10 Homology-Based Modeling of Protein Structure

Zhexin Xiang

10.1 Introduction

10.1.1 Structural Genomics and Homology Modeling

The human genome project has already discovered millions of proteins (http://www.swissprot.com). The potential of the genome project can only be fully realized once we can assign, understand, manipulate, and predict the function of these new proteins (Sanchez and Sali, 1997; Frishman et al., 2000; Domingues et al., 2000). Predicting protein function generally requires knowledge of protein three-dimensional structure (Blundell et al., 1978; Weber, 1990), which is ultimately determined by protein sequence (Anfinsen, 1973). Protein structure determination using experimental methods such as X-ray crystallography or NMR spectroscopy is very time consuming (Johnson et al. 1994). To date, fewer than 2% of the known proteins have had their structures solved experimentally. In 2004, more than half a million new proteins were sequenced that almost doubled the efforts in the previous year, but only 5300 structures were solved. Although the rate of experimental structure determination will continue to increase, the number of newly discovered sequences grows much faster than the number of structures solved (see Fig. 10.1).

Fortunately, many protein sequences are evolutionarily related, and thus can be classified into different families. Proteins in the same families frequently have noticeable similarities and thus share three-dimensional architecture, which allows a structural description of all proteins in a family even when only the structure of a single member is known. This evolutionary relationship provides the rationale for structural genomics, a systematic and large-scale effort toward structural characterization of all proteins, where a representative protein in each family is chosen to be solved experimentally with the rest reliably predicted by a homology modeling method (Goldsmith-Fischman and Honig, 2003; Al-Lazikani et al., 2001a). Fold recognition has also become an important tool that supplements sequence-based methods to detect remote homologues. However, the line between traditional homology modeling and fold recognition has diminished due to the progress in the alignment sensitivity and the increase in database size. *Ab initio* structure methods have made notable progress in recent years and are extremely important, not only for what they can accomplish but also for what they can teach us about protein folding (Bonneau and Baker, 2001). The progress in *ab initio* prediction makes it possible in a few cases to refine homology models to the accuracy of low-resolution X-ray

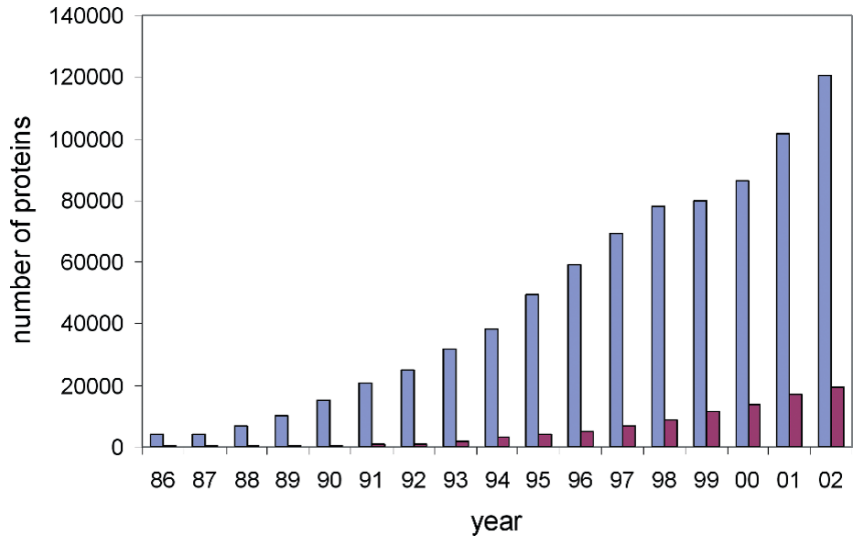

Fig. 10.1 Number of protein sequences and structures available each year. Blue bar denotes the number of protein sequences in SWISS-PROT, red bar is the number of protein structures in PDB.

structures. In fact, if we assume that a native protein structure is at the global free-energy minimum, comparative modeling is a simple scheme to focus the search of conformation space by minimally disturbing those existing solutions, i.e., the experimentally solved structures. The obvious advantage is that the comparative modeling technique relaxes the stringent requirements of force field accuracy and prohibitive conformational searching, because it dispenses with the calculation of a physical chemistry force field and replaces it, in large part, with the counting of identical residues between template and target sequences.

Currently there are about 2 million protein sequences in Swiss-Prot and TrEMBL, but only 7677 protein families have been identified according to the Pfam database (http://pfam.wustl.edu/). This number is strongly dependent on the sequence similarity cutoffs used to cluster the sequence space. If 30% sequence identity cutoff is used, which is generally considered as a threshold for successful homology modeling, statistical estimates place the number somewhere between 10,000 and 30,000 for all proteins in Nature (Liu et al., 2004), but only a fraction of which have distinct spatial arrangements (Brenner et al., 1997). Ninety percent of protein structures deposited today share a similar fold to others already in the PDB. Many of these solved structures are site-directed mutants or inhibitor-bound complexes of previously deposited proteins, and many are related members of families of proteins with similar sequences and closely related three-dimensional structures. Even proteins with no sequence homology can have very similar folds; for example, the sulfate and phosphate binding proteins, the transferrins, and the porphobilinogen deaminase have similar bilobal anion binding structures but not significant sequence identities. Protein topologies such as the four α-helix bundle, the $\alpha\beta$-nucleotide binding motif,

the β-jelly roll, the αβ-barrel, and the β-immunoglobulin domain have been found in a wide range of protein structures (Johnson et al., 1994; Al-Lazikani et al., 1997; Efimov, 1993). A recent study has found that roughly 33% of all proteins have complete sequence coverage to a protein with known structure (Ekman et al., 2005). This kind of protein structure redundancy versus protein sequence variability is the cornerstone of homology modeling algorithms. It is quite likely that homology modeling will assume an increasingly important role in both biological and chemical applications with the advent of structural genomics initiatives around the world.

10.1.2 History of Homology Modeling

Homology modeling techniques became important only after 1990 when hundreds of protein structures had already been deposited in the Protein Data Bank. Early modeling studies in the late 1960s and early 1970s frequently relied on the construction of hand-made wire and plastic models and only later depended on computer software (Cox and Bonanou, 1969; Tometsko, 1970). The first homology model was built simply by copying existing coordinates from a homologous protein and those nonidentical residues were then substituted by reassembling corresponding side chains (Browne et al., 1969). This approach, called rigid-body assembly, is still widely employed today with considerable success, especially when the proteins have sequence identity above 40% (Greer, 1980, 1981). The homology modeling method was pioneered in two studies by Browne et al. (1969) and Greer (1981). Browne et al. (1969) published the first homology model using an X-ray-derived structure as a template. They modeled bovine α-lactalbumin on the three-dimensional structure of hen egg-white lysozyme, where the pair sequence identity was about 39% and only deletions were considered as the polypeptide chain was shortened in α-lactalbumin. Their prediction was proven generally correct later on when the structure of α-lactalbumin was solved (Acharya et al., 1989).

McLachlan and Shotton (1971) modeled alpha-lytic proteinase of Myxobacter 495 based on the structures of both chymotrypsin and elastase, where the sequence identity between these two proteinases was only 18% and the alignment was fragmented by frequent gaps. When the structure of alpha-lytic proteinase was published (Brayer et al., 1979), it became clear that misalignment of the sequence with those of the known 3D structures led to incorrect regions, but portions of both domains were constructed correctly (Delbaere et al., 1979). This model demonstrated the difficulty in aligning sequences of limited similarity and in modeling variable, mainly loop regions.

Greer was the first to demonstrate the importance of modeling variable regions (1981). By abstracting approximate conformations from a family of homologous proteins of known structures, he could distinguish structurally conserved regions, which contain strong sequence homology, and structurally variable regions, which include all the insertions and deletions. By applying the structural distinction to new sequences, erroneous alignments of the sequences are greatly minimized. For each new aligned sequence, the structurally conserved regions can be constructed from

any of the known structures. The construction of variable regions, however, is not straightforward. However, the conformations of loops of just one- or two-residue deletions or insertions can be extrapolated from one of the homologous structures. This approach was further applied to predict the structure of mammalian serine proteases based on a number of proteins from this family, including a variety of blood-serum, intestinal, and pancreatic proteins as well as a closely related bacterial enzyme (Greer, 1981).

Instead of deriving protein backbone structure from only one of its homologues, Taylor (1986) developed a method of generating templates for each part of protein to be modeled based on the conserved patterns observed in the known 3D structures of a family. The conserved templates were derived from a small number of related sequences of the known tertiary structures. The templates were then made more representative by aligning with other sequences of unknown structures. The specificity of the templates was demonstrated by their ability to identify the conserved features in known immunoglobulin and the related sequences but not in other sequences. However, assembling these conserved patterns into a complete structure requires the use of a force field and conformation sampling.

Due to the small number of protein structures available before the 1990s, the comparative modeling technique was not widely successful. The real development of homology modeling began in the mid-1990s with the progress of genome projects and the growth of the number of solved structures in the PDB. With the advent of structural genomics, the importance of homology modeling continues to grow. Although 25 years have passed since Greer's pioneering work on comparative model building of mammalian serine protease in 1981, the basic technique used in today's most advanced modeling programs remains almost the same, i.e., finding the closest homologues as the basis of modeling the query sequence. Recent efforts in comparative modeling have been concentrated on the discovery of distant homologues, the improvement of alignment accuracy, and especially the refinement of models by optimization of empirical energy functions.

10.1.3 Accuracy and Applicability of Homology Modeling

Approximately 57% of all known sequences have at least one domain that is related to at least one protein of known structure (Pieper et al., 2002). The probability of finding a related known structure for a randomly selected sequence from a genome ranges from 30% to 65%, since a few genomes have received more research attention than others (Kelley et al., 2000; Teichmann et al., 1999; Fiser and Sali, 2003). The percentage is steadily increasing because more distinct folds are discovered each year, and because the number of different structural folds that proteins adopt is limited (Irving et al., 2001). Current estimates suggest that there are between 1000 and 5000 folds in the universe of compact globular proteins, with about 200 new folds realized annually from the structure deposition (Brenner et al., 1997). The number of known protein sequences is close to 2 million so far. Over 1.1 million proteins can readily have at least one of their domains reliably predicted with homology modeling

methods. Given the rate of experimental structure determination, approximately 6000 proteins each year, it is arguable that homology modeling has already saved up to hundreds of years of human effort, though homology models often have low quality. In the next 10 years, structural genomics will possibly discover all protein distinct folds in Nature, making comparative modeling applicable to almost any protein sequence (Vitkup et al., 2001). The usefulness of comparative modeling is ever increasing as more proteins can be predicted with higher accuracy. The accuracy of homology modeling depends primarily on the sequence similarity between the target sequence and the template structure.

When the sequence identity is above 40%, the alignment is straightforward, there are not many gaps, and 90% of main-chain atoms can be modeled with an RMSD error of about 1 Å (Sanchez and Sali, 1997). In this range of sequence identity, the structural difference between proteins mainly arises from loops and side chains. When the sequence identity is about 30–40%, obtaining correct alignment becomes difficult, where insertions and deletions are frequent. For sequence similarity in this range, 80% of main-chain backbone atoms can be predicted to RMSD 3.5 Å, while the rest of the residues are modeled with larger errors, especially in the insertion and deletion regions (Harrison et al., 1995; Mosimann et al., 1995; Yang and Honig, 2000; Sauder et al., 2000). Even in correctly aligned regions, loop modeling and side-chain placement pose difficulties (Bower et al., 1997; Rapp and Friesner, 1999). When the sequence similarity is below 30%, the main problem becomes the identification of the homologue structures, and alignment becomes much more difficult. For some sequences where the structures in the family are very conserved in evolution (e.g., kinase family), homology modeling can make predictions as accurate as low-resolution X-ray experiments even if the sequence identity is much less than 30% identity to the template (Yang and Honig, 1999; Petrey et al., 2003).

Even if homology modeling is generally much less accurate than experimental methods, it can still be helpful in proposing and testing hypotheses in molecular biology, such as predictions of ligand binding sites (Zhou and Johnson, 1999; Francoijs et al., 2000), substrate specificities (Jung et al., 2000; De Rienzo et al., 2000), function annotation, protein interaction pathways, and drug design (Nugiel et al., 1995; Sanchez and Sali, 1997). It can also provide starting models for solving structures from X-ray crystallography, NMR, and electron microscopy (Talukdar and Wilson, 1999; Ceulemans and Russell, 2004).

10.2 Procedures in Homology Modeling

Given a protein sequence, successful homology modeling usually consists of the following steps as shown in Fig. 10.2: (1) identify the homologue of known structure from the Protein Data Bank; (2) align the query sequence to the template structure; (3) build the model based on the alignment; (4) assess and refine the model. Each step may involve some errors.

