessed in blind predictions by the CAPRI (Critical Assessment of PRedicted Interactions) experiment. The results show that docking routinely yields good models of the protein-protein complexes that undergo only minor changes in conformation and associate as rigid bodies. In contrast, flexible recognition accompanying large conformation changes in the components remains difficult to simulate, and structural predictions generally yield lower quality models. In recent years, a new challenge has been to predict affinity and to estimate the stability of the complex along with its structure. Over the years, CAPRI has proved to be a strong incentive to develop new flexible docking procedures and more discriminative scoring functions, and it has provided a common ground for discussing methods and questions related to protein-protein recognition.

Keywords Community-wide experiment • Protein-protein recognition • Binding affinity • Protein flexibility • Protein engineering • Docking simulation • Computational biology • DOCK • CAPRI • Rigid body docking • Cube representation • Geometric hashing algorithm • Monte-Carlo • RosettaDock • Template-based docking • CASP • Blind prediction • Flexible docking • Molecular dynamics • Flexibility • Affinity benchmark

J. Janin (✉)
IBBMC, Université Paris-Sud, 91405, Orsay, France
e-mail: joel.janin@u-psud.fr

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_5,

5.1 Introduction

The specific recognition between two proteins is the physical process that governs the construction of the macromolecular machines and assemblies which carry out most biological functions in cells and living organisms. Ubiquitous and essential to life, protein-protein recognition has in recent years become a major subject of study in post-genomic molecular biology, biochemistry, structural biology, and biophysics. When structural data are available, it can also be approached computationally, by docking simulations in which a protein-protein complex is assembled from the component structures. We relate here how protein-protein docking was attempted in the early 1970s, preceding small molecule docking at a time where very few proteins had a known three-dimensional structure, and how it developed into a family of novel algorithms after 1990. Since then, docking algorithms have turned into structural prediction procedures, and their reliability has been tested in the CAPRI blind prediction experiment. An outcome of the test was that the initial model of recognition in which the proteins bind as rigid bodies, progressively evolved into one of flexible recognition. The new paradigm takes into account the structure changes that may accompany the association reaction, and offers estimates of their effect on the stability of the assembly that the reaction produces, and on the specificity of the recognition process.

5.2 An Early History of Protein Docking

The first attempt to model the self-assembly of two proteins concerned trypsin and the bovine trypsin pancreatic inhibitor (BPTI). David Blow of Cambridge, UK, and Robert Huber of Martinsried, Germany, respectively authors of the α-chymotrypsin and BPTI X-ray structures, teamed to build an atomic model of the trypsin/BPTI complex. Their paper (Blow et al. 1972) does not say how they did it, only that “when a model of the relevant part of the inhibitor was compared with the active site of α-chymotrypsin, it was evident that only one mode of binding was possible”. At the time, “model” meant a physical wire model, not one a computer could handle, and no atomic coordinates of the complex remain to assess its accuracy. Beddell et al. (1976) still used a wire model to do molecular modeling at the Wellcome Research Laboratories in Kent, UK. They engineered biphenyl compounds to bind at the DPG (2,3-diphospho-glycerate) site of hemoglobin. Some of the compounds did, and they had the predicted effects on oxygen binding, possibly the first success of structure-based drug design. While the Wellcome scientists had access to hemoglobin atomic coordinates from Pr. Max Perutz, their paper says that “a more accurate representation was needed”, and they chose to build a wire model. Their designs were based on interactions predicted from that model, not computation.

Nevertheless, Perutz’ hemoglobin coordinates had already been used to do molecular modeling in the computer, and more specifically, to dock proteins together. Pr. Cyrus Levinthal of Columbia University, New York, had devised an algorithm that could build a model of the sickle cell fiber from individual hemoglobin molecules (Levinthal et al. 1975). Shoshana Wodak, one of Levinthal's co-authors, joined me in Pr. Georges Cohen's laboratory at the Pasteur Institute in Paris, and we decided together to investigate what computer simulations could tell us about protein-protein recognition. For that purpose, we designed a procedure that generated all the orientations of one protein relative to another, and brought the two surfaces into contact by translation. To gain computer time and memory space, we borrowed from Michael Levitt a simplified protein model that represented each amino acid residue by a sphere of appropriate radius (Levitt 1976). We allowed a degree of penetration between the spheres, and estimated the quality of the fit by the number of intersubunit residue-residue contacts. Our test system was the same trypsin/BPTI complex as in Blow et al. (1972), but by then, Huber's lab had determined a X-ray structure of the complex (Huber et al. 1974), and issued coordinates that could serve to assess the accuracy of the docking models. In the summer of 1976, we were given access to a state-of-the-art computer in Orsay, France – one that was only about 10,000 times slower than a laptop today – during a workshop of the Centre Européen de Calcul Atomique et Moléculaire (CECAM). In about an hour of cpu time, our software (named DOCK like several others after it) generated models of the inhibitor filling the active site of the protease in 2,300 different orientations. To our satisfaction, an orientation close to Huber's X-ray structure showed a good fit, but there were several other that achieved a similar score. In other terms, the procedure had produced a native-like model of the assembly, plus some false positives. We attributed the false positives to the coarse nature of our score, which took into account the geometric complementarity of the two molecular surfaces, but ignored their chemical nature and the physics of their interaction (Wodak and Janin 1978).

Computational biology had no established status in the mid-1970s, and we had a difficult time convincing journal editors that protein-protein docking was more than a futile game. Yet, Levinthal had addressed a related question, protein folding, several years before, and ambitious attempts were already being made to solve it in the computer (Levitt and Lifson 1969; Levitt 1976; Némethy and Scheraga 1977). With rigid molecules, docking is a much simpler problem than protein folding. Whereas folding has thousands of degrees of freedom, docking has only six, and by restricting the search to the active site of trypsin, we had reduced that number to four, which had made the calculation feasible.

The next application of our software was to simulate the allosteric transition of hemoglobin. Hemoglobin is an order of magnitude larger than BPTI, but its twofold symmetry also reduces the search to four degrees of freedom, and the computation was within the reach of extant computers. It was done at the Free University, Brussels, in the summer of 1981, also during a CECAM workshop (Janin and Wodak 1981). We used a much improved version of DOCK to build hemoglobin tetramers from alpha-beta dimers in a range of orientations that covered the T and R quaternary structures described by Perutz. The results showed that the allosteric transition from R to T could not proceed along a linear pathway, due to steric hindrance at the dimer-dimer interface, and it drew an alternative pathway in excellent agreement with the classical description of Baldwin and Chothia (1979).

5.3 Protein-Protein Docking Algorithms

5.3.1 Bound vs. Unbound Docking

The hemoglobin simulation faced even more editorial skepticism than the trypsin/BPTI study. By the time it got published (Janin and Wodak 1985), small molecule docking had come of age in the hands of Kuntz et al. (1982), Goodford (1985), and a few others. Soon, it became an established procedure in drug design, while protein-protein docking remained confidential for over a decade. Meanwhile, computers became orders of magnitude faster, and crystallographers determined many new structures. The latter included a score of protease/inhibitor complexes, and the first antigen/antibody complexes (Janin and Chothia 1990). Cherfils et al. (1991) tested on those complexes the Wodak-Janin algorithm, implemented as a simulating annealing procedure to make the search more efficient. This allowed all six degrees of freedom to be explored, and most importantly, "unbound" docking to be tested for the first time. Unbound docking uses the atomic coordinates of the free proteins, bound docking, coordinates taken from the complex. Bound docking ignores the conformation changes that may accompany association, and it has no predictive value, since the solution must be known in advance. The new study yielded native-like models of all the target complexes, and a majority of those models scored near the top. However, there were many false positives, especially with the unbound proteins, and it was evident that other features than shape complementarity had to be taken into account to identify the correct docking models among all the false positives.

5.3.2 Rigid-Body Docking

The early 1990s were a period of renewed interest in protein-protein docking. Several new algorithms, all based on geometry and shape complementarity, were published almost simultaneously. Connolly (1986) had devised a procedure in which molecular surfaces were described by sets of discrete points; matching critical points (holes and pits) of two surfaces assessed their complementarity, and this could be used for docking. A related method of surface triangulation, independently developed for "computer-vision" by Pr. Haim Wolfson of Tel Aviv University in Israel, was implemented into a docking procedure through a very efficient geometric hashing algorithm (Nussinov and Wolfson 1991; Norel et al. 1994). In Berkeley, California, Jiang and Kim (1991) designed a "cube representation" of proteins specifically for docking. In that model, the surface of the proteins and their interior volume are sampled on a cubic grid, and a docking pose is generated by matching surface cubes while rejecting overlaps between volume cubes. Jiang and Kim made a very important point: docking must be "soft" to allow for minor conformation changes. The cube model, like the residue sphere model of the Wodak-Janin procedure, made for that softness by blurring the atomic details of the protein structures.

The cubic grid representation is an essential element of the FFT correlation docking algorithm published soon afterwards by Katchalski-Katzir et al. (1992) of the Weizmann Institute in Israel. To start with, one picks an orientation of a protein relative to the other, and assigns appropriate weights to grid points of the surface and the interior volume of the two molecules. The correlation between the two sets of weights is used as a score. It may be written as a convolution product, and efficiently computed for all translations at one time thanks to the Fast Fourier Transform (FFT) algorithm. Then, the orientation is changed and the calculation repeated. The method has been very successful, and it has benefited from many developments (Vakser and Aflalo 1994; Gabb et al. 1997; Ritchie and Kemp 2000; Mandell et al. 2001; Heifetz et al. 2002; Chen et al. 2003a). Whereas the original formulation of the algorithm assessed only the geometric complementarity, other molecular features can be encoded as weights on a cubic grid; for instance, an electrostatic interaction energy may be calculated by correlating the electric charges on one protein with the electric field created by the other protein. Electrostatics, hydrophobicity, and a number of other terms may be combined into a scoring function. Each of the Web sites listed in Table 5.1 has its own scoring function, and its own way to calculate its terms as FFT correlations.

5.3.3 Monte-Carlo and Related Docking Algorithms

Albeit “soft”, the FFT correlation and the geometric hashing algorithms explore only the six degrees of freedom of rigid-body docking. Other algorithms developed afterwards handle other variable parameters, dihedral angles for instance, in order to simulate side chain rotations and main chain conformation changes. They take a heuristic approach to the problem, instead of performing an exhaustive search. Monte-Carlo simulated annealing, the choice method in the 1990s, allowed Totrov and Abagyan (1994) to adjust side chain conformations at the same time as the docking search. These authors employed a detailed atomic model and a standard molecular mechanics force field, which was computationally very expensive. Instead, all the later docking procedures based on simulated annealing or related algorithms, proceed in two or more steps. The first step explores the rigid-body parameter space with a simplified protein model and a coarse force field, the second carries out a detailed refinement of the local minima (Fernández-Recio et al. 2002; Zacharias 2003). The RosettaDock procedure (Gray et al. 2003) is a good example: a first Monte-Carlo search is carried out on a low-resolution protein model with residue-level potentials; it identifies many (a thousand or more) candidate solutions, which are refined afterwards using a full-atom model and the Rosetta force field. That force field, optimized on protein data, includes terms for desolvation or rotamer preferences not present in standard force fields. It performs very well in protein folding, its original application, and also in docking, at least when the conformation changes are of limited amplitude (Schueler-Furman et al. 2005).

Table 5.1 Web servers for protein-protein docking

Protein structure and benchmark sets	
Protein Data Bank (PDB)	http://www.rcsb.org/pdb/
CAPRI experiment	http://capri.ebi.ac.uk/
Docking benchmark	http://zlab.bu.edu/zdock/benchmark.shtml
Structure/affinity benchmark	http://bmm.cancerresearchuk.org/~bmmadmin/Affinity
FFT correlation and related docking algorithms	
ClusPro	http://cluspro.bu.edu/login.php
DOT	http://www.sdsc.edu/CCMS/Papers/DOT_sc95.html
FTDOCK	http://www.sbg.bio.ic.ac.uk/docking/ftdock.html
GRAMM-X	http://vakser.bioinformatics.ku.edu/resources/gramm/grammx/
HEX	http://www.loria.fr/~ritchied/hex/
MolFit	http://www.weizmann.ac.il/Chemical_Research_Support//molfit/
ZDOCK	http://zlab.bu.edu/zdock/
Molecular dynamics, Monte-Carlo and related flexible docking algorithms	
ATTRACT	http://www.ibpc.fr/chantal/www/ptools/
HADDOCK	http://www.nmr.chem.uu.nl/haddock/
ICM-DISCO	http://www.molsoft.com/icm_pro.html
RosettaDock	http://graylab.jhu.edu/docking/rosetta/
Geometric hashing and related flexible docking algorithms	
PatchDock	http://bioinfo3d.cs.tau.ac.il/PatchDock
FireDock	http://bioinfo3d.cs.tau.ac.il/FireDock/
SymmDock	http://bioinfo3d.cs.tau.ac.il/SymmDock
FiberDock	http://bioinfo3d.cs.tau.ac.il/FiberDock/
MultiFit	http://salilab.org/multifit/ and http://bioinfo3d.cs.tau.ac.il/
3D-Garden	http://www.sbg.bio.ic.ac.uk/3dgarden
SKE-Dock	http://www.pharm.kitasato-u.ac.jp/bmd/files/SKE_DOCK.html

5.3.4 Template-Based Docking

An alternative to docking is to use a template, and build a model of a protein-protein complex by analogy to one of known structure. When both components of two complexes are close homologs with a high level of sequence identity (40% or more), it is straightforward to model build both the components and their assembly, but the method has a very limited field of application. It can be extended by accepting templates with a low level of sequence identity, or templates that have similar three-dimensional structures irrespective of their sequences, under the assumption that the mode of interaction is conserved (Lu et al. 2002; Sinha et al. 2010; Kundrotas et al. 2012). Although the limits of validity of this assumption are uncertain, genome-wide libraries of model assemblies have been built in this way (Lu et al. 2003; Stein et al. 2011). Templates may also be selected on the basis of the local similarity of the protein surfaces: two surfaces that have a similar geometry and similar physical-chemical features may be expected to make similar interactions (Günther et al. 2007; Keskin et al. 2008), in which case the PDB may already be adequate to represent the diverse architectures observed in nature (Tuncbag et al. 2008; Kundrotas et al. 2012). Here again, the quality of the models remains to be assessed.

5.4 Assessing Docking Predictions: The CAPRI Experiment

5.4.1 CAPRI

By the turn of the century, several docking algorithms had developed into full-fledged prediction procedures (see reviews by Smith and Sternberg 2002; Camacho and Vajda 2002; Halperin et al. 2002). At that time, an entirely new field of application opened, due to the structural genomics (or proteomics) initiatives that accompanied the completion of the human genome sequence. High-throughput X-ray and NMR studies were going to determine the structure of thousands of new proteins that would include the components of many binary or larger assemblies. Docking procedures could in principle build models of these assemblies from the component structures, but should we trust the results at all? The procedures had been thoroughly tested, but most of the unbound docking tests had been done on protease/inhibitor or antigen/antibody complexes, the only ones for which the component structures were available. How would docking perform on new, possibly very different, systems, and how accurate would the models be?

These questions were discussed in Charleston, South Carolina, in June 2001, at a meeting on Modeling Protein Interactions in Genomes organized by Pr. Sandor Vajda and Ilya Vakser, and the conclusion was that a blind prediction experiment should be organized (Vajda et al. 2002). Named CAPRI (Critical Assessment of PRedicted Interactions), the experiment was modeled after CASP (Critical Assessment of Structural Predictions), an older experiment that tests methods to predict a protein fold based on its amino acid sequence (Moult et al. 1995). The targets of CAPRI would be protein-protein complexes, and the prediction start from component structures taken from the Protein Data Bank. The predictors would dock the components, and submit models to the CAPRI Website, to be assessed by comparison with a newly determined, but unpublished, experimental structure of the complex (Janin et al. 2003). A blind prediction of that sort had been done once before, on a β-lactamase in complex with a protein inhibitor. Six participant groups had submitted models of the complex that were close to the X-ray structure (Strynadka et al. 1996). Could that performance be reedited?

An answer came soon after the Charleston meeting. The first round of CAPRI, held in the summer of 2001, had three targets, three complexes whose X-ray structures had just been determined by collaborators of mine, willing to help starting the experiment. Two were viral antigen proteins in complex with monoclonal antibodies, the third, a bacterial protein kinase co-crystallized with its substrate, the small protein HPr. Fifteen predictor groups submitted a total of 193 models, and the CAPRI assessors led by S. Wodak, compared them to the X-ray structures. The assessors found that the submissions contained good models of the two antigen/antibody complexes, but not of the HPr/kinase complex (Méndez et al. 2003). They nevertheless decided that a few of the HPr/kinase models were "acceptable": their geometry was poor, but most of the residues in the contact regions were correctly predicted, which could in principle help designing experiments. Predicting the residues in contact was not a big feat in that case, since the location of the kinase active site and the serine residue phosphorylated on HPr were known from the literature. Moreover, the poor geometry of the models had an obvious origin: in the X-ray structure of the complex, the rotation of a α-helix in the kinase modified the shape of the substrate binding site and the way it bound HPr (Fieulaine et al. 2002). Thus, rigid-body docking was able to locate the correct epitopes on the two viral antigens, and place them correctly at the antibody combining sites, but it failed on HPr/kinase due to a conformation change, albeit one of limited amplitude.

5.4.2 Success and Failure in Blind Predictions

This pattern was repeatedly observed in later prediction rounds (Méndez et al. 2005; Lensink et al. 2007; Lensink and Wodak 2010; Janin 2005, 2010). In the 10 years that followed the Charleston meeting, CAPRI has had 22 rounds, with a total of 43 targets and an average of 45 predictor groups, each submitting ten models of each target. In addition to protein-protein complexes, the targets have been a protein-RNA complex and four oligomeric proteins. For each target, the predictors were given the coordinates of the unbound components, or of an homolog protein that could be used for model building, and they had 3–6 weeks to make their prediction and submit their models. A majority (70%) of the targets obtained good quality models. Almost all those that displayed only small backbone movements did, and in most cases, the good models came from several groups using different docking procedures. Figure 5.1 shows an example. Target T37, drawn here after the X-ray structure of Isabet et al. (2009), is a complex between the G-protein Arf6, a member of the Ras family of small GTPases, and the LZ2 segment of JIP4 (JNK-interacting protein 4), an effector of Arf6. LZ2 was known to form a leucine zipper, and it had to be model built from its amino acid sequence before docking on Arf6. A standard leucine zipper yields a rather accurate model of its structure in the complex, while Arf6 undergoes little change in the interaction. Correspondingly, the submissions contained a number of good quality models of LZ2/Arf6, submitted by nine different groups (Lensink and Wodak 2010).

On the other hand, CAPRI predictions have yielded at best "acceptable" models of the targets in which the backbone changes were large, or the homology models of poor quality. Prediction yielded no valid model at all in six cases. In two, the failure could be traced to misleading biochemical information rather than the structure itself, in the other four, to large conformation changes. Moreover, some of the targets

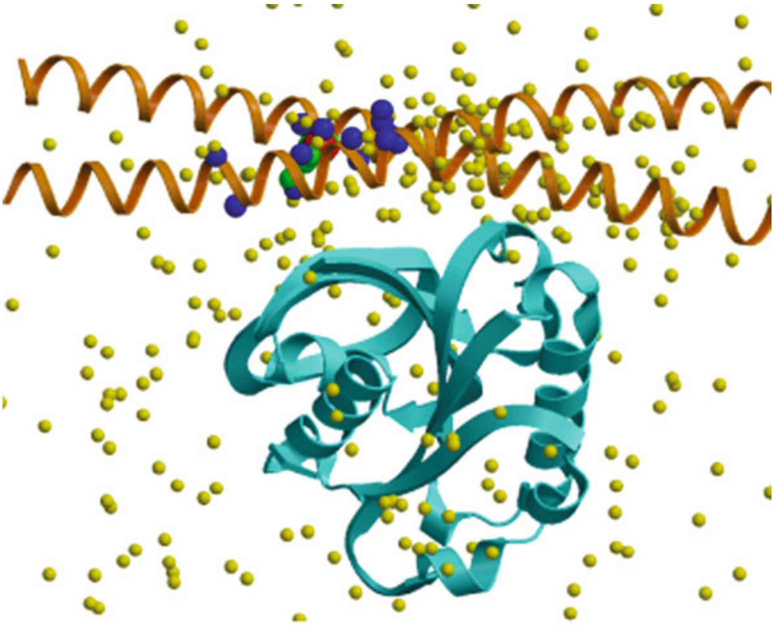


Fig. 5.1 A successful docking prediction. Target T37 was submitted for blind prediction during Round 16 of CAPRI, held in November 2008. The target, a complex of the small GTPase Arf6 with the LZ2 leucine zipper of the JIP4 effector protein, was a gift of Dr. Julie Ménétrey (Institut Curie, Paris). Predictors were given an unbound Arf6 structure, and the amino acid sequence of LZ2, which they had to model build before docking on Arf6. The figure represents the X-ray structure (PDB code 2 W83, Isabet et al. 2009) with Arf6 in cyan, LZ2 in *pink*. The dots are the centers of mass of LZ2 in the models submitted by the 39 predictor groups and the 11 scorer groups who participated in Round 16. The dots are *green* for good quality models, *blue* for "acceptable" models, and *yellow* for incorrect models. All the models can be accessed at http://www.ebi.ac.uk/msd-srv/capri/round16/ (Courtesy of Dr. Marc Lensink (Lille))

were subjected twice for prediction, first with both components unbound, then with the more flexible component in its bound conformation. The second step always yielded much better models; for instance, prediction of the protein/RNA complex failed with the unbound RNA, but there were many good models with the bound RNA, which has a very different conformation (Lensink and Wodak 2010).

5.5 Flexible Docking and the Scoring Experiment

5.5.1 Simulating Conformation Changes and Mechanisms of Recognition

Two very important objectives of CAPRI were to stimulate the development of new methods, and create a forum where they could be discussed and information would spread within the community. The experiment succeeded on both grounds.

It generated lively discussions on line through the CAPRI Website hosted by the European Bioinformatics Institute (Hinxton, UK), and face-to-face during assessment meetings that took place at regular intervals. Moreover, CAPRI is at the origin of most of the progress seen in the last 10 years. The new scoring functions, the methods to model conformation changes and the flexible docking procedures developed since 2001, owe much to the experiment. Docking searches based on simulated annealing and molecular dynamics have been adapted to reproduce loop movements or the rotation of a structural domain about a pre-defined hinge, with good results on some CAPRI targets (Janin 2005; Lensink et al. 2007; Lensink and Wodak 2010). However, a more general solution to the problem of flexible docking is to generate conformers of the two components prior to the search, and assemble them pairwise (Grünberg et al. 2004; Bonvin 2006; Lesk and Sternberg 2008; Dobbins et al. 2008; Ritchie 2008; Zacharias 2010). A recent, and valuable, application of the method is implemented in the MultiFit server (Table 5.1); it allows multiple conformers generated from a X-ray structure to be fitted in electron microscopy images that have a much lower resolution, but may display significant conformation changes relative to the atomic model (Tjioe et al. 2011).

The different approaches to the problem of conformation changes in docking correspond to different possible recognition mechanisms. Rigid-body docking mimics the specific recognition between two proteins that bear complementary surfaces, ready to interact when they come into contact. Monte-Carlo searches with variable dihedral angles simulate an induced fit mechanism in which the components first make a low-stability, low-specificity contact, and then adjust their conformation to optimize their interaction. Docking conformers pairwise closely reproduces conformer selection, an alternative to induced fit. In this mechanism, a minority of the molecules have the conformation that allows rigid-body recognition to start with, and it is the formation of a complex that causes the equilibrium to shift (Grünberg et al. 2004, 2006).

5.5.2 *Scoring in CAPRI*

Docking conformers pairwise multiplies the number of searches, and this is practical only with a fast docking algorithm (Schneidman-Duhovny et al. 2005; Mashiach et al. 2010). In addition to being computationally demanding, the method generates a great number of false positives, and puts a heavy load on the scoring functions. In recent years, CAPRI has been adapted to assess scoring separately from docking. The scoring part of the experiments operates in this way: after a prediction round is completed, the predictor groups are asked to upload a hundred or so of their models, which are merged into a file that may contain a thousand models, issued from ten or more different procedures. The scorer groups download the file and rank the whole set, to make their own ten-model submission. In several cases, target T37 of Fig. 5.1 for instance, the scorers' submissions contained more accurate models than the predictors' ones. These models came from the same docking searches, but the predictors' procedures scored them low, whereas some of the schemes developed by scorer groups adequately identified them as correct (Lensink and Wodak 2010).

A scoring function can be physical-chemical (force fields, solvation energies), or empirical, combining terms from different origins with weights optimized on sets of positive and negative examples. It may include non-structural information derived from the comparison of homologous sequences, from point mutants or other genetic or biochemical experiments. However, such information is often ambiguous, and sometimes misleading. If external information is used to screen models during or after the search, it should be treated as a flexible restraint rather than a rigid constraint. The HADDOCK procedure efficiently incorporates such information into a search algorithm that can also handle data from other sources, NMR experiments for instance (Dominguez et al. 2003; de Vries et al. 2007, 2010; Stratmann et al. 2011).

5.5.3 Flexibility and the Docking Benchmark

Developing scoring functions is an active field of research in many fields of science, but in docking, the main difficulty remains flexibility. The structures deposited in the Protein Data Bank illustrate many kinds of conformation changes, the docking benchmark of Weng and colleagues, also. The benchmark is a set of PDB entries assembled to test docking procedures. It contained only 59 complexes in its first version (Chen et al. 2003b), but now has entries for 176 protein-protein complexes and their unbound components; one-third display significant backbone movements with root-mean-square amplitudes that range from 1.5 to 10 Å (Hwang et al. 2010, and Table 5.1).

