1 A Historical Perspective and Overview of Protein Structure Prediction

John C. Wooley and Yuzhen Ye

1.1 Introduction

Carrying on many different biological functions, proteins are all composed of one or more polypeptide chains, each containing from several to hundreds or even thousands of the 20 amino acids. During the 1950s at the dawn of modern biochemistry, an essential question for biochemists was to understand the structure and function of these polypeptide chains. The sequences of protein, also referred to as their *primary structures*, determine the different chemical properties for different proteins, and thus continue to captivate much of the attention of biochemists. As an early step in characterizing protein chemistry, British biochemist Frederick Sanger designed an experimental method to identify the sequence of insulin (Sanger et al., 1955). He became the first person to obtain the primary structure of a protein and in 1958 won his first Nobel Price in Chemistry. This important progress in sequencing did not answer the question of whether a single (individual) protein has a distinctive shape in three dimensions (3D), and if so, what factors determine its 3D architecture. However, during the period when Sanger was studying the primary structure of proteins, American biochemist Christian Anfinsen observed that the active polypeptide chain of a model protein, bovine pancreatic ribonuclease (RNase), could fold spontaneously into a unique 3D structure, which was later called *native conformation* of the protein (Anfinsen et al., 1954). Anfinsen also studied the refolding of RNase enzyme and observed that an enzyme unfolded under extreme chemical environment could refold spontaneously back into its native conformation upon changing the environment back to natural conditions (Anfinsen et al., 1961). By 1962, Anfinsen had developed his theory of protein folding (which was summarized in his 1972 Nobel acceptance speech): "The native conformation is determined by the totality of interatomic interactions and hence, by the amino acid sequence, in a given environment."

Anfinsen's theory of protein folding established the foundation for solving the protein structure prediction problem, i.e., for predicting the native conformation of a protein from its primary sequence, because all information needed to predict the native conformation is encoded in the sequence. The early approaches to solving this problem were based solely on the thermodynamics of protein folding. Scheraga and his colleagues applied several computer searching techniques to investigate the

free energy of numerous local minimum energy conformations in an attempt to find the global minimum conformation, i.e., the thermodynamically most stable conformation of the protein (Gibson and Scheraga, 1967a,b; Scott et al., 1967). The major challenge for an energy minimization approach to protein structure prediction is that proteins are very flexible; thus, their potential conformation space is too large to be enumerated. [Despite the huge space of possible conformations, that proteins fold reliably and quickly to their native conformation is known as "Levinthal's paradox" (Levinthal, 1968)]. To address this issue, one needs an accurate energy function to compute the energy for a given protein conformation and a rapid computer searching algorithm. The progress of peptide molecular mechanics enabled the development of molecular force fields that described the physical interactions between atoms using Newton's equations of motion. In general, the interactions considered in the force field include covalent bonds and noncovalent interactions, such as electrostatic interactions, the van der Waals interactions, and, sometimes, hydrogen bonds and hydrophobic interactions. The parameters used in these force fields were obtained through experimental studies of small organic molecules. On the other hand, many computational methods developed in the field of optimization theory and mechanics have been applied to the rapid conformation search. These fall into two categories: the molecular dynamics method and the Brownian dynamics (or stochastic dynamics) method. Both methods sample a portion of potential protein conformations and evaluate their free energy. Molecular dynamics samples the conformations by simulating the protein motion based on Newton's equation, starting from an arbitrarily chosen protein conformation. Brownian dynamics, instead, uses Monte Carlo random sampling technique or its derivatives to evaluate protein conformations. Combining various force fields and conformation searching methods, many software packages were developed, such as AMBER (Pearlman et al., 1995), CHARMM (Brooks et al., 1983) and GROMOS (van Gunsteren and Berendsen, 1990), all aimed at using computing simulations to predict the native conformation of proteins.

Despite the great theoretic interest in energy minimization methods, these have not been very successful in practice, because of the huge search space for potential protein conformations. In 1975, Levitt and Warshel used a simplified protein structure representation and successfully folded a small protein [bovine pancreatic trypsin inhibitor, (BPTI), 58 amino acid residues] into its native conformation from an open-chain conformation using energy minimization (Levitt and Warshel, 1975). Little progress, however, has been made since then; the simulation usually takes an unrealistic compute or run time, and the final prediction is not very satisfactory. For instance, in 1998, Duan and Kollman reported a simulation experiment of one small protein (the villin headpiece subdomain, 36 amino acid residues), running on a Cray T3D and then a Cray T3E supercomputer, that took months of computation with the entire machine dedicated to the problem (Duan and Kollman, 1998). Even though the resulting structure is reasonably folded and shows some resemblance to the native structure, the simulated and native structure did not completely match. Currently, energy minimization methods are largely used to refine a low-resolution initial structure obtained by experimental methods or by comparative modeling (Levitt and Lifson, 1969).

At nearly the same time as these energy minimization approaches were developed, computational biochemists were looking for practical approaches to the protein structure prediction problem, which need not and presumably does not "mimic" the protein folding process inside the cell. An important observation was that proteins that share similar sequences often share similar protein structures. Based on this concept, Browne and co-workers modeled the structure of α-lactalbumin using the X-ray structure of lysozyme as a template (Browne et al., 1969). This success opened the whole new area of protein structure prediction that came to be known as *comparative modeling* or *homology modeling*. Many automatic computer programs and molecular graphics tools were developed to speed up the modeling. The potential targets of homologous modeling were also expanded through the rapid development of homologous modeling software and approaches. New technologies, including threading or the assembly of minithreaded fragments, were proposed and have now been successfully applied to many cases for which the target modeled does not have a sequence similar to the template proteins.

In this chapter, we review the history of protein structure prediction from two different angles: the methodologies and the modeling targets. In the first section, we describe the historical perspective for predicting (largely) globular proteins. The specialized methodologies that have been developed for predicting structures of other types of proteins, such as membrane proteins and protein complexes and assemblies, are discussed along with the review of modeling targets in the second section. The current challenges faced in improving the prediction of protein structure and new trends for prediction are also discussed.

1.2 The Development of Protein Structure Prediction Methodologies

1.2.1 Protein Homology Modeling

The methodology for homology modeling (or comparative modeling), a very successful category of protein structure prediction, is based on our understanding of protein evolution: (1) proteins that have similar sequences usually have similar structures and (2) protein structures are more conserved than their sequences. Obviously, only those proteins having appropriate templates, i.e., homologous proteins with experimentally determined structures, can be modeled by homologous modeling. Nevertheless, with the increasing accumulation of experimentally determined protein structures and the advances in remote homology identification, protein homology modeling has made routine, continuing progress: both the space of potential targets has grown and the performance of the computational approaches has improved.

1.2.1.1 First Structure Predicted by Homology Modeling: α-Lactalbumin (1969)

The first protein structure that was predicted by the use of homologous modeling is α-lactalbumin, which was based on the X-ray structure of lysozyme. Browne and

co-workers conducted this experiment (Browne et al., 1969), following a procedure that is still largely used for model construction today. It starts with an alignment between the target and the template protein sequences, followed by the construction of an initial protein model created by insertions, deletions, and side chain replacements from the template structure, and finally finished by the refinement of the model using energy minimization to remove steric clashes.

1.2.1.2 Homology: Semiautomated Homology Modeling of Proteins in a Family (1981)

Greer developed a computer program to automate the whole procedure of homologous modeling. Using this program, 11 mammalian serine proteases were modeled based on three experimentally determined structures for mammalian serine proteases (Greer, 1981). The prediction used in this work was based on the analysis of multiple protein structures from the same protease family. He observed that the structure of a protease could be divided into structurally conserved regions (SCRs) with strong sequence homology, and structurally variable regions (SVRs) containing all the insertions and deletions in order to minimize errors in the query–template alignments significantly. Next, SVRs of the eight structurally unknown proteins were constructed directly from the known structures, based on the observation that a variable region that has the same length and residue character in two different known structures usually has the same conformation in both proteins.

This successful modeling experiment demonstrated that mammalian serine proteases could be constructed semiautomatically from the known homologous structures; both the need for manual inspections using biological intuition and the use of energy force fields were greatly reduced. The whole modeling procedure from this exercise was later implemented in the first protein modeling program, Homology, and integrated into a molecular graphics package InsightII (commercialized by Biosym, now Accelrys). Several important features of Homology, including the identification of modeling template using pairwise sequence alignment in the same protein family, the layout of sequence alignment between target and template protein sequences, and the identification and distinct modeling of conserved and variable regions using multiple structural templates from the same family, have been included in more recently developed homology modeling programs.

1.2.1.3 Composer: High-Accuracy Homology Modeling Using Multiple Templates (1987)

Greer's homology modeling method used multiple protein structures from the same family to define the conserved and variable regions in the target protein. It, however, used only one protein structure as the template to model the target protein. Blundell and co-workers recognized that the structural framework (or the "average" structure) of multiple protein structures from the same family usually resembled the target

protein structure more than any single protein structure did. Based on this concept, they implemented a program called Composer (Sutcliffe et al., 1987), which was later integrated into the protein modeling package Sybyl, which was commercialized by Tripos.

The framework-based protein modeling significantly increased the accuracy of model construction over the previous semiautomatic methods, and hence made modeled protein structures practically useful. However, Composer applies empirical rules for modeling SVRs and the structure of amino acid side chains. As a result, the accuracy of these regions is much lower than the backbone structures in the SCRs. Therefore, the modeling of SVRs (or loops) and side chain placement have become two independent research topics for protein modeling. Many different solutions have been proposed (see Section 1.2.4 for a detailed review).

1.2.1.4 Modeller: Automatic Full-Atom Protein Modeling (1993)

Before 1993, protein modeling was done through a semiautomatic and multistep fashion, including distinct modeling procedure for SCRs, SVRs, and side chains. MODELLER, developed by Sali and Blundell, was the first automatic computer program full-atom protein modeling (Sali and Blundell, 1993). MODELLER computes the structure of the target protein by optimally satisfying spatial restraints derived from the alignment of the target protein sequence and multiple related structures, which are expressed as probability density functions (pdfs) of the restrained structural features. MODELLER facilitates high-throughput modeling of protein targets from genome sequencing project (Sanchez et al., 2000) and remains one of the popular or widely used modeling packages.

1.2.1.5 Other Protein Modeling Programs

SWISS-MODEL is a fully automated protein structure homology-modeling server, which was initiated in 1993 by Manuel Peitsch (Peitsch and Jongeneel, 1993). SWISS-MODEL automates the complete modeling pipeline including homology template search, alignment generation and model construction. It uses ProMod (Peitsch, 1996) to construct models for protein query with an alignment of the query and template sequences. NEST (Petry et al., 2003) realizes model generation by performing operations of mutation, insertion, and deletion on the template structure finished with energy minimization to remove steric clashes. The minimization starts with those operations that least disturb the template structure (which is called an artificial evolution method). The minimization is done in torsion angle space, and the final structure is subjected to more thorough energy minimization. Kosinski et al. (2003) developed the "FRankenstein's monster" approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation; its novelty is that it employs the idea of combination of fragments that are often used by *ab initio* methods.

1.2.2 Remote Homology Recognition/Fold Recognition

All homology-based protein modeling programs rely on a good-quality alignment of the target and the template (of known structure). The identification of appropriate templates and the alignment of templates and target proteins are two essential topics for protein modeling, especially when no close homologue exists for modeling. The power or accuracy of homology modeling benefits from any improvement in the homology detection and target–template(s) alignment. Initially, a sequence alignment algorithm was used to derive target–template(s) alignment. More complicated methods (considering structure information) were later developed to improve the target–template(s) alignment.

1.2.2.1 Threading

The process of aligning a protein sequence with one or more protein structures is often called *threading* (Bryant and Lawrence, 1993). The protein sequence is placed or threaded onto a given structure to obtain the best sequence–structure compatibility. Obviously, the problem of identifying appropriate templates for a given target protein sequence can also be formulated as a *threading problem*, in which the structure in the database that is most compatible to the target sequence will be discerned and distinguished from those that are sufficiently compatible. Evolutionary information has been introduced to improve the sensitivity of homology recognition and to improve the target–template alignment quality, resulting a series sequence–profile and profile–profile alignment programs.

The threading method is able to go beyond sequence homology and identify structural similarity between unrelated proteins; "fold recognition" might be a better term for such cases. Homology recognition is used to detect templates that are homologous to the target with statistically significant sequence similarity; however, with the introduction of the powerful *profile-based* and *profile–profile-based methods*, the boundary between homology and fold recognition has blurred (Friedberg et al., 2004).

The threading-based method is typically classified in a separate category that is parallel to the homology-based modeling and *ab initio* modeling; it can be further divided into two subclasses considering whether or not the target and template have sequence similarity (homology) for quality evaluation purposes (Moult, 2005). However, from a methodology point of view, most threading-based modeling packages borrow similar ideas or even the existing modules from homology-based methods, to model the structure of a template after deriving the target–template alignment.