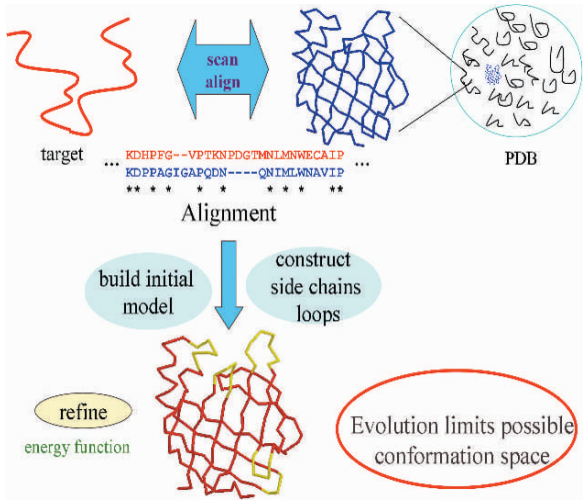

Fig. 10.2 Basic homology modeling protocol. Homology modeling starts by scanning the PDB for sequences similar to the target. The hit of highest sequence similarity is chosen as the template. Model for the target is then built based on the alignment between the target and template sequence, which will be subjected to further refinement.

10.2.1 Homologue Detection and Alignment

Homology modeling starts from selection of homologues with known structures from the PDB. If the query sequence has high sequence identity (>30%) to the structure, the homology detection is quite straightforward which is usually done by comparing the query sequence with all the sequences of the structures in the PDB. This can often be achieved simply with the dynamic programming method (Needleman and Wunsch, 1970) and its derivatives (Smith and Waterman, 1981; Gotoh, 1982). The most popular software is BLAST (Altschul et al., 1997) (http://www.ncbi.nlm.nih.gov/blast/) that searches sequence databases for optimal local alignments to the query. The BLAST program improves the overall speed of searches while retaining good sensitivity by breaking the query and database sequences into fragments, and initially seeking matches between fragments. The matched fragments are then extended in either direction in an attempt to generate an alignment with a score exceeding a particular threshold. The score based on substitution matrices reflects the degree of similarity between the query and the sequence being compared, capable of ranking the quality of each pairwise alignment. The BLAST program functions very well for alignment of sequences with high similarities. But when the sequence identity is well below 30%, homology hits from BLAST are not reliable. A number of alternative strategies have been developed. These include template consensus sequences (Taylor, 1986; Chappey et al., 1991) and profile analysis (Barton and Sternberg, 1990; Suyama et al., 1997; Lolkema and Slotboom, 1998). All these approaches, based

on either multiple sequence or structure alignments, are more sensitive because the consensus sequences are more representative of the sequence family, and the profile reflects the conserved structural or functional preferences.

In the past several years, sequence profile methods have emerged as the primary approach in distant homology detection. Position-specific profile search methods such as PSI-BLAST (Altschul et al., 1997) and hidden Markov models (HMMs) (Krogh et al., 1994), as implemented in the SAM (Karplus et al., 1998) and HMMER (http://hmmer.wustl.edu) packages, have vastly improved the accuracy of sequence alignments and have extended the boundaries of detectable sequence similarity. Sequence profiles methods, e.g., PSI-BLAST, start from performing a pairwise search of the database. The significant alignments are then used by the program to construct a position-specific score matrix (PSSM). This matrix replaces the query sequence in the next round of database searching. The procedure may be iterated until no new significant alignments are found. The profile method can be further improved with information from multiple structure alignment, secondary structure prediction, and solvent accessibility. Since the structural information is more conserved than sequence, it may represent the crucial requirement, in the process of evolution, of residues at specific positions with respect to the stability and function of the structure as a whole. Although a major goal of this effort has been remote homologue detection, an important side benefit has been significant improvement in alignment quality, even at levels of sequence identity for which pairwise alignment methods are known not to work. This, in turn, has had a positive impact on the starting alignments used in homology modeling, and thus has the potential to extend the applicability of homology modeling to increasingly lower levels of sequence similarity. Indeed, perhaps the largest part of recent improvements in homology modeling can be traced directly to improvement in sequence alignment algorithms.

If multiple homologues from the PDB have been identified, the next step is to select one or a few templates that are most appropriate for building a model. Sequence similarity between the query and the template is usually considered as the primary criterion used to choose the best template. Higher sequence similarity often suggests a closer relationship in evolution, thus more conservation of structure, and vice versa. On the other hand, gaps (insertion or deletion) in the alignment have a severe impact on the quality of the model to be built, since gaps are regions where no templates are readily available to guide the model building process. Generally, insertions in the alignment are more difficult to handle than deletions, particularly for insertions of more than 10 residues, because modeling inserted residues is a mini *ab initio* problem. Thus, the second criterion is to choose an alignment that has fewer gaps and short insertions. Moreover, since sequences in the same family often share similar structure and function, templates that can be clustered into the same subfamily as the target are often favored. This can usually be achieved with construction of a multiple alignment and phylogenetic tree (Felsenstein, 1981; Retief, 2000). If the function of the protein to be modeled (target protein) is known, templates with similar functions should also be given more consideration. If possible, the

environment (e.g., ligands, solvent, temperature, pH) at which the template structure was determined and the native environment of the target protein to be modeled should also be properly taken into account. A protein, such as calcium binding protein, can adopt quite different conformations in solvent under different environments (Mishig-Ochiriin et al., 2005). Thus, structures determined in environments similar to the physiological environment of the target are generally preferred. In addition, structures of higher quality are generally used, such as X-ray structures of high resolution and low R-factor, and NMR structures with sufficient constraints. In the end, structures that are most representative of the family should be used if all other criteria are identical. The trait can easily be calculated as the average RMSD of the structure to all other family members with known structures. The hypothesis is that the target protein is more likely to be similar to the most "typical" structure. Instead of relying on a single template, it can be advantageous to select multiple structures from one or several families. Multiple template structures may be aligned with different domains of the target, thus a composite model can be built with each domain based on the best template. It is also useful for modeling variable regions of a structure family, where the segment, which is not conserved, assumes multiple conformations, and the "best" model is assumed to have the lowest value of some empirical energy function. If the sequence identity is too low and there is no clear hit, a better approach is always to make multiple models with each model based on one template. Thus, the best model is determined by physical chemistry- or statistics-based energy or a combination of both (Sippl, 1995; Petrey et al., 2003; also see Chapters 2 and 3).

In homology modeling, one of the most difficult and important tasks is to improve sequence–template alignment. Although profile methods have significantly improved alignment accuracy, manual inspection is often required to further improve the quality if the alignment is well below 30% identity with frequent gaps. This is because the current alignment software usually seeks an alignment of global optimality with an empirical scoring function that may misalign functionally important residues. Manual inspection of the alignment does not necessarily need to have the model actually built, since residue–residue interaction in the target sequence can easily be identified from their corresponding aligned positions in the template structure. There are several general rules to guide alignment tuning. First, charged residues in the target sequence should not be aligned with a buried residue in the template, unless it will form hydrogen bonds or salt bridges with another residue in the target; second, fragments of predicted secondary structures (alpha helix and beta sheet) in the target sequence should be aligned with the fragments of identical secondary structure characterization from the template; third, residues with known important functions, either for protein activity or structural stability, should be aligned with residues of similar functions in the template; fourth, insertions or deletions in the secondary structure regions should be pushed to the loop regions. Manual editing of the alignment is the most tedious part in homology modeling. A misalignment by only one residue position will result in an error of approximately 4 Å in the model because the current homology-modeling algorithms generally cannot recover from errors in the alignment (Fiser and Sali, 2003).

Table 10.1 Comparative modeling programs

Programs	Availability
NEST	http://trantor.bioc.columbia.edu/programs/jackal/
COMPOSER	http://www-cryst.bioc.cam.ac.uk/
Tripos (COMPOSER)	http://www.tripos.com/
CONGEN	http://www.congenomics.com/congen/congen_toc.html
MODELLER	http://guitar.rockefeller.edu/modeller/modeller.html
InsightII (MODELLER)	http://www.accelrys.com/
SWISS-MODEL	http://www.expasy.ch/swissmod/SWISS-MODEL.html
SCHRODINGER	http://www.schrodinger.com
WHATIF	http://swift.cmbi.kun.nl/whatif/
SEGMOD	http://www.bioinformatics.ucla.edu/genemine/
DRAGON	rmunro@nimr.mrc.ac.uk
ICM	http://www.molsoft.com/
3D-JIGSAW	http://www.bmm.icnet.uk/servers/3djigsaw/
Builder	koehl@csb.stanford.edu
PrISM	http://trantor.bioc.columbia.edu/programs/PrISM/index.html

10.2.2 Model Building

Given the alignment between the query sequence and templates, there are generally four methods in model building depending on how the information in the known structures is transferred to the query sequence. In this section we are going to discuss the first three methods, i.e., rigid body assembly, segment matching, and spatial restraint, and leave our own approach (artificial evolution model building) to Section 10.3 for more detailed description. Table 10.1 shows the most widely used model building programs that are publicly available. Most of the programs were based on the rigid body assembly method, and some have been commercialized, e.g., COMPOSER (Sutcliffe et al., 1987a,b) in Tripos and MODELLER (Sali and Blundell, 1993) in InsightII. In addition to model building protocols, the programs also differ from each other in model refinement.

10.2.2.1 Model Building by Rigid Body Assembly

The simplest and most widely used method is called rigid body assembly (also called cut-and-paste method) as shown in Fig. 10.3. This method was initiated by Greer in 1981 and is still widely used, e.g., in the software packages PriSM (Yang and Honig, 1999), Congen (Bruccoleri, 1993), and COMPOSER (Sutcliffe et al., 1987a,b). It starts from identification of the conserved and variable regions of the templates. The identification can often be achieved from the superimposed template structures. Conserved regions are evident from the multiple structure alignment, that is, the RMSDs (root mean square distance) among the fragments are relatively small, and variable regions are usually in loops with frequent gaps in the structural alignment. A

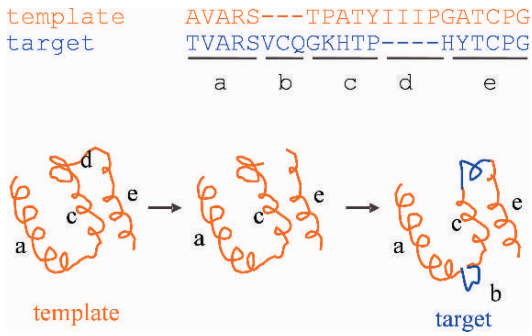

Fig. 10.3 Rigid body assembly with single template. Red and blue denote template and model structure, respectively. Conformations of aligned segments a, c, and e are directly transferred to the model; segment b is a new conformation inserted between a and c, which is obtained from either *ab initio* sampling or database searching; segments c and e are fused following the deletion of segment d.

framework for the superimposed templates can be calculated by averaging the atom coordinates of the structurally conserved regions. The averaging is often weighted based on the sequence similarity of the target sequence to the templates, higher sequence similarity carrying larger weight. The core residues of the target model, i.e., residues aligned with conserved regions of the templates, obtain their main-chain coordinates from the closest conserved segment (in terms of RMSD to the framework) from the template, or from the segment whose template has highest sequence identity to the target. The model is constructed by fitting the core rigid bodies onto the framework. The unconserved, or loop, region is then constructed either with *ab initio* approach, or by searching a database for structures that fit the anchor core regions and have a compatible sequence (Topham et al., 1993). The side chains are modeled based on their intrinsic conformational preferences and on the conformations of the equivalent side chains in the template structures (Sutcliffe et al., 1987a,b). If a single template is chosen, the model construction is straightforward, copying coordinates of aligned residues from the template to the model, and connecting broken segments with database searching or *ab initio* sampling as mentioned earlier. For strong sequence homologues, a single template is sufficient; but for weakly homologous templates, a framework based on weighted averaging over multiple templates is often more reliable.

10.2.2.2 Model Building by Segment Matching

Segment matching (Levitt, 1992), which has been adopted in SegMod software, is based on the finding that most hexapeptide segments of protein structure can be clustered into only 100 structurally different classes (Unger et al., 1989). Homology model construction relies on approximate positions of conserved atoms from the templates as “guiding positions” to calculate the coordinates of other atoms. The guiding positions usually correspond to the atoms of the segments that are conserved

in the alignment between the template structures and the target sequence. They can be calculated by averaging the positions of corresponding atoms in all the template structures with weights based on their sequence similarity to the target. The averaged positions are then fitted by all-atom segments that are obtained by scanning databases of short segments of protein structures, or by a conformation search with restraints of potential energies or geometry rules. The segment matching method can construct both side chain and backbone atoms. If the distance between the conserved positions is too large, there may be no proper segments in the database to cover the missing atoms (usually only segments of five residues have their accessible conformations in the databank) (see Fidelis et al., 1994), thus the *ab initio* method may be the only approach. Segment matching could be considered as an extension of the rigid body assembly method since it scans a database of segments not restricted to those in the template structures. Indeed, segment matching has other applications, such as in side chain and loop modeling, where database scanning, instead of *ab initio* conformation sampling, is employed to identify the best conformers for the prediction.