The complexes of the benchmark are implicated in all sorts of biological processes. Antigen/antibody and enzyme/inhibitor complexes are no longer a majority. Signal transduction and cellular trafficking (exemplified by Arf6 in Fig. 5.1) are well represented, and the protein-protein complexes involved these processes offer many examples of flexible recognition. Conformation changes mediate signal transduction in many ways: they may change the affinity of a protein for a small ligand, another protein or DNA, enhance or inhibit a catalytic activity, the GTPase activity of a G-protein for instance, mask or reveal a group that governs the cellular localization of the protein or its attachment to a membrane. Their variety is immense, comparable in principle to the variety of macromolecular interactions seen in nature, which neither the docking benchmark nor the PDB itself, are close to cover. Moreover, entire classes of interactions are missing: those that involve membrane proteins and intrinsically disordered proteins (IDP), for instance. IDP are implicated in many macromolecular interactions (Dunker et al. 2005, 2008; Tompa et al. 2009), and they undergo disorder-to-order transitions when they interact with other components. Simulating such transitions in the context of docking will remain a challenge for many years.

5.6 Designing Interactions and Predicting Affinity

5.6.1 Engineering Novel Protein-Protein Interactions

Docking can serve other purposes than predicting structures. In Seattle, David Baker, who developed Rosetta, uses docking to engineer novel interactions. The procedure starts by selecting a pair of protein scaffold structures; a coarse-grain docking search identifies candidate complexes; they are computationally mutated at a few interface sites, the modeled mutant complexes are energy-refined, and the top-scoring solutions selected for cloning and expression in yeast. A first experiment aimed to generate a stable interaction between an ankyrin repeat protein and a set of 37 small, structurally diverse, proteins (Karanicolas et al. 2011). A second experiment targeted the stem region of the flu virus hemagglutinin, aiming to mimic the way a neutralizing antibody binds to that epitope (Fleishman et al. 2011a). Both yielded protein constructs that showed reproducible binding, and a round of in vitro evolution was sufficient to improve their affinity to K_d values below nanomolar. Moreover, two co-crystal structures showed that the binding modes had been correctly modeled, although in one, the ligand was oriented 180° away from the model (Karanicolas et al. 2011).

This remarkable piece of protein engineering demonstrates that rational design is now capable to create functional interactions de novo. However, the success rate was low. In the flu hemagglutinin experiment, computational design had culled some 260,000 docking models down to 88 candidate binders derived from 79 different protein scaffolds, but when the constructs were expressed and tested in yeast, only two actually bound (Fleishman et al. 2011a). Nevertheless, the Rosetta force field had predicted about the same binding energies for the designs that failed and for the natural complexes of the Weng docking benchmark. To improve the success rate, a more accurate force field, or a more discriminative scoring function, was clearly required.

5.6.2 The CAPRI Affinity Prediction Experiment

The Seattle group decided to put the question to the CAPRI community: given the structure of a designed complex, can one predict whether it will be stable or not? And they submitted as targets of the scoring experiment a total of 108 designs, including two that bound, during two successive CAPRI rounds held in 2010. The scorers were asked to estimate the affinity of the designed complexes, and rank them along with the complexes of the docking benchmark. When the submissions were analyzed, none of the scorers had ranked the natural complexes significantly above the designs (Fleishman et al. 2011b). Moreover, of the two designs that bound, one had been predicted to be stable by two groups, the other, by no one, a result not far from random. The obvious conclusion of this experiment was that the scoring functions used in docking did not yield reliable binding energies. They had been developed to identify the correct mode of assembly of two proteins known to interact, not to determine whether or not they form a stable complex, and this was beyond their capacity. A parallel study showed a very poor correlation between experimental binding energies and values calculated with several scoring procedures (Kastritis and Bonvin 2010), with the same conclusion that the latter could not predict affinity.

5.6.3 A Structure Affinity Benchmark

The binding energy of a complex, or more correctly its Gibbs free energy of dissociation ΔG_d derived from the equilibrium constant K_d, is a convenient measure of affinity. K_d is known from biophysical measurements in solution for many protein-protein complexes that have been studied by crystallography, and a number of authors have attempted to derive ΔG_d from these structures. The first were Horton and Lewis (1992). They collected data on 16 protein-protein complexes of known structure (mostly protease/inhibitor complexes at that time), and found that a model based on just the size and chemical composition of the interface yielded ΔG_{calc} values that were within 1 or 2 kcal.mol^{-1} of the measured ΔG_{exp}. However, there was an exception: their model predicted a very similar affinity for BPTI binding to trypsin and trypsinogen, whereas the experimental values differed by 10 kcal.mol^{-1}. Horton and Lewis knew the reason why, and their paper discusses it. Trypsinogen, an inactive precursor of trypsin, has flexible surface loops that become ordered when BPTI binds (Bode et al. 1978). As a result, its affinity for the inhibitor is orders of magnitude less than trypsin, where no such change occurs, even though the two complexes with BPTI are nearly identical in structure.

Like trypsin, most of the proteases and inhibitors of the Horton-Lewis set bind as rigid bodies, with no major conformation change to affect their thermodynamic stability. Later studies of the affinity/structure relationship in protein-protein complexes employed larger data sets and more elaborate models of ΔG_{calc}. But as none took into account the structure of the free proteins, they all ignored the role of conformation changes, and also the large effect that experimental conditions, especially pH, can have on K_d. Not surprisingly, the correlation between ΔG_{calc} and ΔG_{exp} was poor in these studies. In addition, errors accumulated in the structure/affinity sets that served to optimize or test the models, as each study re-used data collated by previous ones. Many of the experimental values in the sets were incorrect, some grossly so; for instance, trypsinogen/BPTI and trypsin/BPTI were given the same ΔG_{exp}, a 10 kcal.mol^{-1} error. There was an obvious need for a validated test set, and in 2010, I teamed with three other groups to assemble a benchmark set of binary complexes that would have (a) experimental structures for both the complex and its components; (b) a reliable K_d measured under well-defined conditions. The 176 complexes of the Weng docking benchmark satisfied condition (a). They were an obvious starting point, and we undertook to scan the biochemical literature in search of a K_d for them.

To our great satisfaction, we could locate thermodynamic data for most of the docking benchmark, although some complexes had to be replaced by homologs that also satisfied condition (a). The K_d values, which cover a wide range from 10^{-5} to 10^{-14} M, are derived from either a titration, mostly ITC (isothermal titration calorimetry), or from the binding kinetics (surface plasmon resonance); a few are from enzymic inhibition. The present version of the structure/affinity benchmark comprises 144 complexes, and includes nine pairs that have very similar structures and very different affinities, due to differences in conformation or in sequence. For each entry, the benchmark cites PDB codes for the complex and its components, the K_d and ΔG_d values with the method and experimental conditions of their measurement, and the relevant literature references (Kastritis et al. 2011, and Table 5.1).

5.7 Conclusion

The major achievement of protein-protein docking has been its contribution to our understanding of macromolecular interaction. Docking simulations demonstrate that the shape and chemical complementarity of the molecular surfaces is the major determinant in rigid-body recognition, which is a valid approximation in a number of biological systems. Then, docking has a high predictive value, confirmed by CAPRI and by experiments in which novel interactions are rationally designed *de novo*. However, many processes of great biological importance rely on flexible recognition, in which case the molecular surfaces become complementary only as a result of conformation changes. The CAPRI targets that display flexible recognition have stimulated new developments in the field of docking. Albeit still be far from routine, methods to predict and simulate conformation changes have reached the stage where they can produce useful models, and this has relevance to other fields. In structural biology, much effort is made to fit the atomic resolution structure of assembly components into lower resolution images from cryo-electron microscopy, or an envelope derived from small-angle X-ray scattering, while allowing the structure to change. This is a typical flexible docking problem, to which some docking algorithms have already been applied. In drug design, the target proteins often make other interactions than the one of interest. This may induce conformation changes and allosteric effects that should be taken into account in the design procedure. Similarly, computational biologists may want to study how protein folding is affected by external interactions, in a homodimer for instance. Beyond the structure, we want to understand what governs the specificity of macromolecular recognition and the stability of protein assemblies. This implies that we should be able to model the thermodynamics and the mechanism of the association reaction. The recent attempt to predict affinity within the CAPRI experiment suggests that present force fields are inadequate, and new methods must be developed. The structure/affinity benchmark assembled on this occasion should help biophysicists to correlate function to structure, and remind them that the structure may change as new interactions are formed.

Acknowledgements JJ acknowledges a long term collaboration with Pr. Shoshana Wodak (Free University, Brussels, and University of Toronto), fostered in its early days by Dr. Georges Cohen (Institut Pasteur, Paris), the European Molecular Biology Organisation, and Centre Européen de Calcul Atomique et Moléculaire. The CAPRI experiment owes much to R. Mendez and M. Lensink who perform the assessment together with S. Wodak, to K. Henrick and S. Velenkar who manage its Web site at the European Bioinformatics Institute (Hinxton, UK), to J. Moult, S. Vajda, I. Vakser, M. Sternberg, and L. Ten Eyck, who assist in its management, and to all the structural biologists who have contributed targets.

References

Baldwin J, Chothia C (1979) Haemoglobin: the structural changes related to ligand binding and its allosteric mechanism. J Mol Biol 129:175–220

Beddell CR, Goodford PJ, Norrington FE, Wilkinson S, Wootton R (1976) Compounds designed to fit a site of known structure in human haemoglobin. Br J Pharmacol 57:201–209

Blow DM, Wright CS, Kukla D, Rühlmann A, Steigemann W, Huber R (1972) A model for the association of bovine pancreatic trypsin inhibitor with chymotrypsin and trypsin. J Mol Biol 69:137–144

Bode W, Schwager P, Huber R (1978) The transition of bovine trypsinogen to a trypsin-like state upon strong ligand binding. The refined crystal structures of the bovine trypsinogen-pancreatic trypsin inhibitor complex and of its ternary complex with Ile-Val at 1.9 A resolution. J Mol Biol 118:99–112

Bonvin AM (2006) Flexible protein-protein docking. Curr Opin Struct Biol 16:194–200

Camacho CJ, Vajda S (2002) Protein-protein association kinetics and protein docking. Curr Opin Struct Biol 12:36–40

Chen R, Li L, Weng Z (2003a) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87

Chen R, Mintseris J, Janin J, Weng Z (2003b) A protein-protein docking benchmark. Proteins 52:88–91

Cherfils J, Duquerroy S, Janin J (1991) Protein-protein recognition analyzed by docking simulation. Proteins 11:271–280

Connolly ML (1986) Shape complementarity at the hemoglobin α1 β1 subunit interface. Biopolymers 25:1229–1247

de Vries SJ, van Dijk AD, Krzeminski M, van Dijk M, Thureau A, Hsu V, Wassenaar T, Bonvin AM (2007) HADDOCK versus HADDOCK: new features and performance of HADDOCK2.0 on the CAPRI targets. Proteins 69:726–733

de Vries SJ, van Dijk M, Bonvin AM (2010) The HADDOCK web server for data-driven biomolecular docking. Nat Protoc 5:883–897

Dobbins SE, Lesk VI, Sternberg MJ (2008) Insights into protein flexibility: The relationship between normal modes and conformational change upon protein-protein docking. Proc Natl Acad Sci USA. 105:10390–5

Dominguez C, Boelens R, Bonvin AM (2003) HADDOCK: a protein-protein docking approach based on biochemical or biophysical information. J Am Chem Soc 125:1731–1737

Dunker AK, Cortese MS, Romero P, Iakoucheva LM, Uversky VN (2005) Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J 272:5129–5148

Dunker AK, Silman I, Uversky VN, Sussman JL (2008) Function and structure of inherently disordered proteins. Curr Opin Struct Biol 18:756–764

Fernández-Recio J, Totrov M, Abagyan R (2002) Soft protein-protein docking in internal coordinates. Protein Sci 11:280–291

Fieulaine S, Morera S, Poncet S, Mijakovic I, Galinier A, Janin J, Deutscher J, Nessler S (2002) X-ray structure of a bifunctional protein kinase in complex with its protein substrate HPr. Proc Natl Acad Sci USA 99:13437–13441

Fleishman SJ, Whitehead TA, Ekiert DC, Dreyfus C, Corn JE, Strauch EM, Wilson IA, Baker D (2011a) Computational design of proteins targeting the conserved stem region of influenza hemagglutinin. Science 332:816–821

Fleishman SJ, Whitehead TA, Strauss EM et al., Wodak SJ, Janin J, Baker D (2011b) Community-wide assessment of protein-interface modeling suggests improvements to design methodology. J Mol Biol Accepted manuscript doi:10.1016/j.jmb.2011.09.031

Gabb HA, Jackson RM, Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272:106–120

Goodford PJ (1985) A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. J Med Chem 28:849–857

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

Grünberg R, Leckner J, Nilges M (2004) Complementarity of structure ensembles in protein-protein binding. Structure 12:2125–2136

Grünberg R, Nilges M, Leckner J (2006) Flexibility and conformational entropy in protein-protein binding. Structure 14:683–693

Günther S, May P, Hoppe A, Frömmel C, Preissner R (2007) Docking without docking: ISEARCH–prediction of interactions using known interfaces. Proteins 69:839–844

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

Heifetz A, Katchalski-Katzir E, Eisenstein M (2002) Electrostatics in protein-protein docking. Protein Sci 11:571–587

Horton N, Lewis M (1992) Calculation of the free energy of association for protein complexes. Protein Sci 1:169–181

Huber R, Kukla D, Bode W, Schwager P, Bartels K, Deisenhofer J, Steigemann W (1974) Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. II. Crystallographic refinement at 1.9 A resolution. J Mol Biol 89:73–101

Hwang H, Vreven T, Janin J, Weng Z (2010) Protein-protein docking benchmark version 4.0. Proteins 78:3111–3114

Isabet T, Montagnac G, Regazzoni K, Raynal B, Elkhadali F, Franco M, England P, Chavrier P, Houdusse A, Menetrey J (2009) The structural basis of Arf effector specificity: the crystal structure of ARF6 in a complex with JIP4. EMBO J 28:2835–2845

Janin J (2005) Assessing predictions of protein-protein interaction: the CAPRI experiment. Protein Sci 14:278–283

Janin J (2010) Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6:2351–2362

Janin J, Chothia C (1990) The structure of protein-protein recognition sites. J Biol Chem 265:16027–16030

Janin J, Wodak SJ (1981) Report on the workshop on non-bonded interactions and the specificity of protein association and folding. CECAM, Orsay, France pp 5–64

Janin J, Wodak SJ (1985) Reaction pathway for the quaternary structure change in hemoglobin. Biopolymers 24:509–526

Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJ, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: a critical assessment of PRedicted interactions. Proteins 52:2–9

Jiang F, Kim SH (1991) “Soft docking”: matching of molecular surface cubes. J Mol Biol 219:79–102

Karanicolas J, Corn JE, Chen I, Joachimiak LA, Dym O, Peck SH, Albeck S, Unger T, Hu W, Liu G, Delbecq S, Montelione GT, Spiegel CP, Liu DR, Baker D (2011) A de novo protein binding pair by computational design and directed evolution. Mol Cell 42:250–260

Kastritis PL, Bonvin AM (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, Erratum in: J Proteome Res 10:921–2, 2011

Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J (2011) A structure-based benchmark for protein-protein binding affinity. Protein Sci 20:482–491

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

Keskin O, Nussinov R, Gursoy A (2008) PRISM: protein-protein interaction prediction by structural matching. Methods Mol Biol 484:505–521

Kundrotas PJ, Zhu Z, Janin J, Vakser IA (2012) Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci USA.109:9438–41

Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE (1982) A geometric approach to macromolecule-ligand interactions. J Mol Biol 161:269–288

Lensink MF, Wodak SJ (2010) Docking and scoring protein interactions: CAPRI 2009. Proteins 78:3073–3084

Lensink MF, Méndez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69:704–718

Lesk VI, Sternberg MJ (2008) 3D-Garden: a system for modelling protein-protein complexes based on conformational refinement of ensembles generated with the marching cubes algorithm. Bioinformatics 24:1137–1144

Levinthal C, Wodak SJ, Kahn P, Dadivanian AK (1975) Hemoglobin interaction in sickle cell fibers. I: theoretical approaches to the molecular contacts. Proc Natl Acad Sci USA 72:1330–1334

Levitt M (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107

Levitt M, Lifson S (1969) Refinement of protein conformations using a macromolecular energy minimization procedure. J Mol Biol 46(2):269–279

Lu L, Arakaki AK, Lu H, Skolnick J (2003). Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. Genome Res. 13:1146–54

Lu L, Lu H, Skolnick J (2002) MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 49:350–364

Mandell JG, Roberts VA, Pique ME, Kotlovyi V, Mitchell JC, Nelson E, Tsigelny I, Ten Eyck LF (2001) Protein docking using continuum electrostatics and geometric fit. Protein Eng 14:105–113

Mashiach E, Schneidman-Duhovny D, Peri A, Shavit Y, Nussinov R, Wolfson HJ (2010) An integrated suite of fast docking algorithms. Proteins 78:3197–3204

Méndez R, Leplae R, De Maria L, Wodak SJ (2003) Assessment of blind predictions of protein-protein interactions: current status of docking methods. Proteins 52:51–67

Méndez R, Leplae R, Lensink MF, Wodak SJ (2005) Assessment of CAPRI predictions in rounds 3–5 shows progress in docking procedures. Proteins 60:150–169

Moult J, Pedersen JT, Judson R, Fidelis K (1995) A large-scale experiment to assess protein structure prediction methods. Proteins 23(3):ii–v

Némethy G, Scheraga HA (1977) Protein folding. Q Rev Biophys 10:239–252

Norel R, Lin SL, Wolfson HJ, Nussinov R (1994) Molecular surface complementarity at protein-protein interfaces: the critical role played by surface normals at well placed, sparse, points in docking. J Mol Biol 252:263–273

Nussinov R, Wolfson HJ (1991) Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques. Proc Natl Acad Sci USA 88:10495–10499

Ritchie DW (2008) Recent progress and future directions in protein-protein docking. Curr Protein Pept Sci 9:1–15

Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39:178–194

Schneidman-Duhovny D, Inbar Y, Nussinov R, Wolfson HJ (2005) Geometry-based flexible and symmetric protein docking. Proteins 60:224–231

Schueler-Furman O, Wang C, Bradley P, Misura K, Baker D (2005) Progress in modeling of protein structures and interactions. Science 310:638–642

Sinha R, Kundrotas PJ, Vakser IA (2010) Docking by structural similarity at protein-protein interfaces. Proteins 78:3235–3241

Smith GR, Sternberg MJ (2002) Prediction of protein-protein interactions by docking methods. Curr Opin Struct Biol 12:28–35

Stein A, Mosca R, Aloy P (2011) Three-dimensional modeling of protein interactions and complexes is going 'omics. Curr Opin Struct Biol 21:200–208

Stratmann D, Boelens R, Bonvin AM (2011) Quantitative use of chemical shifts for the modeling of protein complexes. Proteins 79:2662–2670

Strynadka NC, Eisenstein M, Katchalski-Katzir E, Shoichet BK, Kuntz ID, Abagyan R, Totrov M, Janin J, Cherfils J, Zimmerman F, Olson A, Duncan B, Rao M, Jackson R, Sternberg M, James MN (1996) Molecular docking programs successfully predict the binding of a β-lactamase inhibitory protein to TEM-1 β-lactamase. Nat Struct Biol 3:233–239

Tjioe E, Lasker K, Webb B, Wolfson HJ, Sali A (2011) MultiFit: a web server for fitting multiple protein structures into their electron microscopy density map. Nucleic Acids Res 39:W167–W170

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 31:328–335

Totrov M, Abagyan R (1994) Detailed ab initio prediction of lysozyme-antibody complex with 1.6 A accuracy. Nat Struct Biol 1:259–263

Tuncbag N, Gursoy A, Guney E, Nussinov R, Keskin O (2008) Architectures and functional coverage of protein-protein interfaces. J Mol Biol 381:785–802

Vajda S, Vakser IA, Sternberg MJ, Janin J (2002) Modeling of protein interactions in genomes. Proteins 47:444–446

Vakser IA, Aflalo C (1994) Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins 20:320–329

Wodak SJ, Janin J (1978) Computer analysis of protein-protein interaction. J Mol Biol 124:323–342

Zacharias M (2003) Protein-protein docking with a reduced protein model accounting for side-chain flexibility. Protein Sci 12:1271–1282

Zacharias M (2010) Accounting for conformational changes during protein-protein docking. Curr Opin Struct Biol 20:180–186

Chapter 6
Prediction of Protein-Protein Binding Interfaces

Damian Marchewka, Wiktor Jurkowski, Mateusz Banach, and Irena Roterman-Konieczna

Keywords HADDOCK • ZDOCK • RosettaDock • Oil drop • Sequence conservation • Mutagenesis • Epitope mapping • H-D exchange • Crosslinking experiments • Solvated docking • Ambiguous Interaction Restraints (AIRs) • Rotamer packing • Side-chain rotamer probabilities • Monte Carlo-based modeling package • Geometric alignment • Fast Fourier Transform algorithms • CHARMM forcefields • Homodimer • Pair-wise interaction • Lock-key • Fuzzy oil drop

6.1 Introduction

When it comes to regulating protein activity, complexation mechanisms are just as important as ligand binding. Most proteins never exist in isolation – instead they serve as building blocks for more complex systems. Some proteins form multimers to ensure maintain spatial alignment (required e.g. for phase separation in the dual lipid layer and formation of hydrophilic compartments in ion channels (Unwin 2005; Jasti et al.. 2007)); others may require temporary binding of cofactors (e.g. regulation of transcription factors (Huxford et al. 1998)), or are part of complicated protein machinery (e.g. proton-driven rotors in ATP synthases (Boyer 1997; Oster and Wang 1999, 2003)).

D. Marchewka • M. Banach • I. Roterman-Konieczna (✉)
Department of Bioinformatics and Telemedicine, Jagiellonian
University – Medical College, Lazarza 16, 31-530 Cracow, Poland
e-mail: myroterm@cyf-kr.edu.pl

Faculty of Physics, Astronomy and Applied Computer Science, Jagiellonian University,
Reymonta 4, 30-059 Cracow, Poland

W. Jurkowski
Computational Biology Group, Luxembourg Centre for Systems
Biomedicine, University of Luxembourg, Campus Belval 7, avenue des
Hauts-Fourneaux, L-4362 Esch-Belval, Luxembourg

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_6,

In spite of the ever-increasing quality and cost-effectiveness of experimental techniques, *in silico* protein interaction models remain an important tool for the study of biological processes which involve molecular complexation – such as signaling pathways and biological clock systems (Johnson et al. 2008; Bass and Takahashi 2010; Duong et al. 2011; van Gelder et al. 2003). Analysis of protein-protein interactions may be performed on multiple levels of accuracy. On the most basic level it is usually sufficient to determine whether interaction occurs at all under certain conditions. This is done in order to infer the so-called protein-protein interaction (PPI) networks. Such networks can then be refined by predicting protein binding interfaces (to determine their approximate steric relationships). Accurately modeling the 3D structure of the entire complex is the most challenging task, requiring knowledge of molecular interactions on the atomic level.

Over the past two decades many methodologies and algorithms have been developed and applied (with varying success) on each of these three distinct levels.

An objective measure of the accuracy of protein complexation models is provided by the CAPRI (Critical Assessment of Predicted Interactions) challenge (see preceeding chapter). Similarly to CASP contestants are provided with the necessary input data (structures of individual monomers or amino acid sequences – if the structure is easy to predict). Since the goal is to determine subtle details of protein-protein interfaces, prediction quality is dependent not only on the number of correctly modeled contact points, but also on the accuracy of atom positions within the interface zone. Solutions are penalized by the number of steric clashes between interacting chains. Successive editions of the CAPRI challenge are being organized on a regular basis since 2001. Numerous publications are available regarding the challenge itself and its most accurate prediction pipelines (Janin et al. 2003; Janin and Wodak 2007; Janin 2007, 2010a, b; Kastritis et al. 2011).

Docking analysis is a complex process, typically composed of three phases: (1) selection of interface candidates based on experimental data (or predictions) to focus the conformational search; (2) generation of the protein complex via rigid-body docking; (3) ranking and scoring of results. Thus, we limit our description to tools most often used in these sophisticated algorithms.

Our study focuses on prediction of protein interfaces – a task which should enable us to detect protein complexation events and may also serve higher-order structure prediction workflows. Even on this basic level existing tools are incapable of identifying interface residues with consistent accuracy. The varying difficulty of modeling complexation events suggests that many different binding mechanisms come into play and thus many different kinds of interacting residues can be observed (differing with respect to their specificity, sequence conservation, hydrophobicity, etc.)

This chapter presents four differing approaches to prediction of protein-protein binding interfaces by means of molecular docking. We will compare FOD with three state-of-the-art models: HADDOCK, ZDOCK and RosettaDock, each of which is an implementation of rigid-body grid-based docking algorithms. Besides the specifics of formulation and parameterization of force fields, the main difference between these three distinct programs lies in their approach to focusing the conformational space search. ZDOCK allows the user to select specific amino acids while HADDOCK calls for experimental preselection of interface residues. Users of RosettaDock may take advantage of the Rosetta modeling package to predefine docking sites and design a refinement pipeline which consists of global and local search steps.

Other toolkits available on the Internet but not detailed in this chapter include Autodock 3 (Cosconati et al. 2010), eHiTs (Zsoldos et al. 2006), MOE (Feldman and Labute 2010), FlexX (Kramer et al. 1999), ClusPro (Kozakov et al. 2010), GRAMM-X (Tovchigrechko and Vakser 2006), the PatchDock and SymmDock (Schneidman-Duhovny et al. 2005) programs based on shape complementarity principles and symmetry restrictions, as well as Hex which bases on spherical harmonic representations (Macindoe et al. 2010)

6.2 Programs Description

The models applied are presented in alphabetic order.