The concept of the threading approach to protein structure prediction is that in some cases, proteins can have similar structures but lack detectable sequential similarities. Indeed, it is widely accepted that there exist in nature only a limited number of distinct protein structures, called *protein folds*, which a virtually infinite number of different protein sequences adopt. As a result, it is hopeful that it is more sensible comparing the template protein structures with the target protein sequence

than comparing their sequences. Protein threading methods fall into two categories. One kind of method represents protein structures first as a sequence of symbolic *environmental features*, e.g., the secondary structures, the accessibility of amino acid residues, and so on; next, it aligns this sequence of features with the target protein sequence using the classical dynamic programming algorithm for sequence alignment with a special scoring function. The other kind of method is based on a *statistical potential*, i.e., the frequency of observing two amino acid residues at a certain distance, in order to evaluate the compatibility between a protein structure and a protein sequence. Threading approaches have three distinct applications in protein structure prediction: (1) identifying appropriate protein structure templates for modeling a target protein, (2) identifying protein sequences adopting a known protein fold, and (3) accessing the quality of a protein model.

1.2.2.2 3D-profile: Representing Structures by Environmental Features

The pioneering work of Bowie and co-workers on "the inverse protein folding problem" led to a simple method for assessing the fitness of a protein sequence onto a structure, thus laying the foundation of the first kind of protein threading approach. In their work, structural environments of an amino acid residue were simply defined in terms of solvent accessibility and secondary structure (Bowie et al., 1991; Luthy et al., 1992). Statistics of residue–structure environment compatibility (3D-profile) were then computed based on the statistics of the frequency of a particular type of amino acid appearing in a particular structural environment in the collection of known structures. Threading programs using 3D-profile include 123D (Alexandrov et al., 1996), 3D-PSSM (Kelley et al., 2000), and FUGUE (Shi et al., 2001).

1.2.2.3 Statistical Potential Models

An alternative approach to threading is to measure the protein structure–sequence compatibility by a statistical potential model, which represents the preference of two types of amino acids to be at some spatial distance. Sippl proposed the concept of "reverse Boltzmann Principle" to derive a statistical potential, which he called potential of mean force, from a set of unrelated known protein structures (Sippl, 1990; Casari and Sippl, 1992). The basic idea of this energy function is to compare the observed frequency of a pair of amino acids within a certain distance for known protein structures with the expected frequency of this pair of amino acid types in a protein. Bryant and Lawrence first used the term "threading" to describe the approach of aligning a protein sequence to a known structure when they reported a new statistical potential model (Bryant and Lawrence, 1993).

1.2.2.4 Algorithmic Development for Threading Using Statistical Potential

Unlike the 3D-profile approach, statistical potential-based threading approach cannot use the classical dynamic programming approach for structure–sequence comparison. In fact, if pairwise interaction between residues is considered in assessing

the compatibility of sequence and structure, the problem becomes very difficult (specifically, it is an NP-hard problem).

Various algorithms have been developed to address this computational difficulty. Early threading programs used various heuristic strategies to search for the optimal sequence–structure alignment. For example, GenTHREADER (Jones, 1999) and mGenTHREADER (based on the original GenTHREADER method, but adding the PSI-BLAST profile and predicted secondary structure as inputs) adopted a double dynamic programming strategy, which did not treat pairwise interactions rigorously. New threading programs have come to use more rigorous optimization algorithms. For example, PROSPECT (Xu and Xu, 2000) introduced a divide-and-conquer technique, and RAPTOR (Xu et al., 2003) used linear programming.

1.2.2.5 Profile-Based Alignment

Threading is not the only way to improve the sensitivity of (remote) template identification and the quality of template–target alignment. The other kind of method to achieve this goal makes use of multiple sequences from the same protein families to improve the sensitivity of homology detection and to improve the quality of sequence alignment.

Sequence–profile alignment strategy was first used to increase the sensitivity of distant homology detection. The development of Position Specific Iterative BLAST (PSI-BLAST) (and of course the accumulation of protein sequences) boosted the development of profile-based database search for homologies. In PSI-BLAST, a profile (or Position Specific Scoring Matrix, PSSM) is generated by calculating position-specific scores for each position in the multiple alignment constructed from the highest scoring hits in an initial BLAST search. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero. The profile is then used to perform a second BLAST search by performing a sequence–profile alignment and the results are used to refine the profile, and so forth. This iterative searching strategy results in significantly increased sensitivity. PSI-BLAST is now often used as the first step in many studies including the profile–profile alignments. Profile information is also employed in hidden Markov models (HMMs) (Krogh et al., 1994), as implemented in the SAM (Karplus et al., 1998) and HMMER (http://hmmer.wustl.edu), which have vastly improved the accuracy of sequence alignments and sensitivity of homology detection.

Several profile–profile alignment methods have been developed more recently, including FFAS (Rychlewski et al., 2000), COMPASS (Sadreyev et al., 2003), Yona and Levitt's profile–profile alignment algorithm (Yona and Levitt, 2002), a method developed in Sali's group (Marti-Renom et al., 2004), and COACH (using hidden Markov models) (Edgar and Sjolander, 2004). The FFAS program pioneered the profile–profile alignment; it is now used in many modeling pipelines and metaservers. Zhou and Zhou (2005) developed a fold recognition method by combining sequence profiles derived from evolution and from a depth-dependent structural alignment of fragments. A key process for this group of methods is the alignment of the profile of

target and homologies and the profile of structural template and homologies. They share the basic idea of profile–profile alignment but differ in many details, such as the profile calculation, profile–profile matching score, and alignment evaluation. The application of profile-profile alignment in homology detection highly increases the sensitivity of homology detection, even to the level of fold recognition.

1.2.3 *Ab Initio* Protein Structure Prediction

Despite the great success of homology modeling and threading methods, there are still many important target proteins that have no appropriate template (the number of such proteins is expected to be reduced due to the efforts of Structural Genomics, which aims at experimentally determining protein structures from all families and thus with providing new folds). *Ab initio* methods (which predict structures from sequence without using any structural template) are more general in this sense. *Ab initio* approaches are in principal based on Anfinsen's folding theory (Anfinsen, 1973), according to which the native structure corresponds to the global free energy minimum. Successful *ab initio* protein structure prediction methods fall roughly into several broad categories: (a) approaches that start from random/open conformations and simulate the folding process or minimize the conformational energy, (b) segment assembly-based methods as represented by the Rosetta method, and (c) methods that combine the two types of approaches (Samudrala et al., 1999).

1.2.3.1 Protein Folding Simulation and *ab Initio* Structure Prediction

Protein folding simulation and protein tertiary structure prediction are two distinct yet closely coupled problems. The main goal of protein folding simulation is to help characterize the mechanism of protein folding and also the interactions that determine the folding process and serve to specify the native structure; the goal of protein structure prediction is to determine the native structure. The solution of both problems relies on the effectiveness of energy function and conformation search methods utilized. Folding simulation approaches can be applied to predict protein structure *ab initio*, as seen in examples in which "folded" states resembling the native structures were derived. But only very few folding simulation approaches have been widely adopted for protein structure prediction and applied to a large number of predictions.

Molecular dynamics (MD) simulation is a natural approach for simulating protein folding. This approach has a long history and is still widely used; this could be viewed as illustrated most dramatically by IBM's Blue Gene project (http://www.research.ibm.com/bluegene/). However, the computational cost of folding simulations requires that the proteins to be simulated are small and fold ultrafast, even when supported by powerful computing (Duan and Kollman, 1998). Besides, the inadequacy in current potential functions for proteins in solution complicates the problem. The folded state by simulation does not necessarily correspond to the native state of proteins; actually, for current simulations, folding to the stable native state

has not (yet) occurred. Considering these two types of difficulties in fold simulation and *ab initio* prediction of protein structures, many researches have either adopted simplified representation of proteins (including lattice and off-lattice models) to alleviate computational complexity, and/or to apply some conformational constraints to reduce the conformational searching space (e.g., the application of local structures in segment assembly based methods). Doing so improves the efficiency of folding simulation and *ab initio* methods for protein structure prediction.

1.2.3.2 Reduced Models of Proteins and Their Applications

Reduced models of proteins are necessary for easy and unambiguous interpretation of computer simulations of proteins and to obtain dramatic reduction (by orders of magnitude) of the computational costs. Such reduced models are still very important tools for theoretical studies of protein structure, dynamics, and thermodynamics in spite of the enormous increase in computational power (Kolinski and Skolnick, 2004). Simplified representations of protein structures include lattice models, continuous space models (e.g., a protein structure is reduced to the Cα trace and the centroid of side chains), and hybrid models (in which some degrees of conformational freedom are locally discretized). The resolution of lattice models can vary from a very crude shape of the main chain to a resolution similar to that of good experimental structures. Usually, the protein backbone is restricted to a lattice. The side chain, if explicitly treated, could be restricted to a lattice or could be allowed to occupy off-lattice positions. The HP model, proposed by Lau and Dill (1989), is a type of simple lattice model, which only considers two types of residues, hydrophobic and polar in a simple cubic lattice. Lattice models of moderate to high resolutions were also designed to retain more details of actual protein structure, including SICHO (SIde CHain Only) model (Kolinski and Skolnick, 1998), CABS, and "hybrid" 310 lattice model (considering 90 possible orientations of the Cα-trace vectors with off-lattice side chains and multiple rotamers). Reduced representations of proteins were employed in many studies, for example in studies of the cooperativity of protein folding dynamics (Dill et al., 1993) and in the *ab initio* prediction of protein structures (Skolnick et al., 1993).

1.2.3.3 *Ab Initio* Methods Using Reduced Representation of Proteins

Levitt and Warshel made one of the very first attempts to model real proteins using a reduced representation of proteins in 1975 (Levitt and Warshel, 1975). They applied a simplified continuous representation of protein structures with each residue represented as two centers (Cα atom, and the centroid of the side chain) in the simulation of the folding of bovine pancreatic trypsin inhibitor (BPTI), in which BPTI was folded from an open-chain conformation into a folded conformation resembling the crystallographic structure, with a backbone RMSD in the range of 6.5 Å.

Skolnick et al. developed a hierarchical approach to protein-structure prediction using two cycles of the lattice method (the second on a finer lattice), in which reduced representations of proteins are folded on a lattice by Monte Carlo simulation using

statistically derived potentials, and a full-atom MD simulation afterwards (Skolnick et al., 1993; Kolinski and Skolnick, 1994b). This procedure was applied to model the structures of the B domain of staphylococcal protein (60 residues) and mROP (120 residues) (Kolinski and Skolnick, 1994a). Skolnick's group also developed TOUCHSTONE, an *ab initio* protein structure prediction method that uses threading-based tertiary restraints (Kihara et al., 2001). This method employs the SICHO model of proteins to restrict the protein's conformational space and uses both predicted secondary structure and tertiary contacts to restrict further the conformational search and to improve the correlation of energy with fold quality.

Scheraga's group developed a hierarchical approach that is similar to Skolnick's hierarchical method, but uses *off-lattice* simplified representation of proteins in the first steps of the prediction process; namely, one based solely on global optimization of a potential energy function (Liwo et al., 1999). This global optimization method is called Conformational Space Annealing (CSA), which is based on a genetic algorithm and on local energy minimization. Using this method, Liwo et al. built models of RMSD to native below 6 Å for protein fragments of up to 61 residues. This method was further assessed through two blind tests; the results were reported in Oldziej et al. (2005).

In specialized cases, parallel computation allows protein fold simulations using all-atom representation of proteins, and even explicit solvents, at the microsecond level. As described in brief above, a representative example is the folding of HP35, which is a subdomain of the headpiece of the actin-binding protein villin (Duan and Kollman, 1998), which has only 36 residues and folds autonomously without any cofactor or disulfide bond. This simulation was enabled by a parallel implementation of classic MD using an explicit representation of water, and the folded state of HP35 significantly resembles the native structure (but is not identical). But all-atom simulations are still limited and only practical for small ultrafast folding proteins.

1.2.3.4 *Ab Initio* Methods by Segment Assembling

A significant progress in the development of *ab initio* methods was the introduction of conformational constraints to reduce the computational complexity. Several *ab initio* modeling methods have been developed based on this strategy (Zhang and Skolnick, 2004; Lee et al., 2005), which was pioneered in the implementation of the Rosetta method (Simons et al., 1997, 1999a).

The basic idea of Rosetta is to narrow the conformation searching space with local structure predictions and model the structures of proteins by assembling the local structures of segments. The Rosetta method is based on the assumption that short sequence segments have strong local structural biases, and the strength and multiplicity of these local biases are highly sequence dependent. Bystroff et al. developed a method that recognizes sequence motifs (I-SITES) with strong tendencies to adopt a single local conformation that can be used to make local structure predictions (Bystroff and Baker, 1998). In the first step of Rosetta, fragment libraries for each three- and nine-residue segment of the target protein are extracted from

the protein structure database using a sequence profile–profile comparison method. Then, tertiary structures are generated using a Monte Carlo search of the possible combinations of likely local structures, minimizing a scoring function that accounts for nonlocal interactions such as compactness, hydrophobic burial, specific pair interactions (disulfides and electrostatics), and strand pairing (Simons et al., 1999b). A test of Rosetta on 172 target proteins showed that 73 successful structure predictions were made out of 172 target proteins with lengths below 150 residues, with an RMSD < 7 Å in the top five models (Simons et al., 2001). Rosetta has achieved the top performance in a series of independent, blind tests (Moult et al., 1999; Simons et al., 1999a), ever since those for CASP3 (see below for details about the CASP series of workshop). Rosetta has also been further refined and extended to related prediction tasks, namely, docking on predicted interactions (see below).