10.2.2.3 Model Building by Satisfaction of Spatial Restraints

The third group of methods, satisfaction of spatial restraints, was proposed by Havel and Snow (1991) and Sali and Blundell (1993). The method was adopted in one of the most widely used homology model building programs, MODELLER (http://salilab.org). The method starts by generating many restraints for the target protein based on its alignment to the template structure. The restraints are generally obtained by assuming that the distance between two residues in the query model is similar to the distance between the two corresponding aligned residues in the template as shown in Fig. 10.4. The restraints are further supplemented with stereochemical constraints on bond angle, bond length, peptide bond dihedral angle, nonbonded van der Waals clashes, and so on. For weak homologues, additional constraints from experiments, if available, should also be used to increase the model accuracy. This additional information can be obtained from experimental data, for instance, distances between atoms of protein residues as measured by mass spectroscopy (MS) (Chapman, 1996), which uses protein cross-linking reagents as molecular rulers, or by nuclear Overhauser effect (NOE) restraints of NMR spectroscopy. In addition to the hard constraints, a lower and upper bound for each restraint is often provided, with a tight bound for stereochemical restraint and a relatively loose bound for longer distances. The bounds can be best estimated from statistical analysis of the relationships between similar protein structures. By scanning a set of related structures, various correlations can be quantified, such as correlations between two corresponding distances, or between corresponding main-chain dihedral angles. These relationships are expressed as conditional probability density functions and can be used directly as spatial restraints. Probabilities for different values of the equivalent distances and main-chain dihedral angles are calculated from the type of residue considered, from main-chain conformation of an equivalent residue, and from sequence similarity between the two proteins. The spatial restraints and the CHARMM22 force field terms

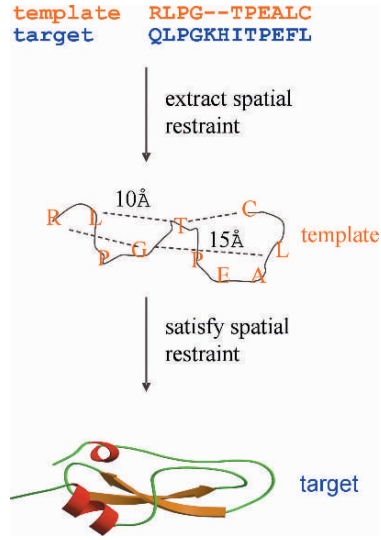

Fig. 10.4 Model building by satisfaction of spatial restraints. Distance restraints to be satisfied by the model are extracted from the template structure based on the alignment between the target sequence and the template. For example, distance between residues L and H (residues 2 and 6) for the model is assumed to be 10 Å, equivalent to the distance between residues L and T (residues 2 and 5) in the template structure.

enforcing proper stereochemistry (Brooks and Karplus, 1983) are combined into an objective function. The model is obtained by optimizing the objective function in Cartesian space. Thus, a proper model should not violate any of the constraints, and have low energy of the objective function. The advantage of the spatial restraint method is that it can use many different types of information about the target sequence including Cα–Cα distance and secondary structure preference. However, for highly homologous sequences, the information is already stored in the template structures, and introducing information derived from other members of the family may degrade the model.

10.2.3 Homology Model Refinement

High-resolution refinement is a difficult task that requires an effective sampling strategy and an accurate energy function. Homology model refinement is primarily focused on tuning alignment and modeling loops and side chains (see Fig. 10.5). Loops are usually the most variable regions of a structure where insertion and deletion often occur. Correct alignment is the most important task for homology modeling, since the errors introduced into the model by misalignment are hard to remove in the later stages of refinement. When the sequence identity is above 40%, errors in the homology structure mainly come from side chains; when the sequence identity is between 30 and 40%, loops and side chains become most problematic. Given a good energy function, loop and side-chain refinement can, in principle, be applied

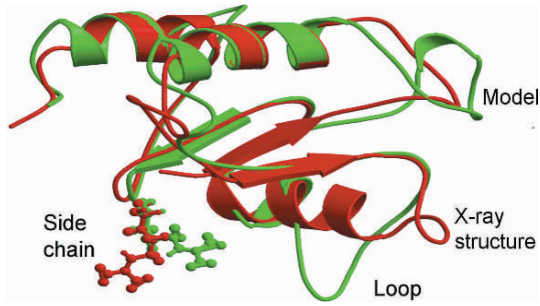

Fig. 10.5 Superimposition of the model for CASP (Critical assessment of techniques for protein structure prediction) Target 113 and its native structure. The root-mean deviation distance is 2.96 Å. The largest RMSDs typically occur in side chains and loops.

repeatedly to relax the backbone closer to native. Refinement on helix and β-sheet can be handled with similar methods as for loops, where proper hydrogen bond constraints should be applied to retain the secondary structure definition (Li et al., 2004). Recent attempts have been made to use physical chemistry energy to refine side chains, loops, and secondary structures, sometimes as a step in choosing the alignment.

10.2.3.1 Loop Prediction

When modeling loops, the basic goal is to predict the conformation of a loop that is fixed at both ends of its protein backbone. A number of methods have been proposed for loop prediction, i.e., *ab initio* methods (Zheng and Kyle, 1996; Rapp and Friesner, 1999; Fiser et al., 2000; Xiang et al., 2002; Jacobson et al., 2004), database-related methods (Li et al., 1999; Wojcik et al., 1999), or a combination of both (Fidelis et al., 1994; van Vlijmen and Karplus, 1997). *Ab initio* methods of loop prediction involve the generation of a large number of randomly chosen candidate conformations and their evaluation with energetic or other criteria. Database methods generate trial conformations based either on sequence relationships to loops of known structure, or on geometric criteria such as the distance between the amino and carboxyl termini of the loop in question. Once loops are generated in this way, energetic criteria are often applied to select the final model.

Clearly, it is important that near-native conformations be present among the trial conformations generated in the first step of loop modeling. Adequate sampling does not appear to be a problem if a large enough number of loops are generated randomly. Indeed, Rapp and Friesner (1999) were able to generate near-native conformations for even a 12-residue loop. However, database methods generate a much smaller number of trial conformations and the lack of a large enough template library to cover the many possible conformations of longer loops (more than five residues not including the stem, or anchoring, residues that are kept fixed) limits their utility for these cases (Fidelis et al., 1994). Using a sequence-dependent database method,

Wojcik et al. (1999) reported an average accuracy of 3.8 Å RMSD for the backbone atoms of an eight-residue loop. Van Vlijmen and Karplus (1997) used CHARMM to optimize initial conformations that were selected from the protein database. They reported improved results for longer loops but their optimization procedure, which involves simulated annealing, effectively extends the range of conformation space searched beyond that provided by the database conformations. In this sense, their approach is closer to *ab initio* loop generation. The accuracy of loop modeling is highly dependent not only on the number of residues in the loop, but also on the distance between the loop stems. Generally, when the distance between the loop stems is shorter, the loop conformation is more like "Ω," and thus has more freedom to move around; therefore, it is more difficult to predict. A database approach is usually more reliable, especially for long loops, if the segment identified from the PDB comes from a protein structure of the same family as the target protein.

Because conformational sampling does not appear to be a problem for loops of less than 12 residues, the quality of the scoring function used to evaluate loop conformations is the major determinant of loop-prediction accuracy. Loop accuracy is usually evaluated in terms of local RMSD (involving the optimal superposition of the predicted and native loop independent of the rest of the structure) or global RMSD (where the RMSD is evaluated with the loop stems kept in place). The latter measure is preferred because the former allows for two loops to be seen as similar, and to have a small RMSD, even if they have very different orientations in the context of the native structure. Rapp and Friesner (1999) used the generalized Born solvation model and the AMBER94 force field to obtain low RMSD values for the two loops they studied. Their approach still needs to be tested on a larger sample size. Fiser et al. (2000) have recently published an extensive *ab initio* study on a data set of 40 loops and also report low RMSD from known structures. Using global RMSD as a criterion, Fiser et al. (2000) reported an accuracy of less than 2 Å for 8-residue loops.

The study of Fiser et al. (2000) utilized a scoring function that included the CHARMM22 force field and statistical preferences taken from protein databases. Scoring functions based entirely on physical chemistry potentials and an accurate solvation model have the potential of identifying the native conformation as lowest in energy, but there are cases where lower energy conformations appear (Smith and Honig, 1994; Steinbach, 2004). One problem may be that most loop prediction approaches seek the lowest energy conformation, thus ignoring conformational entropy effects that will favor broad energy wells. We have recently implemented a procedure called "colony energy" (for detail, see Section 10.3.2) that takes the shape of the energy well into account and yields highly accurate loop prediction (e.g., 1.4 Å global RMSD for eight-residue loops) (Xiang et al., 2002). With crystal environments considered, Jacobson et al. (2004) achieved the best accuracy of 1.0 Å RMSD for eight-residue loops with a computing-intensive approach that combines OPLS all-atom energy function, efficient methods for loop buildup and side-chain optimization, and the hierarchical refinement protocol. Fogolari and Tosatto (2005) demonstrated that molecular mechanics/Poisson–Boltzmann solvent-accessible surface area, if

Table 10.2 Loop modeling program

Programs	Availability
LOOPY	http://trantor.bioc.columbia.edu/programs.html
PLOP	http://francisco.compbio.ucsf.edu/~jacobson/plop_manual/plop_overview.htm
COILS	http://www.ch.embnet.org/software/COILS_form.html
MODELLER (loop module)	http://guitar.rockefeller.edu/modeller/modeller.html
CODA	http://www-cryst.bioc.cam.ac.uk/coda/

combined with the colony energy approach, is very effective in discriminating loop decoys.

Most methodological tests compare predicted loop conformations to known structures, with the backbone conformation of anchoring residues identical to that of the native conformation. This does not properly simulate real modeling conditions under which the backbone of the target protein may not be identical to that of the template. Not surprisingly, loop prediction accuracy degrades as the constraints provided by the loop ends are less accurately defined (Lessel and Schomburg, 1999; Fiser et al., 2000). Table 10.2 shows some loop modeling software that can be easily obtained from the Web. Other loop modeling software only exists as internal components of model building packages listed in Table 10.1. Compared with database scanning methods, most *ab initio* loop prediction programs are very slow.

10.2.3.2 Side-Chain Prediction

The greatest success in the prediction of side-chain conformations has been achieved for core residues where packing constraints significantly simplify the problem. Even for core residues, the accuracy of side-chain prediction degrades when the structure of the backbone is itself not known to a high degree of accuracy. Many side-chain programs are based on rotamer libraries (Ponder and Richard, 1987), which are generally defined in terms of side-chain torsion angles for preferred conformations of a particular side chain. The resolution of rotamer libraries has increased over time and rotamer libraries have been compiled simply by sampling all angles at some given level of resolution (Maeyer et al., 1997). Since backbone conformation changes the frequency of the rotamers, a backbone-dependent rotamer library is often used in side-chain modeling (Dunbrack and Karplus, 1993; Canutescu et al., 2003). The major advantage is to increase computing efficiency, since bad rotamers, e.g., clashing with the backbone, have been automatically removed during construction of the rotamer library. Baker and his co-workers have developed a "solvated rotamer" approach that shows improvement on side-chain packing at protein–protein interfaces (Jiang et al., 2005). This approach extends current side-chain packing methods by using a rotamer library including solvated rotamers with one or more water molecules fixed to polar functional groups in probable hydrogen-bond orientations, together

with a simple energetic description of water-mediated hydrogen bonds. As the number of rotamers increases, however, so does the problem of sampling all possible conformations. There have been a variety of approaches developed to deal with the combinatorial problem in side-chain prediction (Lee and Subbiah, 1991; Lee, 1994; Vasquez, 1996; Dahiyat and Mayo, 1997; Gordon and Mayo, 1999; Samudrala et al., 2000; Kingsford et al., 2005).