6.2.1 Fuzzy Oil Drop Model

The "fuzzy oil drop" model aims to not only recognize protein complexation sites, but also discover the mechanisms which cause proteins to form complexes (Konieczny et al. 2006). Having explained the premise of the model in the previous chapter we will now limit ourselves to a brief recapitulation of its key features. At the core of the model lies the assumption that proteins which undergo folding in an aqueous environment tend to internalize hydrophobic residues while exposing hydrophilic residues on their surfaces. Entropic considerations suggest that a standalone protein molecule should assume a globular form as a result of interaction with water. Once folded, the protein possesses a clearly identifiable hydrophobic core (hence the reference to Kauzmann's "oil drop" concept (Kauzmann 1959)), which can be modeled with a 3D Gauss-like hydrophobicity distribution field. Particularly good agreement with this model can be observed e.g. in fast-folding proteins (Roterman et al. 2011a, b), although it should be noted that most proteins exhibit certain deviations from the "ideal" hydrophobicity distribution. Such deviations are caused by the influence of external factors on the folding process – this includes ligands and other protein molecules which form complexes with the protein in question.

The $\Delta\tilde{H}$ profile is a measure of the discrepancy between the expected hydrophobicity (given by Gauss' distribution – $\tilde{H}t$) and the actual (observed) hydrophobicity for the *i*th aminoacid (or, more specifically, for its effective atom, placed at the geometric center of the amino acid's side chain). Actual hydrophobicity ($\tilde{H}o$) can be determined by calculating hydrophobic interactions between the amino acid and all of its neighbors in a 9 Å radius. According to Levitt (Levitt 1976):

$$\Delta\tilde{H}_i = \tilde{H}t_i - \tilde{H}o_i$$

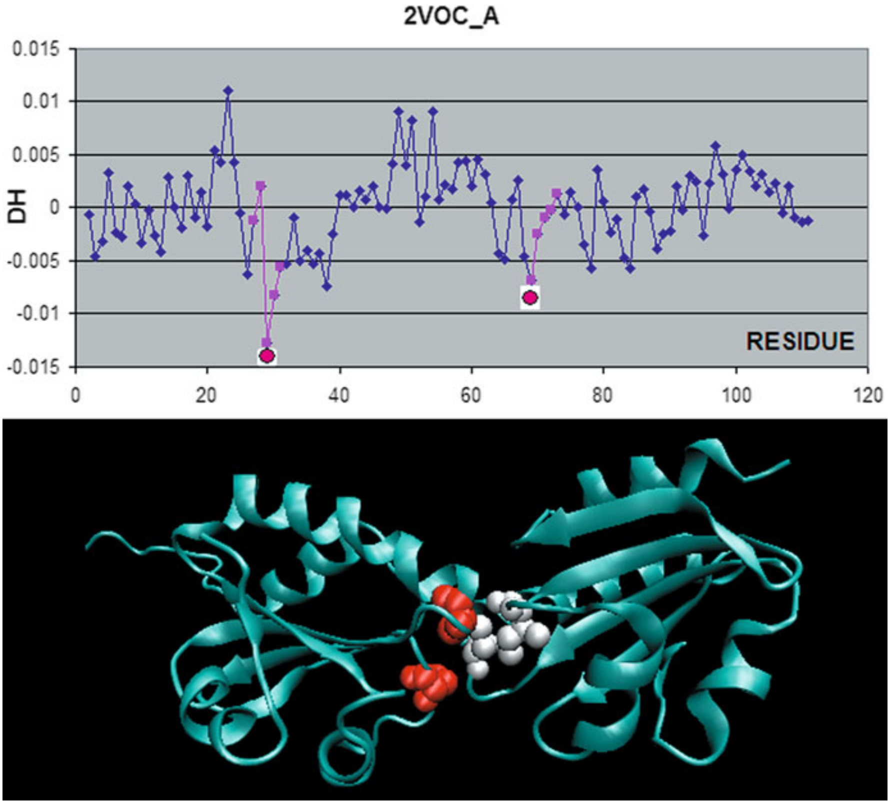


Fig. 6.1 Thioredoxin homodimer complex (2VOC) with a mutated active center, composed of mixed disulfide dimers which resemble the enzyme-substrate reaction intermediate. *Top*: $\Delta\tilde{H}$ profile indicating residues engaged in complexation (according to PDBSum). *Bottom*: 3D ribbon model with CPK representations of local $\Delta\tilde{H}$ profile minima (highlighted in the *top* diagram)

The value of $\Delta\tilde{H}$ expresses irregularities in the distribution of hydrophobicity, which can manifest themselves as either localized deficiencies or excess of hydrophobicity. Deficiencies are typically associated with the presence of cavities in the globular protein body, while excess hydrophobicity, when present on the surface, indicates potential complexation sites. Such sites are naturally attracted to one another – this interaction shields them from the entropically disadvantageous contact with water, ensuring the formation of a stable complex.

Successful prediction of complexation sites via this mechanism is possible e.g. for the thioredoxin A homodimer (2VOC) with a mutated active center composed of mixed disulfide dimers, resembling the enzyme-substrate reaction intermediate (Kouwen et al. 2008). The biological role of this protein is to assist in electron transport. Each monomer contains strongly hydrophobic Cys and Val residues, representing local minima of the $\Delta\tilde{H}$ profile. Both residues are exposed on the protein surface and therefore preferentially attract hydrophobic residues belonging to the complementary monomer (Fig. 6.1). The resulting dimer is a good example of the applicability of our model.

Another example of structural accordance with the presented model is the transmembrane protein discussed in Zobnina and Roterman (2009), where the central hypothesis of the "fuzzy oil drop" model is validated in the context of interaction between proteins and cellular membranes.

6.2.2 HADDOCK – High Ambiguity Driven Biomolecular DOCKing Based on Biochemical and/or Biophysical Information

HADDOCK implements algorithms for systematic search on a grid (Dominguez et al. 2003; de Vries et al. 2007). Its key innovation lies in a heuristic approach to experimental data, including NMR data such as residual dipolar couplings and relaxation (van Dijk et al. 2005a, b, c), sequence conservation, mutagenesis, epitope mapping, H-D exchange or crosslinking experiments to provide distance restraints, diffusion anisotropy (van Dijk et al. 2006a, b), solvated docking (van Dijk and Bonvin 2006) and flexible protein-DNA docking (van Dijk et al. 2006b). If no experimental cues are available, docking is guided by randomly selected patches of tentatively active residues and restrained by molecular center-of-mass criteria.

Prior to running HADDOCK a set of Ambiguous Interaction Restraints (AIRs) has to be generated. This procedure involves distinguishing "active" and "passive" residues. Active residues are those which interact with the target protein while remaining in contact with water. Passive residues are also exposed to water and lie in direct proximity to active ones. The cutoff criterion is not explicit; rather, it acknowledges the structures of important functional groups. Determination is based on NMR data and requires end-user assessment.

HADDOCK also enables identification of protein-protein interaction candidates on the basis of the so-called conservative sequences. This process necessitates further division of residues into "active" and "passive" sets (de Vries et al. 2006). If none of the presented approaches is feasible (e.g. due to the lack of NMR input), identification can proceed by way of modeling solvent-accessible areas.

HADDOCK computations have proven quite successful in several editions of the CAPRI challenge (van Dijk et al. 2005a, b; de Vries et al. 2007).

6.2.3 RosettaDock

RosettaDock works as an extension of the Rosetta structure prediction package. Similarly to other docking algorithms, the first phase of the search involves sampling the rigid-body conformational space with side chains represented by single pseudo-atoms. Docking conformations may be refined in the second (full-atom) phase via small-scale perturbations of the complex and side chain optimization, employing rotamer packing and continuous minimization in order to avoid entrapment in local energy minima (Lyskov and Gray 2008; Gray et al. 2003). Results are ranked on the basis of residue-residue contacts and clashes, as well as the properties of pairwise residue-environment and residue-residue interactions derived from previously registered datasets (Lyskov and Gray 2008). In the refinement phase the goal is to find a structure with the lowest possible free energy. Energy calculations acknowledge van der Waals forces (Gray et al. 2003), orientation-dependent hydrogen bonding (Kortemme et al. 2003), implicit Gaussian solvation (Lazaridis and Karplus 2000), side-chain rotamer probabilities (Dunbrack and Cohen 1997) and electrostatic potentials (Gray et al. 2003).

Input datasets consist of subunit (monomer) structures expressed in the PDB format. It is also necessary to determine a suitable starting structure, free from steric clashes. Based on the results of the CAPRI challenge, approximately 1,000 structures are generated for each pair of interacting proteins. The program then determines the relationship between free energy and RMS-D values for the final structure and for the initial conformation. Success is defined as the ability to locate an area in the RMS-D/free energy space where both parameters are suitably low. The best result (lowest values of both parameters) is treated as the correct final structure.

RosettaDock is therefore a multistart, multiscale Monte Carlo-based modeling package. The end result is highly dependent on the initial structure, which – as already mentioned – should be provided by the user, preferably on the basis of experimental data (e.g. site-directed mutagenesis). In addition, RosettaDock is capable of exploiting structures generated by other global search-oriented software packages, such as ClusPro (http://cluspro.bu.edu/login.php) (Kozakov et al. 2010), GRAMM-X (http://vakser.bioinformatics.ku.edu/resources/gramm/ grammx) (Tovchigrechko and Vakser 2006), HEX (http://hex.loria.fr/) (Macindoe et al. 2010), PatchDock (http://bioinfo3d.cs.tau.ac.il/PatchDock/) (Schneidman-Duhovny et al. 2005) and SymmDock (http://bioinfo3d.cs.tau.ac.il/SymmDock/) (Schneidman-Duhovny et al. 2005).

RosettaDock has been notably successful in blind-prediction studies within the CAPRI challenge (Gray et al. 2003; Lensink et al. 2007).

6.2.4 ZDOCK (http://zdock.bu.edu/)

ZDOCK is a rigid-body simulation toolkit. The docking procedure involves geometric alignment of the surfaces of two molecules, treated as potential intermolecular complexation sites. The "target" molecule is rigid, while the complementary molecule is rotated around its surface using a grid with a predefined density. This systematic search bases on Fast Fourier Transform algorithms (Wiehe et al. 2008; Mintseris et al. 2005, 2007; Chen and Weng 2002, 2003; Mintseris and Weng 2003; Li et al. 2003; Chen et al. 2003a, b, c; Pierce et al. 2005, 2007; Pierce and Weng 2007, 2008). ZDOCK applies a scoring function which acknowledges shape complementarity, electrostatics and pairwise atomic potentials determined on the basis of known protein complexes. A separate scoring package (RDOCK) is provided, enabling ZDOCK to estimate the strength of interactions through CHARMM forcefields (MacKerell et al. 1998; Brooks et al. 2009). The conformational search procedure implemented in ZDOCK is often applied in docking protocols which typically augment it with a scoring scheme. It has been successfully tested in several editions of the CAPRI challenge (Wiehe et al. 2005, 2007; Chen et al. 2003a, b, c; Mintseris et al. 2005).

In our study ZDOCK was used with its default settings. Proteins were docked on a dense 6 Å grid, producing 54,000 unique alignments. In each complex the longer chain (receptor) was immobilized while the other chain was rotated around the receptor molecule. The resulting alignments were subsequently scored using ZDOCK, reflecting surface complementarity, electrostatic energy and statistical data on atomic contact potentials in the interface zone. The highest ranked structures were clustered by applying the Root Mean Square Deviation (RMSD) distance criterion for all heavy atom positions. All structures with RMSD lengths of less than 10.0 Å were clustered together, while outliers were removed from pool. Such ranking-based clustering enables relatively fast extraction of the most representative alignments.

6.3 Results

In order to validate the proposed complexation site identification method, a set of homodimers has been prepared by scanning the PDB database for occurrences of the “HOMODIMER” keyword. Structures which did not consist of exactly two chains, or which occurred in complexes with DNA, were exempted from analysis. In addition, the Needleman-Wunsch (Needleman and Wunsch 1970) alignment algorithm was applied to verify sequence similarity (identity): chains differing by more than 20 amino acids (through substitutions or insertions/deletions) were discarded, resulting in a set of 208 acceptable homodimers. This selection was based upon PDB as it existed in March 2010.

Results produced by the above described programs were evaluated in terms of their accuracy, which can be expressed by four distinct ratios: TP (true positive), FP (false positive), TN (true negative) and FN (false negative). The study was based on the F-measure and MCC criteria presented in the previous chapter.

The “fuzzy oil drop” model identifies residues involved in complexation by searching for local maxima (i.e. deficiencies) and minima (i.e. excesses) of the $\Delta\tilde{H}$ profile, as compared to the idealized 3D Gauss distribution of hydrophobicity. This identification is dependent on a predetermined set of thresholds (cutoff values), which, in our study, was pegged at 80% of the peak value (for minima and maxima of the profile). Thus, residues whose $\Delta\tilde{H}$ values were in excess of 80% of their respective peaks (or troughs), were suspected of involvement in complexation. The fraction of these residues which are actually involved in complexation constitutes the true positive (TP) ratio. Similarly, the fraction of residues which the algorithm suspects of involvement in complexation but which do not actually participate in forming complexes is defined as the false positive (FP) ratio. This operation is repeated for several distinct cutoff values (e.g. 80%), resulting in a set of TP/FP ratios, as well as the corresponding TN/FN ratios, calculated in an analogous manner. ROC curves may be applied to quantitative analysis of the results produced by the "fuzzy oil drop" model for various $\Delta\tilde{H}$ cutoff levels (Fawcett 2006).

Below we discuss the accuracy of results produced by various tools, listed in alphabetical order.

6.3.1 Fuzzy Oil Drop Model

Figure 6.2 presents a summary of results generated by the "fuzzy oil drop" model, in terms of F-measure and MCC criteria. Additional numerical data can be found in Table 6.1. Since the MCC and F-measure values depend on the assumed cutoff level, all calculations assumed a threshold of 80% (meaning that residues are identified as participating in complexation if their corresponding $\Delta\tilde{H}$ values are between 80 and 100% of peak levels).

The identification method based on the "fuzzy oil drop" model, when used to pinpoint a single complexation site, produced the following results: for $\Delta\tilde{H}$ of 80% (−) – 35 and 25% on the F-measure and MCC scales respectively, and for $\Delta\tilde{H}$ of 80% (+) – 37 and 30% on both scales. Due to the large number of low-ranked results (with an aggregate score of 0), the LOWEST category has been omitted. The 80% (−) and 80%(+) denotes the 80% level for local maxima and local minima respectively.

It should be noted that, when applying a cutoff level, the set of true positives (TP) usually includes all amino acids directly adjacent to the target residue, even when these amino acids are not directly engaged in complexation. This makes the FP biased.

According to the model, residues which represent local $\Delta\tilde{H}$ profile minima possess excess hydrophobicity compared to theoretical predictions, while $\Delta\tilde{H}$ profile minima correspond to hydrophobicity deficiencies. Thus, the former can be expected to seek out similarly conditioned residues on the surface of the partner molecule.

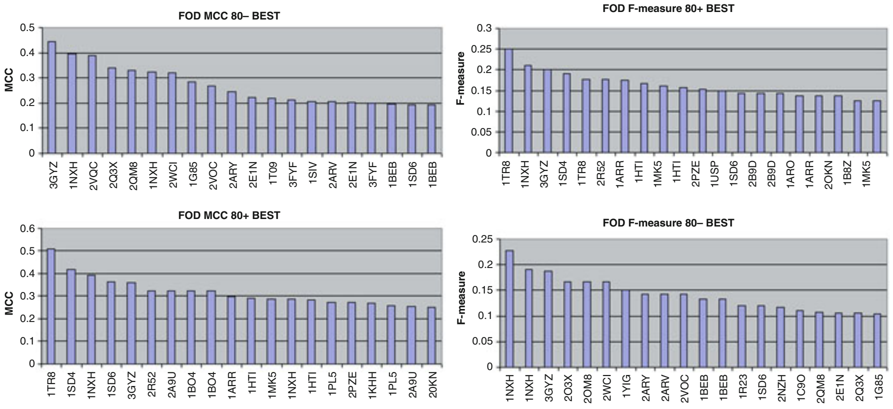


Fig. 6.2 Comparison of F-measure and MCC values for the "fuzzy oil drop" model. The assumed cutoff level was 80% of either the highest or the lowest value of the $\Delta\tilde{H}$ profile

Table 6.1 Summary of most accurate complexation site predictions

$\Delta\tilde{H}$ profile for $\Delta\tilde{H}$ >0			$\Delta\tilde{H}$ profile for $\Delta\tilde{H}$ <0		
PDB ID	Chain	Surface	PDB ID	Chain	Surface
1TR8	A	0.742	1YGA	B	0.82
1TR8	B	0.714	1YGA	A	0.78
1G8M	A	0.609	3GYZ	A	0.56
3CRN	A	0.573	1 G85	A	0.536
2ARV	A	0.550	2QM8	A	0.508
2R52	B	0.543	1 T09	A	0.471
1G8M	B	0.542	1 T09	B	0.453
3GYZ	A	0.527	2E1N	A	0.438
1HDF	B	0.511	2WCI	A	0.436
1HUX	B	0.467	2FJT	A	0.427
1HUX	A	0.457	2FJT	B	0.425
2A9U	B	0.449	2ARV	B	0.405
1 V58	A	0.447	1BFT	A	0.381
3FYF	A	0.442	2QM8	B	0.378
2A9U	A	0.431	1BFT	B	0.375
1FZV	A	0.430	1SD4	A	0.368

The ranking criterion is the surface area bounded by the ROC curve (placed above the diagonal in the TPR/FPR relation graph – see Fig. 6.3) and the corresponding diagonal

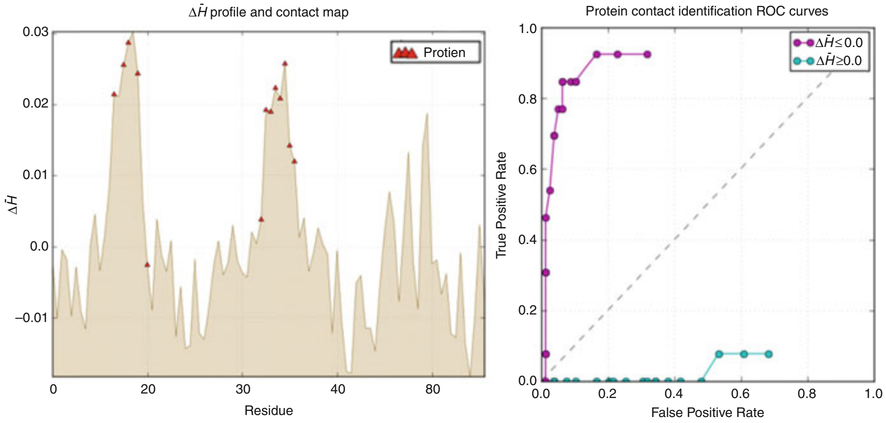


Fig. 6.3 $\Delta\tilde{H}$ profile of the 1TR8 homodimer, indicating residues involved in complexation (*left*). The graph on the *right-hand side* shows the ROC curve corresponding to profile maxima (hydrophobicity deficiencies). The relatively large area bounded by this curve and the diagonal reflects excellent accordance with theoretical predictions and – correspondingly – high accuracy of results

This is true e.g. for the 1TR8 homodimer, which provides a particularly good example of the presented mechanism (Spreter et al. 2005). Figure 6.3 depicts its $\Delta\tilde{H}$ profile, indicating which residues are involved in complexation and presenting the corresponding ROC curve plotted on the FPR/TPR graph, where the surface area bounded by the curve and the diagonal is appropriately large (74% of the unit triangle).

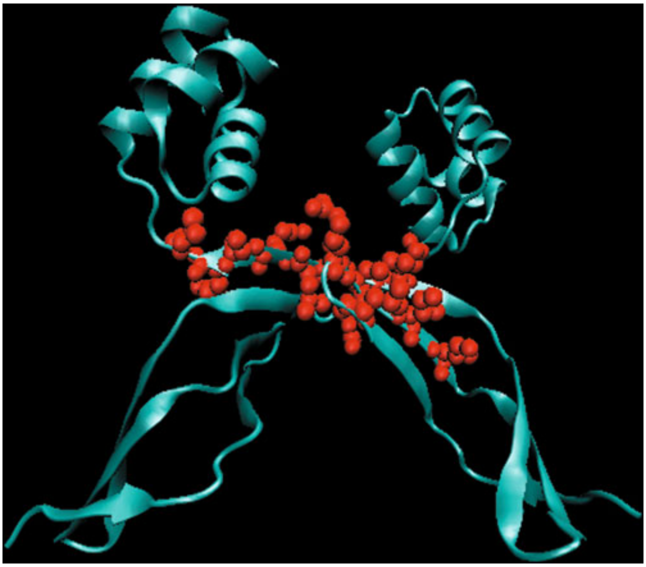


Fig. 6.4 The 1TR8 homodimer. *Red spheres* in the A chain indicate residues which correspond to local $\Delta\tilde{H}$ profile maxima (local hydrophobicity deficiencies). In line with theoretical predictions, these residues are actually involved in complexation

Reviewing the 3D representation of the homodimer reveals that residues which correspond to local $\Delta\tilde{H}$ profile maxima form a docking cavity, anchoring the partner molecule (Fig. 6.4).

The 1YGA homodimer provides an example of complexation mediated by surface residues with excess hydrophobicity. Contact between both monomers enables their surface-bound hydrophobic residues to shield one another from water, resulting in the formation of a stable complex. This phenomenon is illustrated in Fig. 6.5.

Figure 6.6 presents a 3D view of the 1YGA homodimer, indicating which residues correspond to $\Delta\tilde{H}$ profile minima and are actually involved in complexation. Such residues, when present on the surface of the molecule, hint at a potential complexation site (Fig. 6.6).

An interesting phenomenon occurs in the 1SD6 homodimer (Safo et al. 2005). Here, one monomer exposes residues with low hydrophobicity, which, in turn, interact with areas of excess hydrophobicity on the surface of its partner. This homodimer ranks near the top of the F-measure list, both for $\Delta\tilde{H}$ >0 and $\Delta\tilde{H}$ <0. As it turns out, 1SD6 monomers are capable of mutually compensating their deviations from the idealized hydrophobicity distribution, by forming a "lock-and-key-like" configuration (Fig. 6.7).

The list of 51 nonbonding interactions between 1SD6 monomers includes 24 "lock-key" pairs (i.e. pairs where one residue has a positive $\Delta\tilde{H}$ value while the other one corresponds to a negative value of the same profile). We can therefore conclude that, in the case of 1SD6, complexation is dominated by mutual compensation of deviations from the idealized profile. Figure 6.8 further illustrates this situation and in 3-D presentation (Fig. 6.9).

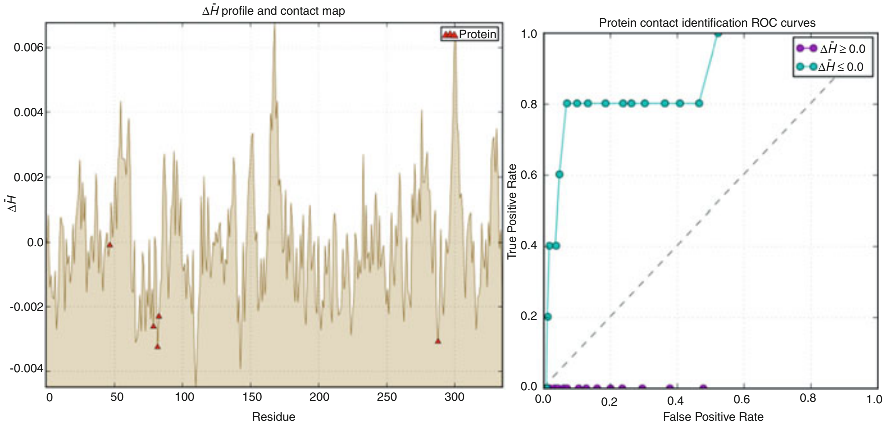


Fig. 6.5 $\Delta\tilde{H}$ profile of the 1YGA homodimer, indicating residues involved in complexation (*left*). The graph on the *right-hand side* shows the ROC curve corresponding to profile minima (excess hydrophobicity) and indicates the relation between TPR and FPR for this homodimer

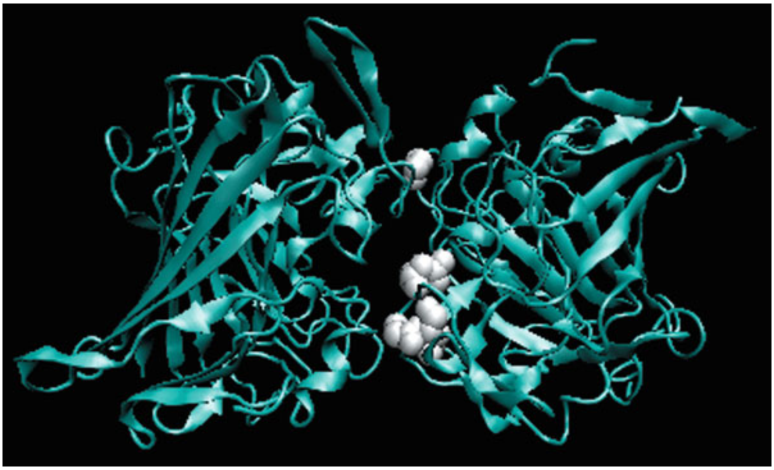


Fig. 6.6 The 1YGA homodimer. *White spheres* in the A chain indicate residues which correspond to local $\Delta\tilde{H}$ profile minima (excess hydrophobicity), such as Val and Cys. The presence of such residues on the molecule surface enables the model to accurately identify complexation sites

6.3.2 HADDOCK

HADDOCK computations were performed using the web service available at http://haddock.science.uu.nl/enmr/services/HADDOCK/haddockserver-easy.html. Input data consisted of PDB files describing "active" amino acids. The Gramm-X preprocessing package (Tovchigrechko and Vakser 2006) was not used. HADDOCK works by optimizing binding energy values in a six-dimensional space (three translations + three rotations). The "passive amino acid" option was skipped.

