Protein Structure Prediction: From Recognition of Matches with Known Structures to Recombination of Fragm....

Chapter · October 2010

DOI: 10.1007/978-1-4419-6889-0_10

CITATION
1

READS
31

3 authors, including:

Michal J Gajda
Max Planck Institute for Biophysical Chemistry
28 PUBLICATIONS **1,461** CITATIONS
SEE PROFILE

Janusz Bujnicki
International Institute of Molecular and Cell ...
442 PUBLICATIONS **10,691** CITATIONS
SEE PROFILE

Chapter 10
Protein Structure Prediction: From Recognition of Matches with Known Structures to Recombination of Fragments

Michal J. Gajda, Marcin Pawlowski, and Janusz M. Bujnicki

Abstract The field of protein structure prediction has been revolutionized by the application of "mix-and-match" methods both in template-based homology modeling, as well as in template-free, "de novo" folding. Automated generation of models that are closer to the native structure of the target protein than the structure of its closest homolog is currently possible by recombination of fragments copied from known protein structures or extracted from alternative starting models. It is also the most successful approach in the cases where the target protein exhibits a novel three-dimensional fold. This chapter is an update of a review article published earlier [Bujnicki JM, Chem Bio Chem 7(1):19–27, 2006 Jan 9, Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission]. It summarizes the recent developments in both template-based and template-free protein structure modeling and compares the available methods for protein structure prediction by recombination of fragments.

10.1 Introduction

The high-resolution three-dimensional structure of a protein is the key to the understanding and manipulation of its biochemical and cellular function. However, the rate of protein structure determination by X-ray crystallography lags behind the rate of determination of new protein sequences. As of January 2010, the National Center for Biotechnology Information's Non-Redundant RefSeq database (Pruitt et al. 2007) contained 9,662,677 sequences, while the Protein Data Bank (Berman et al. 2000) contained only 36,043 protein structures with non-redundant sequences

J.M. Bujnicki (✉)
Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology, Warsaw, Poland; Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Adam Mickiewicz University, Poznan, Poland
e-mail: iamb@genesilico.pl

A. Kolinski (ed.), *Multiscale Approaches to Protein Modeling*,
DOI 10.1007/978-1-4419-6889-0_10,

(50,495 structures total). Currently, the size of the sequence database doubles approximately every 5 years, while the structure database doubles every 8 years. Thus, the gap between the number of known structures and known sequences will continue to widen in the foreseeable future and it is unlikely that it will be ever closed, e.g., the structures will never be solved experimentally for all proteins.

Almost 50 years ago Anfinsen demonstrated that all of the information necessary for RNase A to fold to the native structure is contained in its amino acid sequence (Anfinsen et al. 1961). This finding has been generalized to most globular proteins, suggesting that a protein's structure could be calculated (modeled) based on the knowledge of its sequence and our understanding of the sequence–structure relationships. Thus, the current structural genomics initiative aims to solve experimentally the structures of representative proteins, while the others are hoped to be modeled computationally (Baker and Sali 2001). The theoretical prediction of the native structure of a protein from its amino acid sequence remains, however, one of the most challenging problems in contemporary life sciences.

10.2 Protein Structure Prediction Methods: Classification and Critical Evaluation

Efforts to solve the protein-folding problem have been traditionally rooted in two schools of thought (Fig. 10.1). One is based on the principles of physics, e.g., on the thermodynamic hypothesis formulated by Anfinsen, according to which the native structure of a protein corresponds to the global minimum of its free energy (Anfinsen 1973). Accordingly, physics-based methods model the process of protein folding by simulating the conformational changes and searching for the free-energy minimum. The other school of thought is based on the principles of evolution. After experimental determination of the first handful of protein structures it became clear that evolutionarily related (homologous) proteins usually retain the same three-dimensional fold (i.e., the arrangement and connectivity of elements of secondary structure) despite the accumulation of divergent mutations (Chothia and Lesk 1986). It was also found that structural divergence is much slower than sequence divergence, although these two features are strongly correlated. Thus, methods have been developed to map the sequence of one protein (a target) to the structure of another protein (a template), model the overall fold of the target based on that of the template, and infer how the target structure will change due to substitutions, insertions, and deletions, as compared with the template (reviews: Cohen-Gonsaud et al. 2004; Krieger et al. 2003). Table 10.1 summarizes the key features of methods discussed in this chapter.