Accuracies of about 1 Å RMSD have been reported for core residues in known structures where the backbone has been fixed in the native conformation (Koelh and Delarue, 1994; Vasquez, 1996; Bower et al., 1997; Samudrala and Moult, 1998). A number of studies suggest that further improvements may still be possible. Mendes et al. (1999) found, for example, that the use of an intrinsic torsional potential can improve prediction accuracy. Lovell et al. (2000) reported a novel rotamer library in which internal clashes between side chain and backbone are removed. This library could, in principle, be used to improve prediction accuracy. Xiang and Honig (2001) have shown that using a very detailed rotamer library, which is based on rotamers that use Cartesian coordinates taken from known structures rather than idealized bond lengths and angles, yields RMSD values relative to the native of only 0.62 Å for core residues. This appears to constitute a significant improvement over existing procedures and demonstrates that the combinatorial problem, usually assumed to greatly complicate side-chain prediction, may in fact be of little consequence. This was later confirmed in a more detailed study (Desmet et al., 2002), which showed that local minima for all side-chain prediction may be almost as accurate as the global minimum when evaluated against experimentally determined structures. Improvement on side-chain prediction in recent years has mainly come from better energy functions. Eyal et al. (2004) showed that solvent accessibility and contact surface area are important with regard to the accuracy of side-chain prediction, particularly for modeling buried side chains. Liang and Grishin (2002) have developed a new and simple scoring function for side-chain prediction that consists of the following energy terms: contact surface, volume overlap, backbone dependency, electrostatic interactions, and desolvation energy. The weights of these energy terms were optimized to achieve the minimal average root-mean-square deviation between the lowest energy rotamer and the observed side-chain conformation on a training set of high-resolution protein structures. The derived scoring function combined with a Monte Carlo search algorithm was used to place all side chains onto a protein backbone simultaneously. The average prediction accuracy was 87.9 and 73.2% for the first and second torsion angles correctly predicted to within 40 degrees of native. As is the case for loop prediction, side-chain prediction accuracy depends sensitively on the accuracy to which the backbone conformation is known (Huang et al., 1998). This suggests the possibility of developing procedures where side-chain and backbone conformation can be used iteratively to refine homology models.

Table 10.3 lists some publicly available side-chain prediction programs and the methods they used. Earlier side chain predictions, e.g., RAMP (Samudrala and Moult, 1998), SMD (Tuffery et al., 1993), and CONFMAT (Koelh and Delarue, 1994), were usually based on small rotamer libraries; more recent programs use very

Table 10.3 Side-chain modeling program

Programs	Availability
SCAP	http://trantor.bioc.columbia.edu/programs/jackal/
SCWRL	http://dunbrack.fccc.edu/SCWRL3.php
SMOL	Nikolai.Grichine@UTSouthwestern.Edu
SCCOMP	http://atlantis.weizmann.ac.il/~eyale/
RAMP	http://www.ram.org/computing/ramp/ramp.html
SMD	http://condor.urbb.jussieu.fr/Smd.php
CONFMAT	koehl@csb.stanford.edu
MAXSPROUT	http://www.ebi.ac.uk/maxsprout/

detailed rotamer libraries, e.g., SCAP (Xiang and Honig, 2001), SCWRL (Canutescu et al., 2003), SMOL (Liang and Grishin, 2002). In our recent benchmark study of SCAP, SMOL, and SCWRL, SCAP excelled in prediction for core and surface residues (Xiang et al., to be submitted). For partially buried residues, SMOL performed the best, which was due to its more sufficient conformation sampling and optimized scoring function. SCWRL also performed quite well though not as accurately as the other two, but with much less CPU cost. On a 300-MHz SGI machine, SCWRL is very fast, 3 seconds for each protein, while SMOL needs 11,700 seconds and SCAP needs 361 seconds. Since the test was performed on the native protein backbones, their performance may vary with homology models.

10.2.3.3 Other Improvements to Refinement

Recent improvements to refinement have been mainly achieved by increasing alignment accuracy. Almost all alignment software currently in use has to rely on one of the derivatives of dynamic programming. Although dynamic programming can obtain the global optimal alignment for a given scoring matrix, it cannot account for nonlocal residue–residue interactions. For example, double mutant effects can only be properly estimated if their spatial conformations are both available. As such, a cumbersome but effective method of refining alignment is to build multiple models based on the alternative alignments, with the best alignment corresponding to the model of the lowest physical-chemistry energy. The assumption is that a conformation of lower energy is more likely close to the native state. The method becomes possible due to the availability of more discriminatory energy functions and faster model building tools (Petrey et al., 2003). Tens of thousands of models can be built in a short time with Linux clusters, each based on one variation of alignment. An effective scoring energy can be readily applied to the ensemble of models. The energies of these models can be further minimized with an approach similar to genetic algorithm, i.e., shuffling segments among different models by fixing other parts of protein, where the stems of the segment should have identical residues aligned with the template. Similarly, genetic algorithms are also important tools to increase model quality based on multiple templates. The multiple models, each based on one

template, will be superimposed. Variable regions identified are exchanged and then optimized among different models. In the optimization process, an RMSD restraint can be applied to restrict sampling to a conformational space close to the averaged framework of the original templates. This method has been utilized in the NEST program and produced satisfactory results in CASP6, which will be discussed in more detail in Section 10.3.

Recent research has attempted to use MD simulation to refine models. Lee et al. (2001) used MD simulations with an explicit solvent model to refine Rosetta models followed by scoring with the Poisson–Boltzman/surface area solvation model. Their results showed that native structures could be distinguished energetically from structurally different low-resolution models. Lu and Skolnick (2003) used a combination of local restraints, knowledge-based potentials, and MD approaches that showed promising improvements over previous studies using standard MD methods. Fan and Mark (2004) used classical MD simulations with explicit water to refine homology models. A significant improvement over the model structures has been observed in a number of cases. The results indicate that homology models could be possibly refined with MD simulations on a time scale of tens to hundreds of nanoseconds. Qian et al. (2004) used the principal components of the variation of backbone structures within a homologous family to define a small number of evolutionarily favored sampling directions and showed that model quality can be improved by energy-based optimization along these directions. Li et al. (2004) developed new hierarchical and multiscale algorithms to sample helices and flanking loops, which were evaluated with an all-atom protein force field (OPLS) and a generalized Born continuum solvent model. This method, integrated with a loop and side-chain modeling technique, can potentially be used to refine homology structures iteratively. The next-generation structure modeling algorithms should be able to refine a protein structure closer to the native conformation. The most critical part is to obtain an energy function that is sensitive enough to discriminate near-native conformations from other nonnative folds. Though conformation sampling is also difficult, computer clusters allow more thorough sampling of states around the original models.

10.2.4 Model Assessment

All models built by homology will have errors as discussed in the previous section. Verification of the model, and estimation of the likelihood and magnitude of errors has become one of the most important steps in advancing the state of the art of homology modeling. Errors of the model are usually estimated either from the energy of the model, or from the resemblance of a given characteristic of the model to real structures. The most critical component is the development of a scoring function that is capable of distinguishing good from bad models.

Scoring functions used for the evaluation of protein models generally fall into two broad categories. “Statistical” effective energy functions (Sippl, 1995) are based on the observed properties of amino acids in known structures, and have been widely used in fold recognition and homology modeling applications. A variety of statistical

criteria have been used successfully to discriminate between deliberately misfolded and native structures. Most of them are directly or indirectly based on the analysis of contacts, either interresidue contacts, interatom contacts, or contacts with solvent. For example, preferential distributions of polar and apolar residues inside or outside of a protein can be used to detect completely misfolded models (Baumann et al., 1989); solvation potentials can detect local errors as well as complete misfolds (Holm and Sander, 1992); packing rules have been implemented for structure evaluation (Gregoret and Cohen, 1990). Residue or atom contacts are discriminative because they are energetically favored, and many real structures cannot tolerate too many unfavorable interactions. Thus, for a model to be correct, only a few infrequently observed atomic contacts are allowed. However, bond angles and bond lengths, though powerful in checking the quality of experimental structures, are usually less useful for the evaluation of models because these factors have already been considered appropriately in the model building stage (Fiser and Sali, 2003).

Physical effective energy functions (Lazaridis and Karplus, 1999a) are based on a direct evaluation of the solvation free energy of a protein. It has been demonstrated that such a direct evaluation of the conformational free energy can be at least as successful as statistically based scoring functions in distinguishing the native structure of a protein from an incorrectly folded decoy, although generally at greater computational cost (Janardhan and Vajda, 1998; Vorobjev et al., 1998; Lazaridis and Karplus, 1999b; Petrey and Honig, 2000). A distinct advantage of such physically derived functions is that they are based on well-defined physical interactions, thus making it easier to learn and to gain insight from their performance. Moreover, the success in CASP (Critical Assessment of Protein Structure Prediction) of *ab initio* methods based on purely physical chemistry methods (Lee et al., 1999) suggests that our understanding of the forces that drive protein stability may have reached the point where it can be translated into widely applicable computational tools. One of the major drawbacks of accurate physical chemical description of the folding free energy of a protein is that the treatment of solvation required usually comes at a significant computational expense. Fast solvation models such as the generalized Born (Still et al., 1990) and SCP-ISM (Hassan et al., 2000), together with a variety of simplified scoring schemes (Huang et al., 1995; Petrey and Honig, 2000), may prove to be extremely useful in this regard.

A number of freely available programs can be used to verify homology models as shown in Table 10.4. They generally belong to one of two categories. The first category (e.g., PROCHECK and WHATIF) checks for proper protein stereochemistry, such as symmetry checks, geometry checks (e.g., chirality, bond lengths, bond angles, torsion angles), and structural packing quality; the second category (e.g., VERIFY3D and PROSAII) checks the fitness of sequence to structure, and assigns a score for each residue fitting its current environment. A new graphics software called GRASP2 is also useful in model assessment (Petrey and Honig, 2003). The software can display alignments and template structures simultaneously for assessment of the alignment quality. For example, insertions or deletions can be mapped to the structures to verify that they make sense geometrically. Where residue substitutions

Table 10.4 Model assessment program

Programs	Availability
PROCHECK	http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html
WHATCHECK	http://www.sander.embl-heidelberg.de/whatcheck/
ProSaII	http://www.came.sbg.ac.at
VERIFY3D	http://www.doe-mbi.ucla.edu/Services/Verify_3D/
ERRAT	http://www.doe-mbi.ucla.edu/Services/Errat.html
ANOLEA	http://www.fundp.ac.be/pub/ANOLEA.html
AQUA	http://www.nmr.chem.uu.nl/users/jurgen/Aqua/server/
Probe	http://kinemage.biochem.duke.edu/software/probe.php
SQUID	http://www.ysbl.york.ac.uk/~oldfield/squid/
PROVE	http://www.ucmb.ulb.ac.be/UCMB/PROVE
ProQ	http://www.sbc.su.se/~bjorn/ProQ
GRASP2	http://trantor.bioc.columbia.edu/programs.html

occur, the user can verify that structural features such as hydrophobic packing are maintained and that active-site residues and other features of the target identified from the literature are conserved. The manual inspection should be combined with existing programs to further identify problems in the model.

10.3 Homology Modeling with JACKAL

A new set of homology modeling tools have been developed that are publicly distributed in the JACKAL package (http://trantor.bioc.columbia.edu/programs.html). JACKAL integrates knowledge-based and physics-based methods for protein structure prediction and refinement. At the heart of our approach to structure prediction and refinement is the use of the colony energy concept (see Section 10.3.2). The purpose of JACKAL is to automate the process of structure prediction, from template identification and alignment tuning to model building, refinement and structure verification. JACKAL contains the following major components: NEST for model building and refinement; SCAP for side-chain modeling (Xiang and Honig, 2001); LOOPY for loop prediction (Xiang et al., 2002); AUTOALIGN for alignment tuning; CONREF for model refinement. The core of JACKAL is the NEST program, which, based on our newly developed artificial evolution algorithm (Fig. 10.6), attempts to build models by simulating the natural process of structural evolution from the template structure to the target model. SCAP and LOOPY are used for residue mutation and insertion/deletion, respectively.

10.3.1 Model Building with Artificial Evolution Algorithm

Given an alignment between the query and template sequence, the alignment can be broken down into a list of operations such as residue mutation, insertion, or deletion

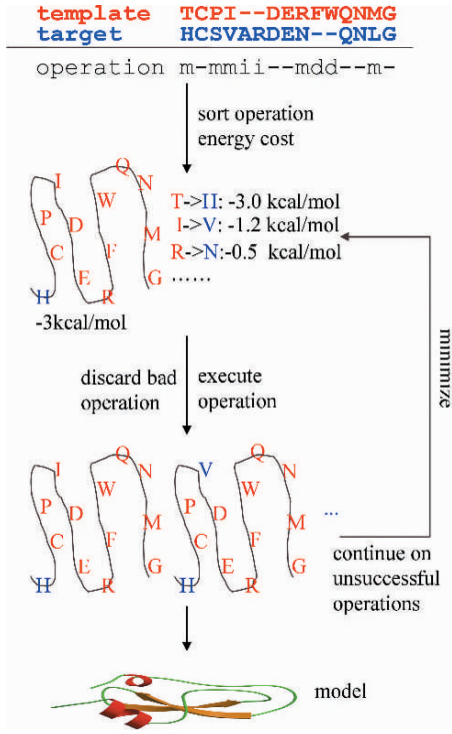

Fig. 10.6 Model building with artificial evolution. m, i, and d denote mutation, insertion, and deletion, respectively.