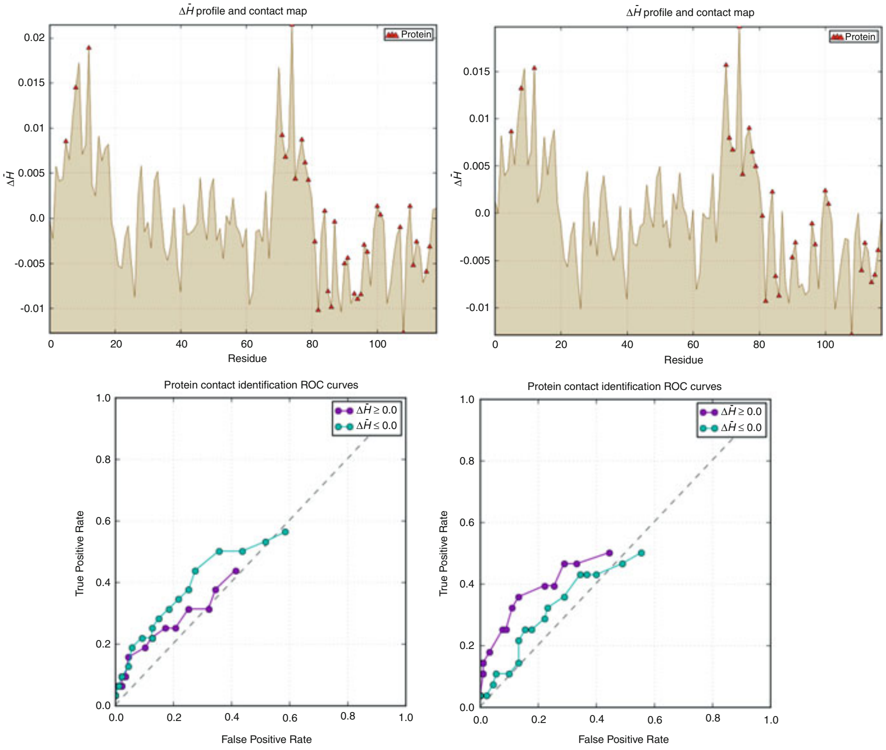


Fig. 6.7 $\Delta\tilde{H}$ profiles for both chains of the 1SD6 homodiner, indicating residues which represent excess hydrophobicity as well as local hydrophobicity deficiencies. In this case areas of excess hydrophobicity appear to be paired with areas of low hydrophobicity of the surface of the partner molecule. This suspicion is further confirmed by analysis of the 3D representation of 1SD6

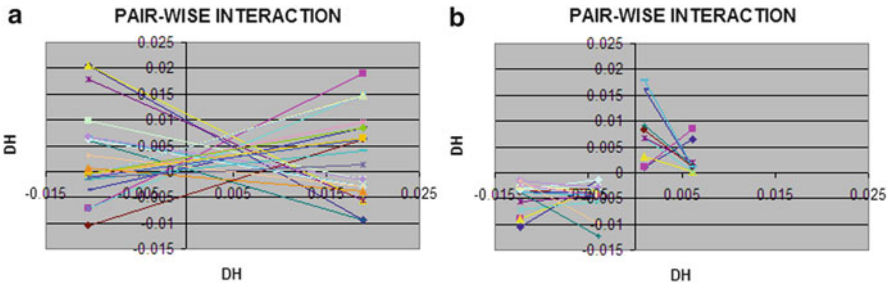


Fig. 6.8 Pairwise interactions in the 1SD6 homodimer. (**a**) Pairs of residues with complementary (*mirrored*) deviations from the idealized profile (positive and negative values of $\Delta\tilde{H}$); (b) Pairs where both participating residues have identically signed $\Delta\tilde{H}$ values. Note that interaction between residues with negative $\Delta\tilde{H}$ values indicates shared hydrophobicity excess, while pairing of residues with positive $\Delta\tilde{H}$ values corresponds to shared hydrophobicity deficiencies. Scale is preserved between figures to enable quantitative comparisons

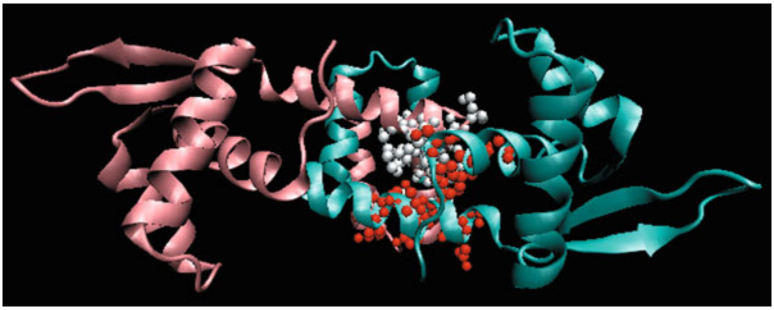


Fig. 6.9 The 1SD6 homodimer. Local $\Delta\tilde{H}$ profile maxima (*red*) are indicated for the A chain (*turquoise*), while local minima (*white*) are highlighted in the B chain (*salmon*). In this case, interacting residues mutually compensate their respective deviations from the idealized hydrophobicity profile

For each protein dimer HADDOCK generated ten possible conformations. From these, the top three structures were further processed to determine which residues form the complexation site (this step was based on custom scripts). Figure 6.10 illustrates the BEST and LOWEST solutions produced by HADDOCK, while Fig. 6.11 ranks the BEST and LOWEST structures in terms of their F-measure and MCC scores. Of note is the high similarity of both rankings, differing only by a single item.

6.3.3 RosettaDock

The starting structure for RosettaDock was generated using Gramm-X (Protein-Protein Docking Web Server v.1.2.0, Center for Bioinformatics of the University of Kansas, http://vakser.bioinformatics.ku.edu/resources/gramm/grammx). This software package applies FFT in its search for optimal structures. Input data consists of the PDB protein structure and the numerical positions of amino acids which belong to its complexation site (according to PDBsum). Gramm-X produces a ranked list of structures, the topmost of which was selected as input for RosettaDock.

RosettaDock is based on a local docking algorithm which seeks out all conformations in the vicinity of a user-specified starting point. All structural translations are performed with a step of approximately ±3 Å in the longitudinal plane and ±8 Å in the lateral plane. Rotations apply a step of 8° and cover the entire 360° angle around the geometric centers of participating structures. A total of 1,000 independent simulations were performed for random conformations (i.e. random steric adjustments). Results were ranked according to their total free energy and the highest ranked structure was selected for further processing. Its complexation site was thoroughly mapped (by applying the PDBsum distance criterion) using custom scripts. Complexation site mapping relied on the distance between individual monomers, with cutoff distances derived from PDBsum.

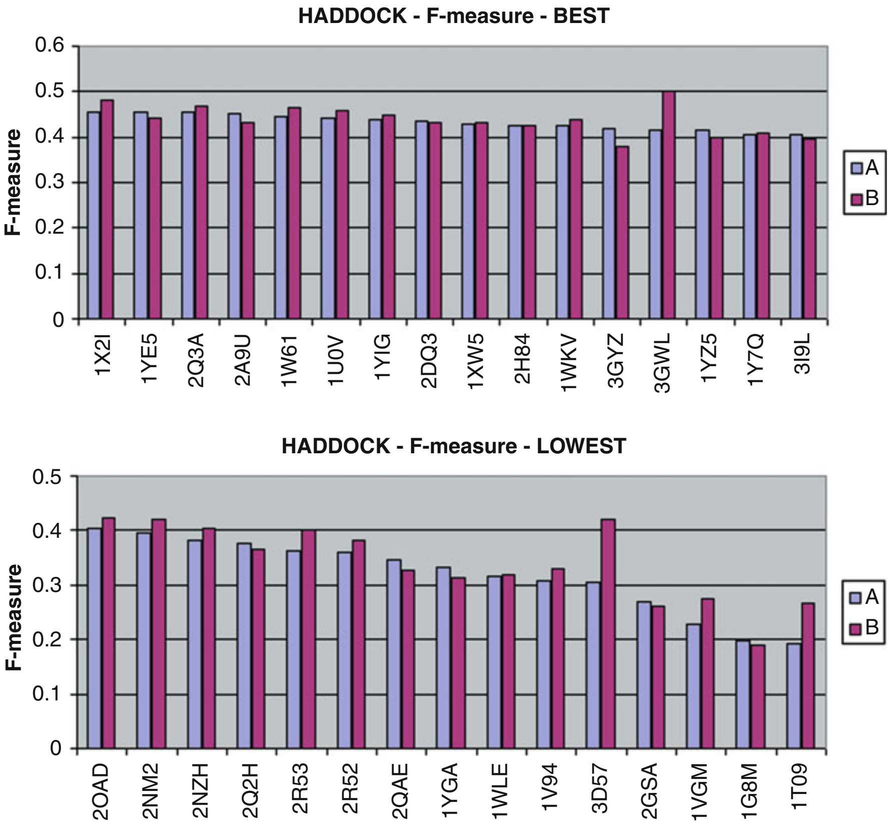


Fig. 6.10 Best (*top*) and worst (*bottom*) solutions produced by HADDOCK and ranked according to the F-measure criterion. A, B – chain identification

It is worth noting that over 60% of complexes produced by ROSETTA-Dock achieved a score higher than 0.8 on the MCC scale, while 71% obtained a score higher than 0.4 on the F-measure scale (Figs. 6.12 and 6.13).

6.3.4 ZDOCK

The ZDOCK service deployed at http://zdock.umassmed.edu/ can be used to perform computations. Input consists of structural descriptions of two protein molecules. The user should also specify which molecule is to act as a receptor (the remaining molecule is treated as the ligand and subjected to dynamic docking). From a computational standpoint, it is advisable to select the smaller molecule as the ligand. Within the ligand, certain residues can be excluded from the binding site (this usually involves a handful of residues that are on the opposite side of where the

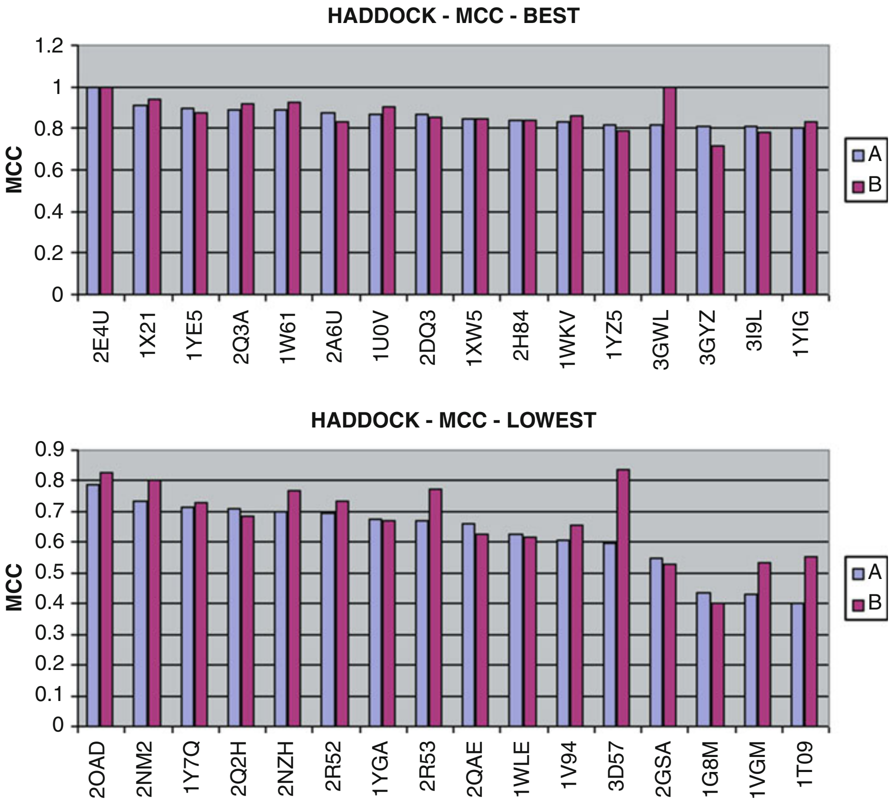


Fig. 6.11 Best (*top*) and worst (*bottom*) solutions produced by HADDOCK and ranked according to the MCC criterion. 16 solutions are depicted for each case. *Colored bars* indicate differences between individual protein chains (designated *A* and *B*)

interface is expected to be found). For a single pair of proteins with properly established sets of candidate residues calculations take approximately 5 min. The program is capable of generating up to 2,000 different complexes, although the list may be restricted by the user – in our study we requested a list of 10 structures, ranked according to their fitness scores (Fig. 6.14).

6.4 Comparative Analysis

The goal of comparative analysis is to identify complexes which specifically demonstrate the properties of a given computational tool and its underlying theoretical model. In order to highlight differences between models we have focused on the best and worst structures produced by each application (according to F-measure and MCC criteria). Table 6.2 presents a list of complexes which ranked among the

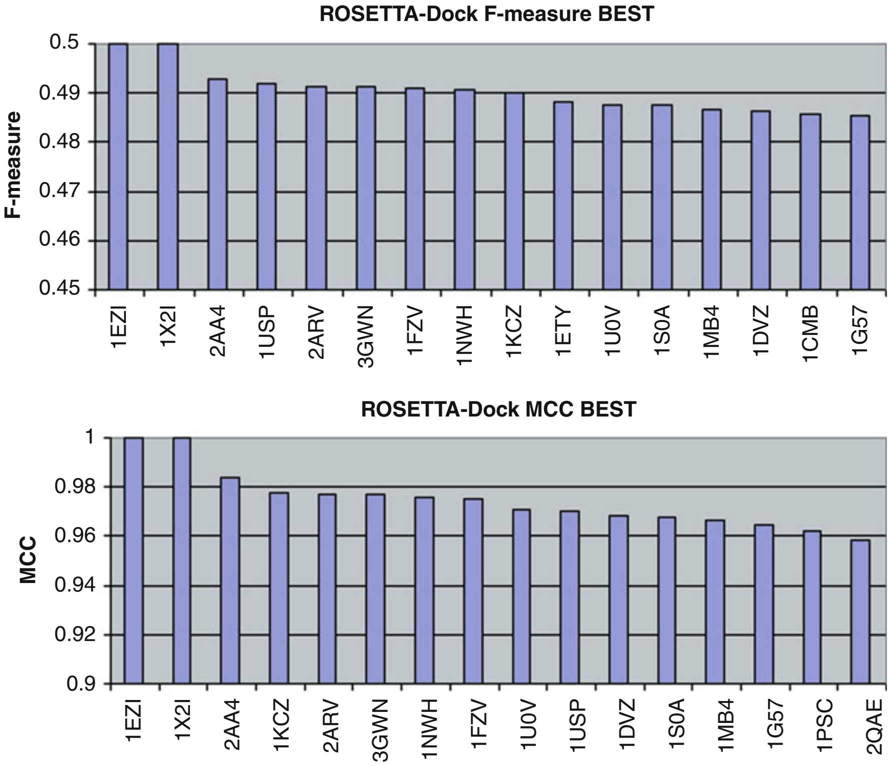


Fig. 6.12 Best solutions produced by RosettaDock, ranked according to F-measure (*top*) and MCC (*bottom*) criteria

top 10 or bottom 10 solutions (usually scoring 0 on either the F-measure or the MCC scale in the latter case). A more detailed analysis of each complex is required to determine the reasons behind these extreme values.

The summary presented in Table 6.2 enables us to study the specific features of each complexation model. Due to the fact that the hydrophobic core model is markedly different from all other software packages, we will focus our analysis on the best and worst results produced by this model. This decision is also conditioned by the procedures employed by other toolkits: unlike HADDOCK, RosettaDock or ZDOCK, the "fuzzy oil drop" model does not generate a large number of candidate structures, nor does it involve clustering. Instead, it produces a single $\Delta\tilde{H}$ profile which it then uses to pinpoint a specific complexation site. In the case of the "fuzzy oil drop" model, it is also possible to determine the actual causes of successes and failures in predicting complexation sites. For these reasons our comparative analysis will focus on the proteins listed in the first row and the first column of Table 6.2.

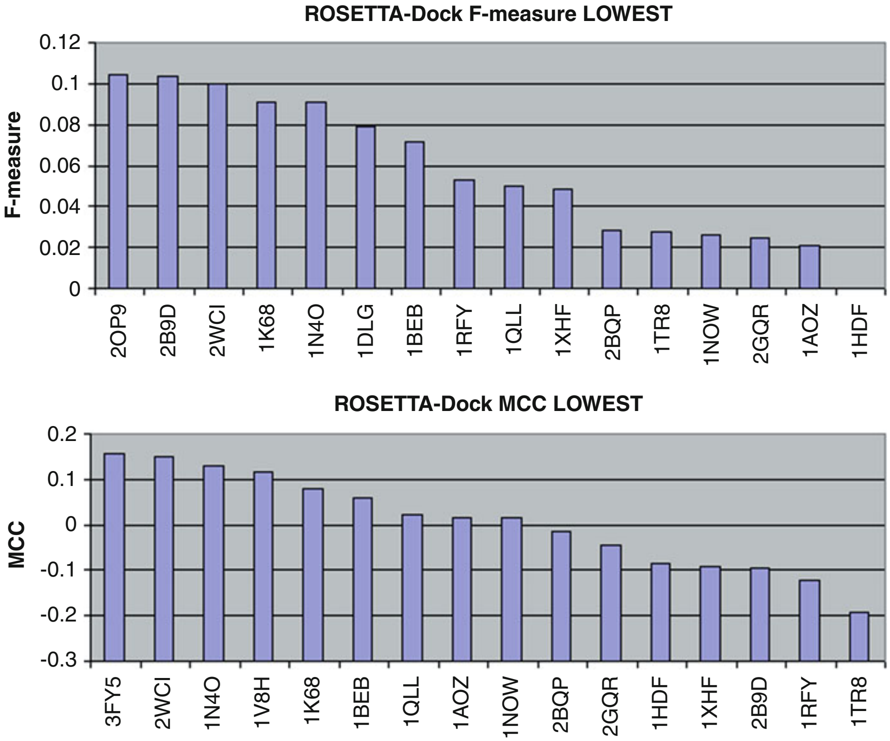


Fig. 6.13 Worst solutions produced by RosettaDock, ranked according to F-measure (*top*) and MCC (*bottom*) criteria. Only chains A are shown

The "fuzzy oil drop" model identifies complexation sites as specific deformations in the protein's hydrophobic core associated with the presence of residues whose actual hydrophobicity values diverge from theoretical predictions. When the core is perturbed by more than one external molecule (for instance by a protein and a ligand), it becomes difficult to distinguish one distortion from the other. Thus, accurate prediction of ligand binding and protein complexation sites depends on measuring the relative significance of each factor.

For the sample protein designated 1G8M (transferase, hydrolase – crystal structure of avian atic, a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis – EC 2.1.2.3, EC 3.5.4.10) (Greasley et al 2001) the "fuzzy oil drop" model was able to correctly identify the complexation site (by locating residues which represent local maxima of the $\Delta\tilde{H}$ profile). However, this protein is also capable of binding a ligand (specifically, $C_{10}H_{14}N_5O_8P$ – Guanosine-5′-monophosphate). Identifying this ligand's binding pocket would likely prove difficult as the deformation

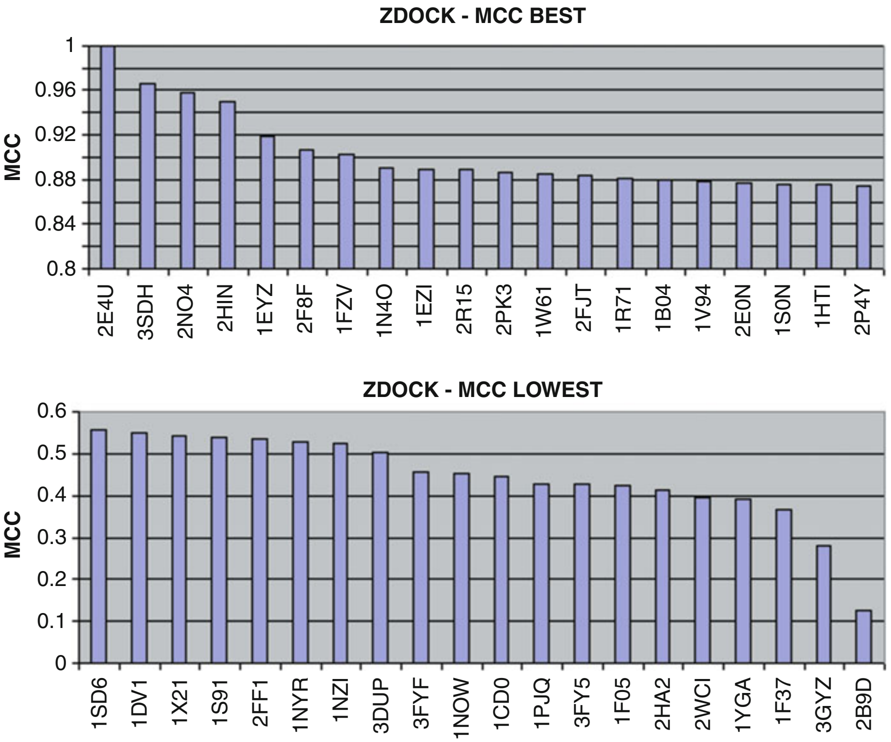


Fig. 6.14 Best (*top*) and worst (*bottom*) results produced by ZDOCK and ranked according to the MCC criterion. Only chain A is shown

Table 6.2 Comparison of correct and incorrect solutions, indicating the validity of various models

Lowest → Best ↓	FOD	HADDOCK	RosettaDock	ZDOCK
FOD		1SD6 1 T09 2QM8 3FYF	1BEB 1 T09 1TR8	1G8M 1 T09 1YGA 2R52
HADDOCK	3SDH / LL 2QX0 / LL		1QLL 2BQP 2GQR	1 V94 2E4U 3D57
RosettaDock	1DVZ / LL			
ZDOCK	1X2I 2FJT 2Q3A 3GWL / LL 3GYZ / LL			

"LL" stands for "*large ligand*", which, in the case of the "fuzzy oil drop" model, plays a key role in shaping the molecule's hydrophobic core, outstripping the influence of complexation events

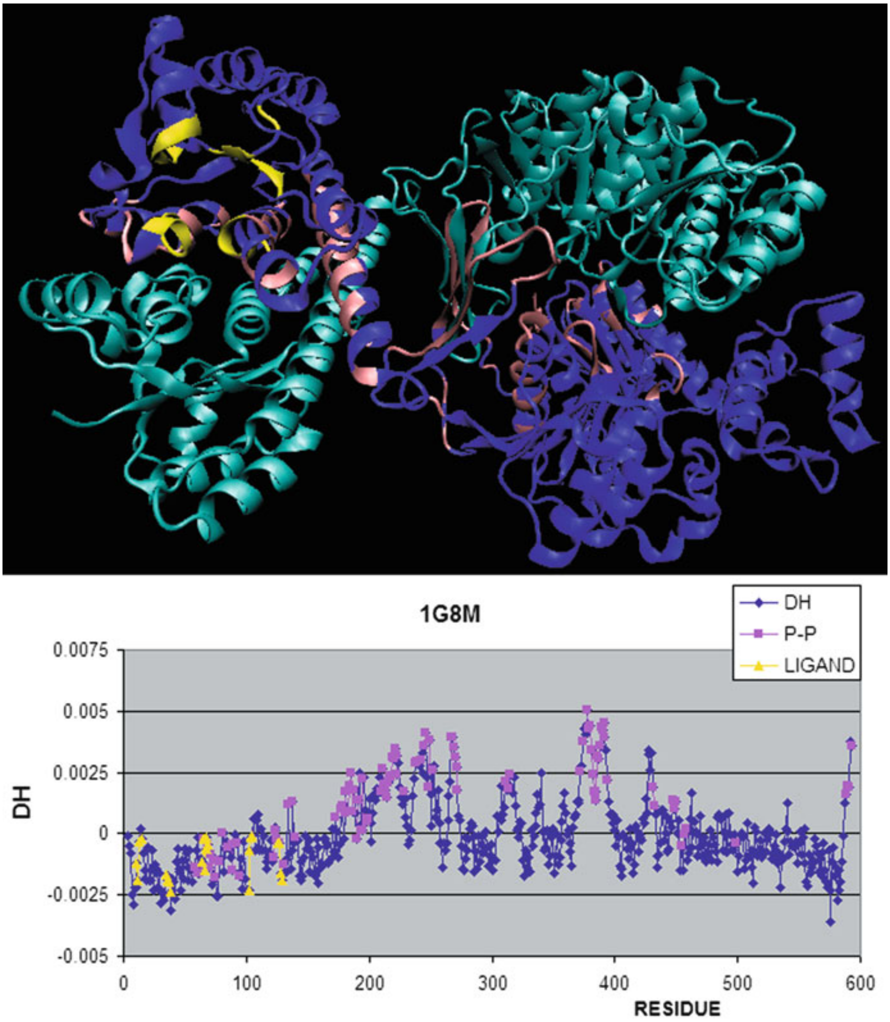


Fig. 6.15 The 1G8M homodimer. *Top*: 3D representation, showing complexation (*dark blue*) and ligand binding (*yellow*) areas. *Bottom*: $\Delta\tilde{H}$ profile with the corresponding residues highlighted (same colors as above). Protein complexation is largely mediated by residues with high $\Delta\tilde{H}$ values, while the binding pocket consists of residues with average $\Delta\tilde{H}$ values, suggesting that this pocket may be difficult to identify by using the "fuzzy oil drop" model

triggered by the ligand is far less pronounced than the one caused by protein complexation. It should be noted that other tools described in this chapter ran into serious problems when trying to model the 1G8M complex, most likely due to the relatively large surface area of its complexation interface (Fig. 6.15).