Zhang and Skolnick developed TASSER, a threading template assembly/refinement approach, for *ab initio* prediction of protein structures (Zhang and Skolnick, 2004). The test of TASSER on a comprehensive benchmark set of 1489 single-domain proteins in the Protein Data Bank (PDB) with length below 200 residues showed that 990 targets could be folded by TASSER with an RMSD $<$ 6.5 Å in at least one of the top five models. The fragments used for assembly in TASSER are derived in a different way than in Rosetta. Specifically, the fragments or segments are excised from the threading results, and thus are generally much longer (about 20.7 residues on average) than the segments used by Rosetta (which are 3–9 residues).

1.2.4 Modeling of Side Chains and Loops

We review the modeling of side chains and loops as a separate section because these are two main problems that both homology modeling and *de novo* methods face, and because they differ more among protein homologues than do the backbone and protein cores. Yet, the conformation of side chains and loops may carry very important information for understanding the function of proteins.

There are mainly two classes of computational approaches to building the loop structures: knowledge-based methods and *ab initio* methods. Knowledge-based methods build the loop structures using the known structures of loops from all proteins in the structure database, whether or not they are from the same family as the target protein (Sucha et al., 1995; Rufino et al., 1997). This approach is based on the principle that the plausible conformations of loops within a certain length cannot be that many, i.e., must be limited. Assuming a sufficient variety of known protein structures, almost all plausible loop structures should be represented by at lease one protein structure in the database. In fact a library of plausible loop structures for a given loop size has been constructed (Donate et al., 1996; Oliva et al., 1997). Typically, for a given loop in the target protein, the selection of the optimal template structure usually relies on the similarity of the anchor regions (i.e., the flanking residues around the loop) between template loop structure and

the modeled core structure of the target, and the compatibility of the template loop structure with the core structure as measured by a residue level empirical scoring function (van Vlijmen and Karplus, 1997). *Ab initio* methods build loop structures from scratch (Moult and James, 1986; Pedersen and Moult, 1995; Zheng and Kyle, 1996). Recently, methods that combine knowledge-based and *ab initio* methods for better loop modeling have been introduced (Deane and Blundell, 2001; Rohl et al., 2004). MODELLER (Sali and Blundell, 1993) uses a different methodology from the above, which builds both core and loop regions by optimally satisfying spatial restraints derived from the target–template alignment.

Similarly, side-chain conformations can be predicted from similar structures and from steric or energetic considerations (Vasquez, 1996). The construction of side-chain rotamers and the development of powerful conformation searching algorithms (such as Dead End Elimination, DEE) (Desmet et al., 1992) and the mean force field-based method (Lee, 1994; Koehl and Delarue, 1995) contributed to the success of side-chain conformation prediction. Rotamer libraries are generally defined in terms of side-chain torsional angles for preferred conformations of a particular side chain. Ponder and Richards set up the first rotamer library (Ponder and Richards, 1987). A backbone-dependent rotamer library was later constructed and used for side-chain prediction (SCWRL) (Dunbrack and Karplus, 1993; Canutescu et al., 2003). Wang et al. developed a rapid and efficient method for sampling off-rotamer side-chain conformations through torsion space minimization; this starts from discrete rotamer libraries supplemented with side-chain conformations taken from the unbound structures. This approach has been used to improve side chain packing in protein–protein docking.

1.2.5 Modeling Structural Differences

Mutation data are an important source of information in the study of the functions of proteins; similarly, analyzing the differences among protein families is one way to study their function and functional specificity. It is therefore very important to study the detailed structural differences associated with mutations and sequence differences among families. For example, homology modeling (Lee, 1995) and molecular dynamics (MD) were used for studying the consequences of mutations (see the section "Molecular Dynamics Simulations of Membrane Proteins").

Baker's group tried to model structural differences based on comparative modeling by free-energy optimization along principal components of natural structural variation, which serves to improve the accuracy of protein modeling (Qian et al., 2004). In comparative modeling, an issue has been that a given protein model is frequently more similar to the template(s) used for modeling than to the target protein's native structure. In principle, energy-based minimization might help to improve the resolution of models. However, in practice, energy-based refinement of comparative models generally leads to degradation rather than improvement in model quality. The work of Baker's group (Qian et al., 2004) led to an improved use of energy-based minimization, through restricting the search space along the evolutionarily favored

direction and thereby avoiding the false attractors that might lead the minimization to wrong answers.

There are numerous limits within current efforts, and considerable effort is still required to improve the methods for predicting the structures resulting from mutations and the modeling of structural difference within families. The reasons underlying the difficulties include our inability to model protein structures in fine resolution despite the strict requirements for quality in modeling of the structural differences. Indeed, "modeling of the structure of a single mutation" and "modeling structure changes associated with specificity changes within protein families" were identified as two of the three modeling challenges as viewed by a community meeting in 2005 [see the summary from CASP6 (Moult et al., 2005), which is a summary from the sixth in a series of structure prediction meetings described below].

1.2.6 Novel Communitywide Activities to Improve Prediction and Demonstrate Value

CASP (Critical Assessment of Structure Prediction) is a communitywide experiment with the primary aim of assessing the effectiveness of modeling methods. CASP deserves special recognition in any consideration of the role of modeling/computational methods for biology, since the meeting/process has transformed the level of recognition (for modeling studies) coming from experimentalists; CASP has become a model for all computational biology communities and an exemplar for evaluating techniques or methods beyond software/the approaches of scientific computing. In light of these competitions and the overall efforts in the field, the general status for high-resolution refinement of protein structure models and overall progress in modeling has been reviewed in depth recently (Misura and Baker, 2005; Schueler-Furman et al., 2005b).

CASP was first held in 1994 and six CASP meetings were held through 2004; the most recent meeting was held in 2006 (as the 7th Community Wide Experiment on Critical Assessment of Techniques for Protein Structure Prediction). The key feature of CASP is that participants make *blind* predictions of structures. CASP has monitored since 1994 the progress of protein modeling (covering all categories of modeling methods). Also it provides a good arena for testing the performance of newly developed modeling methods. The prediction season, during a cycle, begins in spring and all predictions are due at the end of the summer. The essential aspect is that experimentalists make lists available of what they are likely to solve during this time period and agree not to release their structures, when obtained, until after the deadline for predictions. Establishing this clear process solved the longstanding assertions about structure prediction being based on previously known information.

How well one does in CASP has become important—some would say too important—as a metric for research in the field. As a consequence, as well as CASP, which is a manual method in which any amount of scientific knowledge and any

collection of algorithms can be employed, an automated prediction approach has been added, to test the state of computational prediction schemes rather than the participants' insight into protein structure. This is the Critical Assessment of Fully Automated Structure Prediction (CAFASP). Besides using automated approaches for the competition, numerous protein prediction servers have been introduced for the community, including, for example, PROSPECT-PSPP (Guo et al., 2004) and Robetta (Kim et al., 2004). Other aspects of large-scale prediction servers are described below (Section 1.2.7). Interestingly, services, such as EVA, have also been created to monitor the quality or performance of the numerous prediction servers, and provide continuous, fully automatic, and statistically significant analysis of such servers (Koh et al., 2003).

CASP is now organized by the Protein Structure Prediction Center. The Center's goal is to help advance the methods of identifying protein structure from sequence. The Center has been organized to provide the means of objective testing of these methods via the process of blind prediction. In addition to support of the CASP meetings, their goal is to promote an objective evaluation of prediction methods on a continuing basis. Some of the recent successes in CASP have been described previously.

A very powerful related community scheme looks at the nature of macromolecular interactions or docking, Critical Assessment of PRedicted Interactions (CAPRI), which grew up directly from the successes of CASP, where this new chain of meetings was launched after 1996. While few of the proteins identified through major genome sequencing efforts will ever have their structure solved, since proteins actually carry out biological processes as larger, multimeric or even heterologous complexes, characterizing the structure of proteins in native complexes is more important, and even fewer of those complexes will ever be experimentally determined, due to the greater inherent difficulties in doing so. To test what are therefore essential computational methods, the starting points for predicting the structures of protein complexes ("docked" proteins) are the independently solved structures of the constituents of a protein complex, whose 3D structure is unknown, and against which the community's algorithms and approaches can be tested. For example, an in-depth evaluation of certain docking algorithms in early CAPRI rounds (3, 4, and 5) has been provided (Wiehe et al., 2005); of particular value has been the introduction of benchmarks for analysis, such the Protein-Protein Docking Benchmark 2.0 (Mintseris et al., 2005), which provides a platform for evaluating the progress of docking methods on a wide variety of targets. An extension of the very successful Rosetta approach to the challenges of predicting the structure of complexes is RosettaDock, which uses real-space Monte Carlo minimization on both rigid body and the side chain degrees of freedom in order to find the lowest free energy arrangement of two docked protein structures; more recently, this has been extended to take into account backbone flexibility and employed very successfully in more recent CAPRI competitions (Schueler-Furman et al., 2005a). More details about docking approaches in general are discussed below.

1.2.7 Protein Modeling Metaservers

Several protein modeling metaservers have appeared since 2001, including Pcon (a neural-network–based consensus predictor) (Lundstrom et al., 2001), Structure Prediction Meta Server (Bujnicki et al., 2001), 3D-Jury (Ginalski et al., 2003), GeneSilico protein structure prediction metaserver (Kurowski and Bujnicki, 2003), and 3D-SHOTGUN (Daniel, 2003). These automatic servers collect models from other servers and use that input to produce consensus structures. According to the assessment performed via CASP, protein modeling metaservers perform generally better than other single modeling methods; their performance is even close to that of human experts. [The noteworthy progress between CASP4 and CASP5 was partly due to the effective use of metaservers (Moult, 2005).] However, some CASP participants have worried that the increasing successes of metaservers might discourage researchers from developing new prediction methods. This seems a small worry, in light of the various objectives for improved modeling methods and the potential impact from delivering more accurate, high-throughput genome annotation to enhanced drug discovery. Of course, there is a community goal, to seek improved tools and validation of the overall approach in the eyes of experimentalists, and the many personal goals, to seek to make the best contribution possible. As a consequence, the larger worry, under the current environment for CASP itself, is that it is hard to dissect the individual computational contributions to prediction and ascertain progress and what tools to choose since considerable manual or intellectual intervention is inevitably involved in order to achieve the highest validated successes in prediction. This difficulty is among the factors that led to the introduction of automated approaches, including metaservers, in the first place.

1.3 A Shift in the Focus for Protein Modeling

In recent years, the efforts in genome sequencing have been enormously successful. Hundreds of whole or complete microbial genomes and dozens of eukaryotic plant and animal genomes have been sequenced, and many more genome projects are underway. In contrast to the quickly increasing number of predicted protein sequences (open reading frames or ORFs) that are deposited in the community database, Genbank, the number of proteins whose architecture has been solved increases much more slowly. This continues despite the advances in structure determination techniques and the effects of the (National Institutes of Health, NIH) Protein Structure Initiative in the United States and Structural Genomics Projects worldwide. Therefore, more modeling per se as well as improved computational modeling of protein structures is of crucial importance to keep pace with the advances of genome sequencing and functional genomics, that is, our ability to predict the structure of newly discovered or predicted proteins has to increase greatly in order for the community to be able to characterize and utilize fully the extraordinary delivery of new sequence information. Accordingly, the focus of modeling has shifted in recent

years, from modeling of monomers to modeling of simple protein–protein complex and even the modeling of large protein assemblies; that is, the focus has moved from small-scale modeling to large-scale modeling (and even genome-scale efforts at comprehensive modeling). In this section, we will focus on a discussion of modeling of different targets. Also, we will discuss specific methods that have already been developed and those that are emerging to deal with the various requirements, which are different from the methods discussed above. (These methods are mostly for modeling soluble, single-domain globular proteins.)

1.3.1 Modeling of Membrane Proteins

Membrane proteins play a central role in many cellular and physiological processes. Any aspect of cell activity is regulated by extracellular signals that are recognized and transduced inside the cell via different classes of plasma membrane receptors. It is estimated that integral membrane or transmembrane (TM) proteins make up about 20–30% of the proteome (Krogh et al., 2001). They are essential mediators of material and information transfer across cell membranes. Identifying these TM proteins and deciphering their molecular mechanisms is of great importance for understanding many biological processes. In addition, membrane proteins are of particular importance in biomedicine, because they are the targets of a large number of pharmacologically and toxicologically active substances, and are directly involved in their uptake, metabolism, and clearance. Membrane proteins can be loosely associated on the surface of the lipid bilayer (peripheral membrane proteins) or embedded (integral membrane protein, e.g., bacteriorhodopsin). The prediction and analysis of membrane proteins largely involves a focus on integral membrane proteins.