Accordingly, methods for protein structure prediction have been divided into two classes: "de novo" modeling, in principal applicable to all types of proteins, including those for which no appropriate templates are available, and "comparative (homology) modeling" (CM), in which the target sequence must be aligned to an evolutionarily related, experimentally solved template structure. In this context it is

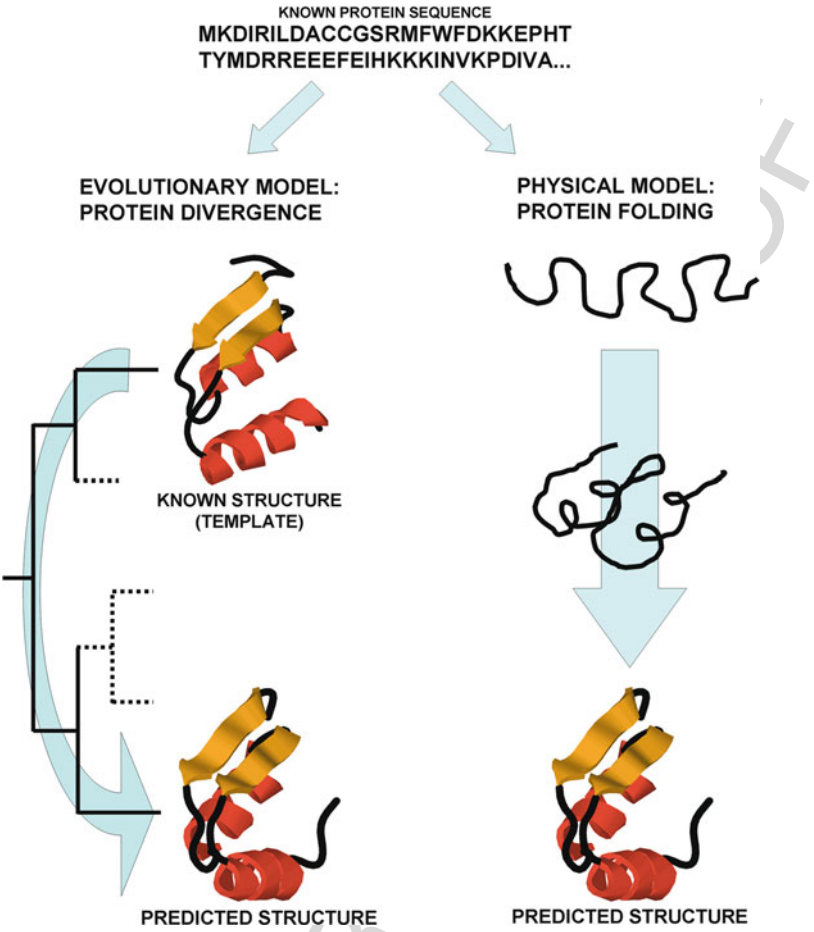

Fig. 10.1 The "evolutionary" and "physical" approaches for protein structure prediction**.** Given the amino acid sequence, a simulation of either protein evolution or protein folding is carried out, according to quantitative models of either divergence of sequences and structures or physical interactions within the molecule and between the molecule and the solvent. (Bujnicki, 2006, Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

worthwhile to remind that protein structure can be described by a hierarchical system, with levels corresponding to primary sequence (covalent bonding of amino acids), secondary structure (segments of recurring arrangement of amino acids consecutive in the sequence), tertiary structure (mutual arrangement of secondary structures in a protein domain), and quaternary structure (mutual arrangement of domains within a multi-domain protein or different subunits in a multi-protein complex). Knowledge-based methods are usually effective on all levels of this hierarchy, first because evolutionary forces tend to preserve all aspects of protein structure. And if homology is present, it often manifests itself at all levels of hierarchy.