The set of homodimers for which the "fuzzy oil drop" model generates incorrect predictions includes the 2Q3A (immune system) protein (Bertini et al. 2004). This protein does not bind any additional ligands and moreover, its hydrophobicity profile is highly consistent with idealized values (O/T = 0.137; O/R = 0.173 – see Fig. 6.16 for details), suggesting that complexation does not significantly distort the structure of each monomer's hydrophobic core. We call this phenomenon *static complexation*: since the hydrophobic core is not affected, the "fuzzy oil drop" model cannot make accurate predictions regarding the complexation interface. Accordingly, cases where the presence of an external protein molecule triggers significant deformations in the protein's core (such as in 1G8M) are referred to as *dynamic complexation*. This distinction provides a strong indication that complexation mechanisms may take on many different forms. Active complexation appears to be an example of chaperone-like activity (where the complementary molecule acts as the chaperone); however in 2Q3A individual monomers evolve separately and develop a stable tertiary structure which includes a well-ordered hydrophobic core.

The 1X21 protein, which is a fragment of a helix-hairpin-helix DNA binding domain in a larger hydrolase molecule, does not constitute a functional group on its own (Nishino et al. 2005). Nevertheless, its structure exhibits a certain ordering of hydrophobicity, approximating the idealized "fuzzy oil drop" model. For this protein, O/T = 0.143 while O/R = 0.177, indicating a situation similar to 2Q3A (where the dimer emerges through static aggregation of two molecules without distorting their respective cores). For this reason, analysis of the $\Delta\tilde{H}$ profile (i.e. its minima and maxima) is not sufficient to identify the residues involved in complexation.

The 1DVZ protein (hormone/growth factor – human transthyretin in complex with o-trifluoromethylphenyl antranilic acid) binds a potential drug which (according to theoretical predictions) should help prevent the buildup of amyloidogenic plaque (Klabunde et al. 2000). Analysis of its $\Delta\tilde{H}$ profile suggests that the presence of a ligand does not result in substantial structural changes in the transthyretin molecule; however section 69–80 diverges from the theoretical optimum (in terms of hydrophobicity) and introduces an element of instability which may, in turn, destabilize the entire molecule. Section 69–80 appears to be connected with the ability of transthyretin to attach additional molecules, facilitated by rapid structural changes. While somewhat speculative, this conclusion is justified: studies suggests that "divergent" sections frequently participate in complexation processes (Fig. 6.17).

3SDH (cooperative dimeric hemoglobin from the blood clam *Scapharca inaequivalvis*) is another example of a protein where the "fuzzy oil drop" does not provide sufficient data to pinpoint complexation sites (Royer 1994). This protein has been studied both in its unliganded (deoxy) and carbon monoxide (CO) liganded states.

In the case of 3SDH, the presence of a large ligand (heme) dominates the activity of hemoglobin. Lys 96 and Phe 97 participate in two important processes: binding heme and facilitating intersubunit communication. The "fuzzy oil drop" model correctly singles out these residues as belonging to a local hydrophobicity maximum. Other residues which form the complexation interface are somewhat less pronounced on the $\Delta\tilde{H}$ graph – which is why the graph itself is not sufficient to accurately model the complexation site. Figure 6.18 (top) also highlights Ile 114, Trp 135 and Leu 138 (white spheres), which – despite being strongly hydrophobic – generate a localized hydrophobicity deficiency, resulting from the relatively loose packing of residues in their region and suggesting a potential interaction site.

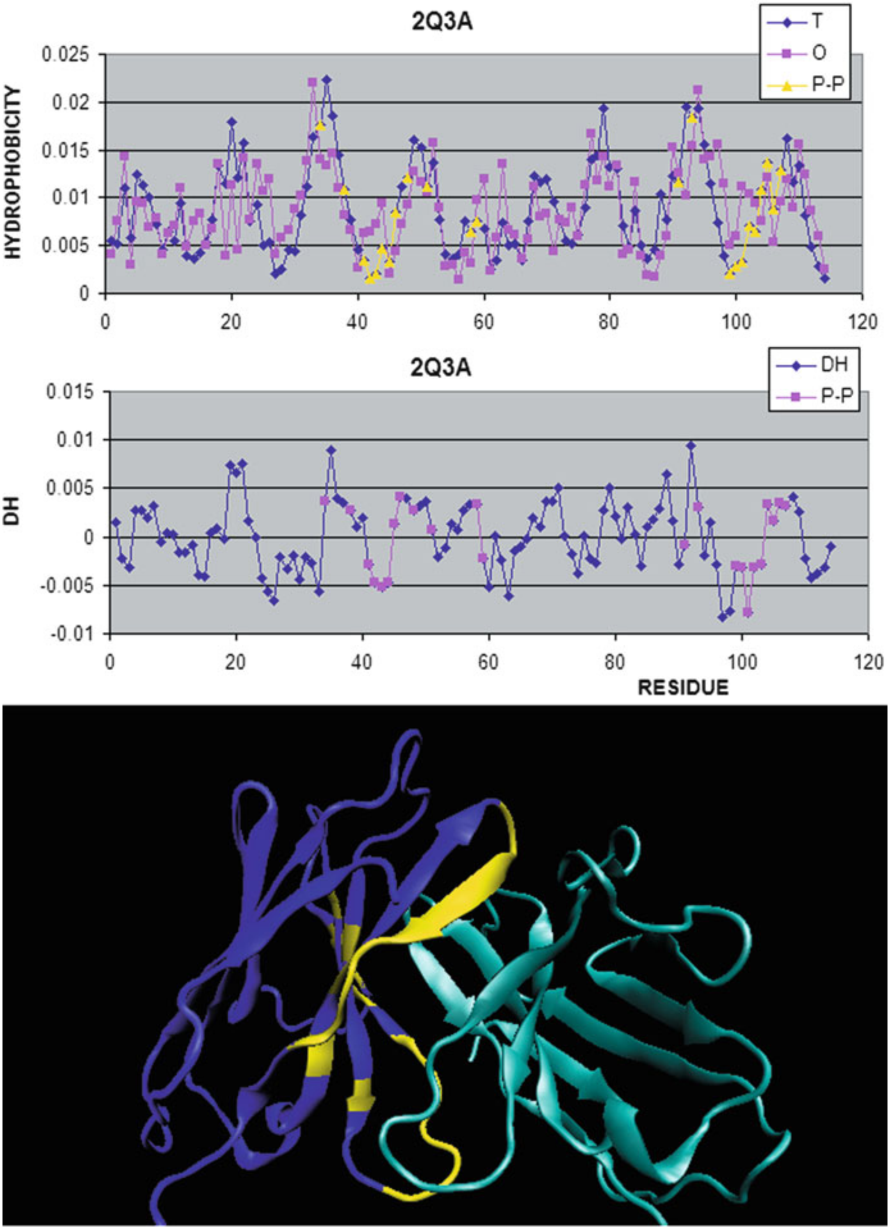


Fig. 6.16 The 2Q3A homodimer and its hydrophobicity profile. *Top*: theoretical (T) and observed (O) values, with residues engaged in P-P interaction marked in *yellow* (residues belonging to the complementary monomer are marked in *pink*). *Bottom*: 3D representation of the resulting homodimer, with the complexation interface marked in *yellow*

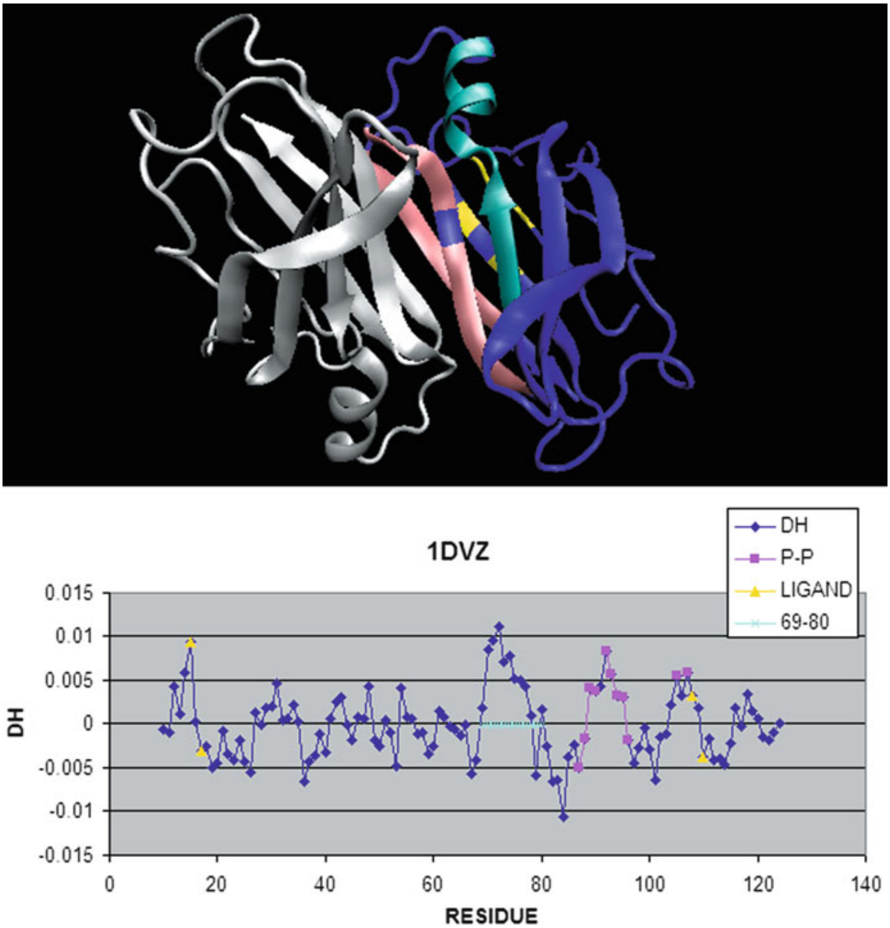


Fig. 6.17 The 1DVZ protein: 3D representation (*top*) with ligand binding residues marked in *yellow* and complexation interface marked in *pink*. The cyan-colored fragment (69–80) corresponds to the greatest divergence between the observed (O) and theoretical (T) distribution of hydrophobicity in the $\Delta\tilde{H}$ profile (*bottom*)

The summary of structural analysis of the proteins discussed in Table 6.2 are given in Table 6.3.

No particular correlation can be found for characteristics of protein-protein complexes as shown in Table 6.3. The HADDOCK program was the best for complexation interface characterized by the lowest number of hydrophobic residues engaged in complexation.

The "fuzzy oil drop" model identifies complexation sites as specific deformations in the protein's hydrophobic core associated with the presence of residues whose actual hydrophobicity values diverge from theoretical predictions. When the core is perturbed by more than one external molecule (for instance by a protein and

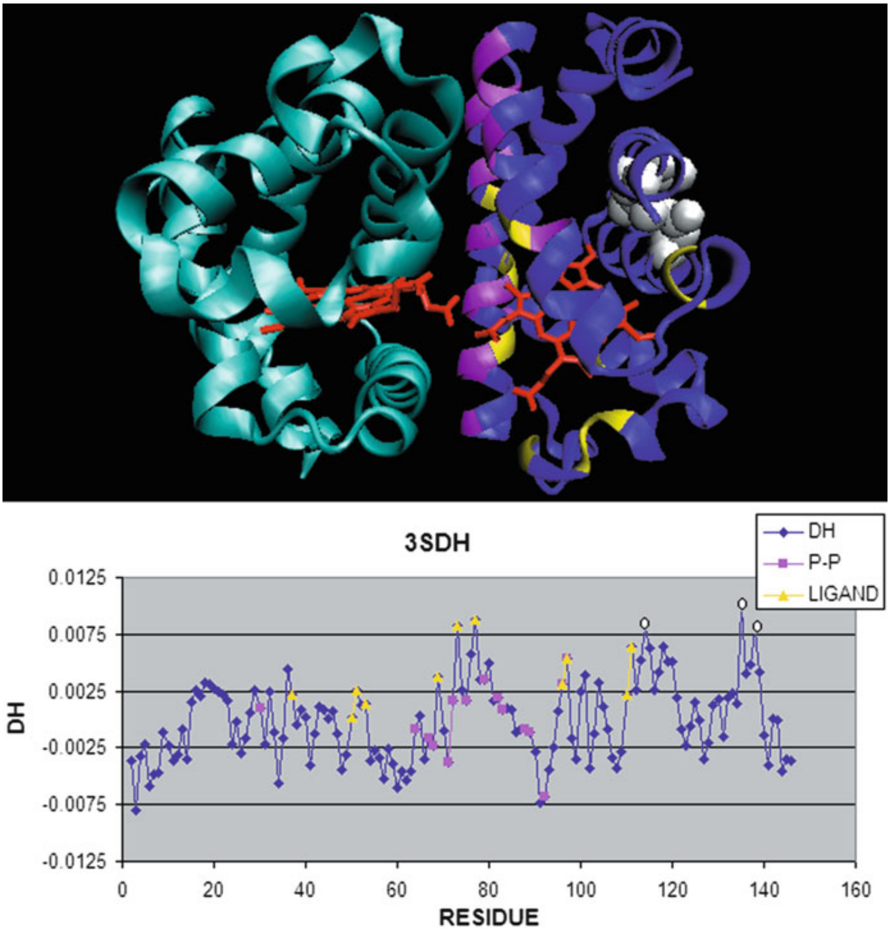


Fig. 6.18 The 3SDH protein: 3D representation (*top*) with ligand binding residues marked in *yellow* and the complexation interface marked in *magenta*. Ligands are given in *red*. Local $\Delta\tilde{H}$ profile maxima are marked in *yellow*. All colors correspond to the *bottom* graph, which shows the actual $\Delta\tilde{H}$ profile for 3SDH

a ligand), it becomes difficult to distinguish one distortion from the other. Thus, accurate prediction of ligand binding and protein complexation sites depends on measuring the relative significance of each factor.

For the sample protein designated 1G8M (transferase, hydrolase – crystal structure of avian atic, a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis – EC 2.1.2.3, EC 3.5.4.10) (Greasley et al. 2001) the "fuzzy oil drop" model was able to correctly identify the complexation site (by locating residues which represent local maxima of the $\Delta\tilde{H}$ profile). However, this protein is also capable of binding a ligand (specifically, $C_{10}H_{14}N_5O_8P$ – Guanosine-5′-monophosphate). Identifying this ligand's binding pocket would likely prove difficult as the deformation triggered by

Table 6.3 Comparison of correct and incorrect solutions, indicating the validity of various models

Lowest →				
Best ↓	FOD	HADDOCK	RosettaDock	ZDOCK
FOD		101.30 7.88 5.26 16.53	71.63 7.70 3.48 11.16	71.25 5.40 5.14 12.88
HADDOCK	67.75 4.90 2.29 12.40		46.63 5.47 2.26 8.28	39.34 2.59 2.39 9.00
RosettaDock	103.51 15.79 5.26 15.79			
ZDOCK	109.49 5.64 8.70 19.38			

The values given in each cell of the table are as follows: averaged number of non-bonding contacts per residue, averaged number of H-bonds contacts per residue, averaged number of hydrophobic residues engaged in protein-protein interaction per residue. The values are calculated for proteins as shown in Table 6.2. Values calculated according to the data available in PDBSum database

the ligand is far less pronounced than the one caused by protein complexation. It should be noted that other tools described in this chapter ran into serious problems when trying to model the 1G8M complex, most likely due to the relatively large surface area of its complexation interface.

6.5 Summary

The programs discussed in this chapter have been selected to showcase various means of identifying protein-protein complexation sites. Such analysis cannot be called "blind prediction" since it relies on user-picked starting structures and potential zones of interest. The aim of our study was to assess the validity and accuracy of each algorithm for a large set of sample proteins. A secondary goal was to divide protein complexes into subgroups: since some proteins form rather peculiar complexes, the "lock and key" abstraction is not always applicable.

Even when potential complexation interfaces are suggested by the user, the presented tools do not always produce correct results. This is most likely due to deficiencies in their conformation space search algorithms.

The "fuzzy oil drop" package attempts to link the final state of the complex with the mechanisms which govern the folding process. We assume that an undisturbed molecule folding in an aqueous environment generates a regular hydrophobic core, as is evident in fast-folding (Roterman et al. 2011a, b) and antifreeze proteins (Banach et al. 2012). The presence of an external ligand (Bryliński et al. 2007a, b), a partner molecule or a membrane (Zobnina and Roterman 2009) deforms the core of the emerging protein in a way which ensures its specificity. Note that even though fully folded proteins may encounter a wide variety of potential ligands in their environment, they are usually highly specific with regard to the molecules they bind with. Some researchers even postulate that the presence of a ligand is an essential factor in the polypeptide chain folding process (Brylinski et al. 2006, 2007a). As highlighted in our analysis, the "fuzzy oil drop" model is capable of acknowledging such factors and explains how proteins are conditioned to perform their intended biological role. This is especially important in enzymes, where localized deformations of the hydrophobic core seem to correspond to active sites of hydrolases (Prymula et al. 2011).

Genomics-scale analysis of protein complexes suggests that, when it comes to determining the biological profiles of proteins, complexation is frequently as important as interaction with ligands. A noteworthy presentation of current progress in studying protein complexation mechanisms can be found in Fleishman et al. (2011)

Acknowledgements The presented research was carried out in 1995–2011 and funded by a series of grants from the Jagiellonian University Medical College. The Academic Computing Center CYFRONET AGH Krakow provided computational support. We would like to express our gratitude to Piotr Nowakowski of CYFRONET AGH for valuable editorial remarks. Technical support provided by Anna Zaremba-Śmietańska is also gratefully acknowledged.

References

Banach M, Prymula K, Jurkowski W, Konieczny L, Roterman I (2012) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J Mol Model 18(1):229–237

Bass J, Takahashi JS (2010) Circadian integration of metabolic and energetics. Science 330:1349–1354

Bertini I, Calderone V, Fragai M, Luchinat C, Mangani S, Terni B (2004) Crystal structure of the catalytic domain of human matrix metalloproteinase 10. J Mol Biol 336:707–716

Boyer PD (1997) The ATP synthase–a splendid molecular machine. Annu Rev Biochem 66:717–749

Brooks BR, Brooks CL III, Mackerell AD Jr, Nilsson R, 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

Brylinski M, Konieczny L, Roterman I (2006) Hydrophobic collapse in (in silico) protein folding. Comput Biol Chem 30(4):255–267

Bryliński M, Konieczny L, Roterman I (2007a) Is the protein folding an aim-oriented process? Human haemoglobin as example. Int J Bioinform Res Appl 3(2):234–260

Bryliński M, Prymula K, Jurkowski W, Kochańczyk M, Stawowczyk E, Konieczny L, Roterman I (2007b) Prediction of functional sites based on the fuzzy oil drop model. PLoS Comput Biol 3(5):e94

Chen R, Weng Z (2002) Docking unbound proteins using shape complementarity, desolvation, and electrostatics. Proteins 47:281–294

Chen R, Weng Z (2003) A novel shape complementarity scoring function for protein-protein docking. Proteins 51:397–408

Chen R, Li L, Weng Z (2003a) ZDOCK: an initial-stage protein-docking algorithm. Proteins 52:80–87

Chen R, Mintseris J, Janin J, Weng Z (2003b) A protein-protein docking benchmark. Proteins 52:88–91

Chen R, Tong W, Mintseris J, Li L, Weng Z (2003c) ZDOCK predictions for the CAPRI challenge. Proteins 52:68–73

Cosconati S, Forli S, Perryman AL, Harris R, Goodsell DS, Olson AJ (2010) Virtual screening with AutoDock: theory and practice. Expert Opin Drug Discov 5(6):597–607

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

Dominguez C, Boelens R, Bonvin AMJJ (2003) HADDOCK: a protein-protein docking approach based on biochemical and/or biophysical information. J Am Chem Soc 125:1731–1737

Dunbrack RL Jr, Cohen FE (1997) Bayesian statistical analysis of protein side-chain rotamer preferences. Protein Sci 6:1661–1681

Duong HA, Robles MS, Knutti D, Weitz CJ (2011) A molecular mechanism for circadian clock negative feedback. Science 332:1436–1439

Fawcett T (2006) An introduction to ROC analysis. Pattern Recognit Lett 27:861–874

Feldman HJ, Labute P (2010) A novel method for measuring protein pocket similarity was devised, using only the α carbon positions of the pocket residues. J Chem Inf Model 50(8):1466–1475

Fleishman SJ, Whitehead TA, Strauch EM, Corn JE, Qin S, Zhou HX, Mitchell JC, Demerdash ON, Takeda-Shitaka M, Terashi G, Moal IH, Li X, Bates PA, Zacharias M, Park H, Ko JS, Lee H, Seok C, Bourquard T, Bernauer J, Poupon A, Azé J, Soner S, Ovali SK, 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 AM, 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 SY, 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(2):289–302

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

Greasley SE, Horton P, Ramcharan J, Beardsley GP, Benkovic SJ, Wilson IA (2001) Crystal structure of a bifunctional transformylase and cyclohydrolase enzyme in purine biosynthesis. Nat Struct Biol 8:402–406

Huxford T, Huang DB, Malek S, Ghosh G (1998) The crystal structure of the IkappaBalpha/NF-kappaB complex reveals mechanisms of NF-kappaB inactivation. Cell 95:759–770

Janin J (2007) The targets of CAPRI rounds 6–12. Proteins 69(4):699–703

Janin J (2010a) The targets of CAPRI rounds 13–19. Proteins 78(15):3067–3072

Janin J (2010b) Protein-protein docking tested in blind predictions: the CAPRI experiment. Mol Biosyst 6(12):2351–2362

Janin J, Wodak S (2007) The third CAPRI assessment meeting Toronto, Canada, April 20–21, 2007. Structure 15(7):755–759

Janin J, Henrick K, Mount J, Eyck LT, Sternberg MJE, Vajda S, Vakser I, Wodak SJ (2003) CAPRI: A Critical Assessment of PRedicted Interactions. Proteins 52:2–9

Jasti J, Furukawa H, Gonzales EB, Gouaux E (2007) Structure of acid-sensing ion channel 1 at 1.9-Å resolution and low pH. Nature 449:316–323

Johnson CH, Egli M, Stewart PL (2008) Structure insight into a circadian oscillator. Science 322:697–701

Kastritis PL, Moal IH, Hwang H, Weng Z, Bates PA, Bonvin AM, Janin J (2011) A structure-based benchmark for protein-protein binding affinity. Protein Sci 20(3):482–491

Kauzmann W (1959) Some factors in the interpretation of protein denaturation. Adv Protein Chem 14:1–63

Klabunde T, Petrassi HM, Oza VB, Raman P, Kelly JW, Sacchettini JC (2000) Rational design of potent human transthyretin amyloid disease inhibitors. Nat Struct Biol 7:312–321

Konieczny L, Bryliński M, Roterman I (2006) Gauss-function-based model of hydrophobicity density in proteins. In Silico Biol 6(1–2):5–22

Kortemme T, Morozov AV, Baker D (2003) An orientationdependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. J Mol Biol 326:1239–1259

Kouwen TR, Andréll J, Schrijver R, Dubois JY, Maher MJ, Iwata S, Carpenter EP, van Dijl JM (2008) Thioredoxin a active-site mutants form mixed disulfide dimers that resemble enzyme-substrate reaction intermediates. J Mol Biol 379:520–534

Kozakov D, Hall DR, Beglov D, Brenke R, Comeau SR, Shen Y, Li K, Zheng J, Vakili P, Paschalidis IC, Vajda S (2010) Achieving reliability and high accuracy in automated protein docking: cluspro, PIPER, SDU, and stability analysis in CAPRI rounds 13–19. Proteins 78:3124–3130

Kramer B, Rarey M, Lengauer T (1999) Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. Proteins 37(2):228–241

Lazaridis T, Karplus M (2000) Effective energy functions for protein structure prediction. Curr Opin Struct Biol 10:139–145

Lensink MF, Mendez R, Wodak SJ (2007) Docking and scoring protein complexes: CAPRI 3rd edition. Proteins 69:704–718

Levitt M (1976) A simplifed representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104:59–107

Li L, Chen R (joint first authors), Weng Z (2003) RDOCK: refinement of rigid-body protein docking predictions. Proteins 53, 693–707

Lyskov S, Gray JJ (2008) The RosettaDock server for local protein–protein docking. Nucleic Acids Res 36(Web Server issue):W233–W238

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

MacKerell ADJ, Bashford D, Bellot M, Dunbrack RLJ, Evenseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WEI, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M (1998) All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 102:3586–3616

Mintseris J, Weng Z (2003) Atomic contact vectors in protein-protein recognition. Proteins 53:629–639

Mintseris J, Wiehe K, Pierce B, Anderson R, Chen R, Janin J, Weng Z (2005) Protein-protein docking benchmark 2.0: an update. Proteins 60(2):214–216

Mintseris J, Pierce B, Wiehe K, Anderson R, Chen R, Weng Z (2007) Integrating statistical pair potentials into protein complex prediction. Proteins 69(3):511–520