Membrane proteins account for less than 1% of the known high-resolution protein structures (White, 2004), despite their importance in essential cellular functions. Solving the structure of a membrane protein remains challenging and no high-throughput methods, or even general methods, have been developed. In the first instance, structure determination of membrane proteins remains a challenge because of difficulties in expressing sufficient quantities of protein and in manipulating the protein *in vitro* with an artificial environment mimicking some attributes of the *in situ* environment. Even when these challenges are met, there are remaining difficulties in obtaining ordered crystals for analysis by X-ray crystallography. NMR remains the modality of choice for structural analysis of membrane proteins but cannot readily tackle larger proteins and requires substantive quantities of material. Given the challenges for crystallographic analysis, membrane proteins were inevitably listed as “lower priority” or “avoided” targets for the Structural Genomics Centers, during the early phase of the Protein Structure Initiative. Research funding has even included set-aside opportunities to address the challenges of characterizing the biophysical properties and structure of membrane proteins. However, no demonstrated method yet exists to deliver a pipeline for high-throughput structure determination of membrane proteins.

Given the relatively and absolutely (!) small number of known, high-resolution membrane protein structures, computational methods are very important in predicting the structures of membrane protein, and in this case especially, if a prediction could be said to "determine" the structure, computational methods would have a huge impact on fundamental biology and biomedicine, and on applied life sciences research around drug targets. Most of the tools used for analyzing and predicting the structure of soluble, nonmembrane proteins can also be used for this important class. That is, many secondary structure prediction methods from primary sequences based on statistical methods, physicochemical methods, sequence pattern matching, and evolutionary conservation can also be applied for modeling the structures of membrane proteins, as can the conventional 3D structure prediction methods, including homology modeling techniques. At the same time, due to the limited number of known structures of membrane proteins, the application of homology modeling in predicting membrane protein structures remains very limited.

In the absence of a high-resolution 3D structure (experimental or computational), an important cornerstone for the functional analysis of any membrane protein is an accurate topology model. A topology model describes the number of TM spans and the orientation of the protein relative to the lipid bilayer. The secondary structure of a membrane-spanning segment can be an α-helix or a β-strand, but a TM β-strand usually has fewer residues than an α-helix. Nearly all TM β-strand proteins are found in prokaryotes, and belong to only a few protein families. Generally, integral membrane transporters of the inner membrane consist largely of α-structures, and they traverse the membrane as α-helices, whereas those of the outer membranes consist largely of ß-barrels. Because of this, many methods have been developed to focus on the prediction of transmembrane α-helices. These methods are mainly based on the special properties of membrane proteins (Chen and Rost, 2002), such as differences in amino-acid compositions in cytoplasmic and extracellular regions (positive-inside rule) (Heijne, 1986), the hydrophobic/hydrophilic patterns of TM regions (Kyte and Doolittle, 1982), and the minimum length of TM regions.

1.3.1.1 Methods for Topology Model Prediction (α-Helix Membrane Proteins)

One of the earliest and still most widely practiced methods for identification of membrane regions is hydropathy analysis, which uses a sliding-window approach to calculate the average hydrophobicity of an amino-acid position. By definition, hydrophobicity is the property of being water-repellent. Rose first introduced the concept of hydrophobicity analysis as a means of identifying chain turns in soluble proteins in 1978 (Rose, 1978), and in 1982, Kyte and Doolittle developed the first hydropathy scale (KD hydropathy scale, or KD scale), which is widely used by many prediction programs for evaluating the hydrophobicity of a protein along the amino acid sequence (Kyte and Doolittle, 1982). In practice, there are multiple ways to quantify the hydrophobicity of amino acids. Indeed, to date, more than 100 hydrophobicity scales have been published in the literature. These were either

derived experimentally based on the free energy of transfer or empirically calculated based on surface accessibility.

The use of more complex processing of the hydrophobicity scale (and in combination with other physicochemical parameters) helped to improve the performance of membrane protein prediction. An early effort used discriminant analysis to classify membrane proteins as integral or peripheral and to estimate the odds that the classification is correct (Klein et al., 1985). TopPred (von Heijne, 1992) combines hydrophobicity analysis with the positive-inside rule and achieves better performance than using hydrophobicity alone. The Dense Alignment Surface (DAS) method optimizes the use of hydrophobicity plots by assessing sequence similarities between segments of the query protein and known transmembrane segments (Cserzo et al., 1997). For making predictions, the SOSUI method combines four physicochemical parameters: KD scale, amphiphilicity, relative and net charges, and protein length (Hirokawa et al., 1998). TMFinder combines segment hydrophobicity and the nonpolar phase helicity to predict TM segments (Deber et al., 2001).

A more general strategy is to infer the statistical preference of amino acids in membrane proteins from unknown membrane proteins (since consecutive residues have preferences for certain secondary structure states), and then to use the derived preference (instead of hydrophobility) for prediction. This strategy can be used for general secondary structure prediction for globular proteins and, upon considering different states, for membrane proteins. Methods developed following this strategy include MEMSAT, SPLIT, TMAP, and TMpred (for a review see Chen and Rost, 2002).

Many advanced methods have been developed employing statistical preferences and machine learning methods, including neural networks (NN; e.g., PHDhtm), hidden Markov models [HMM; e.g., HMMTOP (Tusnady and Simon, 1998) and TMHMM (see below)], and SVM (e.g., SVMtm—see below) for membrane protein prediction. Rost et al. (1995) developed a neural network system for predicting the locations of TM helices in integral membrane proteins using evolutionary information as input. TMHMM (Krogh et al., 2001) embeds a number of statistical preferences and rules into a hidden Markov model to optimize the prediction of the localization of TM helices and their orientation. It incorporates hydrophobicity, charge bias, helix lengths, and grammatical constraints (i.e., cytoplasmic and noncytoplasmic loops have to alternate) into one model for which algorithms for parameter estimation and prediction already exist. TMHMM achieved highly accurate performance: it correctly predicts 97–98% of the TM helices, and discriminates between soluble and membrane proteins with both specificity and sensitivity better than 99% (but the accuracy drops when signal peptides are present). This high degree of accuracy makes it possible to use this method to predict integral membrane proteins reliably from numerous genomes. Based on this prediction across a wide collection of complete genomes, an estimate has been made that 20–30% of all genes in most genomes encode membrane proteins, which is in agreement with previous estimates. A more recent method SVMtm (Yuan et al., 2004) applies support vector machines to predict transmembrane segments; various sequence coding schemes

(including three different hydropathy scales and 21-UNIT) (Rost et al., 1995) were tested.

1.3.1.2 Methods for Topology Model Prediction (β-Strand Membrane Proteins)

β-Strand TMs lack a clear pattern in their membrane-spanning strands, making them different from the α-helical membrane proteins, which have hydrophobic segments and the positive-inside rule. Predictions made for TM β-strands are currently less successful than those for TM α-helices. An early method developed in 1995 used Gibbs motif sampling to detect bacterial outer membrane protein repeats; these were then used in searching for outer membrane proteins (Neuwald et al., 1995).

One of the key structural determinants of β-barrel membrane proteins is a pattern of β-barrel dyad repeats. β-Barrel proteins of known 3D structure share two physicochemical properties (i.e., hydrophobicity and amphipathicity): most of the TM strands correspond to a peak of hydrophobicity, but the hydrophobic values of these peaks are generally not as high as those of the TM α-helices of cytoplasmic integral membrane proteins. Most of the TM β-strands exhibit peaks of amphipathicity caused by the alternating hydrophilic residues located inside the barrel and the hydrophobic residues located outside the barrel. These two physicochemical properties laid the basis for many software programs aimed at β-barrel TM proteins. The β-Barrel Outer Membrane protein Predictor (BOMP) program (Berven et al., 2004) combines two independent methods for identifying the possible integral outer membrane proteins and also a filtering mechanism to remove false positives; it was designed to predict whether a protein sequence specifically from Gram-negative bacteria is an integral β-barrel outer membrane protein (80% accuracy and 88% sensitivity achieved when applied to *E. coli* K12 and *S. typhimurium*). Similar to predictions for α-helix TM proteins, statistical preferences and machine learning methods (NN in BBF, OM_Topo_predict and TMBETA-NET; HMM in BETA-TM, BIOSINO-HMM, HMM-B2TMR, PRED-TMBB, and ProfTMB) have also been introduced to improve the prediction of β-barrel TM proteins. BBF, Beta-Barrel Finder (Zhai and Saier, 2002), is a program based on physicochemical properties (both hydropathy and amphipathicity), which uses NNs to identify TM β-barrel proteins in *E. coli*. TBBPred (Natt et al., 2004) uses both NNs and SVMs for predicting TM β-barrel regions.

1.3.1.3 Molecular Dynamics Simulations of Membrane Proteins

MD simulations are widely used in studying the structures of membrane proteins such as the conformational dynamics of the receptors, the functions (such as open or closed states) of ion channels (Giorgetti and Carloni, 2003), and the receptor and ligand interactions. The simulations enable us to extrapolate from the essentially static (time- and space-averaged) structure revealed by X-ray diffraction to a more dynamic picture of the behavior of a membrane protein in a more realistic environment that mimics a small patch of the membrane. The first MD simulation of a biological

process was the 1976 simulation of the primary event in rhodopsin (Warshel, 1976). MD simulations were next applied to the earliest simulations of enzymatic reactions and electron transfer reactions and then simulations of proton translocations and ion transport in proteins (see the review by Warshel, 2002). MD simulations have been employed in a number of studies on outer membrane proteins, in order, for example, to probe protein and solvent dynamics in relationship to permeation mechanisms in porins (Tieleman and Berendsen, 1998), to explore possible pore-gating mechanisms in OmpA (Bond et al., 2002), and to examine the role of calcium binding and dimerization in the catalytic mechanism of OMPLA (Baaden et al., 2003). MD simulations can also be used to assess whether any *mutation* in a protein has an effect on the structure and function of the protein before more time-consuming experiments have been performed; for example, this has been done with the computational alanine scanning of human growth hormone-receptor complex (Huo et al., 2002) and the study of TM domain mutants of Vpu from HIV-1 along with the consequences of these mutations on its structure (Candler et al., 2005).

1.3.1.4 Modeling and Simulation of GPCR (G-Protein-Coupled Receptor)

GPCRs constitute the largest family of signal transduction membrane proteins, which mediate the cellular responses to a variety of bioactive molecules, including biogenic amines, amino acids, peptides, lipids, nucleotides, and proteins. The GPCRs play a crucial role in many essential physiological processes as diverse as neurotransmission, cellular metabolism, secretion, cell growth, immune defense, and differentiation. GPCRs are also (not surprisingly) the most common targets for the drugs currently used in clinics and for the wealth of drug candidates that high-throughput methods are expected to deliver in the immediate future. Extensive computational analysis (see the review by Fanelli and DeBenedetti, 2005), which includes predicting families and subfamilies of GPCRs from sequences, 3D structure modeling, and MD simulation of the consequences of mutants, has been done for GPCRs; a dedicated database was created for GPCRs, the G-protein-coupled receptor database (GPCRDB) at http://www.gpcr.org/7tm.

1.3.1.5 Global Topology Analysis of the *E. coli* Membrane Proteome

A study that deserves special mention is the global topology analysis of the *E. coli* inner membrane proteome by Daley et al. (2005). This is the first reported large-scale prediction of membrane proteins in combination with large-scale experiments. Their work exploits the observation that topology prediction can be greatly improved by constraining it with an experimentally determined reference point, such as the location of a protein's C-terminus; an estimate is that at least ten percentage points in overall accuracy in whole-genome predictions can be gained in this way (Melen et al., 2003). Using C-terminal tagging with the alkaline phosphatase and green fluorescent protein, they determined the locations of the C-termini (either periplasmic or cytoplasmic) for 601 inner membrane proteins. Then, by constraining topology predictor TMHMM with these data, they derived high-quality topology models for

these proteins; this research provides a firm foundation for future functional studies of this and other membrane proteomes.

1.3.2 Modeling of Multiple-Domain Proteins

Domain fusion/shuffling is one of the most important events in the evolution of modern proteins (Patthy, 1999; Kriventseva et al., 2003). The majority of proteins, especially in higher organisms, are built from multiple domains (modules), which can be found in various contexts in different proteins. Such domains usually form stable three-dimensional structures, even if excised from a complete protein, and perform the same or similar molecular functions as parts of the protein.

The identification of domain boundaries is critical for both experimental and computational (including *ab initio* and comparative modeling) protein structure determination. NMR spectrometry has a length limitation in solving protein structures, and X-ray diffraction requires high-quality crystals and thus may fail or may have regions lacking in detail due to flexible linker regions between domains. As a consequence, there are inevitably fewer structures deposited in PDB that can be used as structural templates for modeling multiple-domain proteins. *Ab initio* prediction methods also encounter huge difficulties in predicting large, multidomain proteins due to the exceptional computational barrier in exploring conformational space and for the determination of domain–domain interactions. Consequently, current *ab initio* structure prediction methods can only model structures of relatively small size and do so at worse resolution than obtained by homology modeling. Both experimental and computational approaches to protein structure determination would benefit significantly from predicted domain assignments.