(Fig. 10.6 shows an alignment of 9 operations, i.e., 5 mutations, 2 insertions, and 2 deletions). Supposing the template to be the "parent structure," it would take Nature billions of years for the template structure to evolve into the target structure. It is unlikely that Nature would finish the daunting task in one step. Instead, a more probable scenario is for Nature to evolve into the target structure via multiple steps with minimal changes to the template structure at each step. Accordingly, building a target model could be considered a process of evolving the template structure based on the alignment so that changes are carried out step by step, each step on one operation. Each operation, i.e., mutation, deletion, or insertion, will disturb the template structure and thus involve an energy cost, either positive or negative. The model building starts from the operation with the least energy cost and so on. Each operation is followed by a slight energy minimization to remove atom clashes. The final structure is then subjected to more thorough energy minimization. The order for the first round of operations does not have to be determined by actually calculating the energy cost for each operation; instead, it can be conveniently estimated empirically. For example, amino-acid mutation is generally easier in evolution than insertion and deletion. As such, mutation operations on residues that are on the protein surface are usually performed first followed by mutation of buried small-sized residues and so on. The operation is considered successful if it does not cause a significant energy penalty

(less than 5 kcal/mol) to the structure; otherwise the operation is discarded and will return to the waiting list. Insertion or deletion of multiple residues is considered as a group of operations, each operating on one residue. The operation starts from the middle residue of the segment with deletion preferred over insertion, since the structural effects of deletion are more reliably predicted. Similarly, operations with more than 5 kcal/mol energy cost (an empirical cutoff that can easily be modified) would also be considered unsuccessful and returned to the waiting list for the next round of operations. The next round of operations actually works on the waiting list, starting from the operation of the least energy cost that has been calculated from the previous round, but with a doubled energy cutoff, e.g., operation with energy penalty more than 10 kcal/mol would be considered unsuccessful for the second round. A number of rounds (less than five rounds in total) would finally accomplish the evolution of the template structure to the model, which will be followed by a series of model refinements.

NEST is heavily dependent on our previous progress in side-chain and loop modeling, i.e., the SCAP (Xiang and Honig, 2001) and LOOPY (Xiang et al., 2002) program. Both SCAP and LOOPY have been integrated into the NEST code, though they also exist independently in the JACKAL package. In the case of mutation, the residue in the template structure first has its side chain changed to the corresponding one in the target sequence, followed by several steps of minimization of the new side chain. We have adopted a simple conformational sampling strategy for side-chain modeling. Side-chain modeling is first carried out with all other parts of protein fixed. The complete rotamer conformations for the side chain, which has been compiled from 646 nonredundant high-resolution protein chains, will be assembled onto the backbone. The rotamer with the lowest colony energy (see Section 10.3.2 for a description of the colony energy concept) will be selected as the final conformation. However, if the rotamer of the lowest conformation energy participates in a hydrogen bond, the conformation energy is used instead of the colony energy because entropic effects generally do not favor hydrogen bonding, and an accurate balance between hydrogen-bonding energy and entropy is difficult to achieve in a simplified force field. If the best rotamer has positive energy, neighboring side chains contacting with the rotamer will then be subjected to minimization. For each of the neighboring side chains including the one that has just been mutated, a similar strategy, that is, sampling all possible rotamers with evaluation based on colony energy, will be performed. The minimization procedure starts from the first residue to the last in the neighboring list until all the side-chain conformations retain the same rotamer on further iteration. If the energy of the side chain for the mutant residue is larger than 5 kcal/mol, the mutation will be considered unsuccessful, thus the mutation operation will be returned to the waiting list, and all other affected residues associated with this operation will be restored to their previous configurations.

An algorithm similar to LOOPY is used to minimize regions affected by insertion or deletion. A segment of five to eight residues that covers the residue under consideration is used in the minimization process. In order not to introduce large

disturbance to the conserved region, the segment window usually slides to one particular direction depending on the location where the residue has to be handled. For example, a larger part of the segment should be assigned to one side of the residue under consideration if it has lower sequence alignment similarity than the other side; it is advisable for the segment to avoid the helix or β-sheet region in order to keep the regular secondary structure intact. If the insertion or deletion is in the helix or β-sheet, the segment will try to cover as many residues as possible in loops. The segment should overlap with at least one residue on either side of the operation. A shorter segment should be used if the loop has fewer than five residues. The segment is refined by sampling alternative conformations. If the operation is in a loop, random conformations would be generated; otherwise, 50% of the conformations would be generated randomly, and the other 50% would be generated that equivalently extends or shortens the regular secondary structures. In other words, conformational sampling is performed with insertion or deletion pushed to the nearest loop region. The backbone of each conformation is minimized using "direct tweak," a novel energy minimization algorithm that minimizes all torsion angle freedoms of a segment without dislodging the end residues. The "direct tweak" algorithm was achieved by combining conventional energy minimization in torsion-angle space with a set of chain-closure constraints that were based on the random tweak algorithm (Shenkin et al., 1987). Pairs of segments with RMSD greater than 2 Å are then combined (i.e., for an eight-residue segment, the first four residues in one segment joined with the last four in another to form a new segment which is fused in the middle with a segment closure procedure) to generate new segments. This results in a set of the original segments plus all the newly fused segments. Side chains are then assembled onto each of the segments, and the colony energy for each segment is calculated. The lowest 30% survive and the procedure is repeated until a single segment remains. In all of the above steps, no more than 200 segments are retained. The operation is successful only when the energy increase of the segment is less than 5 kcal/mol.

10.3.2 Physical-Chemical Energy and Colony Energy Method

The energy function used in NEST can be expressed as the following terms (for more detailed discussion of energy functions, see Chapters 2 and 3):

$$\Delta E = \Delta E_{\text{vw}} + \Delta E_{\text{torsion}} + \Delta E_{\text{hbond}} + \Delta E_{\text{hydro}}, \tag{10.1}$$

$$\Delta E_{\text{vw}} = 61.66\,\eta\,\exp(-2r^2) * (1/r - 1.12/r^{0.5}), \tag{10.2}$$

$$\Delta E_{\text{hbond}} = \min\left(0, [-16 + 12\Omega]\,\cos(\theta_{\text{DHA}})\cos(1.5\,\theta_{\text{HAC}})/d_{\text{HA}}^3\right), \tag{10.3}$$

$$\text{if } 2\,\text{Å} < d_{\text{HA}} < 3\,\text{Å}, \theta_{\text{DHA}} > 90^\circ, \text{ and } \theta_{\text{HAC}} > 60^\circ;$$

$$\text{else } \Delta E_{\text{hbond}} = 0.$$

ΔE_{vw}, $\Delta E_{\text{torsion}}$, ΔE_{hbond}, and ΔE_{hydro} are van der Waals, torsion, hydrogen bond, and hydrophobic energy, respectively. The van der Waals energy is evaluated with a

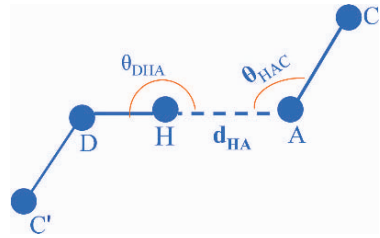

Fig. 10.7 Hydrogen bond. D, A is hydrogen bond donor and acceptor, respectively. H is the proton, and C is carbon atom.

modified expression that fits the CHARMM van der Waals curve but with repulsive term softened to reduce sensitivity to small changes in atomic positions. In Eq. (10.2), η is the energy at the minimum of the potential function and is chosen to correspond to the minimum in the van der Waals potential of the CHARMM22 force field between the two interacting atoms, and r is the ratio of the interatomic distance and the sum of the van der Waals radii of two interacting atoms. The hydrophobic energy is calculated based on the solvent-accessible surface area with the coefficient of 0.025 kcal/mol/Å^2. Here hydrogen-bond energy E_{hbond} was calculated using Eq. (10.3) (see Fig. 10.7), where Ω is the ratio of the solvent-accessible surface area (SASA) of the residue in the protein and the SASA of the same conformation for the residue isolated in solution. D is the hydrogen donor, H is the polar hydrogen, A is the hydrogen acceptor, and C is the carbon atom bonded to A. θ_{DHA} and θ_{HAC} are the angles defined by the coordinates of the respective atoms, and d_{HA} is the distance between atoms H and A. Although the value of θ_{HAC} depends on whether the atomic orbital of the acceptor is sp2 or sp3, the θ_{HAC} angle is nevertheless close to 120°. Since the rotamer library is discretized, we relaxed the standard requirement that θ_{HAC} should be larger than 90° (McDonald and Thornton, 1994). E_{hbond} is defined to assume its minimum value when d_{HA} is 2 Å and θ_{DHA} is 180°. The minimum E_{hbond} values for completely buried and completely exposed side chains are -2 and -0.5 kcal/mol, respectively, representative of experimental data for hydrogen bonds (Efimov and Brazhnikov, 2003).

For each operation (mutation, deletion, and insertion), sufficient conformation sampling is usually performed. The mechanical energy for each sampled conformation is evaluated with Eq. (10.1). NEST does not assume the best prediction to be that of lowest mechanical energy; instead, a new energy term called "colony energy" is used to evaluate all candidates (see Fig. 10.8), and the conformation with the lowest colony energy will be chosen as the prediction (Xiang et al., 2002). For an operation with N sampled conformations, the colony energy of rotamer i, ΔG_i, is calculated as

$$\Delta G_i = -RT * \ln\left[\sum_j \exp(-E_j/(RT) - \beta(\text{RMSD}_{ij}/\text{RMSD}_{\text{avg}})^\gamma)\right], \qquad (10.4)$$

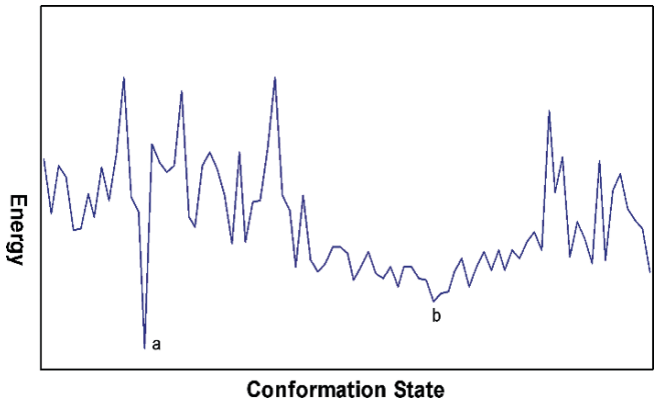

Fig. 10.8 One-dimensional schematic of the sampled conformations. Although conformation a is the global minimum of the mechanical energy, conformation b is structurally similar to many conformations at local minimum, and may possess lower colony energy than conformation a depending on the values of β and γ.

where R is the gas constant, T is absolute temperature, and E_i is the mechanical energy [Eq. (10).1] of the conformation i in the ensemble that has been sampled for the operation. The sum is for all conformations in the ensemble, i.e., j ranges from 1 to N including i. RMSD_{ij} is the root-mean-square distance between conformations i and j. RMSD_{avg} is the average of RMSD between any two conformations in the ensemble for a given operation. The parameter β is set to $-\ln(1/2)$. The ranges of conformational energies and three-dimensional structures sampled in a particular application call for the use of γ values that balance the conformational-energy and RMSD-based factors appropriately. Results obtained with the training data set suggested an optimal value for γ would be 1 and 3 for side-chain and loop modeling, respectively, (Xiang et al., 2002). Equation (10.4) approximates entropic effects by favoring those conformations found in regions of configuration space that are visited most frequently.