Needleman S, Wunsch C (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Nishino T, Nishino T, Komori K, Ishino Y, Morikawa K (2005) Structural and functional analyses of an archaeal XPF/Rad1/Mus81 nuclease: asymmetric DNA binding and cleavage mechanisms. Structure (Camb) 13:1183–1192

Oster G, Wang H (1999) ATP synthase: two motors, two fuels. Structure 7:R67–R72

Oster G, Wang H (2003) Rotary protein motors. Trends Cell Biol 13:114–121

Pierce B, Weng Z (2007) ZRANK: reranking protein docking predictions with an optimized energy function. Proteins 67(4):1078–1086

Pierce B, Weng Z (2008) A combination of rescoring and refinement significantly improves protein docking performance. Proteins 72(1):270–279

Pierce B, Tong W, Weng Z (2005) M-ZDOCK: a grid-based approach for C_n symmetric multimer docking. Bioinformatics 21(8):1472–1476

Pierce B, Phillips AT, Weng Z (2007) Structure prediction of protein complexes. In: Xu Y, Xu D, Liang J (eds) Computational methods for protein structure prediction and modeling volume 2: structure prediction. Springer, New York, pp 109–134

Prymula K, Jadczyk T, Roterman I (2011) Catalytic residues in hydrolases: analysis of methods designed for ligand-binding site prediction. J Comput Aided Mol Des 25(2):117–133

Roterman I, Konieczny L, Jurkowski W, Prymula K, Banach M (2011) Two-intermediate model to characterize the structure of fast-folding proteins. J Theor Biol 283(1):60–70

Royer WE Jr (1994) High-resolution crystallographic analysis of a co-operative dimeric hemoglobin. J Mol Biol 235:657–681

Safo MK, Zhao Q, Musayev FN, Robinson H, Scarsdale N, Archer GL (2005) Crystal structures of the BlaI repressor from *Staphylococcus aureus* and its complex with DNA: insights into transcriptional regulation of the bla and mec operons. J Bacteriol 187:1833–1844

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

Spreter T et al (2005) The crystal structure of archaeal nascent polypeptide-associated complex (NAC) reveals a unique fold and the presence of a ubiquitin-associated domain. J Biol Chem 280:15849–15854

Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34:W310–W314

Unwin N (2005) Refined structure of the nicotinic acetylcholine receptor at 4 Å resolution. J Mol Biol 346:967–989

van Dijk ADJ, Bonvin AMJJ (2006) Solvated docking: introducing water into the modelling of biomolecular complexes. Bioinformatics 22:2340–2347. doi:10.1093/bioinformatics/btl395

van Dijk ADJ, Boelens F, Bonvin AMJJ (2005a) Data-driven docking for the study of biomolecular complexes. FEBS J 272:293–312. doi:10.1111/j.1742-4658.2004.04473.x

van Dijk ADJ, Fushman D, Bonvin AMJJ (2005b) 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, de Vries SJ, Dominguez C, Chen H, Zhou H-X, Bonvin AMJJ (2005c) Data-driven docking: HADDOCKs adventures in CAPRI. Proteins 60:232–238. doi:10.1002/prot.20563

van Dijk ADJ, Kaptein R, Boelens R, Bonvin AMJJ (2006a) 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 Dijk M, van Dijk ADJ, Hsu V, Boelens R, Bonvin AMJJ (2006b) Information-driven protein-DNA docking using HADDOCK: it is a matter of flexibility. Nucleic Acids Res 34:3317–3325. doi:10.1093/nar/gkl412

van Gelder RN, Herzog ED, Schwartz WJ, Taghert PH (2003) Circadian rhythms: in the loop at last. Science 300:1534–1535

Wiehe K, Pierce B, Mintseris J, Tong WW, Anderson R, Chen R, Weng Z (2005) ZDOCK and RDOCK performance in CAPRI rounds 3, 4, and 5. Proteins 60(2):207–213

Wiehe K, Pierce B, Tong WW, Hwang H, Mintseris J, Weng Z (2007) The performance of ZDOCK and ZRANK in rounds 6–11 of CAPRI. Proteins 69(4):719–725

Wiehe K, Peterson MW, Pierce B, Mintseris J, Weng Z (2008) Protein-protein docking: overview and performance analysis. Methods Mol Biol 413:283–314

Zobnina V, Roterman I (2009) Application of the fuzzy-oil-drop model to membrane protein simulation. Proteins 77(2):378–394

Zsoldos Z, Reid D, Simon A, Sadjad BS, Johnson AP (2006) eHiTs: an innovative approach to the docking and scoring function problems. Curr Protein Pept Sci 7(5):421–435

Chapter 7
Support for Cooperative Experiments in e-Science: From Scientific Workflows to Knowledge Sharing

Adam S.Z. Belloum, Reginald Cushing, Spiros Koulouzis, Vladimir Korkhov, Dmitry Vasunin, Victor Guevara-Masis, Zhiming Zhao, and Marian Bubak

Keywords e-Science • Workflow • Storage resources • Uniform interface • Geographically distributed computing resources • Knowledge sharing • Virtual organization • Material analysis of complex surface • Preprocessing phase • Experimentation process • Hyper-spectral data cube • Multivariate data analysis • Integrated application • Life-cycle • Searching workflows • Reproducibility of results • Interoperability between different workflows • Dissemination • Workflow support • Virtual data system • Shared repository • Process flow template • Executable operations • Pegasus • Proxy-service • Single nucleotide polymorphism • Individual therapy

7.1 Introduction

The term e-Science describes computational and data-intensive science. It has become a complementary experiment paradigm alongside the traditional in vivo and in vitro experiment paradigms. e-Science opens new doors for scientists and with it,

A.S.Z. Belloum (✉) • R. Cushing • S. Koulouzis • V. Guevara-Masis • Z. Zhao
The Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands
e-mail: a.s.z.belloum@uva.nl; r.cushing@uva.nl; s.koulouzis@uva.nl; z.zhao@uva.nl

V. Korkhov • D. Vasunin
Faculty of Applied Math and Control Processes, St. Petersburg
State University, Saint Petersburg, Russia
e-mail: vkorkhov@gmail.com; dvasunin@gmail.com

M. Bubak
AGH University of Science and Technology Krakow, Poland and the Informatics Institute,
University of Amsterdam, Amsterdam, The Netherlands
e-mail: bubak@agh.edu.pl; M.T.Bubak@uva.nl

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6_7,

it exposes a number of challenges such as how to organize huge datasets and coordinate distributed execution. For these challenges, a plethora of technologies and innovations have come together to enable e-Science (Foster and Kesselman 2006). Nowadays, complex scientific experiments designed following the e-Science paradigm are preformed using geographically distributed instruments, data and computing resources. The newly designed scientific experiments are costly, time-consuming, and multidisciplinary. Complex scientific experiments not only require access to geographically distributed hardware and software resources, but also extensive support to foster best practices, dissemination, and re-use.

Recently, Scientific Workflow Management Systems (SWMS) have become part of the science infrastructure in realizing e-Science, owing to their intuitive approach in prototyping experiments while concealing the complexity of the underlying middleware. SWMS are also instrumental in research collaborations since knowledge about experiments and data is easily shared through systems. This paradigm of designing, executing and sharing experiments enables scientists to focus on problem solving within their domain whilst intricate knowledge about underlying resources and workflow execution is hidden behind the SWMS. In essence, SWMS strive to bridge the knowledge gap between computational sciences and the myriad of distributed computing technologies. To date, many workflow systems have been developed and vary considerably in terms of workflow modeling, scheduling and targeted resources (Chin et al. 2002; McClatchey and Vossen 1997). The central component in a SWMS is the workflow. A workflow can be described as a connected graph which abstractly represents the flow of an experiment whereby vertices represent the activities and the edges represent dependencies between activities. The graph orchestrates the execution of such activities across the needed resources according to the application flow description.

New technologies such as grids and, recently, clouds allow the coordination and sharing of unprecedented quantities of geographically distributed computing and storage power by groups of trusted users within Virtual Organizations (Pang 2001). Such environments have made it possible to design and build global distributed collaborations involving large numbers of scientists and resources, and make data and computing-intensive scientific experiments feasible (Hey and Trefethen 2002). Within the e-Science community workflow management systems have been adopted as the main approach to designing and simulating complex systems (Chin et al. 2002; McClatchey and Vossen 1997). A Scientific Workflow Management System explicitly models the dependencies between scientific experiment processes.

This chapter describes a way to build a workflow management system for e-Science which provides support for the different phases of the lifecycle of a typical e-Science experiment. The presented results originate from the Virtual Laboratory for e-Science (VL-e) project,[1] which aims is to realize an e-Science framework where scientists from different domains can share their knowledge and resources, and perform domain-specific research. In this project complex

[1] www.vl-e.nl

applications from six scientific domains have been considered: food informatics, medical diagnosis and imaging, biodiversity research, bioinformatics, high energy physics, and telescience.

The goal shared by all the applications developed within the VL-e project is to take advantage of recent achievements in building large scale computing infrastructures and information systems. Regardless of the scientific domain, in terms of computing and information management similar requirements can be identified for applications such as developing models for predicting late-year bird migration volumes (for the purposes of ensuring airspace safety) (van Belle et al. 2007), visualizing high resolution correlated multi-spectral images (Broersen et al. 2007), or developing interactive visualization tools for fused functional magnetic resonance imaging (Blaas et al. 2007). However, it is more challenging to identify, given such a large collaboration, common characteristics in term of methods, techniques and tools and to abstract support for these features and requirements into a shared framework, which avoids redundancy in performing similar tasks across different e-Science domains by promoting exchange of resources.

The rest of the chapter is organized as follows: Section 7.2 presents a typical application use case which is used throughout the chapter to map the concepts introduced to a concrete example. Section 7.3 describes the different phases composing the lifecycle of a complex e-Science experiment. Section 7.4 presents a survey of the state of art in the field of workflow management systems. This survey focuses upon three main points: design, execution, and dissemination and sharing. Section 7.5 describes an approach to constructing an e-Science framework: it describes in detail the main components and tools developed to achieve this vision; specifically the Process Flow Template (PFT) to describe the logic of the experiment, ontology-based tools (OWT) to automate the generation of the PFT data structure, the workflow management system (WS-VLAM) to execute workflows on geographically distributed computing and storage resources, a bus-like architecture (Workflow Bus) to allow the design of a meta-workflow composed of an application created in multiple workflow management systems, and, finally, a service to optimize data sharing across workflows composed from web services.

7.2 Motivation – A Typical e-Science Application

Scientific experimentation often involve applications which are data-intensive, CPU- intensive. Moreover, some applications may require access to special devices. They usually have similar requirements concerning the use of computing resources and the ability to execute legacy or third-party applications, to perform parameter analysis (parameter sweep), or the automation of repetitive tasks (job farming). Typical e-Science applications have a set of common requirements:

- on-demand access to computing resources through a uniform interface,
- on-demand access to storage resources,

- the ability to execute software components written in a variety of programming languages on geographically distributed computing resources,
- access to knowledge and support for sharing.

Most of the requirements listed here concern the execution phase of an e-Science experiment. As science usually follows an iterative approach, the complexity of experiments grows over time. Early designs are aimed at developing simple prototypes to assess news ideas, and technologies. Once scientific application begin to mature, the design phase becomes complex and the need for access to existing knowledge and support for sharing is required. It then becomes important to provide support for the design phase by allowing users within the same domain (or even across multiple domains) to share expertise, reuse one another's tools etc. To facilitate knowledge transfer either within a single scientific domain or across domains, design and dissemination support is important. This is especially true as large projects operate as Virtual Organizations (VO) where knowledge transfer across VOs is restricted according to dynamic VO access policies. To facilitate the sharing of resources, a common framework in which all scientists can perform their experiments is needed.

Throughout this chapter we use the virtual material analysis laboratory as an example of a typical multi-physics scientific experiment. The Material Analysis of Complex Surface (MACS) experiments attempts to identify and determine the elements that comprise complex surfaces, regardless of the nature of the sample (Fig. 7.1). The approach followed in the MACS lab experiment is generic and can be easily applied to other application areas such as art conservation and restoration (e.g. analysis of binding media and organic pigments in old master paintings), biomedical science (e.g. identification of arteriosclerotic deposits in mice), and medical research (e.g. studies of trace elements in brain tissues) (Frenkel et al. 2001). Like most scientific applications, the MACS lab experiment, as can be seen in Fig. 7.1, consists of three phases: preprocessing, experimentation and analysis of results.

The preprocessing phase includes a number of procedures which have to be followed to extract the sample to be used in the experiment. To reach this goal scientists compare various techniques and protocols and select the most appropriate ones. Once the sample is produced, it has to be treated in order to fulfill the requirements of the device used in the material analysis process. It should be noted that, as the MACS lab is currently at its first design iteration, literature currently provides the main source of knowledge.

The experimentation process is performed with a set of specialized hardware devices. Two devices are used in these experiments: The Fourier Transformed Infra-Red imaging spectrometer and a 4 MeV Nuclear Microprobe. The FTIR is a non-dispersive infrared imaging spectrometer coupled to an infrared microscope used to examine the infrared radiation absorbed by complex surfaces. The 4 MeV Nuclear Microprobe has a spatial resolution in the sub-micrometer range and is capable of identifying trace elements on a high-sensitivity surface.

The analysis of results: the outcome of the experiment process is a set of data files containing the experiment results. This data set consists of a stack of images known

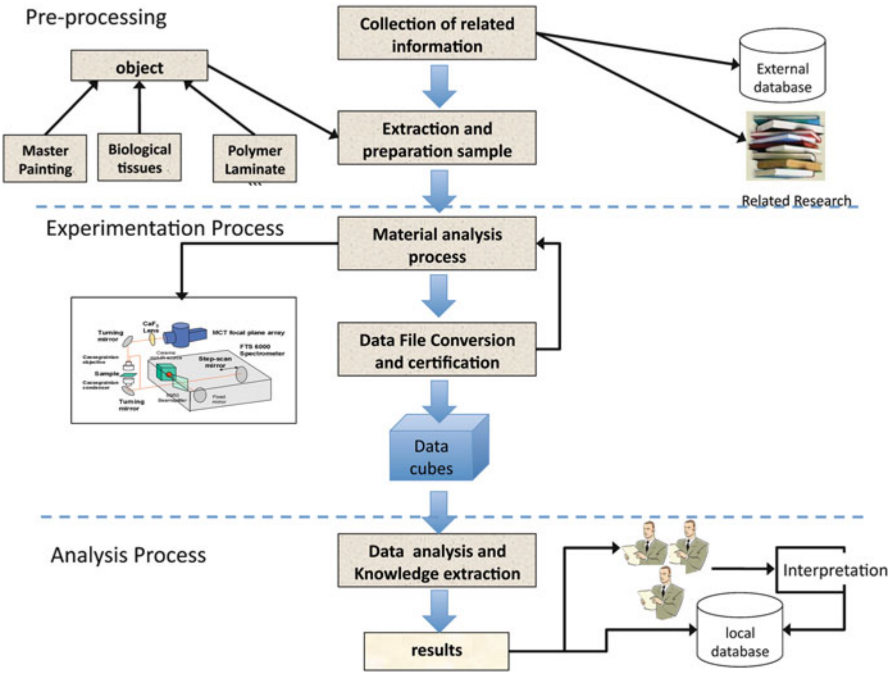


Fig. 7.1 The Material Analysis of Complex Surface (MACS) experiments aim at identifying the elements that compose complex surfaces such as paintings, biological tissues or polymer laminates. The MACS lab experiments can be decomposed into three phases: the pre-processing phase for the extraction of and preparation of samples, the experimentation phase where the sample is processed using special devices, and the analysis phase where the collected experimental data is analysed by scientists (Frenkel et al. 2001)

as a hyper-spectral data cube. A quality control process is carried out to certify that the generated data complies with certain standards.

The large amount of data produced by these devices makes the analysis phase longer and more time-consuming than the experiment phase itself. For example, the size of one single data cube can range from 16 to 100 Mbytes. Up to 20 data cubes can be generated each day. Thus, a set of analysis tools needs to be integrated into the application to facilitate the work of the scientists, e.g. correlation analysis, multivariate data analysis (PCA, pLS) and others.

7.3 Life-Cycle of e-Scientific Experiments

Traditional scientific methodology uses the empirical cycle as a guide for experiments; e-Science is no different, but in its particular incarnation of the empirical cycle some phases require extra emphasis. When scientific experiments are modeled

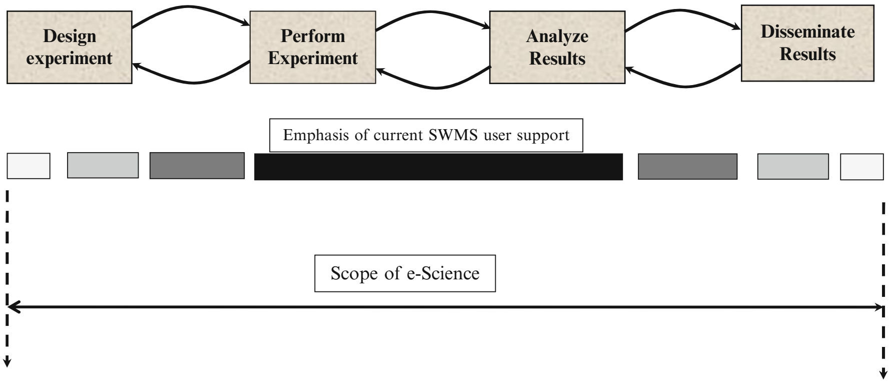


Fig. 7.2 e-Science experiment phases with indications of current and desired support. The gray-scale represents the focus of the existing workflow management systems: most of the support currently offered is associated with the execution and result analysis phases (*black bar*). Dissemination and design phases are less frequently supported (*gray bars*)

following a workflow approach, the latter encode the logic of the experimentation processes and become an important resource to promote knowledge transfer among scientists. A workflow represents a reliably repeatable sequence of tasks composing an e-Science application. It describes the pattern of activity enabled by a systematic organization of resources (Taylor et al. 2007). One aspect of e-Science in particular, the sharing of resources, places greater demands than usual on the experiment validation methodology. The reason for this is that scientists conducting these complex experiments often do not have the required expertise to solve all the problems facing them. It is common that they use third-party components and therefore need extra assurance to make sure that they are using these components in the proper way and these third party components are behaving as expected. There are different views on how the e-Science lifecycle can be defined. A commonly accepted definition is that the development of a scientific experiment has different activities or tasks performed at different times (Jacobs and Humphrey 2004; Humphrey and Hamilton 2004). These activities belong to different phases of a typical e-Science application lifecycle, which can be divided into four phases: *Design*, *Execution*, *Analysis*, and *Dissemination* (Fig. 7.2).

Design: an iterative process which requires discovering the resources that can be used, for instance through semantic search. Interoperation between the discovered resources needs to be established and a proper methodology associated with these resources needs to be considered. This problem becomes more challenging in the context of scientific experimentation where the scientists are continuously defining hypotheses, collecting data, running experiments, revising hypotheses, and publishing results. Multidisciplinary and geographically distributed teams of scientists need to be able to locate, construct, execute and maintain such workflows (De Roure et al. 2007). To design a successful workflow, modeling and composition tools are not sufficient – we also require tools for deriving knowledge and validating models (Miles et al. 2007a). Support for the design of scientific workflows faces a number of challenges:

- semantic support for annotating and searching workflows,
- reproducibility of results, configurations, and the runtime conditions of the experiments,
- interoperability between different workflows when they are used in one experiment.

Execution: most of the support provided by current SWMSs is dedicated to execution. This step of a scientific experiment is concerned with the execution of tasks comprising the workflow, which can be entirely computational, but can also involve interaction or manual steps that have to be performed by the user. Specific requirements from the applications include access to statistical toolkits as well as simple access to parallel computation. Experiment execution support also involves the delegation and automation of non-scientific and redundant tasks to the framework such as the staging of software components constituting the experiment and the data sets needed for the experiments, the search for the appropriate and available computing resources, and monitoring of experiment progress (Mayer et al. 2006; Olabarriaga et al. 2007). Interactive execution control is often required to allow scientists to steer the execution path and tune component parameters, in particular for the calibration phases of scientific experiments where the scientist is still experimenting with various parameter sets.

Analysis: this step of the scientific experimentation is the most delicate as it re-quires a lot of scientific knowledge which is often difficult to model, and thus calls for interaction with and among the scientists. It focuses on checking whether the output of a workflow complies with the theory or expected results. During execution, monitoring progress and steering execution between different paths are basic human activities that control the experiment. There are several scenarios which re- quire collaborative control between scientists. For instance, control of the processes in a complex experiments requires analysis of results, which often represents a joint effort by several scientists. Besides, a complex experiment consists of more than one workflow and these workflows might be shared and modified by geographically distributed scientists at the same time. As such, controlling an entire complex experiment may require input from all these scientists (Humphrey and Hamilton 2004).

Dissemination: Traditionally, dissemination is achieved through scientific publication. It consists in making the resources and workflows themselves available for use by others, providing specific metadata about the circumstances in which the results were created (also known as provenance), as well as sharing the knowledge associated with the proper use of resources, and providing abstracted versions of successful workflows (Miles et al. 2007b). Contemporary means of communicating and sharing scientific results (i.e. by publishing papers) fall short of the requirements of modern computational sciences, as such results do not easily lend themselves to verification and reuse. In order to shift focus from secondary sources (i.e. publications) to the actual data, algorithms, and workflows used in scientific research, a unified collaborative framework has to be developed, which will enable users of various HPC infrastructures, data repositories and virtual laboratories to publish and directly reference their data, and workflows. Science had always a social aspect, as far as collaboration among scientists is considered e.g. peer-review and scholarly communication. This aspect of science has led to quick adoption of new social tools which facilitate and accelerate the communication aspects across the scientific community. Research expertise can be propagated to others and reinvention avoided, and community curation of data and methods is becoming a powerful and acceptable way of validation within the scientific community.

The outcome of scientific experiments as well as the methodology, lessons, and tools obtained can be considered as societal contributions. The publication of the results necessarily includes provenance data which describes experiment steps, execution conditions, input data, interactions, activities performed to control the execution, and the analysis of results (Miles et al. 2007a). An environment which allows scientists to jointly annotate data, and compose documents is therefore necessary in experiments involving scientists located at various institutions worldwide.

Mapping three of the four phases of the lifecycle, as described in this section, to the MACS lab experiment introduced in Sect. 7.2 is straightforward. The preprocessing, experimentation and analysis steps can be respectively mapped to *Design*, *Execution*, and *Analysis* phases. The *Dissemination* phase in the traditional scientific approach is done mainly through publications in journals and participation in international conferences. With the emergence of the Web 2.0 approach, often known as “social web”, new ways of dissemination become possible which go beyond traditional publications. Published results can be more easily reproduced; tools and workflow can be shared among scientists worldwide etc. A good example of such scientific dissemination is the myExperiment site (http://myexperiment.org) which makes it easy to find, use and share scientific workflows, and to build communities. This sharing model is flexible enough to support various aspects of the lifecycle management of scientific workflows (De Roure et al. 2009; De Roure and Goble 2009).

7.4 Related Work on e-Science Frameworks

In the last decade the field of scientific workflow management systems and virtual laboratories has attracted interest of the scientific community. A large number of research projects worldwide focus on the development of workflow frameworks which can improve the lifecycle of scientific experiments. To gain insight on how *workflow support* is provided in popular SWMS, we present in this section a review of the state of art in the field of scientific workflow management systems. This study covers several well-known systems, namely ASKALON (Fahringer et al. 2005; Qin et al. 2006; Deelman et al. 2009), GridNexus (Brown et al. 2005), Grid Workflow Execution Service (GWES) (Hoheisel 2006a; Bubak and Unger 2006), ICENI (Mayer et al. 2006), Karajan (von Laszewski and Hategan 2005), Kepler (Ludäscher et al. 2009; Altintas et al. 2004; Buyya et al. 2000), Pegasus (Callaghan et al. 2009; Deelman et al. 2009), Taverna (Oinn et al. 2002) and Triana (Taylor 2006; Taylor et al. 2005; Harrison et al. 2008; Churches et al. 2006). We review theses systems in light of their *Design features, Execution support and Support for Interoperability, and Results dissemination and information management support*.

7.4.1 Design Features

Since workflows are intuitively depicted as graphs, it is no surprise that most SWMS offer a graphical workflow composer for building such graphs. Although there seems to be a consensus as to the basic notion for a workflow vertex being a computation task and an edge being a data and/or control flow, different SWMS tend to differ on the actual graph modeling. Common differences are cyclic versus acyclic and stateful versus stateless graph models. SWMSs such as ASKALON, Taverna, Karajan model workflows based on Directed Acyclic Graphs (DAG). Since pure DAGs do not model conditional branches and loops, such systems augment the DAG model to include these primitive control structures. ASKALON and Karajan also include advanced control structures such as parallel constructs. Some systems such as GWES do away with DAG modeling altogether and instead use Petri nets. Petri nets differ from DAGs by modeling control flow and, most importantly, model workflow state through the use of token transitions. Furthermore, Petri nets well understood properties such as deadlock and conflict can aid in model analyses and optimization. Kepler goes a step further and allows different types of models, which it achieves through directors. This notion of different Models of Computation (MoC) allows for greater flexibility as many other SWMS only allow one MoC. Common MoC in Kepler are: *Process Network, Dataflow, Discrete Events, Synchronous/Reactive*. It is most often the case that a graphical workflow composition is synthesized to an XML-based language, which facilitates sharing and reusability. Such languages include AWGL, JXPL, GWorkflowDL and Scufl used by ASKALON, GridNexus, GWES and Taverna respectively.