One way of dealing with the multidomain problems is to model the structures of domains of a protein separately and then, if possible, to assemble the domains together. Any computational methods for protein structure modeling can be applied to model structures of individual domains, including comparative modeling, threading, and *ab initio* methods. Special issues in modeling multidomain proteins involve the first step of domain dissection (Contreras-Moreira and Bates, 2002) and the last step of predicting spatial arrangement of constituent domains (Inbar et al., 2003), which will be discussed in this section.

1.3.2.1 Domain Assignment of Proteins

Although the concept of domains as structural components of proteins has been around for years, ever since studies conducted by Wetlaufer (1973) and Rossman and Liljas (1974), and is now well accepted, its definition is full of ambiguities. Caution in using current domain assignments was recommended by Veretnik et al., who systematically assessed the consistency of current domain assignments by investigating six methods [three "human-expert methods"—authors' annotation, CATH, and SCOP—and three "fully-automated methods"—DALI (Holm and Sander 1994), DomainParser (Guo et al., 2003), and PDP (Alexandrov and Shindyalov, 2003)].

Their survey of the consistency of domain assignment also indicated where additional work is needed in domain assignment, including the assignment of the domain boundaries and the assignment of small domains. Nevertheless, significant advances have been made in the domain assignment of proteins (with or without structure information), which hence will be discussed here.

For a protein whose structure is known, domain assignments can usually be done manually by experts, or by automatic programs, or by a combination of both. Earlier, when there were only a few known protein structures, simple visual inspection of protein structures was quite adequate (Wetlaufer, 1973). The SCOP database is one of the widely cited databases of protein domains; the database represents largely the results of human experts, who have been assisted by computer visualization, and in particular, considered evolutionary information (Murzin et al., 1995). Fully automatic methods, i.e., those that are run without intervention and are not affected by human subjectivity in terms of consistently following criteria, are becoming ever more important in order to keep pace with the current accumulation of experimental structures of proteins. One early automatic method was developed by Wodak and Janin, who used surface area measurements based on atomic positions to give a quantitative definition of the structural domains in proteins (Wodak and Janin, 1981). The incorporation of secondary structure information (DOMAK) (Siddiqui and Barton, 1995) or information on hydrophobic cores (DETECTIVE) (Swindells, 1995) has subsequently been shown to enhance the automatic domain assignment. Another program, PUU, was based on achieving/expecting maximal interactions within each unit but minimal interaction between units (or domains) (Holm and Sander, 1994). The CATH database (Orengo et al., 1997) uses three algorithms (DETECTIVE, PUU, and DOMAK) for domain decomposition as a first step in the assignment process, followed by an expert's inspection. The VAST algorithm, which is used for structure neighboring in the Entrez system, is a fully automatic method that splits protein chains at points between secondary structure elements (SSEs) when the ratio of intra- to interdomain contacts exceeds a certain threshold (Madej et al., 1995).

In the absence of a known protein structure (as would be the case for any protein structure modeling), algorithms developed to predict domain boundaries have been based on sequence information, multiple sequence alignments, and/or homology modeling. Early approaches to domain boundary prediction relied on information theory (Busetta and Barrans, 1984) and used statistical potentials (Vonderviszt and Simon, 1986). Later prediction methods took into account the information of (predicted) secondary structure (DomSSEA), sequence conservation (Guan and Du, 1998; George and Heringa, 2002a; Rigden, 2002), or both, in order to improve domain assignment from sequences. DomSSEA (Marsden et al., 2002) uses a fold recognition approach, based on aligning predicted secondary structure of a query protein sequence to the assigned secondary structure of known structures, and then transferring the SCOP assigned domains from the best fold match to the query protein. SnapDRAGON is a suite of programs developed to predict domain boundaries based on the consistency observed in a set of alternative *ab initio* 3D models generated for a given protein multiple sequence alignment (George and Heringa, 2002b).

Many of the Open Reading Frames (ORFs) or predicted protein sequences discovered along the genome of any fully sequenced bacterial organism have been found not to be conserved across organisms; indeed, nearly half of all new ORFs appear to be unique. In these cases, algorithms that do not rely on sequence conservation have been applied to assign domains. The Domain Guess by Size (DGS) algorithm makes predictions based on observed domain size distributions (Wheelan et al., 2000). Galzitskaya and Melnik (2003) developed a method based on the assumption that the domain boundary is conditioned by amino acid residues with a small value of side chain entropy, which correlates with the side chain size. The Armadillo program (Dumontier et al., 2005) uses an amino acid index, called the domain linker propensity index (DLI) and derived from the amino acid composition of domain linkers using a nonredundant structure data set, to convert a protein sequence to a smoothed numeric profile from which domains and domain boundaries may be predicted. In general, most approaches predict the number of domains accurately, but only a few predict the domain boundaries well; prediction of domain boundaries only has a moderate sensitivity of ∼50–70% for proteins with single domains, and does considerably worse (∼30%) for multidomain proteins. Multidomain proteins are also harder to study experimentally. Thus, the proteins from eukaryotes are more difficult for both experimental and computational analysis.

1.3.2.2 Modeling of Domain–Domain Interactions

Considering the similarity of domain–domain interaction (folding) and protein–protein interaction (binding), docking techniques that have been developed for modeling protein–protein complexes (see Section 1.3.3) have been applied to build the model of multidomain proteins by docking separate structures of domains together. Unfortunately, few advances have been made; this is especially the case for predicting the spatial arrangement of protein domains.

Xu et al. (2001) modeled the structure of vitronectin by first modeling its C-terminal and central domains and then modeling the interaction of these two domains using GRAMM, a docking technique. In this work, the threading program PROSPECT was used to find the structure template for modeling and to generate the sequence–structure alignment, which was used as input for the program MODELLER to create the models. Experimental data were also used to guide the docking of the central and C-terminal domains by GRAMM.

Inbar et al. developed CombDock, a combinational docking algorithm, for protein structure prediction via combinatorial assembly of substructural units (building blocks/domains) (Inbar et al., 2003). Three steps are involved in this algorithm to predict the structure of a protein sequence: a dissection into fragments and the assignment of their structures; the assembly of the fragments into an overall structure of the protein sequence; and the prediction of the spatial arrangement of the assigned structures and then the completion and refinement of highly ranked predicted arrangements. The combinatorial assembly of domains is formulated as the problem of finding the spanning tree in a graph (where each substructure is a vertex, and an

edge between two vertices presents the interaction of the two substructures), and a heuristic polynomial solution to this computational hard problem has been provided.

Jones et al. used a similar strategy, i.e., domain docking and microdomain folding, to model complete chains of selected CASP6 targets. Their method, called FRAGFOLD-MODEL, generates models of a complete chain by "docking" domains together by searching possible linker peptide conformations. To this end, a genetic algorithm or simulated annealing can be used for the conformational search (Jones et al., 2005).

1.3.3 Modeling of Protein Complexes

Modeling of protein complexes is far less successful than the modeling of protein monomers. At the same time, obtaining accurate models for complexes is of increasing importance because of their functional importance—in general, proteins act as part of large macromolecular assemblies. Indeed, experimental work routinely extends the known scale (number of proteins) of interactions in any given functional pathway, and the impact on our thinking about molecular processes is now expanding rapidly with the advent of improved and larger-scale identification of protein–protein interactions. A few docking programs have been developed since the late 1970s for predicting the structures of protein–protein complexes. Recent developments include the usage of computational models for docking, the combination of experimental data in computational docking, and the combination of homology modeling and cryoEM data to model large complex structures; these are discussed below.

1.3.3.1 Modeling Protein Complexes by Docking

Most docking methods consist of a global (or stochastic) search of translational and rotational space followed by refinement of the best predictions. The relative performance depends on the conformational searching ability and on the efficiency of complex evaluation. Very often, docking programs treat proteins as a "rigid body" during the first step and use a simple and "soft" energy function to evaluate the potential complex, and subsequently use more fine evaluation in the second step of refinement, during which some programs also consider the flexibility of the side chains, but few consider the flexibility of the backbone as well.

The first computational protein docking tools were developed in the late 1970s. Greer and Bush (1978) introduced a grid-based measure of complementarity between molecules, and used it to score interfaces between hemoglobin subunits. An early docking study by Wodak and Janin (1981) used a simplified protein model with one sphere per amino acid, which they used to dock BPTI to trypsin. The search involved rotating BPTI and varying its center-of-mass distance with trypsin. Newer programs use more complex shape-complementarity and an energy function to evaluate the complex models, and use a more rigorous definition of conformational space to improve the docking performance. A significant improvement in the conformational space search has been the use of the fast Fourier transform (FFT) to perform

correlations in grid-based translational searching (Katchalski-Katzir et al., 1992). FFT is employed by several commonly used docking programs, including GRAMM, FTDock, and ZDOCK. These programs use the same strategy for the conformational search (FFT), but may use different scoring functions and do use different details for overall operation: FTDock's scoring of complexes is based on shape complementarity and on favorable electrostatic interactions (Gabb et al., 1997); GRAMM (Vakser, 1995) implements docking at different resolutions to account for the inaccuracy of input structures; ZDOCK (Chen et al., 2003) combines pairwise shape complementarity with desolvation and electrostatics for complex scoring. Other techniques including geometric hashing (Fischer et al., 1995), stochastic searches such as Monte Carlo search [e.g., RossetaDock (Gray et al., 2003)], or a genetic algorithm (e.g., Gardiner et al., 2001) have also been used for the conformational search step in docking.

Most of the existing docking programs adopt the "rigid-body" strategy while neglecting the conformational changes during binding. For complexes of an enzyme with its inhibitor, the conformational changes might be small and can be compensated by using some "soft" scoring to evaluate the potential complex, which is a key consideration of evaluating the potential complex for unbound docking. (Bound docking uses the structures from the complex structure as input, which obviously has little predictive use, whereas unbound docking uses the structures from the individually crystallized subunits as input.) However, for other types of complexes, the conformational change may be greater. Due to the huge conformational space for protein structures, "flexible" protein–protein docking remains a challenge, despite the advances in incorporating the flexibility of receptors in protein–ligand docking (Jones et al., 1997; Alberts et al., 2005).

1.3.3.2 Data-Driven Docking Approaches

Applying proper constraints to the conformational space during docking can significantly improve the computation speed and the accuracy of docking (van Dijk et al., 2005). The constraints can be derived via many methods, both experimentally and computationally. NMR data have been used in combination with docking methods and in different ways in order to generate information about protein–protein complexes. For example, diamagnetic chemical shift changes and intermolecular pseudocontact shifts were combined with restrained rigid-body molecular dynamics to solve the structure of the paramagnetic plastocyanin–cytochrome f complex (Ubbink et al., 1998). Intermolecular NOEs and residual dipolar couplings (RDCs) were combined to solve the structure of the EIN–HPr complex (Clore, 2000). TreeDock (Fahmy and Wagner, 2002) enumerates the search space at a user-defined resolution subject to the condition that a pair of atoms, one from each molecule, are always in contact, which can be in principle derived from NMR chemical shift perturbation or mutagenesis data. Dominguez et al. (2003) developed an approach called HADDOCK (High Ambiguity Driven protein–protein Docking), which makes use of biochemical and/or biophysical interaction data such as chemical shift perturbation data resulting

from NMR titration experiments or mutagenesis data. The data are transformed as Ambiguous Interaction Restraints (AIRs) between all residues shown to be involved in the interaction to drive the docking process.

1.3.3.3 Integrating Homology Modeling and EM Density Map for Modeling Protein Assemblies

With the advances in functional genomics, experimental methods allow us to determine on a large scale what the partners are for pairwise protein–protein interactions (by yeast two-hybrid system and protein chips) and what the constituents are among large protein assemblies [by tandem-affinity purification (TAP) and mass spectrometry]. Accordingly, the need to model protein–protein interactions and protein assemblies is increasing. The integration of high-resolution structures/models and the electron microscopy (EM) density map is an exciting advance for modeling a large protein–protein complex and assemblies. The basic idea is to fit known high-resolution structures into low-resolution structures of large complexes that are determined by EM to obtain the refined structure of large complexes. This technique has been applied in solving the structures of large biological machines/macromolecular complexes, such as viruses, ion channels, ribosomes, and proteasomes. In cases where no experimental high-resolution structures are available, computational models of the individual proteins may instead be used in fitting. In addition, intermediate-resolution cryo-EM density maps are helpful for improving the accuracy of comparative protein structure modeling in those cases for which no template for modeling can be found by a sequence-based search or a threading method. In fact, the application of EM density maps in structure modeling started quite early; for example, a model for the structure of bacteriorhodopsin was originally generated based on high-resolution electron cryomicroscopy (Henderson et al., 1990). An explosion of joint EM/crystallographic studies in the mid-1990s followed the development of strategies for generating pseudo-atomic-resolution models of macromolecular complexes by combining the data from high-resolution structures of components with lower-resolution EM data for the entire complex (Baker and Johnson, 1996).