10.3.3 Model Refinement with JACKAL

Model refinement in JACKAL is performed in two steps. The first step is to increase alignment quality, and the second step is to directly refine the model itself. A meta server is usually used (http://bioinfo.pl/meta/) to identify as many prospective templates as possible. In the absence of a unanimous template identified by all servers, all possible hits will be considered. For example, if multiple templates are identified but all servers point to the same structural family, all structures in the PDB from that family should be used as possible templates. For each template identified, a number of alignments are obtained either from different servers or from alternative alignments based on a particular alignment protocol. Because model building can be done rapidly using NEST, the ensemble of sequence alignments is readily converted

to an ensemble of 3D model structures. AUTOALIGN can be used on the ensemble of models to repeatedly improve model quality using genetic algorithms. Specifically, all the models are superimposed based on sequence alignment and the regions of high variability are identified. For each model, AUTOALIGN tries all possible conformations on variable regions with corresponding segments from other models. The resultant candidates are then clustered and ranked using the colony energy. The conformation of the lowest colony energy is chosen as the best choice. The process can be repeated until a stable model is derived.

For each model, unaligned regions corresponding to gaps in the sequence alignment are modeled using the independent LOOPY program with a similar approach discussed above but at a more sufficient conformation sampling and energy minimization. Specifically, 2000 initial conformations are randomly sampled and filtered against the consensus secondary-structure predictions from the meta server. The 2000 conformations are then energy-minimized using our fast "direct tweak" method, and the 300 conformations of lowest energy are kept. An additional 300 are obtained from a fragment database using sequence similarity, secondary structure, and end-point geometry. The 600 conformations are subjected to additional energy minimization, and the conformation of lowest colony energy is selected. Side chains are modeled with the independent SCAP program, where the initial conformation starts from the NEST output. The final model will be further optimized using the CONREF module that refines the model with restraints. The restraints include backbone hydrogen bonds and main-chain framework of the template structure, i.e., an energy penalty would be applied if the sampled structure breaks an existing hydrogen bond or deviates significantly (more than 2 Å) from the original model. This is to guarantee that sampling only visits conformations close to the templates.

10.3.4 Comparison with Other Homology Modeling Software

Homology modeling has been widely used in structure prediction, and many homology modeling tools are available (Table 10.1). Given the same alignment and template, it was generally believed there were no major differences between the best modeling programs. However, a recent study by Wallner and Elofsson (2005) has shown that some programs performed better than others. In their study, a benchmark of six different homology modeling programs—MODELLER, SEGMOD/ENCAD (Levitt, 1992), SWISS-MODEL (Schwede et al., 2003), 3D-JIGSAW (Bates et al., 2001), NEST, and BUILDER (Koehl and Delarue, 1996)—is presented. Their study concluded that no single modeling program consistently outperformed the others in all tests. However, it is quite clear that three modeling programs, MODELLER, NEST, and SEGMOD/ENCAD, perform better than the others. Detailed analysis of these homology modeling programs revealed some interesting differences. For example, using a 1.4-GHz AMD XP processor, NEST needs 17 s on average to build a model, while SEGMOD needs 6 s, and MODELLER needs 43 to 430 s in MODELLER6v2 and MODELLER6v2–10, respectively; MODELLER,

SWISS-MODEL, and BUILDER produce more models that do not converge compared to the other programs; in terms of stereochemistry (bond lengths, bond angles, and side-chain planarity), 3D-JIGSAW, BUILDER, and SWISS-MODEL created more residues with bad chemistry for difficult targets, while the other modeling programs showed a fairly constant number of bad residues at all sequence identities. For sequence identities below 40%, all modeling programs manage to bridge some gaps and build some loops correctly or incorrectly; therefore, accordingly, some models are better or worse than the template. In this region the MODELLER programs, NEST, SEGMOD/ENCAD, and SWISS-MODEL, improved 20% of the models. Only NEST rarely made the models worse, while all other programs deteriorated at least 5% of the models. The authors also found that NEST had more of its models "among best" than the other programs; thus, selecting a model from NEST is almost always a good choice.

10.4 Application of Homology Modeling

Homology modeling is often an efficient way to obtain information about proteins of interest. Compared with *ab initio* protein folding, homology modeling is more accurate and reliable. The quality of a homology model is directly correlated with the sequence similarity between target and template. Though a homology model is not perfect, it is still very useful in a wide spectrum of applications where information about 3D conformation of a protein is required.

Highly homologous models with sequence identity above 50% to the templates often have RMSD from the crystal structure around 2 Å, which is roughly comparable to a medium-resolution X-ray structure except for some gapped regions. Models at this level of accuracy can often be used to study a wide range of biological activities that require the knowledge of conformations of individual residues, such as studying catalytic mechanism (Zhou and Johnson, 1999; Francoijs et al., 2000; Xu et al., 2005; Fischer et al., 2005; Kim et al., 2005), designing and improving ligands (Wang and Hampson, 2005; Niv and Weinstein, 2005), predicting protein partners (Orban et al., 2005), solving X-ray structures with molecular replacement (Cupp-Vickery et al., 2003; Schwarzenbacher et al., 2004), refining NMR structures (Skolnick et al., 1997 and Kim et al., 2004) and defining antibody epitopes (Oakhill et al., 2005). In the middle of the accuracy level are the models based on approximately 35% sequence identity, corresponding to 85% of Cα atoms modeled within 3.5 Å of their crystal positions. Though conformations for most side chains have significant errors, fortunately, the active and binding sites are frequently more conserved and are thus modeled more accurately (Sanchez and Sali, 1998; Hassan et al., 2005). Medium-resolution models can be used to improve protein function prediction based on sequence alone (Burley and Bonanno, 2002; Shakhnovich et al., 2003), because ligand binding is more determined by the 3D configurations of active-site residues than by sequence. They can also be used to construct site-directed mutants with

altered or destroyed binding capacity, or design proteins with added disulfide bonds for extra stability, which in turn could test hypotheses about the sequence–structure–function relationships (Ivanenkov et al., 2005; Campillo et al., 2005). For models of low accuracy with sequence identity less than 25%, they sometimes have less than 50% of their Cα atoms within 3.5 Å of their correct positions (Fiser and Sali, 2003). Nevertheless, such models still have the correct fold and even knowing only the fold of a protein is frequently sufficient to predict its approximate biochemical function (Al-Lazikani et al., 2001b). Evaluation of models in this low range of accuracy can be used for confirming or rejecting a match between remotely related proteins (Sanchez and Sali, 1998).

Xu et al. (2005) recently used homology modeling to study HSP90 and kinase Erbb1 interaction. The molecular chaperone Hsp90 modulates the function of specific cell signaling proteins. Although targeting Hsp90 with the antibiotic inhibitor geldanamycin (GA) may be a promising approach for cancer treatment, little is known about the determinants of Hsp90 interaction with its client proteins. Previous studies have shown that Erbb1 binds with HSP90 while Erbb2, having 82% sequence identity to Erbb1, does not. The crystal structure of Erbb1 has been solved to a resolution of 2.6 Å, which was used as the template structure to build the homology model for Erbb2. By superimposing the 3D conformations of Erbb1 and Erbb2, a loop within the N lobe of the kinase domain of Erbb2 was identified that determines Hsp90 binding. Further detailed analysis of the Erbb1 crystal structure and Erbb2 model identified a single residue difference (Gly745 on Erbb1 versus Asp778 on Erbb2) that may account for their different interaction with HSP90. The analysis implied that the amino acid sequence of the loop determines the electrostatic and hydrophobic character of the protein's surface, which in turn governs interaction with Hsp90. The hypothesis was later confirmed by a number of carefully designed mutagenesis experiments.

Another study used low-resolution comparative models to annotate protein functions (Al-Lazikani et al., 2001). Janus kinases (JAKs) are a family of nonreceptor protein tyrosine kinases involved in signaling cascades initiated by various cytokines, interferons, and growth factors (Schindler and Darnell, 1995). There are four human JAK proteins: JAK1–3 and TYK2. JAKs share seven main regions of homology, termed JAK-homology domains JH1–7, numbered from the C to the N terminus. JH1 is the C-terminal protein kinase domain, and JH2 is a kinaselike domain whose precise function remains unclear. JH3–7 play a role in receptor interactions. There has been considerable uncertainty as to whether JAKs contain SH2 domains. Application of homology modeling and other sequence profile analysis method strongly indicates that the Janus family of nonreceptor protein tyrosine kinases contains SH2 domains. One of the Janus kinases, human TYK2, has an SH2 domain that contains a histidine instead of the conserved arginine at the key phosphotyrosine-binding position, βB5. Calculations of the pK_a values of the βB5 arginines in a number of SH2 domains and of the βB5 histidine in a homology model of TYK2 suggest that this histidine is likely to be neutral around pH 7, thus indicating that it may have lost the ability to bind phosphotyrosine.

10.5 Summary

Protein-structure prediction has fascinated the scientific community for decades; it is a problem simple to define but difficult to solve. The dream seems more and more attainable with the explosion of sequence and structural information and because of computational advances in many different areas. These include pure sequence analysis, structure-based sequence analysis, conformational analysis of proteins, and the understanding of the energetic determinants of protein stability. Homology modeling has become a widely used tool, and fold recognition has been shown to extend the limits of detection of sequence search methods. The advent of structural genomic initiatives is certain to spur the development of a host of new computational methods aimed at detecting new relationships between sequence, structure, and function. Continued progress in *ab initio* modeling, combined with ever-increasing databases, makes it possible to further refine homology models to higher accuracy. Such models will provide the basis for a more detailed analysis of structure and function relationships than has been available in the past and will provide powerful tools for the analysis of experimental data and for the design of new experiments.

Despite past progress, much remains to be done. A major problem that still plagues structure prediction by homology is that the structure of the target protein may differ significantly from the closest available template. Unlike the rapid advances made in experimental structure determination, progress in homology structure prediction has been incremental as illustrated at the recent CASP (Critical Assessment of Methods for Structure Prediction of Proteins, http://www.forcasp.org) competitions. Reliability of these homology modeling methods depends critically on the level of sequence identity between the modeling target and the template. When sequence identity is 30% or higher, backbone atoms are usually correctly modeled. The majority of the errors come from side-chain and loop placement during refinement with roughly 3–4 Å RMSD compared to high-resolution crystal structures. When the sequence identity drops below 30%, misalignment happens frequently and model quality suffers dramatically. To increase the utilization and value of the computational models in biomedical research, and to reduce the need for still costly experimental structure determination, significant improvement in the reliability and accuracy of modeling techniques is needed by the research community. There are two immediate goals that have to be addressed in the homology modeling community. The first scientific goal is to expand the modeling coverage to more distantly related proteins that exhibit as low as 10% identity to any known structures. The quality of these models should be close to X-ray structures or high-resolution NMR structures with less than 2 Å RMSD for backbone and side-chain atoms. Significant improvement of modeling methods is needed to push the modeling coverage to remote homologues of existing structures without much compromise on quality. This is both an alignment problem and a refinement problem. Future progress on this issue will depend on advances in the energetic evaluation of structures and the evolutionary analysis of sequences, and the integration of these two fields. The second goal is to achieve the standard of high-resolution X-ray crystal structure quality for

comparative models that are based on known structures with higher homology (30% sequence identity) to the modeling targets. This is predominantly a high-accuracy refinement problem, although substantial improvement of alignment methods is also required. The aim is to acquire the ability to reliably produce computational models with highly accurate placement of both backbone and side-chain atoms, and to significantly reduce the need for experimental structure determinations for close homologues of known structures.

Further Reading

Bates, P. A., Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. 2001. Enhancement of protein modelling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Struct. Funct. Genet. Suppl.* 5:39–46.

Fan, H., and Mark, A.E. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci.* 13:211–220.

Koehl, P., and Delarue, M. 1996. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* 6:222–226.

Levitt, M. 1992. Accurate modeling of protein conformation by automatic segment matching. *J. Mol. Biol.* 226:507.

Li, X., Jacobson, M.P., and Friesner, R.A. 2004. High resolution prediction of protein helix positions and orientations. *Proteins* 55:368–382.

Sali, A., and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 31:3381–3385.

Tang, C.L., Xie, L., Koh, I.Y.Y., Posy, S., Alexov, E., and Honig, B. 2003. On the role of structural information in remote homology detection and sequence alignment: New methods using hybrid sequence profiles. *J. Mol. Biol.* 334:1043–1062.

Xiang, Z.X., Csoto, C., and Honig, B. 2002. Evaluating configurational free energies: The colony energy concept and its application to the problem of protein loop prediction. *Proc. Natl. Acad. Sci. USA* 99:7432–7437.

Xiang, Z.X., and Honig, B. 2001. Extending the accuracy limit of side-chain prediction. *J. Mol. Biol.* 311:421–430.

Acknowledgments

I thank Drs. Peter Steinbach, Cinque Csoto, and Jan Norberg for their many useful comments and critical reading of the manuscript. No endorsement by the U.S. Government should be inferred from the mention of trade names, software packages, commercial products, or organizations.