Scientific workflows are collaborative in nature and hence workflow sharing becomes an important feature in an SWMS. Such collaborative features are included in Kepler and Taverna, which support semantic queries for components that can be reused in new workflows. Taverna also includes a browser which provides navigational capabilities over data stored in the myGrid (Stevens et al. 2003) information repository (MIR), which include experimental designs, experiment results, and intermediate data.

7.4.2 Execution Support

Some of the characteristics that differentiate SWMS execution support from each other and from other systems include the types of resources onto which the workflow can be mapped and the method of enactment which describes the way workflow tasks are scheduled and executed.

A common denominator amongst supported distributed resources is the Globus[2] grid middleware. SWMSs such as ASKALON, GWES, Karajan, Pegasus and Triana readily support Globus due to its widespread use within the scientific community. Some other SWMSs such as Pegasus and Triana are not only bound to one type of resource but can interoperate with other middleware suites. A number of the studied SWMS have the ability to use web/WSRF services as workflow components allowing scientists to have access to a wide range of services which are publicly available such as the BioCatalogue which currently hosts around 1,600 services.[3] Service-oriented resources are exploited through SWMS such as Kepler, Taverna, Triana, GridNexus and GWES.

Workflow enactment engines differ considerably between SWMSs though a typical engine would consist of a scheduler which interprets the workflow graph to deduce data/control flow dependencies and a runtime manager which takes care of controlling individual tasks during runtime. ASKALON uses different schedulers such as Heterogeneous Earliest Finish Time (HEFT), a genetic algorithm and a myopic just-in-time algorithm. Similarly, GWES implements several schedulers. A simple scheduler is the least-used scheduler, which chooses resources that had not been used for the longest time. Another scheduler makes use of the Globus Monitoring and Discovery System (MDS) so as to make a more informed deployment decision. Karajan supports late binding, hence deferring the decision of how a task should be executed until the task is actually mapped to a resource. Pegasus uses DAGMan as its scheduler. DAGMan is a directed acyclic graph meta-scheduler for the Condor job scheduler.[4] Triana schedules groups of components where each group can be piped or parallel. Early Taverna versions (1.x) used a heavily modified version of the Freefluo[5] web-service orchestration engine while later versions (2.x) replaced Freefluo for a new orchestration engine. In Kepler, directors represent enactment engines for different MoCs.

Interoperability at runtime between SWMS is sometimes needed to take advantage of the unique features of existing systems. Recent developments in the field and the adoption of service-oriented architecture standards by most systems should allow for interoperability. Triana takes this one step further by offering the possibility to publish its (sub)workflows as a grid/web services. Kepler pursues another approach which consists in wrapping the resources of other SWMSs, e.g., the Nimrod engine to improve its parameter sweeping support (Abramson et al. 2009).

[2] Globus Middleware: http://www.globus.org/

[3] http://www.biocatalogue.org/services

[4] http://www.cs.wisc.edu/condor/dagman

[5] FreeFluo: http://freefluo.sourceforge.net

7.4.3 Results Dissemination

Although most of the presented systems already provide visualization tools to view and analyze experiment results, we found limited support for sharing analysis results with other (future) users. Virtual Data systems used in Pegasus as well as provenance support in Taverna allow for sharing and reuse of experiment results, while in other systems collaborative sharing of experimental results has to be performed manually.

Kepler and ICENI increase reusability and dissemination. In Kepler a new component is being developed and which allow adding provenance information for each workflow. This new component is attached to the workflow and can be used later as a logbook or to search for experiments. Similar developments are taking place in both Triana and Taverna. In ICENI a backend component called application mapper is used to find the specific appropriate workflow.

While some support for design and dissemination exists, this support is incomplete especially when it comes to sharing resources. While there is no system that can even match all the computational requirements of the limited number of applications, trying to create one SWMS for all e-Science applications seems unfeasible. In the next section we will present our own attempts at workflow support and explain how our approach may help create a framework where sharing between all e-Science applications becomes practical.

7.5 An e-Science Virtual Laboratory

In this section we describe the main components composing the architecture of the e-Science virtual laboratory, which helps scientists to develop CPU- and data-intensive complex applications and allow them to use a distributed and complex computing infrastructure. The architecture follows a service-oriented approach; the main components are either simple web services or WSRF compliant services.[6] This approach has a major benefit that is the virtualization of the complex and distributed computing and storage infrastructure (Fig. 7.3). It also allows building a loosely coupled system, something that is highly required in a dynamic and non-reliable environment such as the grid.

7.5.1 Process Flow Template (PFT)

This initial prototype of the e-Science framework introduces a two-level abstraction approach where the Process Flow Template helps enforce best practices and increases the reusability within a given domain by (Kaletas 2004; Afsarmanesh

[6] http://www.oasis-open.org/committees/tchome.php?wg abbrev=wsrf

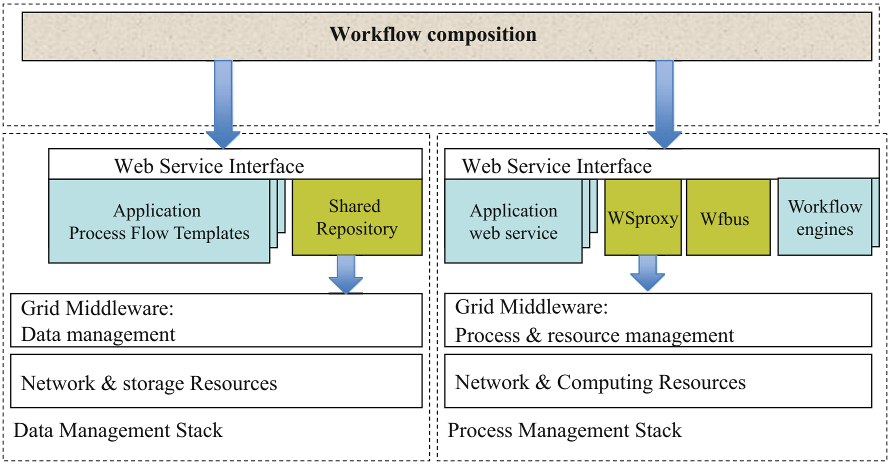


Fig. 7.3 The architecture of the e-science virtual laboratory developed in the VL-e project. A service-oriented architecture has been adopted to glue all the components needed to support the lifecycle of e-Science applications. The components are decomposed into two categories: components for data management and components for process management

et al. 2001; Belloum et al. 2003). A PFT is defined by application domain experts; it captures the expertise and knowledge of the experts and is meant to transfer this knowledge to other scientists. A PFT provides context-sensitive assistance for novice users performing complex studies, helping them avoid mistakes and increasing the efficiency and accuracy of their experiments (Kaletas et al. 2002; Frenkel et al. 2001). PFTs cover both the experiment design and the dissemination phase defined in Sect. 7.3. In the scope of experiment design, PFTs allow domain experts to define the steps of the experiments, the data structure, and the infrastructure needed to complete the experiment and how it should be used. PFTs can be seen as a tool to create templates which determine how experiments should be performed. There are three basic building blocks of PFTs that contain metadata on:

- **Objects**, where objects can be either physical objects or bulk data in mass storage,
- **Manual Operations** which are non-computational in nature and can be performed personally by the scientists,
- **Executable operations** which are specified in a separate specific executable workflow (outside of the PFT).

Using the example of the MACS lab introduced in Sect. 7.2, we describe how PFTs can be applied in a real-world scenario. The PFT has been used to describe the complex flow involved in the data analysis, in this particular case preparation of a sample, surface analysis, visual inspection and possibly re-analysis. A simplified version of such a data flow is shown in Fig. 7.4. Typical steps in a surface-analysis study are represented by the MACS lab PFT.

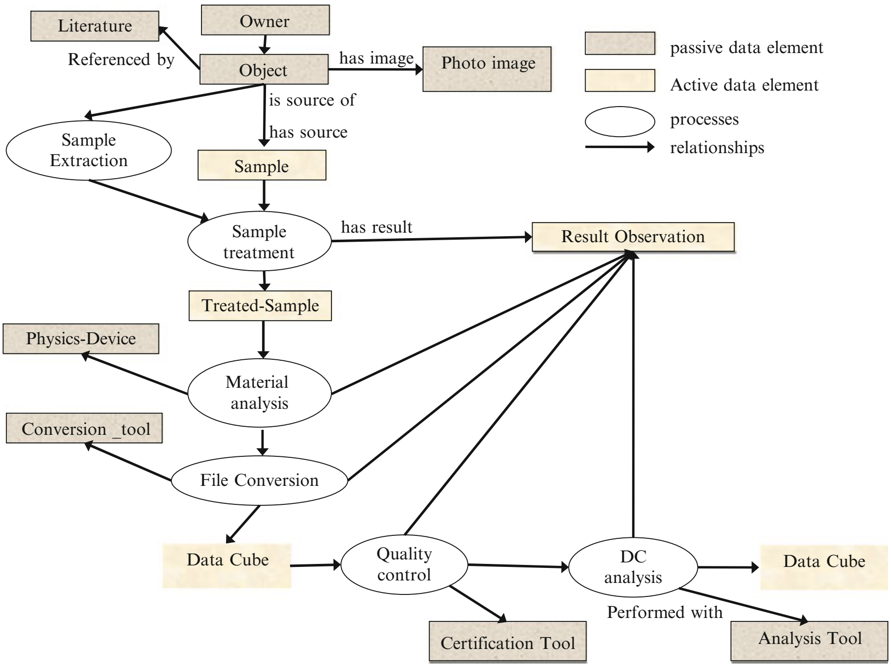
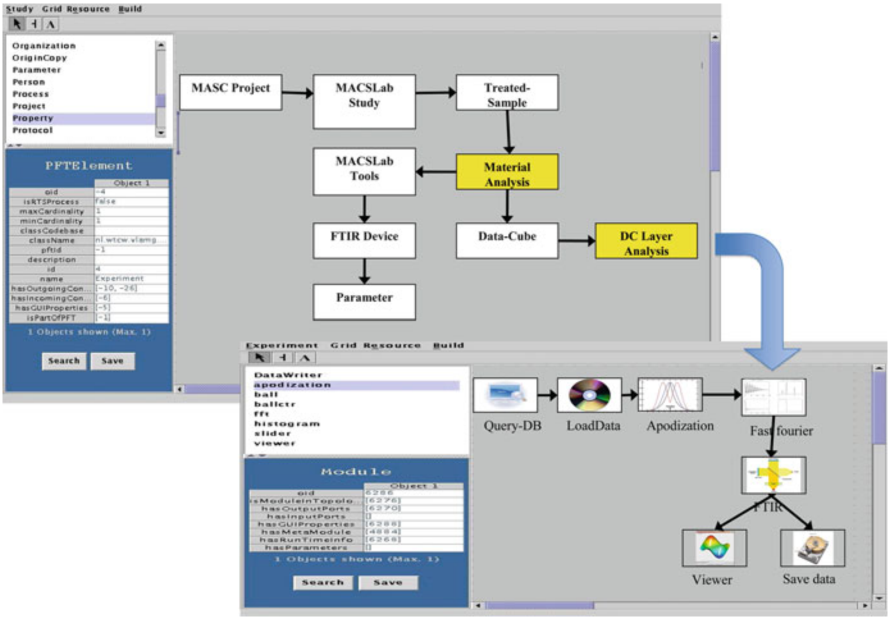


Fig. 7.4 The PFT of the MACS lab experiments describes the flow of processes involved in data analysis. The hatched rectangles represent either physical objects or bulk data in mass storage. Hatched ovals represent operations, both executable and manual (Frenkel et al. 2001)

At a later stage, less experienced users can customize a given PFT such as in the MACS lab example, to perform a specific experiment. The user then creates an instance of the PFT and modifies some parameters, defines input and output data sources, attaches a new application workflow etc. PFTs enforce best practice by providing to the domain expert a view of the flow of a given experiment. By using PFTs the end users are assured of performing the right tasks at the right time. However, they retain the possibility to customize the PFT to suit their own situation. Even when not yet implemented, users should be able to combine different PFTs or even work on subparts of a given PFT.

Because PFT and the instances created by the end users are saved in application domain databases, they can be accessed via a simple query interface by all members of a Virtual organization based on the view management policy (Kaletas 2004; Afsarmanesh et al. 2001) (Fig. 7.5). In the current implementation the query interface is limited to a search based on the PFT and the instance name; the idea being to develop this query interface to support more sophisticated queries based on semantic techniques. The MACS lab PFT was demonstrated at the iGrid 2002 conference where it was applied to the analysis of a sample (paint chip) from a painting in the FTIR spectrometer and subsequent analysis of the measurement data. Using the PFT scientists were able to customize the original PFT to collect experimental



	Object 1
oid	-4
isRTSProcess	false
maxCardinality	1
minCardinality	1
classCodebase	
className	nl.wtcw.vlamp...
pftid	-1
description	
id	4
name	Experiment
hasOutgoingCon...	[-10, -26]
hasIncomingCon...	[-6]
hasGUIProperties	[-5]
isPartOfPFT	[-1]

1 Objects shown (Max. 1)

	Object 1
oid	6266
isModuleInTopolo...	[6276]
hasOutputPorts	[6270]
hasInputPorts	[]
hasGUIProperties	[6288]
hasMetaModule	[4884]
hasRunTimeInfo	[6268]
hasParameters	[]

1 Objects shown (Max. 1)

Fig. 7.5 This excerpt from the MACS Lab PFT shows both metadata descriptions (*white boxes*) and automated experiment steps executed on the distributed infrastructure (*grey boxes*). The inset shows an associated experiment workflow for DC analysis

information about the painting, sampling spot and device parameters. The PFT, together with the descriptive fields, is stored in a database and can be subject to queries (Hendrikse et al. 2003).

PFTs offer a clear separation between the information needed for at the design phase of an experiment and the information needed at the execution time. Such separation will reduce the amount of information the scientists have to deal with for large-scale experiments. A similar abstraction also exists in Pegasus (Deelman et al. 2004). The main difference between these abstract workflows and PFTs is that Pegasus workflows can be made executable by mapping them to the underlying resources, while PFTs provide the context of the experiment. PFTs have the advantage of being decoupled from the execution of the experiments and, consequently, can be used to point to experiments executed in different workflow systems.

7.5.2 *Ontology-Based Approach*

Experience has shown that a successful e-Science framework has to abstract the scientists from all information technology issues. In the first approach to PFT a database expert is needed to define the required underlying data structure. A more practical approach will be to allow domain experts to describe the PFTs at a very high level of abstraction and develop a system which automatically/semi- automatically creates the underlying structure. Ontology-based approaches are aiming at such an abstraction, allowing them to operate on the basis of custom concepts. Olingo is an ontology-based approach we have developed to generate the underlying data structures for PFTs. Olingo is developed in two flavors: a Web portal (Olingo Web Tool (OWT)) and a plugin for the Protege ontology editor.[7]

We use an ontology-based approach because ontologies are extensible mechanisms typically used for describing, via human-readable text, the domain of disclosure, which enables common understanding among scientists and software applications. Although ontology provides an abstract view, it is suitable as input for the generation of underlying structures. Thus, to avoid both schematic and semantic heterogeneity, Olingo targets the generation of a number of explicit output formats for the application cases from different scientific domains. It is the link between ontologies and data structures. Application domain experts may use Olingo during the design phase to generate the data structures needed for the creation of PFTs, which will subsequently be applied during the execution and the analysis phase of the experiment by less experienced users.

It is evident that, in some simple and specific cases, a basic model can be used by a proprietary generator to create the database schema that is later integrated within the scientific application. Usually implemented in an ad-hoc fashion, through scripts or parser generators, this approach does not scale well for large-scale scientific projects. OWT is a Web application that, based on an ontology, contributes to generate the underlying data structures for PFTs. Ontologies are extensible mechanisms typically used for describing, via human-readable text, the domain of disclosure, which enables common understanding among scientists and software applications. OWT targets the generation of a number of explicit output formats for the application cases from different scientific domains. It is the link between ontologies and data structures.

Olingo produces five different output formats, including (1) relational schema with a data definition language for relational databases, (2) Java classes providing the source code of data structures, (3) XML Schema with a specification for XML documents, (4) mapping files for two different frameworks (Castor and Hibernate) that support persistence of Java objects.

7.5.3 Workflow Management System

The WS-VLAM workflow management system[8] is composed of a workflow editor and a workflow enactor. The workflow editor completes the PFT by providing support for the execution phase of an experiment. In the current implementation, the execution model of the workflow is data-driven. Workflow components are

[7] http://protege.stanford.edu/

[8] www.science.uva.nl/gvlam/wsvlam

scheduled on geographical and grid-enabled computational resources. When data is available for processing, it is passed/streamed directly to workflow components enabling concurrent execution. The engine supports enacting workflow components written in a number of programming and scripting languages (Java, C^{++}, python, SWIG), and also allows access to RPC-style web services and workflows (Korkhov et al. 2007a, b). The system has been used to prototype a number of application (Inda et al. 2008; Zudilova-Seintra et al. 2002; Leguy et al. 2009). Once an application workflow is designed, it can be attached to PFT and become available for scientists who can create multiple instances to meet their specific requirements such as setting the different parameters and new input data sets. WS-VLAM offers not only an intuitive way for the creation and execution of application workflows, but also provides seamless access to the underlying complex grid-enabled infrastructure. WS-VLAM has the ability to (1) interact and monitor the workflow at runtime, (2) automatic redirection of the graphical output to the end-user default screen, (3) easily adapt/change the application workflow to meet user-specific needs, and (4) run workflows in batch mode.

The *DC Analysis* workflow for the MACS lab experiment was one of the first scientific workflow to be ported to the earlier versions of WS-VLAM 1.5. Through this workflow, scientists are able to query a database containing detailed information about all data produced by the sample treatment process. The (raw) data sets corresponding to query results are retrieved and piped into an apodisation routine (also called a tapering function). The apodised data sets are subsequently submitted to a fast Fourier transform and calibration procedure. The results are piped to a data viewer for visualisation and a multivariate data analysis module for extraction of principal components (Hendrikse et al. 2003).

7.5.4 The Workflow Bus

From the state of art study presented in Sect. 7.4 it is clear that, at least in the near future, a unique workflow management for e-Science is unlikely to emerge. All of the presented systems have their advantages and disadvantages; moreover, most of them are building small communities of users around themselves. It is evident that at a certain point, in order to continue to promote sharing and reusability, there will be no other way but to bridge these systems to allow scientists to re-assemble workflows in different systems. To achieve this goal, a meta-execution framework is needed for integrating different workflows, coordinating the execution of different enactors and moving data around. This approach will become more feasible once most of the workflow management systems have adopted a service-oriented architecture where the engine and enactors are implemented as standalone services. The basic idea of a workflow bus (Zhao et al. 2006) is to wrap a number of popular and relatively mature legacy SWMSs as federated components, and to loosely couple them as one meta-workflow system using a software bus. The workflow bus is an interactive workflow environment, which provides an agent-based wrapper for integrating legacy workflows or components and for coordinating their behavior in a meta-workflow. The runtime behavior of a legacy workflow is modeled as a scenario and wrapped as an agent called a *scenario manager*. The scenarios are coupled through a *meta-workflow*, called a *study*. At runtime, each study has an administrator agent (the *study manager*), which manages the organization information of *scenario managers*. As a runtime infrastructure, the workflow bus provides basic services for interpreting and scheduling meta-workflows, for orchestrating plugged legacy workflow engines, for passing and distributing data between workflow engines, and for supporting user interaction with workflows. From the system level point of view, features from different systems are then aggregated and integrated as one meta-system.

In the context of the workflow bus, the interface of a legacy workflow is modeled as a set of ports, which have a number of properties: *read (input)* or *write (output), media, type*, and *access*. These properties indicate where the content of the port is hosted, and the type of the port (abstract types only have access to data references, while concrete types describe the location where actual data is stored). The workflow bus provides a schema to specify the interface and other meta-information related to the legacy workflow, such as access point of the original workflow and its execution requirements. The description, namely *scenario description*, can be interpreted by the scenario manager; at runtime a scenario manager generates the port stubs, and is able to search for a suitable workflow engine according to the execution requirements described in the description (Zhao et al. 2007).

One of the use cases which have motivated the development of the workflow bus is the MACS lab experiment. As described by the MACS lab PFT, this experiment is composed of five processes, from which only one has been designed *DC analysis* as a Workflow using the WS-VLAM workflow management system. The remaining processes required features which are not provided by WS-VLAM and thus where developed using third-party systems. The workflow bus approach was then proposed as a means to coordinate the execution of the entire flow comprising the MACS lab experiment.

Figure 7.6 shows how the workflows composing the MACS Lab experiment which are developed in various systems can be executed through the workflow bus: first, the workflows are wrapped as scenarios, then the scenario manager (scenario 6) connects them, and allows the intermediate data to be assigned to the appropriate workflow.

7.5.5 The ProxyService

Rapid adoption of the Service Oriented Architecture has led to development of a huge number of web services which can be accessed remotely to perform scientific calculations. Bioinformatics is a good example of a scientific field which has seen an explosion of the number of available web services (Peachey et al. 2003). Web services offer an appealing paradigm for developing scientific

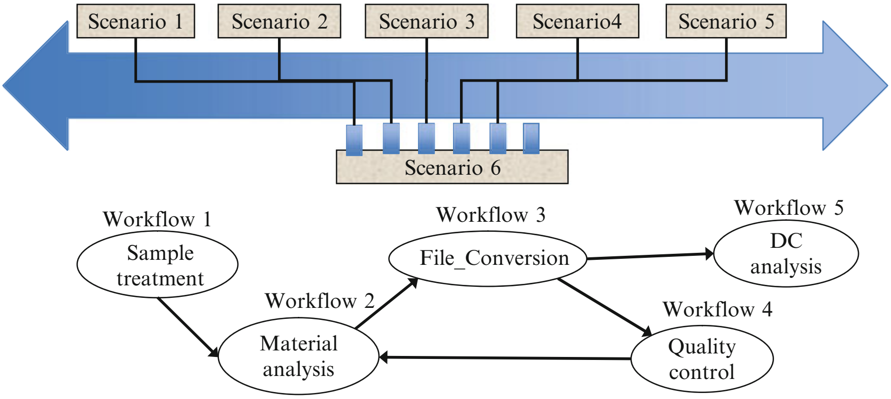


Fig. 7.6 The workflow bus allow the execution of application workflows across multiple SWMS. The runtime behavior of a legacy workflow is modeled as a scenario. The scenarios of legacy workflows are coupled through a meta-workflow. At runtime, an administrator agent manages the delivery of information between the scenarios

applications, by providing interoperability and flexibility in a large-scale distributed environment. Through the use of XML-based protocols (SOAP) and interfaces (WSDL), web services can expose the entirety or selected parts of any application in a language-independent fashion across heterogeneous platforms. Moreover, these features enable them to be combined in a workflow so that more complex operations may be achieved (Zhang et al. 2006). Currently, two approaches apply to workflow implementations: Service Orchestration and Service Choreography.

In Service Orchestration the process is always controlled by a workflow engine, so all invocations (and replies) are made by (and to) that workflow. On the other hand, choreography is more collaborative, because it describes the message exchange among interacting, yet independent web services.[9] Regardless of the architecture chosen, any workflow implementation is faced with a data transport problem which can be summarized in the following way: (1) In service orchestration, all data go through the workflow engine before being delivered to a consuming web service. This practice not only makes data delivery inefficient, but also causes failures in workflow execution due to the data burden workflow engines have to carry. (2) Data transfers are made through SOAP, a protocol unfit for large-scale data transfers (Daly et al. 2005). Curing data through SOAP is problematic for many scientific applications, since the success of the paradigm has caused an abundance of web services to be deployed. As applications started to produce more data, these services fail to scale with the increased data demands. (3) Third-party file transfer is suitable for transferring large data sets, but is restricted to files. This results in unnecessary intermediate transfers that slow down workflow execution and place excessive demands on storage resources.

[9] Web Service Choreography Interface: http://www.w3.org/TR/wsci/

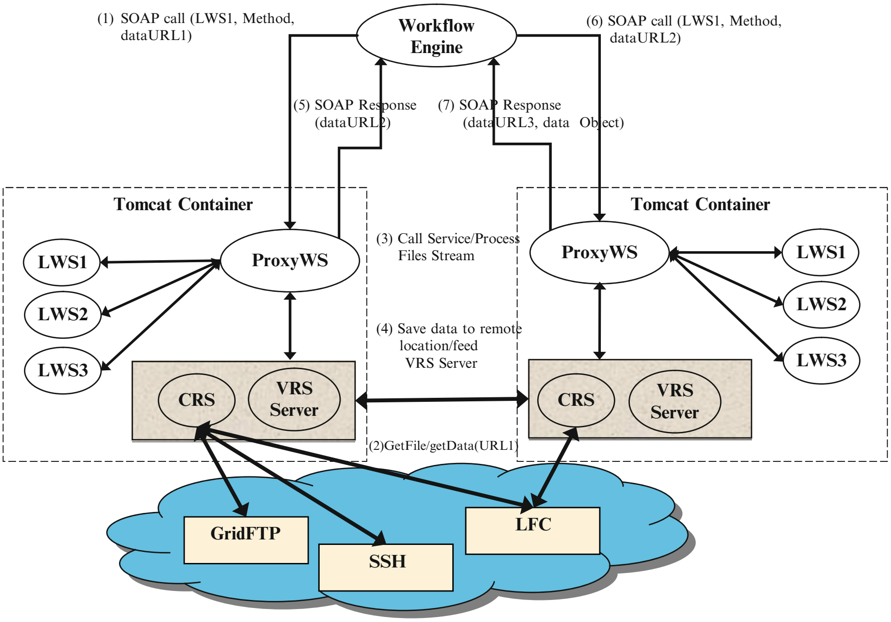

(1) SOAP call (LWS1, Method, dataURL1)
Workflow Engine
(6) SOAP call (LWS1, Method, dataURL2)
(5) SOAP Response (dataURL2)
(7) SOAP Response (dataURL3, data Object)
Tomcat Container
LWS1
LWS2
LWS3
ProxyWS
CRS
VRS Server
(3) Call Service/Process Files Stream
(4) Save data to remote location/feed VRS Server
(2)GetFile/getData(URL1)
Tomcat Container
ProxyWS
LWS1
LWS2
LWS3
CRS
VRS Server
GridFTP
SSH
LFC

Fig. 7.7 The ProxyWS architecture. With the use of the VRS and the VRSServer, web services may access data from remote locations, or other web services which helps ameliorate the data isolation problems of Web services used in scientific workflows

To address these problems we introduced ProxyWS: a web service that is able to access data from remote resources (GridFTP, LFC, etc.) using the Virtual Resource System (VRS), a Java API used by the VBrowser[10] to provide a single access point to grid-enabled systems. Additionally, ProxyWS is able to transport larger volumes of data produced by both legacy and new web service implementations. For ProxyWS to be able to provide better data transfers to legacy web services, it has to be deployed in the same Axis-based container, just like a normal web service. This enables clients to make proxy calls to ProxyWS instead of a legacy web service (Fig. 7.7). Consequently, ProxyWS returns a SOAP message containing a URI referring to the data location, which might be any remote or local data resource, as long as it is supported by the VRS instead of the actual data. For new implementations, ProxyWS is used as an API that can create data streams from remote data resources and other web services using ProxyWS. This allows web services producing large amounts of data to be connected in a data pipeline, something that could optimize workflow execution (Koulouzis et al. 2008). Thus, with the introduction of ProxyWS we have a centralized control flow with all the benefits of distributed data flow for new and legacy web services.