Electron cryomicroscopy (cryo-EM) can image complexes in their physiological environment and does not require large quantities of the sample. Cryo-EM also provides a means of visualizing the membrane proteins *in situ*, as opposed to the usually artificial hydrophobic environments used for crystallizing membrane proteins. Structures of large macromolecular complexes can now be visualized in different functional states at intermediate resolution (6–9 Å) (Chiu et al., 2005). The corresponding cryo-EM maps are generally still insufficient for atomic structure determination on their own. However, one can fit atomic-resolution structures of the components of the assembly (e.g., protein domains, whole proteins, and any subcomplexes) into the lower-resolution density of the entire assembly. In early applications, researchers employed mainly "visual docking" to position the protein components in the envelopes derived from low-resolution data (Schroder et al., 1993). More recently, computational programs were developed to obtain quantitative

means to fit the data. Wriggers et al. used topology-representing neural networks (TNN) to vector-quantize and to correlate features within the structural data sets to generate pseudo-atomic structures of large-scale protein assemblies by combining high-resolution data with volumetric data at lower resolution (Wriggers et al., 1998, 1999). Roseman et al. developed a fitting procedure that uses a real-space density-matching procedure based on local correlation of the density derived from the atomic coordinates of protein components and the density of the EM map (Roseman, 2000). Ceulemans and Russell developed 3SOM (Ceulemans and Russell, 2004) for finding the best fit through surface overlap maximization.

In cases where experimentally determined atomic-resolution structures of assembly components are not available, or the induced fit severely limits their usefulness in the reconstruction of the complex, it may be possible to get useful models of the components by comparative protein structure modeling (or homology modeling) (see the review by Topf and Sali, 2005). The number of models that can be constructed with useful accuracy, at least comparable to the resolution of the cryo-EM maps, is almost two orders of magnitude greater than the number of available experimentally determined structures, which indicates the huge potential for employing models in fitting EM maps (Topf and Sali, 2005).

Moreover, EM maps can be used to improve modeling in some cases (Topf and Sali, 2005). In cases where a structural homologue of the target component cannot be detected by sequence-based or threading search methods, it is possible to use the EM map (if the resolution is better than ~12 Å) for fold assignment of the constituting proteins: at ~12 Å resolution, it is usually possible to recognize boundaries between the individual components in the complex, while secondary structure features, such as long α-helices and large β-sheets, can begin to be identified at ~10 Å resolution, and short helices and individual strands at ~4 Å (Chiu et al., 2002). For example, Jiang et al. (2001) developed the Helixhunter program, which is capable of reliably identifying helix position, orientation, and length using a five-dimensional cross-correlation search of a three-dimensional density map followed by feature extraction; its results can in turn be used to probe a library of secondary structure elements derived from the structures in the PDB. This readily provides for the structure-based recognition of folds containing α-helices. They also developed the Foldhunter program, which uses a six-dimensional cross-correlation search that allows a probe structure to be fitted within a region or component of a target structure. These two methods have been successfully tested with simulated structures modeled from the PDB at resolutions of 6–12 Å. In cases where the fold of the protein component is known, the density maps can be useful in selecting the best template structures for comparative modeling, since a more accurate model fits the EM density map more tightly (Topf et al., 2005). Topf et al. (2005) developed a method for finding an optimal atomic model of a given assembly subunit and its position within an assembly by fitting alternative comparative models (created by MODELLER from different sequence alignments between the modeled protein and template structures) into a cryo EM map, using Foldhunter (Jiang et al., 2001) or Mod-EM (a density fitting module of MODELLER).

1.3.4 Large-Scale Modeling

In the era of many fully sequencing genomes and a focus on systemwide, integrated biological research from proteomics to metabolomics, the introduction of an "omics" for structural biology, that is, structural genomics, was a natural development, and one that reflected the maturity of structural biology as well as the need to obtain structures in order to annotate genomes fully and obtain explicit insight into the information implicit in genome sequences. The most often stated goal of structural genomics is to provide "structural coverage" of protein space by solving enough structures that all known proteins could be accurately modeled (Brenner, 2001; Vitkup et al., 2001). In the United States, the efforts of structural genomics groups also resulted in the launch of the NIH-funded Protein Structure Initiative, which currently supports four large structural genomics centers, as well as a larger number of smaller, technology-focused or specialized centers.

The success of structural genomics will, by definition, rely on both experimental structure determination and computational approaches. A question therefore raised is to ascertain to what extent have the high throughput and comprehensive aspects of genomics and the pipelines for structure determination reached efforts on computational structure prediction. Threading and comparative modeling methods have already been applied on a genomic scale. For example, ModPipe was developed for modeling known protein sequences using the comparative modeling program MODELLER on a larger scale; the models are deposited in a comprehensive database of comparative models, ModBase (Sanchez et al., 2000) (as of July 05, 2005, the database had 3,094,524 reliable models or fold assignments for domains in 1,094,750 proteins). The Web interface to the database allows flexible querying for the models, the fold assignments, sequence–structure alignments, and assessments of models of interest. Automation and large-scale modeling with *de novo* methods lag behind those of comparative modeling methods, because of the relatively poor quality of the models produced, and the relatively large amount of computer time required. Nevertheless, Rosetta initiated the successful use of large-scale modeling calculations done with ab initio methods (Baker and Sali, 2001).

1.3.4.1 Structure Modeling for Structural Genomics

It is clear that the eventual success of structural genomics will be brought about by the growing synergy between experimental structure determination and computational approaches, including the comparative modeling and *ab initio* fold prediction methods (see the review by Friedberg et al., 2004). The efficiency of using comparative modeling will be determined by the advances of distant homology detection and fold recognition algorithms, while the efficiency of using *ab initio* methods will be largely determined by the improvement of the quality of models and the reduction in computing time.

Predictions done through comparative modeling and *ab initio* methods can compensate each other and thus play a particularly important role for structural

genomics; namely, target selection and modeling of the structure of proteins that are not selected for experimental determination (see the review by Baker and Sali, 2001). Of course, even with the worldwide initiatives in high-throughput structural determination, the structures for the vast majority of the proteins in nature will (at most) only be modeled and will never be determined by experiment.

Structural genomics as conducted to date generally omits several groups of proteins since they are considered to be very difficult targets; these largely excluded proteins include the membrane proteins (despite some focused attention on membrane proteins), and those with disordered structures that may fold only in the presence of appropriate interaction partners (Bracken et al., 2004). These "special" proteins, nevertheless, constitute a large portion of the whole proteome, for example membrane proteins constitute 20–30% of the proteome (Krogh et al., 2001). Achieving large-scale experimental and/or computational structural determination of these proteins would be as important as for any other proteins, and certainly as important as for those proteins already within the structural genomics scope.

1.3.4.2 Large-Scale Modeling of Human (Disease-Related) Proteins

Disease-related proteins are of great research interest for both experimental and computational scientists. Their high value for research in biomedicine and clinical medicine, and potentially in health care, stems from the fact that they provide a molecular picture of disease processes, which is a necessary prerequisite for rational drug development. Thousands of genes (proteins) have already been identified as associated with various diseases in humans. Computational modeling will play a more important role in predicting the structure of eukaryotic proteins than that of prokaryotic proteins, since eukaryotic proteins are more difficult to carry through a crystallography pipeline, and fewer, as a consequence, are likely ever to be determined experimentally. Several efforts have been carried out in order to model the structures of human proteins. For example, generated models are extensively used for studying the human disease proteins in association with SNP data. LS-SNP is a resource providing large-scale annotation of coding nonsynonymous SNPs (nsSNPs) based on multiple information sources (including structural models) in human proteins (Karchin et al., 2005). Yip et al. created the Swiss-Prot variant page and the ModSNP database for sequence and structure information on human protein variants (Yip et al., 2004). Ye et al. specifically created models of all human disease-related proteins collected in Swiss-Prot and studied the spatial distribution of disease-related nsSNPs on the models (Ye et al., 2006). These analyses provided some explanation for nsSNPs with known effects (harmful or neutral), and might in turn provide a basis for predicting the effects of nsSNPs.

1.3.4.3 Genome-Scale Modeling of Complexes

Proteins function via interactions with other macromolecules, and most cellular processes are carried out by multiprotein complexes. The identification and analysis of the components of these complexes provides insight into how the ensemble of

expressed proteins (the proteome) is organized into functional units. Large-scale identifications of protein–protein interactions in many genomes are now possible due to the genome-scale discovery approaches for identifying interacting proteins; these methods include the yeast two-hybrid system and protein chips, which have been very widely employed. Using an approach with more potential for quantitative information, Gavin et al. (2002) used tandem-affinity purification (TAP) and mass spectrometry in a large-scale approach to characterize multiprotein complexes in *S. cerevisiae*.

Employing biophysical and computational methods for studying protein–protein interactions and complexes from a structural perspective would be similarly important. A significant step toward understanding how proteins assemble has been taken by Aloy et al. (2004). Starting from the large set of identified complexes of yeast by TAP (Gavin et al. 2002), they screened the complexes using low-resolution EM images. These images were used to assemble and validate models (see the "Integrating Homology Modeling and EM Density Map for Modeling Protein Assemblies" section). They also predicted links between complexes and provide a higher-order, structure-based network of connected molecular machines within the cell. The network they derived currently gives the most complete view available for complexes and their interrelationships.

1.4 Summary

From understanding single molecules, to a simple complex, to large assemblies to the biological networks, we are moving toward an understanding of life. Structure information (derived experimentally or computationally) helps us to understand the mechanisms by which the biochemical processes of cells occur and provides insight beyond chemical architecture—mechanism implications, for example, in suggesting features about evolution of function. In turn, structure prediction has made an increasing number of contributions to our understanding of biology [which has been described elsewhere both in detail and eloquently (Petrey and Honig, 2005)]. Advances have been achieved in computational predictions of structure at each level, and each advance brings new potential to impact our understanding of biology. Yet, challenges remain. We can expect that the computational challenges will be more daunting at a network level, characterizing the metabolic pathways, signal transduction cascades, and genetic circuits through which protein interactions determine cellular and organismic function; existing methods need improvement or new methods need to be developed that must deal with individual proteins, complexes, and sophisticated dynamic networks that connect them. The remainder of this book deals with contemporary efforts toward those advances. The structure-based network derived by Aloy et al. provides a useful initial framework for further studies. "Its beauty is that the whole is greater than the sum of its parts: Each new structure can help to understand multiple interactions. The complex predictions and the associated network will thus improve exponentially as the numbers of structures and interactions

increase, providing an ever more complete molecular anatomy of the cell" (Aloy et al. 2004)

References

Alberts, I. L., Todorov, N. P. and Dean, P. M. 2005. Receptor flexibility in de novo ligand design and docking. *J. Med. Chem.* 48:6585–6596.

Alexandrov, N. N., Nussinov, R., and Zimmer, R. M. 1996. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.* pp. 53–72.

Alexandrov, N., and Shindyalov, I. 2003. PDP: Protein domain parser. *Bioinformatics* 19:429–430.

Aloy, P., Bottcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A.-C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. 2004. Structure-based assembly of protein complexes in yeast. *Science* 303:2026–2029.

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

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H. J. 1961. The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl. Acad. Sci., USA* 47:1309–1314.

Anfinsen, C. B., Redfield, R. R., Choate, W. I., Page, J., and Carroll, W. R. 1954. Studies on the gross structure, cross-linkages, and terminal sequences in ribonuclease. *J. Biol. Chem.* 207:201–210.

Baaden, M., Meier, C., and Sansom, M. S. P. 2003. A molecular dynamics investigation of mono and dimeric states of the outer membrane enzyme OMPLA. *J. Mol. Biol.* 331:177–189.

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

Baker, T. S., and Johnson, J. E. 1996. Low resolution meets high: Towards a resolution continuum from cells to atoms. *Curr. Opin. Struct. Biol.* 6:585–594.

Berven, F. S., Flikka, K., Jensen, H. B., and Eidhammer, I. 2004. BOMP: A program to predict integral β-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. *Nucleic Acids Res.* 32(Web Server Issue):W394–W399.

Bond, P. J., Faraldo-Gomez, J. D., and Sansom, M. S. P. 2002. OmpA: A pore or not a pore? Simulation and modeling studies. *Biophys. J.* 83:763–775.

Bowie, J. U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253:164–170.

Bracken, C., Iakoucheva, L. M., Romero, P. R., and Dunker, A. K. 2004. Combining prediction, computation and experiment for the characterization of protein disorder. *Curr. Opin. Struct. Biol.* 14:570–576.

Brenner, S. E. 2001. A tour of structural genomics. *Nature Rev. Genet.* 2:801–809.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M. 1983. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comp. Chem.* 4:187–217.

Browne, W. J., North, A. C., Phillips, D. C., Brew, K., Vanaman, T. C., and Hill, R. L. 1969. A possible three-dimensional structure of bovine alpha-lactalbumin based on that of hen's egg-white lysozyme. *J. Mol. Biol.* 42:65–86.

Bryant, S. H., and Lawrence, C. E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16:92–112.