Table 10.1 Summary of key features of methods discussed in this chapter. This table lists all prediction methods discussed in this chapter, along with their distinguishing features. It is intended as a reference and guide for contrasting approaches

Method	Type	Search strategy	Evaluation/selection of models	Input and/or fragment library
SWISS-MODEL	CM	Superposition of templates	NA	CM templates, loops from PDB
PCONS5 (PROQ)	FR/CM meta-selector	Superposition of models	Statistical potential (PROQ)	FR models
3D-JURY	FR/CM meta-selector	Superposition of models	NA	FR models
3D-SHOTGUN	FR/CM fragment splicer	Superposition and recombination of models	NA	FR models
FRANKENSTEIN3D	FR/CM fragment splicer	Superposition and recombination of models	Statistical meta-potential (MetaMQAP)	FR models
GS-KudlatyPred	FR/CM/de novo fragment splicer	Recombination of supersecondary structure fragments form any kind of initial models	Statistical meta-potential (MetaMQAPcons)	FR/CM/de novo models
In silico protein recombination	CM fragment splicer	Superposition and recombination of models, local realignment	Statistical potential	Comparative models with similar folds
GENETIC ALGORITHM	CM alignment splicer	Recombination of alignments, local realignment	Statistical potential	alternative target-template alignments
ROSETTA	De novo fragment splicer	Monte Carlo-simulated annealing	Physical energy function with elements of a statistical potential, clustering	3 and 9 aa fragments from PDB
GINZU/ROBETTA	FR/CM/de novo fragment splicer	Merging of domain models, Monte Carlo-simulated annealing	FR score and statistical potential	3 and 9 aa fragments from PDB, FR models
SIMFOLD	De novo fragment splicer	Multi-canonical ensemble Monte Carlo	Physical energy function	4–9 aa fragments from PDB
PROFESY	De novo fragment splicer	Conformational space annealing	Physical energy function	15 aa fragments from PDB

Table 10.1 (continued)

Method	Type	Search strategy	Evaluation/selection of models	Input and/or fragment library
FRAGFOLD	De novo fragment splicer	Simulated annealing or genetic algorithm	Statistical potential	Supersecondary structures and 3–5 aa fragments
UNDERTAKER	De novo fragment splicer	Genetic algorithm	Statistical potential	Fragments of FR models, and 1–4 aa and 9–12 aa fragments from PDB
ABLE	De novo fragment splicer	Monte Carlo-simulated annealing, iterated with restraints from previous rounds	Physical energy function with elements of a statistical potential	Individual residues
TASSER	FR/CM/de novo fragment splicer	Replica Exchange Monte Carlo (on a lattice)	Statistical potential, clustering	FR models
I-TASSER	FR/CM/de novo fragment splicer	Replica Exchange Monte Carlo (on a lattice), iterative enrichment of fragments with templates analogous to ab initio models from previous rounds	Statistical potential, clustering	FR models, SCOP folds most similar to ab initio models
FRANK/CABS	FR/CM/de novo fragment splicer	Replica Exchange Monte Carlo (on a lattice)	Statistical potential	FR models
ZAM	De novo fragment splicer	zipping and assembly from smaller fragments	Physical energy function with elements of a statistical potential	Individual residues
SALAMI	De novo fragment splicer	Bayesian fragment picking, torsion angle dynamics	Steric clashes, clustering	Alphabet of 300 fragments each 6 residues long
Structural descriptors	FR	NA	NA	Descriptors (groups of >2 fragments; >3 residues long)

Further, in homology-based modeling errors at different level of structural hierarchy are largely independent from each other (e.g., it is not difficult to correctly predict a protein fold without getting all secondary structures correctly and without any consideration of the quaternary structure). This has been demonstrated in the course of Critical Assessment of Techniques for Protein Structure Prediction (CASP), as many modelers have generated models with correct folds while completely disregarding the quaternary structure, and with significant errors with respect to secondary structure alignment (Moult et al. 2005, 2007, 2009). On the other hand, "de novo" methods typically require accurate prediction at low levels of hierarchy in order to correctly predict higher levels. Their advantage, however, is that they are independent on the modeling of the primary sequence (they do not attempt to model sequence alignment between the target sequence and some other sequence).