References

Acharya, K.R., Stuart, D.I., Walker, N.P., Lewis, M., and Phillips, D.C. 1989. Refined structure of baboon alpha-lactalbumin at 1.7 A resolution. Comparison with C-type lysozyme. *J. Mol. Biol.* 208:99–127.

Al-Lazikani, A., Jung, J., Xiang, Z.X., and Honig, B. 2001a. Protein structure prediction. *Curr. Opin. Struct. Biol.* 5:51–56.

Al-Lazikani, B., Lesk, A.M., and Chothia, C. 1997. Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol.* 273:927–948.

Al-Lazikani, B., Sheinerman, F., and Honig, B. 2001b. Combining multiple structure and sequence alignments to improve sequence detection and alignment: Application to the SH2 domains of Janus Kinases. *Proc. Natl. Acad. Sci. USA* 98:14796–14801.

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

Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* 181:223–230.

Barton, G.J., and Sternberg, M.J. 1990. Flexible protein sequence patterns. A sensitive method to detect weak structural similarities. *J. Mol. Biol.* 212:389–402.

Bates, P.A., Kelley, L.A., MacCallum, R.M., and Sternberg, M.J.E. 2001. Enhancement of protein modelling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins Struct. Funct. Genet. Suppl.* 5:39–46.

Baumann, G., Froemmel, C., and Sander, C. 1989. Polarity as a criterion in protein design. *Protein Eng.* 2:329–334.

Blundell, T.L., Bedarkar, S., Rinderknecht, E., and Humble, R.E. 1978. Insulin-like growth factor: A model for tertiary structure accounting for immunoreactivity and receptor binding. *Proc. Natl. Acad. Sci. USA* 75:180–184.

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

Bower, M., Cohen, F.E., and Dunbrack, R.L., Jr. 1997. Homology modeling with a backbone- dependent rotamer library. *J. Mol. Biol.* 267:170–184.

Brayer, G.D., Delbaere, L.T., and James, M.N. 1979. Molecular structure of the alpha-lytic protease from Myxobacter 495 at 2.8 angstroms resolution. *J. Mol. Biol.* 131:743–775.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1997. Population statistics of protein structures: Lessons from structural classifications. *Curr. Opin. Struct. Biol.* 7:369–376.

Brooks, B.R., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamic calculations. *J. Comput. Chem.* 4:187–217.

Browne, W.J., North, A.C.T., Phillips, D.C., Brew, K., Vanaman, T.C., and Hill, R.C. 1969. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* 42:65.

Bruccoleri, R.E. 1993. Application of systematic conformational search to protein modeling. *Mol. Simulat.* 10:151–174.

Burley, S.K., and Bonanno, J.B. 2002. Structuring the universe of proteins. *Annu. Rev. Genomics Hum. Genet.* 3:243–262.

Campillo, N.E., Antonio Paez, J., Lagartera, L., and Gonzalez, A. 2005. Homology modelling and active-site-mutagenesis study of the catalytic domain of the pneumococcal phosphorylcholine esterase. *Bioorg. Med. Chem.* 13:6404–6413.

Canutescu, A.A., Shelenkov, A.A., and Dunbrack, R.L., Jr. 2003. A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci.* 12:2001–2014.

Ceulemans, H., and Russell, R.B. 2004. Fast fitting of atomic structures to low-resolution electron density maps by surface overlap maximization. *J. Mol. Biol.* 338:783–793.

Chapman, J.R. 1996. Mass spectrometry. Ionization methods and instrumentation. *Methods Mol. Biol.* 61:9–28.

Chappey, C., Danckaert, A., Dessen, P., and Hazout, S. 1991. MASH: An interactive program for multiple alignment and consensus sequence construction for biological sequences. *Comput. Appl. Biosci.* 7:195–202.

Cox, R.A., and Bonanou, S.A. 1969. A possible structure of the rabbit reticulocyte ribosome. An exercise in model building. *Biochem. J.* 114:769–774.

Cupp-Vickery, J.R., Urbina, H., and Vickery, L.E. 2003. Crystal structure of IscS, a cysteine desulfurase from Escherichia coli. *J. Mol. Biol.* 330:1049–1059.

Dahiyat, B.I., and Mayo, S.L. 1997. Probing the role of packing specificity in protein design. *Proc. Natl. Acad. Sci. USA* 94:10172–10177.

Delbaere, L.T., Brayer, G.D., and James, M.N. 1979. Comparison of the predicted model of alpha-lytic protease with the x-ray structure. *Nature* 279:165–168.

De Rienzo, F., Fanelli, F., Menziani, M.C., and De Benedetti, P.G. 2000. Theoretical investigation of substrate specificity for cytochromes P450 IA2, P450 IID6 and P450 IIIA4. *J. Comput. Aided. Mol. Des.* 14:93–116.

Desmet, J., Spriet, J., and Lasters, I. 2002. Fast and Accurate Side-chain Topology and Energy Refinement (FASTER) as a new method for protein structure optimization. *Proteins* 48:31–43.

Domingues, F.S., Koppensteiner, W.A., and Sippl, M.J. 2000. The role of protein structure in genomics. *FEBS Lett.* 476:98–102.

Dunbrack, R.L., Jr., and Karplus, M. 1993. Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.* 230:543–574.

Efimov, A.V. 1993. Standard structures in proteins. *Prog. Biophys. Mol. Biol.* 60:201–239.

Efimov, A.V., and Brazhnikov, E.V. 2003. Relationship between intramolecular hydrogen bonding and solvent accessibility of side-chain donors and acceptors in proteins. *FEBS Lett.* 554:389–393.

Ekman, D., Bjorklund, A.K., Frey-Skott, J., and Elofsson, A. 2005. Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *J. Mol. Biol.* 348:231–243.

Eyal, E., Najmanovich, R., McConkey, B.J., Edelman, M., and Sobolev, V. 2004. Importance of solvent accessibility and contact surfaces in modeling side-chain conformations in proteins. *J. Comp. Chem.* 25:712–724.

Fan, H., and Mark, A.E. 2004. Refinement of homology-based protein structures by molecular dynamics simulation techniques. *Protein Sci.* 13:211–220.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17:368–376.

Fidelis, K., Stern, P.S., Bacon, D., and Moult, J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng.* 7:953–960.

Fischer, A.J., Rockwell, N.C., Jang, A.Y., Ernst, L.A., Waggoner, A.S., Duan, Y., Lei, H., and Lagarias, J.C. 2005. Multiple roles of a conserved GAF domain tyrosine residue in cyanobacterial and plant phytochromes. *Biochemistry* 22:15203–15215.

Fiser, A., Gian Do, R., and Sali, A. 2000. Modeling of loops in protein structures. *Protein Sci.* 9:1753–1773.

Fiser, A., and Sali, A. 2003. Comparative protein structure modeling. In *Protein Structure* (D. Chasman, Ed.) New York, Dekker, pp. 167–206.

Fogolari, F., and Tosatto, S.C. 2005. Application of MM/PBSA colony free energy to loop decoy discrimination: Toward correlation between energy and root mean square deviation. *Protein Sci.* 14:889–901.

Francoijs, C.J., Klomp, J.P., and Knegtel, R.M. 2000. Sequence annotation of nuclear receptor ligand-binding domains by automated homology modeling. *Protein Eng.* 13:391–394.

Frishman, D., Goldstein, R.A., and Pollock, D.D. 2000. Protein evolution and structural genomics. *Pac. Symp. Biocomput.* 12:3–5.

Goldsmith-Fischman, S., and Honig, B. 2003. Structural genomics: Computational methods for structure analysis. *Protein Sci.* 12:1813–1821.

Gordon, D.B., and Mayo, S.L. 1999. Branch-and-terminate: A combinatorial optimization algorithm for protein desgin. *Structure Fold. Des.* 7:1089–1098.

Gotoh, O. 1982. An improved algorithm for matching biological sequences. *J. Mol. Biol.* 162:705–708.

Greer, J. 1980. Model for haptoglobin heavy chain based upon structural homology. *Proc. Natl. Acad. Sci. USA* 77:3393–3397.

Greer, J. 1981. Comparative model-building of the mammalian serine protease. *J. Mol. Biol.* 153:1027.

Gregoret, L.M., and Cohen, F.E. 1990. Novel method for the rapid evaluation of packing in protein structures. *J. Mol. Biol.* 211:959–974.

Harrison, R.W., Chatterjee, D., and Weber, I.T. 1995. Analysis of six protein structures predicted by comparative modeling techniques. *Proteins* 23:463–671.

Hassan, S.A., Gracia, L., Vasudevan, G., and Steinbach, P.J. 2005. Computer simulation of protein–ligand interactions: Challenges and applications. *Methods Mol. Biol.* 305:451–492.

Hassan, S.A., Guarnieri, F., and Mehler, E.L. 2000. A general treatment of solvent effects based on screened coulomb potentials *J. Phys. Chem. B* 104:6478.

Havel, T.F., and Snow, M.E. 1991. A new method for building protein conformations from sequence alignments with homologues of known structure. *J. Mol. Biol.* 217:1–7.

Holm, L., and Sander, C. 1992. Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* 225:93–105.

Huang, E.S., Koehl, P., Levitt, M., Pappu, R.V., and Ponder, J.W. 1998. Accuracy of side-chain prediction upon near-native protein backbones generated by ab initio folding methods. *Proteins* 33:204–217.

Huang, E., Subbiah, S., and Levitt, M. 1995. Recognizing native folds by the arrangement of hydrophobic and polar residues. *J. Mol. Biol.* 252:709–720.

Irving, J.A., Whisstock, J.C., and Lesk, A.M. 2001. Protein structural alignments and functional genomics. *Proteins* 42:378–382.

Ivanenkov, V.V., Meller, J., and Kirley, T.L. 2005. Characterization of disulfide bonds in human nucleoside triphosphate diphosphohydrolase 3 (NTPDase3): Implications for NTPDase structural modeling. *Biochemistry* 44:8998–9012.

Jacobson, M.P., Pincus, D.L., Rapp, C.S., Day, T.J.F., Honig, B., Shaw, D.E., and Friesner, R.A. 2004. A hierarchical approach to all-atom loop prediction. *Proteins: Struct. Funct. Genet.* 55:351–367.

Janardhan, A., and Vajda, S. 1998. Selecting near-native conformations in homology modeling: The role of molecular mechanics and solvation terms. *Protein Sci.* 7:1772–1780.

Jiang, L., Kuhlman, B., Kortemme, T.A., and Baker, D. 2005. A "solvated rotainer" approach to modeling water-mediated hydrogen bonds at protein–protein interfaces. *Proteins* 58:893–904.

Johnson, M.S., Srinivasan, N., Sowdhamini, R., and Blundell, T.L. 1994. Knowledge-based protein modeling. *Crit. Rev. Biochem. Mol. Biol.* 29:1–68.

Jung, J.W., An, J.H., Na, K.B., Kim, Y.S., and Lee, W. 2000. The active site and substrates binding mode of malonyl-CoA synthetase determined by transferred nuclear Overhauser effect spectroscopy, site-directed mutagenesis, and comparative modeling studies. *Protein Sci.* 9:1294–1303.

Karplus, K., Barrett, C., and Hughey, R. 1998. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* 14:846–856.

Kelley, L.A., MacCallum, R.M., and Sternberg, M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299:499–520.

Kim, D.E., Chivian, D., and Baker, D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* 32:W526–531.

Kim, C.G., Watts, J.A., and Watts, A. 2005. Ligand docking in the gastric H(+)/K(+)-ATPase: Homology modeling of reversible inhibitor binding sites. *J. Med. Chem.* 48:7145–7152.

Kingsford, C.L., Chazelle, B., and Singh, M. 2005. Solving and analyzing side-chain positioning problems using linear and integer programming. *Bioinformatics* 21:1028–1036.

Koelh, P., and Delarue, M. 1994. Application of a self-consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* 239:249–275.

Koehl, P., and Delarue, M. 1996. Mean-field minimization methods for biological macromolecules. *Curr. Opin. Struct. Biol.* 6:222–226.

Krogh, A., Brown, M., Mian, I., Sjolander, K., and Haussler, D. 1994. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* 235:1501–1531.

Lazaridis, T., and Karplus, M. 1999a. Effective energy function for proteins in solution. *Proteins* 35:133–152.

Lazaridis, T., and Karplus, M. 1999b. Discrimination of the native from misfolded protein models with an energy function including implicit solvation. *J. Mol. Biol.* 288:477–487.

Lee, C. 1994. Predicting protein mutant energetics by self-consistent ensemble optimization. *J. Mol. Biol.* 236:918–939.

Lee, C., and Subbiah, S. 1991. Prediction of protein side-chain conformation by packing optimization. *J. Mol. Biol.* 217:373–388.