Bujnicki, J. M., Elofsson, A., Fischer, D., and Rychlewski, L. 2001. Structure prediction meta server. *Bioinformatics* 17:750–751.

Busetta, B., and Barrans, Y. 1984. The prediction of protein domains. *Biochim. Biophys. Acta Protein Struct. Mol. Enzymol.* 790:117–124.

Bystroff, C., and Baker, D. 1998. Prediction of local structure in proteins using a library of sequence-structure motifs. *J. Mol. Biol.* 281:565–577.

Candler, A., Featherstone, M., Ali, R., Maloney, L., Watts, A., and Fischer, W. B. 2005. Computational analysis of mutations in the transmembrane region of Vpu from HIV-1. *Biochim. Biophys. Acta Biomembranes* 1716:1–10.

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.

Casari, G., and Sippl, M. J. 1992. Structure-derived hydrophobic potential: Hydrophobic potential derived from X-ray structures of globular proteins is able to identify native folds. *J. Mol. Biol.* 224:725–732.

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.

Chen, C. P., and Rost, B. 2002. State-of-the-art in membrane protein prediction. *Appl. Bioinformatics* 1:21–35.

Chen, R., Li, L., and Weng, Z. 2003. ZDOCK: An initial-stage protein-docking algorithm. *Proteins* 52:80–87.

Chiu, W., Baker, M. L., Jiang, W., Dougherty, M., and Schmid, M. F. 2005. Electron cryomicroscopy of biological machines at subnanometer resolution. *Structure* 13:363–372.

Chiu, W., Baker, M. L., Jiang, W., and Zhou, Z. H. 2002. Deriving folds of macromolecular complexes through electron cryomicroscopy and bioinformatics approaches. *Curr. Opin. Struct. Biol.* 12:263–269.

Clore, G. M. 2000. Accurate and rapid docking of protein–protein complexes on the basis of intermolecular nuclear Overhauser enhancement data and dipolar couplings by rigid body minimization. *Proc. Natl. Acad. Sci. USA* 97:9021–9025.

Contreras-Moreira, B., and Bates, P. A. 2002. Domain Fishing: A first step in protein comparative modelling. *Bioinformatics* 18:1141–1142.

Cserzo, M., Wallin, E., Simon, I., von Heijne, G., and Elofsson, A. 1997. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: The dense alignment surface method. *Protein Eng.* 10:673–676.

Daley, D. O., Rapp, M., Granseth, E., Melen, K., Drew, D., and von Heijne, G. 2005. Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science* 308:1321–1323.

Daniel, F. 2003. 3D-SHOTGUN: A novel, cooperative, fold-recognition meta-predictor. *Proteins Struct. Funct. Genet.* 51:434–441.

Deane, C. M., and Blundell, T. L. 2001. CODA: A combined algorithm for predicting the structurally variable regions of protein models. *Protein Sci.* 10:599–612.

Deber, C. M., Wang, C., Liu, L.-P., Prior, A. S., Agrawal, S., Muskat, B. L., and Cuticchia, A. J. 2001. TM Finder: A prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales. *Protein Sci.* 10:212–219.

Desmet, J., Maeyer, M. D., Hazes, B., and Lasters, I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature* 356:539–542.

Dill, K. A., Fiebig, K. M., and Chan, H. S. 1993. Cooperativity in protein-folding kinetics. *Proc. Natl. Acad. Sci. USA* 90:1942–1946.

Dominguez, C., Boelens, R., and Bonvin, A. M. J. J. 2003. HADDOCK: A protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* 125:1731–1737.

Donate, L. E., Rufino, S. D., Canard, L. H., and Blundell, T. L. 1996. Conformational analysis and clustering of short and medium size loops connecting regular secondary structures: A database for modeling and prediction. *Protein Sci.* 5:2600–2616.

Duan, Y., and Kollman, P. A. 1998. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. *Science* 282:740–744.

Dumontier, M., Yao, R., Feldman, H. J., and Hogue, C. W. V. 2005. Armadillo: Domain boundary prediction by amino acid composition. *J. Mol. Biol.* 350:1061–1073.

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

Edgar, R. C., and Sjolander, K. 2004. COACH: Profile–profile alignment of protein families using hidden Markov models. *Bioinformatics* 20:1309–1318.

Fahmy, A., and Wagner, G. 2002. TreeDock: A tool for protein docking based on minimizing van der Waals energies. *J. Am. Chem. Soc.* 124:1241–1250.

Fanelli, F., and DeBenedetti, P. G. 2005. Computational modeling approaches to structure–function analysis of G protein-coupled receptors. *Chem. Rev.* 105:3297–3351.

Fischer, D., Lin, S. L., Wolfson, H. L., and Nussinov, R. 1995. A geometry-based suite of molecular docking processes. *J. Mol. Biol.* 248:459–477.

Friedberg, I., Jaroszewski, L., Ye, Y., and Godzik, A. 2004. The interplay of fold recognition and experimental structure determination in structural genomics. *Curr. Opin. Struct. Biol.* 14:307–312.

Gabb, H. A., Jackson, R. M., and Sternberg, M. J. E. 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol.* 272:106–120.

Galzitskaya, O. V., and Melnik, B. S. 2003. Prediction of protein domain boundaries from sequence alone. *Protein Sci.* 12:696–701.

Gardiner, E. J., Willett, P., and Artymiuk, P. J. (2001). Protein docking using a genetic algorithm. *Proteins Struct. Funct. Genet.* 44:44–56.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* 415:141–147.

George, R. A., and Heringa, J. 2002a. Protein domain identification and improved sequence similarity searching using PSI-BLAST. *Proteins* 48:672–681.

George, R. A., and Heringa, J. 2002b. SnapDRAGON: A method to delineate protein structural domains from sequence data. *J. Mol. Biol.* 316:839–851.

Gibson, K. D., and Scheraga, H. A. 1967a. Minimization of polypeptide energy, I. Preliminary structures of bovine pancreatic ribonuclease S-peptide. *Proc. Natl. Acad. Sci. USA* 58:420–427.

Gibson, K. D., and Scheraga, H. A. 1967b. Minimization of polypeptide energy. II. Preliminary structures of oxytocin, vasopressin, and an octapeptide from ribonuclease. *Proc. Natl. Acad. Sci. USA* 58:1317–1323.

Ginalski, K., Elofsson, A., Fischer, D., and Rychlewski, L. 2003. 3D-Jury: A simple approach to improve protein structure predictions. *Bioinformatics* 19:1015–1018.

Giorgetti, A., and Carloni, P. 2003. Molecular modeling of ion channels: Structural predictions. *Curr. Opin. Chem. Biol.* 7:150–156.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. 2003. Protein–protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. *J. Mol. Biol.* 331:281–299.

Greer, J. 1981. Comparative model-building of the mammalian serine proteases. *J. Mol. Biol.* 153:1027–1042.

Greer, J., and Bush, B. L. 1978. Macromolecular shape and surface maps by solvent exclusion. *Proc. Natl. Acad. Sci. USA* 75:303–307.

Guan, X., and Du, L. 1998. Domain identification by clustering sequence alignments. *Bioinformatics* 14:783–788.

Guo, J.-T., Ellrott, K., Chung, W. J., Xu, D., Passovets, S., and Xu, Y. 2004. PROSPECT-PSPP: An automatic computational pipeline for protein structure prediction. *Nucleic Acids Res.* 32(Suppl. 2):W522–525.

Guo, J. T., Xu, D., Kim, D., and Xu, Y. 2003. Improving the performance of DomainParser for structural domain partition using neural network. *Nucleic Acids Res.* 31:944–952.

Heijne, V. 1986. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* 5:3021–3027.

Henderson, R., Baldwin, J. M., Ceska, T. A., Zemlin, F., Beckmann, E., and Downing, K. H. 1990. Model for the structure of bacteriorhodopsin based on high-resolution electron cryo-microscopy. *J. Mol. Biol.* 213:899–929.

Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* 14:378–379.

Holm, L., and Sander, C. 1994. Parser for protein folding units. *Proteins* 19:256–268.

Huo, S., Massova, I., and Kollman, P. A. 2002. Computational alanine scanning of the 1:1 human growth hormone–receptor complex. *J. Comp. Chem.* 23:15–27.

Inbar, Y., Benyamini, H., Nussinov, R., and Wolfson, H. J. 2003. Protein structure prediction via combinatorial assembly of sub-structural units. *Bioinformatics* 19(Suppl. 1):i158–i168.

Jiang, W., Baker, M. L., Ludtke, S. J., and Chiu, W. 2001. Bridging the information gap: Computational tools for intermediate resolution structure interpretation. *J. Mol. Biol.* 308:1033–1044.

Jones, D. T. 1999. GenTHREADER: An efficient and reliable protein fold recognition method for genomic sequences. *J. Mol. Biol.* 287:797–815.

Jones, D. T., Bryson, K., Coleman, A., McGuffin, L. J., Sadowski, M. I., Sodhi, J. S., and Ward, J. J. 2005. Prediction of novel and analogous folds using fragment assembly and fold recognition. *Proteins* 61(Suppl. 7):143–151.

Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. 1997. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* 267:727–748.

Karchin, R., Diekhans, M., Kelly, L., Thomas, D. J., Pieper, U., Eswar, N., Haussler, D., and Sali, A. 2005. LS-SNP: Large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics* 21:2814–2820.

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

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. 1992. Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA* 89:2195–2199.

Kelley, L. A., MacCallum, R. M., and Sternberg, M. J. E. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* 299:501–522.

Kihara, D., Lu, H., Kolinski, A., and Skolnick, J. 2001. TOUCHSTONE: An ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc. Natl. Acad. Sci. USA* 98:10125–10130.

Kim, D. E., Chivian, D., and Baker, D. 2004. Protein structure prediction and analysis using the Robetta server. *Nucleic Acids Res.* 32(Suppl. 2):W526–531.

Klein, P., Kanehisa, M., and DeLisi, C. 1985. The detection and classification of membrane-spanning proteins. *Biochim. Biophys. Acta Prot. Struct. Mol. Enzymol.* 815:468–476.

Koehl, P., and Delarue, M. 1995. A self consistent mean field approach to simultaneous gap closure and side-chain positioning in homology modelling. *Nat. Struct. Biol.* 2:163–170.

Koh, I. Y. Y., Eyrich, V. A., Marti-Renom, M. A., Przybylski, D., Madhusudhan, M. S., Eswar, N., Grana, O., Pazos, F., Valencia, A., Sali, A., and Rost, B. 2003. EVA: Evaluation of protein structure prediction servers. *Nucleic Acids Res.* 31:3311–3315.

Kolinski, A., and Skolnick, J. 1994a. Monte Carlo simulation of protein folding. II. Application to protein A, ROP, and crambin. *Proteins* 18:353–366.

Kolinski, A., and Skolnick, J. 1994b. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins* 18:338–352.

Kolinski, A., and Skolnick, J. 1998. Assembly of protein structure from sparse experimental data: An efficient Monte Carlo model. *Proteins* 32:475–494.

Kolinski, A., and Skolnick, J. 2004. Reduced models of proteins and their applications. *Polymer* 45:511–524.

Kosinski, J., Cymerman, I. A., Feder, M., Kurowski, M. A., Sasin, J. M., and Bujnicki, J. M. 2003. A "FRankenstein's monster" approach to comparative modeling: Merging the finest fragments of Fold-Recognition models and iterative model refinement aided by 3D structure evaluation. *Proteins* 53(S6):369–379.

Kriventseva, E. V., Koch, I., Apweiler, R., Vingron, M., Bork, P., Gelfand, M. S., and Sunyaev, S. 2003. Increase of functional diversity by alternative splicing. *Trends Genet.* 19:124–128.

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

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* 305:567–580.

Kurowski, M. A., and Bujnicki, J. M. 2003. GeneSilico protein structure prediction meta-server. *Nucleic Acids Res.* 31:3305–3307.

Kyte, J., and Doolittle, R. F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157:105–132.

Lau, K. F., and Dill, K. A. 1989. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules* 22:3986–3997.

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

Lee, C. 1995. Testing homology modeling on mutant proteins: Predicting structural and thermodynamic effects in the Ala98→Val mutants of T4 lysozyme. *Fold Des.* 1:1–12.

Lee, J., Kim, S.-Y., and Lee, J. 2005. Protein structure prediction based on fragment assembly and parameter optimization. *Biophys. Chem.* 115:209–214.

Levinthal, C. 1968. Are there pathways for protein folding? *J. Chem. Phys.* 65: 44–45.

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

Levitt, M., and Warshel, A. 1975. Computer simulation of protein folding. *Nature* 253:694–698.

Liwo, A., Lee, J., Ripoll, D. R., Pillardy, J., and Scheraga, H. A. 1999. Protein structure prediction by global optimization of a potential energy function. *Proc. Natl. Acad. Sci. USA* 96:5482–5485.

Lundstrom, J., Rychlewski, L., Bujnicki, J., and Elofsson, A. 2001. Pcons: A neural-network-based consensus predictor that improves fold recognition. *Protein Sci.* 10:2354–2362.