[10] http://www.vl-e.nl/vbrowser

The ProxyWS approach has helped optimize a number of applications where the adoption of a service-oriented architecture has led to a decrease in overall performance, like the *Service-Oriented Visualisation applied to medical data analysis* (Zudilova-Seintra et al. 2002), and the *indexing and Name Entry Recognition services* developed in Adaptive Information Disclosure project.[11] ProxyWS enables exposure of data-intensive applications as web services without loss of performance. Experiments such as MACS Lab and, more specifically, the *DC analysis* workflow can thus be exposed as a set of web services which can be easily used by a wider scientific community.

7.6 Summary

In this chapter, we discussed some of the workflow-related issues in e-Science. First, we analyzed the requirements for supporting different application domains, and then discussed the necessity and importance of developing a unified framework for e-Science. We illustrated this vision by describing the approach followed in the context of the VL-e Project to develop such a framework.

The emergence of grid environments gives scientists new ambitions to tackle more complex and large-scale problems, which can lead to new methodologies in scientific research and problem solving. The core idea of e-Science is to allow scientists representing different domains to share resources and knowledge and to collaborate in their research. In this chapter, we highlighted two issues pertinent to developing a collaboration environment for e-science: Modeling workflows on the level of the entire scientific experiment lifecycle enables knowledge transfer between scientists where successful experiment results and templates can be applied for new problems as reusable resources.

Our state of the art analysis showed that the development of many SWMSs are highly application-driven. The specifics of different application domains result in different workflow models and different styles of user support, which limits the opportunities for sharing and interoperability. The other lesson to be learned from the survey is that while practical support exists in some form for all the stages of the e-Science experiment workflow, methodological support is lacking. A simplistic approach to address this problem will likely fail; more advanced solutions have to be investigated with the aid of recent achievements in semantics, Web 2.0, and ontologies.

One of the conclusion of the report from the NSF/Mellon Workshop on Scientific and Scholarly Workflow organized 2007 (Klingenstein et al. 2007) is that there will be no single SWMS which is usable for all e-Science experiments. The absence of standards in the field of SWMS seriously limits the sharing of resources between diverse e-Science applications across scientific fields and various SWMSs.

A potential approach to lower the interoperability problem is through the use of service-oriented architectures. Standards such as WSDL for publishing services

[11] http://www.adaptivedisclosure.org

enable SWMS to use one another's services. However, being able to use services across multiple SWMSs is just part of the solution as currently there are no common standards for sharing knowledge associated with the proper use of services, nor is there a standard for sharing executable workflows. In this chapter we addressed the issue of knowledge sharing. It is clear that, given the current state of art, knowledge sharing is very much integrated into or dependent upon the executable workflow description. We think that knowledge should only be loosely coupled to the system which generates it, and we propose PFTs as a means to achieve this goal. Through this loose coupling PFTs can easily be implemented on top of different SWMSs. Diverse e-Science applications which require different SWMS for their execution can, by using this construct, share knowledge about the proper use of their web/grid services more easily.

7.6.1 Prospective Usage Applications

The VL-e framework has been created based on requirements extracted from six different scientific domains. In this chapter we have given an example of scientific application developed using the VL-e framework. However, the generic aspects of this framework make it applicable to various scientific domains including practical medicine (Leguy et al. 2011; Inda et al. 2008; Koulouzis et al. 2010). The computer aided drug design makes possible the creation of new chemical compounds which can work as the modificators of the "target" molecules creating the complexes with them. The presence of the ligand – potential drug – may correct the improper activity of the protein which is the source of the pathological process in patients body. This is why the VL-e may be implemented into the practical medicine.

The new challenge for practical medicine is the individual therapy. The traditional diagnostics and therapy stops when particular pathology has been recognized and particular procedure is applied for therapy. So far there arę some clinical paths for the group of patients representing similar symptoms. In post-genomic era, when the SNP (single nucleotide polymorphism) has been recognized and identified the individual therapy is expected. The drug, which can work successfully in one case may be useless in the other due to structural changes in proteins being the results of SNP. This is why the individually created drugs addressed to particular "target" protein should be applied in therapeutic procedures. New drug design takes time. The structure of proteins influenced by SNP should be generated and the possible ligand binding cavities should be recognized as well as possible protein-protein complexation areas. This time period should be as short as possible to make the therapy successful. The availability of tools (and methods) described in his book are the milestones in respect to these expectations.

The generation of mutation-modified structure of target molecule and identification of ligand binding cavities (protein-protein complexation) are necessary for correction malfunctioning proteins in the human body. The feature of individual therapy is applicable to AIDS therapy. HIV virus is characteristic by its very frequent mutations making its proteins unrecognizable for drugs applied in the therapy so far. The monitoring of its mutations and consequent structural changes with the procedures of new drug creation addressed against these modified proteins of HIV virus may significantly speed up the therapeutic processes. The computer-based tools presented in his book when introduced to practical medicine will speed up significantly the therapeutic processes making them individually addresses against the "target" in the form as it appears in the patients body. Identification of ligand binding sites in proteins as well as recognition of potential protein-protein complexation area is the basis for individually designed therapy.

Acknowledgments This work was carried out in the context of the Virtual Laboratory for e-Science project (www.vl-e.nl) supported by the BSIK grant of the Dutch Ministry of Education, Culture and Science and ICT innovation program of the Ministry of Economic Affairs. Currently this research is supported by the COMMIT project (WP8 and WP9). The authors are also grateful to Piotr Nowakowski for his remarks.

References

Abramson D, Bethwaite B, Enticott C, Garic S, Peachey T, Michailova A, Amirriazi S, Chitters R (2009) Robust workflows for science and engineering. In: MTAGS '09: proceedings of the 2nd workshop on many-task computing on grids and supercomputers, pp 1–9. ISBN 978-1-60558-714-1. doi: http://doi.acm.org/10.1145/1646468.1646469

Afsarmanesh H, Kaletas E, Benabdelkader A, Garita C, Hertzberger LO (2001) A reference architecture for scientific virtual laboratories. J Future Gen Comput Syst (Elsevier) 17(8):999–1008. ISSN 0167-739X

Altintas I, Berkley C, Jaeger E, Jones M, Ludascher B, Mock S (2004) Kepler: an extensible system for design and execution of scientific workflows. Scientific and Statistical Database Management, Proceedings of the 16th international conference on 21–23 June 2004, pp 423–424

Belloum ASZ, Groep DL, Hendrikse ZW, Hertzberger LO, Korkhov V, de Laat CTAM, Vasunin D (2003) VLAM-G: a grid-based virtual laboratory. Future Gen Comput Syst 19(2):209–217. ISSN 0167-739X. doi: http://dx.doi.org/10.1016/S0167-739X(02)00147-4

Blaas J, Botha CP, Majoie C, Nederveen A, Vos FM, Post FH (2007) Interactive visualization of fused fMRI and DTI for planning brain tumor resections. In: Cleary KR, Miga MI (eds), Proceedings of the SPIE medical imaging 2007, vol 6509

Broersen A, van Liere R, Heeren RMA (2007) Parametric visualization of high resolution correlated multi-spectral features using PCA. In: Eurographics/IEEE-VGTC symposium on visualization, pp 203–210 doi:10.2312/VisSym/EuroVis07/203-210.http://www.eg.org/EG/DL/WS/VisSym/EuroVis07/203-210.pdf

Brown JL, Ferner CS, Hudson TC, Stapleton AE, Vetter T, Carland A, Martin J, Martin A, Rawls A, Shipman WJ, Wood M (2005) GridNexus: a grid services scientific workflow system. Int J Comput Inf Sci (IJCIS) 6:72–82

Bubak M, Unger S (eds)(2006) K-WfGrid - The Knowledge-based Workflow System for Grid Applications, Proceedings of CGW'06, Vol. II, ISBN 978-83-915141-8-4, available at: http://www.cyfronet.pl/cgw06/

Buyya R, Abramson D, Giddy J (2000) Nimrod/G: an architecture for a resource management and scheduling system in a global computational grid. In: 14–17 May 2000, pp 283–289

Callaghan S, Deelman E, Gunter D, Juve G, Maechling P, Brooks C, Vahi K, Milner K, Graves R, Field E, Okaya D, Jordan T (2009) Scaling up workflow- based applications. J Comput Syst Sci 76(6):428–446. ISSN 0022–0000. doi: DOI:10.1016/j.jcss.2009.11.005

Chin G Jr, Ruby Leung L, Schuchardt K, Gracio D (2002) New paradigms in problem solving environments for scientific computing. In: Proceedings of the international conference of Intelligent user interface, San Francisco pp 39–46. http: //www.iuiconf.org/02pdf/2002-001-0004.pdf

Churches D, Gombas G, Harrison A, Maassen J, Robinson C, Shields M, Taylor IJ, Wang I (2006) Programming scientific and distributed workflow with triana services. Concurr Comput Pract Exp 18:1021–1037

Daly J, Forgue MC, and Harakawa Y (2005) Three-Part Solution Leads to Better Web Services Performance, W3C issues three web services recommendations. W3C press release. http://www.w3.org/2005/01/xmlp-pressrelease.html

De Roure D, Goble C (2009) Software design for empowering scientists. IEEE Softw 26(1):88–95. ISSN 0740–7459

De Roure D, Goble C, Stevens R (2007) Designing the myExperiment virtual research environment for the social sharing of workflows. e-Science and grid computing, international conference on 10–13 Dec 2007, pp 603–610. doi: http://doi.ieeecomputersociety.org/10.1109/E-SCIENCE.2007.29

De Roure D, Goble C, Stevens R (2009) The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Gen Comput Syst 25:561–567

Deelman E, Blythe L, Gil Y, Kesselman C, Mehta G, Patil S, Su MH, Vahi K, Livny M (2004) Pegasus: mapping scientific workflows onto the grid. In: European across grids conference, pp 11–20

Deelman E, Gannon D, Shields M, Taylor I (2009) Workflows and e-science: an overview of workflow system features and capabilities. Future Gen Comput Syst 25:528–540. doi:DOI:10.1016/j.future.2008.06.012

Fahringer T, Prodan R, Duan R, Nerieri F, Podlipnig S, Qin J, Siddiqui M, Truong HL, Villazon A, Wieczorek M (2005) ASKALON: A grid application development and computing environment. In: GRID'05: proceedings of the 6th IEEE/ACM international workshop on grid computing, pp 122–131.ISBN 0-7803-9492-5. doi: http://dx.doi.org/10.1109/GRID.2005.1542733

Foster I, Kesselman C (2006) Scaling system-level science: scientific exploration and IT implications. Computer 39:31–39

Frenkel A, Afsarmanesh H, Eijkel GB, Hertzberger LO (2001) Information management for material science applications in a virtual laboratory. In: Proceedings of the 12th international conference on database and expert systems applications DEXA 2001, Munich, Germany, 3–7 September 2001

Harrison A, Taylor I, Wang I, Shields M (2008) WS-RF workflow in triana. Int J High Perform Comput Appl 22(3):268–283.ISSN 1094–3420. doi: http://dx.doi.org/10.1177/1094342007086226

Hendrikse ZW, Belloum ASZ, Jonkergouw PMR, Eijkel GB, Heeren MR, Hertzberger LO, Korkhov V, de Laat CTAM, Vasunin D (2003) Evaluating the VLAM-G toolkit on the DAS-2. Future Gen Comput Syst 19:815–824

Hey AJG, Trefethen AF (2002) The UK e-science core programme and the grid. Future Gen Comput Syst 18(8):1017–1031. ISSN 0167-739X. doi: http://dx.doi. org/10.1016/S0167-739X(02)00082-1

Hoheisel A (2006a) User tools and languages for graph-based grid workflows: Research articles. concurrency and computation: Pract Exp 18(10):1101–1113. ISSN 1532–0626. doi: http://dx.doi.org/10.1002/cpe.v18:10

Humphrey C, Hamilton E (2004) Is it working? Assessing the value of the Canadian data liberation initiative. Bottom Line 17(4):137–146. ISSN 0888-045X. doi: 10.1108/08880450410567428

Inda MA, van Batenburg MF, Roos M, Belloum AZS, Vasunin D, Wibisono A, van Kampen AHC, Breit TM (2008) SigWin-detector: a grid-enabled workflow for discovering enriched windows of genomic features related to DNA seąuences. BMC Res Notes 8(1):63. doi:10.1186/1756-0500-1-63

Jacobs JA, Humphrey C (2004) Preserving research data. Commun ACM 47(9):27–29. ISSN 0001–0782. doi: http://doi.acm.org/10.1145/1015864.1015881

Kaletas EC (2004) Scientific information management in collaborative experimentation environments. Ph.D., thesis, Universiteit van Amsterdam, May 2004

Kaletas E.C, Afsarmanesh H, Hertzberger L.O (2002) Virtual laboratories and virtual organizations supporting biosciences. In: Camarinha-Matos LM (ed), Proceedings of PRO-VE'02, in collaborative business ecosystems and virtual enterprises,pp 469–480. ISBN 1-4020-7020-9

Klingenstein K, Gannon D et al. (2007) Improving interoperability, sustainability and platform convergence in scientific and scholarly workflow. Technical report, internet2. https:// spaces.internet2.edu/display/SciSchWorkflow/Home;˜jsessionid=1043AF3CE8F678446D63C7A2B0 50F895

Korkhov V, Vasyunin D, Wibisono A, Guevara-Masis V, Belloum A, de Laat C, Adriaans P, Hertzberger LO (2007a) WS-VLAM: towards a scalable workflow system on the grid. In: WORKS '07: proceedings of the 2nd workshop on workflows in support of large-scale science, pp 63–68. ISBN 978-1-59593-715-5. doi: http://doi.acm.org/10.1145/1273360.1273372

Korkhov V, Wibisono A, Vasyunin D, Belloum ASZ (2007b) Interactive dataflow driven workflow engine. Sci Program 15:173–188

Koulouzis S, Meij E, Marshall MS, Belloum A (2008) Enabling data transport between web services through alternative protocols and streaming. In: ESCIENCE'08: proceedings of the 2008 fourth IEEE international conference on e-Science, pp 400–401. ISBN 978-0-7695-3535-7. doi: http://dx.doi.org/10.1109/eScience.2008.127

Koulouzis S, Zudilova-Seinstra E, Belloum A (2010) Data transport between visualization web Services for medical image analysis. International conference on computational science (ICCS 2010), Proccedia computer Science, Vol1, Issue 1, page 1727-1736:doi: http://dx.doi.org/10.1016/j.procs.2010.04.194

Leguy CA, Bosboom EM, van de Vosse MN (2009) A global sensitivity analysis of a 1D wave propagation model of the arms arterial tree. In: International meeting of the French Society of Hypertension, Paris, France, 17–18 December 2009

Leguy CA, Bosboom EM, Belloum AS, Hoeks AP, van de Vosse FN (2011) Global sensitivity analysis of a wave propagation model for arm arteries. J Med Eng Phys 33:1008–1016

Ludäscher B, Altintas I, Bowers S, Cummings J, Critchlow T, Deelman E, De Roure D, Freire J, Goble C, Jones M, Klasky S, McPhillips T, Podhorszki N, Silva C, Taylor I, Vouk M (2009) Scientific process automation and workflow management. In: Arie Shoshani ,Doron Rotem (eds), Scientific data management, computational science series, chapter 13. Chapman & Hall. http://daks.ucdavis.edu/˜ludaesch/Paper/ch13-preprint.pdf

Mayer A, McGough S, Furmento N, Cohen J, Gulamali M, Young L, Afzal A, Newhouse S, Darlington J (2006) ICENI: an integrated grid middleware to support e-Science. Compon Models Syst Grid Appl 17(4):109–124. ISSN 978-0-387-23351-2 (Print) 978-0-387-23352-9 (Online). doi: 10.1007/0-387-23352-0_7

McClatchey R, Vossen G (1997) Workshop on workflow management in scientific and engineering applications report. SIGMOD Rec 26(4):49–53. ISSN 0163–5808. doi: http://doi.acm.org/10.1145/271074.271087

Miles S, Groth P, Branco M, Moreau L (2007a) The requirements of using provenance in e-science experiments. J Grid Comput 5(1):1–25. ISSN 1570–7873. doi: 10.1007/s10723-006-9055-3

Miles S, Wong SC, Fang W, Groth P, Zauner KP, Moreau L (2007b) Provenance- based validation of e-science experiments. Web Semant 5(1):28–38.ISSN 1570–8268. doi: http://dx.doi.org/10.1016/j.websem.2006.11.003

Oinn T, Greenwood M, Addis MJ, Alpdemir NM, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin DJ, Li P, Lord P, Pocock MR, Senger M, Stevens R, Wipat A, Wroe C (2002) Taverna: lessons in creating a workflow environment for the life sciences. J Concurr Comput: Pract Exp http://eprints.ecssoton.ac.uk/10908/

Olabarriaga SD, Snel JG, Botha CP, Belleman RG (2007) Integrated support for medical image analysis methods: from development to clinical application. IEEE Trans Inf Technol Biomed 11:47–57

Pang L (2001) Understanding virtual organizations. Inf Syst Control J 6:603–610

Peachey, T., Abramson, D., Lewis, A., Kurniawan, D. and Jones, R. "Optimization using Nimrod/O and its Application to Robust Mechanical Design", PPAM 2003, Fifth International Conference on Parallel Processing and Applied Mathematics, Czestochowa, Poland, Lecture Notes in Computer Science, Volume 3019 / 2004, pp. 730–737, ISBN: 3-540-21946-3, September 7–10, 2003

Qin J, Fahringer T, Pllana S (2006) UML based grid workflow modeling under ASKALON. In: proceedings of 6th Austrian-Hungarian workshop on distributed and parallel systems, 21–23 September 2006

Stevens RS, Robinson AJ, Goble CA (2003) MyGrid: personalised bioinformatics on the information grid. Bioinform J 19:302–304

Taylor I (2006) Triana generations. In: e-Science '06: proceedings of the second IEEE international conference on e-Science and grid computing p 143. ISBN 0-7695-2734-5. doi: http://dx.doi.org/10.1109/E-SCIENCE.2006.146

Taylor I, Wang I, Shields M, Majithia S (2005) Distributed computing with Triana on the grid: research articles. concurrency and computation: Pract Exp 17(9):1197–1214.ISSN 1532–0626. doi: http://dx.doi.org/10.1002/cpe.v17:9

Taylor IJ, Deelman E, Gannon BD, Shields M (2007) Workflows for e-science: scientific workflows for grids. Springer, London. ISBN 978-1-84628-519-6

Van Belle J, Shamoun-Baranes J, Van Loon E, Bouten W (2007) An operational model predicting autumn bird migration intensities for flight safety. J Appl Ecol 44:864–874. doi:10.1111/j.1365-2664.2007.01322.x

Von Laszewski G, Hategan M (2005) Workflow concepts of the java CoG Kit. J Grid Comput 3:239–258

Zhang J, Altintas I, Tao J, XLiu X, Pennington DD, Michener WK (2006) Integrating data grid and web services for e-science applications: a case study of exploring species distributions. e-Science, 0:31.doi: http://doi.ieeecomputersociety.org/10.1109/E-SCIENCE.2006.90

Zhao Z, Booms S, Belloum A, de Laat C, Hertzberger LL (2006) VLE-WFBus: a scientific workflow bus for multi e-science domains. In: proceedings of the 2nd IEEE international conference on e-science and grid computing, pp 11–19, 4–6 December 2006

Zhao Z, Belloum A, de Laat C, Adriaans P, Hertzberger LO (2007) Distributed execution of aggregated multi domain workflows using an agent framework. In: Scientific workflow in the proceedings of the IEEE international conference web service, 9 July 2007

Zudilova-Seintra E, Yang N, Axner L, Wibisono A, Vasunin D (2002) Service-oriented visualisation applied to medical data analysis. Future Gen Comput Syst 18(8):1017–1031.ISSN 0167-739X. doi: http://dx.doi.org/10.1016/S0167-739X(02)00082-1

Glossary

PFT (Process Flow Template) is a concept that helps to enforce best practices, and increase the re-usability within a given scientific domain

VO (Virtual Organizations), in Grid computing, refers to a dynamic set of individual and/or institutions defined around a set of resource-sharing rules and conditions

SWMS Scientific Workflow Management System

VL-e (Virtual Laboratory for e-Science) is a Dutch research project with the aim to bridge the gap between the technology push of the high performance networking and the Grid and the application pull of a wide range of scientific experimental applications

tele-science merges advanced solutions for remote instrumentation (via Telemicroscopy), distributed data computation and storage, and transparent access to federated databases of cell structure

MACS (Material Analysis of Complex Surface) experiments try to identify and determine the elements that compose complex surfaces, regardless of the nature of the sample

DAG (Directed Acyclic Graphs) is a directed graph with no directed cycles

MDS (Globus Monitoring and Discovery System)is the information services component of the Globus Toolkit and provides information about the available resources on the Grid and their status

HEFT (Earliest Finish Time) is equal to the Early Start Time of a given task plus the duration of this task

OWT (Olingo Web Tool) is a tool to automate the generation of the PFT data structure

SOAP (Simple Object Access Protocol) is a protocol specification for exchanging structured information in the implementation of Web Services in computer net- works

VRS (Virtual Resource System) a Java API used which provide a single access point to the Grid resources

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6,

WSDL (Web Service Description Language) is an XML format for describing network services as a set of endpoints operating on messages containing either document-oriented or procedure-oriented information

HPC (High Performance Computing) uses supercomputers and computer clusters to solve advanced computation problems

CPU (Central processing unit)the part of a computer (a microprocessor chip) that does most of the data processing

MoC (Models of Computing) is the definition of the set of allowable operations used in computation and their respective costs

MIR myGrid information repository

SWIG (Simplified Wrapper and Interface Generator)s a software development tool that connects programs written in C and C++ with a variety of high-level programming language

RPC (Remote Procedure Call) is an inter-process communication that allows a computer program to cause a subroutine or procedure to execute in another address space (commonly on another computer on a shared network) without the programmer explicitly coding the details for this remote interaction

PCA (Principal component analysis) involves a mathematical procedure that transforms a number of possibly correlated variables into a number of uncorrelated variables called principal components, related to the original variables by an orthogonal transformation

XML (Extensible Markup Language) is a set of rules for encoding documents in machine-readable form

GridFTP (Grid-enabled File Transport Protocol) is an extension of the standard File Transfer Protocol (FTP) for use with Grid computing

LFC (Logical File Catalog) has been developed by LCG (http://lcg.web.cern.ch/LCG/) to resolve problems with the EDG Replica Manager File Catalog

WSRF (Web Service Reference Framework) is a family of OASIS-published specifications for web services. Major contributors include the Globus Alliance and IBM

ASKALON is a Grid application development and computing environment

AWGL (Abstract Grid Workflow Language) for describing Grid workflow applications at a high level of abstraction

JXPL is an XML-based Scripting Language for Workflow Execution in a Grid Environment

GWorkowDL (Generic Workflow Description Language) is a generic description language for workflows in distributed environments. This software package contains the XML Schema as well as Java tools for creating, parsing, and editing GWorkflowDL documents

Karajan is a workflow specification language and execution engine, being developed within the Java CoG Kit

Pegasus project encompasses a set of technologies the help workflow-based applications execute in a number of different environments including desktops, campus clusters, grids, and now clouds

Triana is an open source problem solving environment developed at Cardiff University that combines an intuitive visual interface with powerful data analysis tools

Kepler is dedicated to furthering and supporting the capabilities, use, and awareness of the free and open source, scientific workflow application

ICENI (Imperial College e-Science Networked Infrastructure) is a collection of grid middleware used for providing and coordinating grid services for e-Science applications

VBrowser (Virtual Resource Browser) is intended as single frontend to the Grid. This is the main frontend from the VL-e Toolkit and most users from the VL-e Toolkit will only use this Graphical User Interface to access their Grid resources

Index

I. Roterman-Konieczna (ed.), *Identification of Ligand Binding Site and Protein-Protein Interaction Area*, Focus on Structural Biology 8, DOI 10.1007/978-94-007-5285-6,