Lee, J., Liwo, A., Ripoll, D., Pillardy, J., and Scheraga, H. 1999. Calculation of protein conformation by global optimization of a potential energy function. *Proteins* 37 (Suppl. 3):204–208.

Lee, M.R., Baker, D., and Kollman, P.A. 2001. 2.1 and 1.8 Å average CαRMSD structure predictions on two small proteins, HP-36 and S15. *J. Am. Chem. Soc.* 123:1040–1046.

Lessel, U., and Schomburg, D. 1999. Importance of anchor group positioning in protein loop prediction. *Proteins* 37:56–64.

Levitt, M. 1992. Accurate modeling of protein coformation by automatic segment matching. *J. Mol. Biol.* 226:507.

Li, W., Liu, Z., and Lai, L. 1999. Protein loops on structurally similar scaffolds: Database and conformational analysis. *Biopolymers* 49:481–495.

Li, X., Jacobson, M.P., and Friesner, R.A. 2004. High resolution prediction of protein helix positions and orientations. *Proteins* 55:368–382.

Liang, S.D., and Grishin, N.V. 2002. Side-chain modeling with an optimized scoring function. *Protein Sci.* 11:322–331.

Liu, X., Fan, K., and Wang, W. 2004. The number of protein folds and their distribution over families in nature. *Proteins* 54:491–499.

Lolkema, J.S., and Slotboom, D.J. 1998. Estimation of structural similarity of membrane proteins by hydropathy profile alignment. *Mol. Membr. Biol.* 15:33–42.

Lovell, S.C., Word, J.M., Richardson, J.S., and Richardson, D.C. 2000. The penultimate rotamer library. *Proteins* 40:389–408.

Lu, H., and Skolnick, J., 2003. Application of statistical potentials to protein structure refinement from low resolution *ab initio* models. *Biopolymers* 70:575–584.

Maeyer, M.D., Desmet, J., and Lasters, I. 1997. All in one: A highly detailed rotamer library improves both accuracy and speed in the modelling of sidechains by dead-end elimination. *Fold. Des.* 2:53–66.

McDonald, I.K., and Thornton, J.M. 1994. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* 238:777–793.

McLachlan, A.D., and Shotton, D.M. 1971. Structural similarities between alpha-lytic protease of Myxobacter 495 and elastase. *Nat. New. Biol.* 229:202–205.

Mendes, J., Baptista, A., Carrondo, M., and Soares, C.M. 1999. Improvement of side-chain modeling in proteins with the self-consistent mean field theory method based on an analysis of the factors influencing prediction. *Biopolymers* 50:111–131.

Mishig-Ochiriin, T., Lee, C.H., Jeong, S.Y., Kim, B.J., Choi, C.H., Yim, H.S., and Kang, S.O. 2005. Calcium-induced conformational changes of the recombinant CBP3 protein from Dictyostelium discoideum. *Biochim. Biophys. Acta* 1748:157–164.

Mosimann, S., Meleshko, R., and James, M.N. 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins* 23:301–317.

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

Niv, M.Y., and Weinstein, H. 2005. Flexible docking procedure for the exploration of peptide binding selectivity to known structures and homology models of PDZ domains. *J. Am. Chem. Soc.* 127:14072–14079.

Nugiel, D.A., Voss, M.E., Brittelli, D.R., and Calabrese, J.C. 1995. An approach to the design of novel cognitive enhancers using molecular modeling and X-ray crystallography. *Drug Des. Discov.* 12:289–295.

Oakhill, J.S., Sutton, B.J., Gorringe, A.R., and Evans, R.W. 2005. Homology modelling of transferrin-binding protein A from Neisseria meningitidis. *Protein Eng. Des. Sel.* 18:221–228.

Orban, T., Kalafatis, M., and Gogonea, V. 2005. Completed three-dimensional model of human coagulation factor va. Molecular dynamics simulations and structural analyses. *Biochemistry* 44:13082–13090.

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

Petrey, D., and Honig, B. 2003. GRASP2: Visualization, surface properties, and electrosttics of macromolecular structures and sequences. *Methods Enzymol.* 374:492–509.

Petrey, D., Xiang, X., Tang, C.L., Xie, L., Gimpelev, M., Mitors, T., Soto, C.S., Goldsmith-Fischman, S., Kernytsky, A., Schlessinger, A., Koh, I.Y.Y., Alexov, E., and Honig, B. 2003. Using multiple structure alignments, fast model

building, and energetic analysis in fold recognition and homology modeling. *Proteins Struct. Funct. Genet.* 53:430–435.

Pieper, U., Eswar, N., Ilyin, V.A., Stuart, A., and Sali, A. 2002. ModBase, a database of annotated comparative protein structure models. *Nucleic Acids Res.* 30:255–259.

Ponder, J.W., and Richard, F.M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequence for different structure classes. *J. Mol. Biol.* 193:775–791.

Qian, B., Ortiz, A.R., and Baker, D. 2004. Improvement of comparative model accuracy by free-energy optimization along principal components of natural structural variation. *Proc. Natl. Acad. Sci. USA* 101:15346–15351.

Rapp, C.S., and Friesner, R.A. 1999. Prediction of loop geometries using a generalized Born model of solvation effects. *Proteins* 35:173–183.

Retief, J.D. 2000. Phylogenetic analysis using PHYLIP. *Methods Mol. Biol.* 132:243–258.

Sali, A., and Blundell, T.L. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234:779–815.

Samudrala, R., Huang, E.S., Koehl, P., and Levitt, M. 2000. Constructing side chains on near-native main chains for *ab initio* protein structure prediction. *Protein Eng.* 7:453–457.

Samudrala, R., and Moult, J. 1998. Determinants of side chain conformational preferences in protein structures. *Protein Eng.* 11:991–997.

Sanchez, R., and Sali, A. 1997. Comparative protein structure modeling as an optimization problem. *J. Mol. Struct. (Theochem)* 398–399:489–496.

Sanchez, R., and Sali, A. 1998. Large-scale protein structure modeling of the Saccharomyces cerevisiae genome. *Proc. Natl. Acad. Sci. USA* 95:13597–13602.

Sauder, J.M., Arthur, J.W., and Dunbrack, R.L., Jr. 2000. Large-scale comparison of protein sequence alignment algorithms with structure alignments. *Proteins* 40:6–22.

Schindler, C., and Darnell, J.E., Jr. 1995. Transcriptional responses to polypeptide ligands: The JAK-STAT pathway. *Annu. Rev. Biochem.* 64:621–651.

Schwarzenbacher, R., Godzik, A., Grzechnik, S.K., and Jaroszewski, L. 2004. The importance of alignment accuracy for molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.* 60(Pt. 7):1229–1236.

Schwede, T., Kopp, J., Guex, N., and Peitsch, M.C. 2003. SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* 31:3381–3385.

Shakhnovich, B.E., Harvey, J.M., Comeau, S., Lorenz, D., DeLisi, C., and Shakhnovich, E. 2003. ELISA: Structure function inferences based on statistically significant and evolutionarily inspired observations. *BMC Bioinformatics* 4:34.

Shenkin, P.S., Yarmush, D.L., Fine, R.M., Wang, H.J., and Levinthal, C. 1987. Predicting antibody hypervariable loop conformation. I. Ensembles of random conformations for ringlike structures. *Biopolymers* 26:2053–2085.

Sippl, M. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235.

Skolnick, J., Kolinski, A, and Ortiz, A.R. 1997. MONSSTER: A method for folding globular proteins with a small number of distance restraints. *J. Mol. Biol.* 265:217–241.

Smith, K.C., and Honig, B. 1994. Evaluation of the conformational free energies of loops in proteins. *Proteins* 18:119–132.

Smith, T.F., and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195–197.

Steinbach, P.J. 2004. Exploring peptide energy landscapes: A test of force fields and implicit solvent models. *Proteins* 57:665–677.

Still, W., Tempczyk, A., Hawley, R., and Hendrickson, T. 1990. Semi-analytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* 112:6127–6129.

Sutcliffe, M.J., Haneef, I., Carney, D., and Blundell, T.L. 1987a. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng.* 1:377–384.

Sutcliffe, M.J., Hayes, F.R., and Blundell, T.L. 1987b. Knowledge based modelling of homologous proteins, Part II: Rules for the conformations of substituted sidechains. *Protein Eng.* 1:385–392.

Suyama, M., Matsuo, Y., and Nishikawa, K. 1997. Comparison of protein structures using 3D profile alignment. *J. Mol. Evol.* 44 (Suppl. 1):S163–173.

Talukdar, A.S., and Wilson, D.L. 1999. Modeling and optimization of rotational C-arm stereoscopic X-ray angiography. *IEEE Trans. Med. Imaging.* 18:604–616.

Taylor, W.R. 1986. Identification of protein sequence homology by consensus template alignment. *J. Mol. Biol.* 188:233–258.

Teichmann, S.A., Chothia, C., and Gerstein, M. 1999. Advances in structural genomics. *Curr. Opin. Struct. Biol.* 9:390–399.

Tometsko, A.M. 1970. Computer approaches to protein structure. II. Model building by computer. *Comput. Biomed. Res.* 3:690–698.

Topham, C.M., McLeod, A., Eisenmenger, F., Overington, J.P., Johnson, M.S., and Blundell, T.L. 1993. Fragment ranking in modelling of protein structure. Conformationally constrained environmental amino acid substitution tables. *J. Mol. Biol.* 229:194–220.

Tuffery, P., Etchebest, C., Hazout, S., and Lavery, R. 1993. A critical comparison of search algorithms applied to the optimization of protein side-chain conformations. *J. Comput. Chem.* 14:790–798.

Unger, R., Harel, D., Wherland, S., and Sussman, J.L. 1989. A 3-D building blocks approach to analyzing and predicting structure of proteins. *Proteins* 5:355–373.

Van Vlijmen, H.W., and Karplus, M. 1997. PDB-based protein loop prediction: Parameters for selection and methods for optimization *J. Mol. Biol.* 267:975–1001.

Vasquez, M. 1996. Modeling side-chain conformation. *Curr. Opin. Struct. Biol.* 6:217–221.

Vitkup, D., Melamud, E., Moult, J., and Sander, C. 2001. Completeness in structural genomics. *Nat. Struct. Biol.* 8:559–566.

Vorobjev, Y., Almagro, J., and Hermans, J. 1998. Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. *Proteins* 32:399–413.

Wallner, B., and Elofsson, A. 2005. All are not equal: A benchmark of different homology modeling programs. *Protein Sci.* 14:1315–1327.

Wang, M., and Hampson, D.R. 2005. An evaluation of automated in silico ligand docking of amino acid ligands to Family C G-protein coupled receptors. *Bioorg. Med. Chem.* 14:2030–2039.

Weber, I.T. 1990. Evaluation of homology modeling of HIV protease. *Proteins* 7:172–184.

Wojcik, J., Mornon, J.P., and Chomilier, J. 1999. New efficient statistical sequence-dependent structure prediction of short to medium-sized protein loops based on an exhaustive loop classification. *J. Mol. Biol.* 289:1469–1490.

Xiang, Z.X., Csoto, C., and Honig, B. 2002. Evaluating configurational free energies: The colony energy concept and its application to the problem of protein loop prediction. *Proc. Natl. Acad. Sci.* 99:7432–7437.

Xiang, Z.X., and Honig, B. 2001. Extending the accuracy limit of side-chain prediction. *J. Mol. Biol.* 311:421–430.

Xiang, Z., Steinbach, P., Jacobson, M.P., Friesner, R.A., and Honig, B. Prediction of side-chain conformations on protein surfaces (to be submitted).

Xu, W.P., Yuan, X.T., Xiang, Z.X., Mimnaugh, E., Marcu, M., and Neckers, L. 2005. Surface charge and hydrophobicity determine ErbB2 binding to the Hsp90 chaperone complex. *Nat. Struct. Mol. Biol.* 12:120–126.

Yang, A.S., and Honig, B. 1999. Sequence to structure alignment in comparative modeling using PrISM. *Proteins* 37(S3):66–72.

Yang, A.S., and Honig, B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.* 301:665–678.

Zheng, Q., and Kyle, D.J. 1996. Accuracy and reliability of the scaling-relaxation method for loop closure: An evaluation based on extensive and multiple copy conformational samplings. *Proteins* 24:209–217.

Zhou, Y., and Johnson, M.E. 1999. Comparative molecular modeling analysis of 5-amidinoindole and benzamidine binding to thrombin and trypsin: Specific H-bond formation contributes to high 5-amidinoindole potency and selectivity for thrombin and factor Xa. *J. Mol. Recognit.* 12:235–241.