Luthy, R., Bowie, J. U., and Eisenberg, D. 1992. Assessment of protein models with three-dimensional profiles. *Nature* 356:83–85.

Madej, T., Gibrati, J.F., and S.H. Bryant 1995 'Threading a database of protein cores.' *Proteins* 32:289–306.

Marsden, R. L., McGuffin, L. J., and Jones, D. T. 2002. Rapid protein domain assignment from amino acid sequence using predicted secondary structure. *Protein Sci.* 11:2814–2824.

Marti-Renom, M. A., Madhusudhan, M. S., and Sali, A. 2004. Alignment of protein sequences by their profiles. *Protein Sci.* 13:1071–1087.

Melen, K., Krogh, A., and von Heijne, G. 2003. Reliability measures for membrane protein topology prediction algorithms. *J. Mol. Biol.* 327:735–744.

Mintseris, J., Wiehe, K., Pierce, B., Anderson, R., Chen, R., Janin, J., and Weng, Z. 2005. Protein–protein docking benchmark 2.0: An update. *Proteins* 60:214–216.

Misura, K. M. S., and Baker, D. 2005. Progress and challenges in high-resolution refinement of protein structure models. *Proteins* 59:15–29.

Moult, J. 2005. A decade of CASP: Progress, bottlenecks and prognosis in protein structure prediction. *Curr. Opin. Struct. Biol.* 15:285–289.

Moult, J., Fidelis, K., Tramontano, A., Rost, B., and Hubbard, T. 2005. Critical assessment of methods of protein structure prediction (CASP)—Round VI. *Proteins* 61(S7):3–7.

Moult, J., Hubbard, T., Fidelis, K., and Pedersen, J. T. 1999. Critical assessment of methods of protein structure prediction (CASP): Round III. *Proteins*(Suppl. 3):2–6.

Moult, J., and James, M. N. G. 1986. An algorithm for determining the conformation of polypeptide segments in proteins by systematic search. *Proteins* 1:146–163.

Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247:536–540.

Natt, N. K., Kaur, H., and Raghava, G. P. 2004. Prediction of transmembrane regions of β-barrel proteins using ANN- and SVM-based methods. *Proteins* 56:11–18.

Neuwald, A. F., Liu, J. S., and Lawrence, C. E. 1995. Gibbs motif sampling: Detection of bacterial outer membrane protein repeats. *Protein Sci.* 4:1618–1632.

Oldziej, S., Czaplewski, C., Liwo, A., Chinchio, M., Nanias, M., Vila, J. A., Khalili, M., Arnautova, Y. A., Jagielska, A., Makowski, M., Schafroth, H. D., Kazmierkiewicz, R., Ripoll, D. R., Pillardy, J., Saunders, J. A., Kang, Y. K., Gibson, K. D., and Scheraga, H. A. 2005. Physics-based protein-structure prediction using a hierarchical protocol based on the UNRES force field: Assessment in two blind tests. *Proc. Natl. Acad. Sci. USA* 102:7547–7552.

Oliva, B., Bates, P. A., Querol, E., Aviles, F. X., and Sternberg, M. J. 1997. An automated classification of the structure of protein loops. *J. Mol. Biol.* 266:814–830.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. 1997. CATH—a hierarchic classification of protein domain structures. *Structure* 5:1093–1108.

Patthy, L. 1999. *Protein Evolution*. Malden, MA, Blackwell Science.

Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. R., Cheatham, T. W., DeBolt, S., Ferguson, D., Seibel, G., and Kollman, P. 1995. AMBER, a computer program for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to elucidate the structures and energies of molecules. *Comput. Phys. Commun.* 91:1–41.

Pedersen, J., and Moult, J. 1995. Ab initio structure prediction for small polypeptides and protein fragments using genetic algorithms. *Proteins* 23:454–460.

Peitsch, M. C. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* 24:274–279.

Peitsch, M. C., and Jongeneel, V. 1993. A 3-dimensional model for the CD40 ligand predicts that it is a compact trimer similar to the tumor necrosis factors. *Int. Immunol.* 5:233–238.

Petrey, D., and Honig, B. 2005. Protein structure prediction: Inroads to biology. *Mol. Cell* 20:811–819.

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.

Ponder, J. W., and Richards, F. M. 1987. Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural 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(43):15346–15351.

Rigden, D. J. 2002. Use of covariance analysis for the prediction of structural domain boundaries from multiple protein sequence alignments. *Protein Eng.* 15:65–77.

Rohl, C. A., Strauss, C., Chivian, D., and Baker, D. 2004. Modeling structurally variable regions in homologous proteins with Rosetta. *Proteins* 55:656–677.

Rose, G. D. 1978. Prediction of chain turns in globular proteins on a hydrophobic basis. *Nature* 272:586–590.

Roseman, A. M. 2000. Docking structures of domains into maps from cryo-electron microscopy using local correlation. *Acta Crystallogr. Sect. D Biol. Crystallogr.* 56 (Pt 10):1332–1340.

Rossman, M. G., and Liljas, A. 1974. Recognition of structural domains in globular proteins. *J. Mol. Biol.* 85:177–181.

Rost, B., Casadio, R., Fariselli, P., and Sander, C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* 4:521–533.

Rufino, S. D., Donate, L. E., Canard, L. H. J., and Blundell, T. L. 1997. Predicting the conformational class of short and medium size loops connecting regular secondary structures: Application to comparative modelling. *J. Mol. Biol.* 267:352–367.

Rychlewski, L., Jaroszewski, L., Li, W., and Godzik, A. 2000. Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* 9:232–241.

Sadreyev, R. I., Baker, D., and Grishin, N. V. 2003. Profile–profile comparisons by COMPASS predict intricate homologies between protein families. *Protein Sci.* 12:2262–2272.

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

Samudrala, R., Xia, Y., Huang, E., and Levitt, M. 1999. Ab initio protein structure prediction using a combined hierarchical approach. *Proteins* 37(S3):194–198.

Sanchez, R., Pieper, U., Mirkovi, N., de Bakker, P. I. W., Wittenstein, E., and Ali, A. (2000). MODBASE, a database of annotated comparative protein structure models. *Nucleic Acids Res.* 28:250–253.

Sanger, F., Thompson, E. O., and Kitai, R. 1955. The amide groups of insulin. *Biochem. J.* 59:509–518.

Schroder, R. R., Manstein, D. J., Jahn, W., Holden, H., Rayment, I., Holmes, K. C., and Spudich, J. A. 1993. Three-dimensional atomic model of F-actin decorated with Dictyostelium myosin S1. *Nature* 364:171–174.

Schueler-Furman, O., Wang, C., and Baker, D. 2005a. Progress in protein–protein docking: Atomic resolution predictions in the CAPRI experiment using RosettaDock with an improved treatment of side-chain flexibility. *Proteins* 60:187–194.

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

Scott, R. A., Vanderkooi, G., Tuttle, R. W., Shames, P. M., and Scheraga, H. A. 1967. Minimization of polypeptide energy, III. Application of a rapid energy minimization technique to the calculation of preliminary structures of gramicidins. *Proc. Natl. Acad. Sci. USA* 58:2204–2211.

Shi, J., Blundell, T. L., and Mizuguchi, K. 2001. FUGUE: Sequence–structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.* 310:243–257.

Siddiqui, A. S., and Barton, G. J. 1995. Continuous and discontinuous domains: An algorithm for the automatic generation of reliable protein domain definitions. *Protein Sci.* 4:872–884.

Simons, K. T., Bonneau, R., Ruczinski, I., and Baker, D. 1999a. Ab initio protein structure prediction of CASP III targets using ROSETTA. *Proteins* 37(S3):171–176.

Simons, K. T., Kooperberg, C., Huang, E., and Baker, D. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.

Simons, K. T., Ruczinski, I., Kooperberg, C., Fox, B. A., Bystroff, C., and Baker, D. 1999b. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins* 34:82–95.

Simons, K. T., Strauss, C., and Baker, D. 2001. Prospects for ab initio protein structural genomics. *J. Mol. Biol.* 306:1191–1199.

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

Skolnick, J., Kolinski, A., Brooks, C. L., III, Godzik, A., and Rey, A. 1993. A method for predicting protein structure from sequence. *Curr. Biol.* 3:414–423.

Sucha, S., Dubose, R. F., March, C. J., and Subhashini, S. 1995. Modeling protein loops using a {phi}(i+1), {psi}(i) dimer database. *Protein Sci.* 4:1412–1420.

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

Swindells, M. B. 1995. A procedure for detecting structural domains in proteins. *Protein Sci.* 4:103–112.

Tieleman, D. P., and Berendsen, H. J. 1998. A molecular dynamics study of the pores formed by Escherichia coli OmpF porin in a fully hydrated palmitoyloleoylphosphatidylcholine bilayer. *Biophys. J.* 74:2786–2801.

Topf, M., Baker, M. L., John, B., Chiu, W., and Sali, A. 2005. Structural characterization of components of protein assemblies by comparative modeling and electron cryo-microscopy. *J. Struct. Biol.* 149:191–203.

Topf, M., and Sali, A. 2005. Combining electron microscopy and comparative protein structure modeling. *Curr. Opin. Struct. Biol.* 15:578–585.

Tusnady, G. E., and Simon, I. 1998. Principles governing amino acid composition of integral membrane proteins: Application to topology prediction. *J. Mol. Biol.* 283:489–506.

Ubbink, M., Ejdeback, M., Karlsson, B. G., and Bendall, D. S. 1998. The structure of the complex of plastocyanin and cytochrome f, determined by paramagnetic NMR and restrained rigid-body molecular dynamics. *Structure* 6:323–335.

Vakser, I. A. 1995. Protein docking for low-resolution structures. *Protein Eng.* 8:371–377.

van Dijk, A. D. J., Boelens, R., and Bonvin, A. M. J. J. 2005. Data-driven docking for the study of biomolecular complexes. *FEBS J.* 272:293–312.

van Gunsteren, W. F., and Berendsen, H. J. C. 1990. Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.* 29:992–1023.

van Vlijmen, H. W. T., 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.

von Heijne, G. 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 225:487–494.

Vonderviszt, F., and Simon, I. 1986. A possible way for prediction of domain boundaries in globular proteins from amino acid sequence. *Biochem. Biophys. Res. Commun.* 139:11–17.

Warshel, A. 1976. Bicycle-pedal model for the first step in the vision process. *Nature* 260:679–683.

Warshel, A. 2002. Molecular dynamics simulations of biological reactions. *Acc. Chem. Res.* 35:385–395.

Wetlaufer, D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc. Natl. Acad. Sci. USA* 70:697–701.

Wheelan, S. J., Marchler-Bauer, A., and Bryant, S. H. 2000. Domain size distributions can predict domain boundaries. *Bioinformatics* 16:613–618.

White, S. H. 2004. The progress of membrane protein structure determination. *Protein Sci.* 13:1948–1949.

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

Wodak, S. J., and Janin, J. 1981. Location of structural domains in protein. *Biochemistry* 20:6544–6552.

Wriggers, W., Milligan, R. A., and McCammon, J. A. 1999. Situs: A package for docking crystal structures into low-resolution maps from electron microscopy. *J. Struct. Biol.* 125:185–195.

Wriggers, W., Milligan, R. A., Schulten, K., and McCammon, J. A. 1998. Self-organizing neural networks bridge the biomolecular resolution gap. *J. Mol. Biol.* 284:1247–1254.

Xu, D., Baburaj, K., Peterson, C. B., and Xu, Y. 2001. Model for the three-dimensional structure of vitronectin: Predictions for the multi-domain protein from threading and docking. *Proteins* 44:312–320.

Xu, J., Li, M., Kim, D., and Xu, Y. 2003. RAPTOR: Optimal protein threading by linear programming. *J. Bioinform. Comput. Biol.* 1:95–117.

Xu, Y., and Xu, D. 2000. Protein threading using PROSPECT: Design and evaluation. *Proteins* 40:343–354.

Ye, Y., Li, Z., and Godzik, A. 2006. Modeling and analyzing three-dimensional structures of human disease proteins. *Pac. Symp. Biocomput.* (Maui).

Yip, Y. L., Scheib, H., Diemand, A. V., Gattiker, A., Famiglietti, L. M., Gasteiger, E., and Bairoch, A. 2004. The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. *Hum. Mutat.* 23:464–470.

Yona, G., and Levitt, M. 2002. Within the twilight zone: A sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.* 315:1257–1275.

Yuan, Z., Mattick, J. S., and Teasdale, R. D. 2004. SVMtm: Support vector machines to predict transmembrane segments. *J. Comp. Chem.* 25:632–636.

Zhai, Y., and Saier, M. H. J. R. 2002. The β-barrel finder (BBF) program, allowing identification of outer membrane β-barrel proteins encoded within prokaryotic genomes. *Protein Sci.* 11:2196–2207.

Zhang, Y., and Skolnick, J. 2004. Automated structure prediction of weakly homologous proteins on a genomic scale. *Proc. Natl. Acad. Sci. USA* 101:7594–7599.

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, H., and Zhou, Y. 2005. Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments. *Proteins* 58:321–328.