The "de novo" approach can be further subdivided into "ab initio" methods, i.e., those based exclusively on the physics of the interactions within the polypeptide chain and between the polypeptide and the solvent (Scheraga 1996), and "knowledge-based" methods that utilize statistical potentials based on the analysis of recurrent patterns in the known protein structures and sequences (Kolinski 2004). De novo methods may utilize different representations of the protein chain, frequently employing either coarse-grained models (see the chapter by Liwo et al in this volume) or fragments of experimentally solved structures (from parts of the side chain or the backbone, to individual residues, to groups of residues up to the size of a few elements of secondary structure).

The CM approach can also be subdivided into two main trends. One is to model the structure by copying the coordinates of the template (both the backbone and the side chains) in the aligned core regions, which can also include "averaging" over coordinates of multiple templates. The variable regions are modeled by taking fragments with similar sequence from a database of previously observed loops, followed by replacing the mutated side chains with rotamers that satisfy the stereochemical criteria, and (optionally) limited energy optimization, as implemented in SWISS-MODEL (Peitsch 1995). The other possibility is to use the distance and torsion angles and interatomic distances from the aligned regions of the template(s) as modeling restraints, which permits the use of information from multiple, possibly conflicting structures. This approach also requires to idealize the geometry and packing of the entire chain by satisfying stereochemical constraints derived from the database of protein structures, as implemented in MODELLER (Sali and Blundell 1993). The CM approach has been also extended to "fold-recognition" (FR), where one attempts to identify a template with a similar fold that does not need to exhibit significant sequence similarity to the target (e.g., the target and the template may or may not be homologous, but they need to share the common fold) (Godzik et al. 1992; Jones et al. 1992). While the early FR methods relied mostly on the "threading" approach, i.e., evaluation of protein energy as the sum of pairwise residue–residue interactions based on physical or statistical potentials, nearly all contemporary FR methods are based mostly (or exclusively) on sequence comparisons and are tuned to detect distantly related homologs rather than unrelated structural analogs (reviews: Cymerman et al. 2004; Ginalski et al. 2005).

Another way of subdividing the "comparative" approach is into orthodox (traditional) methods that use entire proteins (or domains) as templates and methods that use different (not necessarily related) structures or their fragments to model different parts of the target sequence (which are discussed in this review). With the decreasing size of fragments, the latter type of comparative methods blends with de novo methods that represent protein structures with fragments of known structures.

Recently, a new generation of protein structure prediction methods has been developed that combine comparative modeling with de novo modeling. Typically, an initial model or its significant part is modeled by comparative approach, based on the general prediction of the three-dimensional fold, and then the entire structure or its part is "refined" by de novo methods, often in connection with evaluation of local quality by Model Quality Assessment Methods (MQAPs) (Kryshtafovych and Fidelis 2009). This review describes examples of all the above-mentioned approaches.

In order to objectively assess the abilities (and inabilities) of different methods for protein structure prediction, Moult and coworkers organized the biennial Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction (CASP) to compare computationally predicted protein structures with the "golden standard" of experimentally determined ones. The first assessment experiment (CASP1) was held in 1994 and revealed that computational methods for protein structure prediction perform quite poorly, those based on physics and evolution alike (Moult et al. 1995). Since then, the progress in the field of protein structure prediction has been significant, especially in the "template/knowledge-based" category (e.g., CM and FR), in part due to the improvement in methodology but mostly because of the rapid growth of databases and accumulation of new potential template structures as well as numerous new sequences that can serve as convenient "evolutionary intermediates" in the homology searches. Nonetheless, it appears that in the recent years there has been little progress, if any, in the ability of both comparative and de novo methods (Moult et al. 2009).

10.3 "Meta" Approaches to Template-Based Prediction

The series of CASP experiments have shown that the combined use of human expertise and automated methods can often result in successful predictions. This has became especially clear in the cases of very remote homology, where most FR methods return predictions with scores indicating the lack of statistical significance and correct models are "buried" among a number of incorrect models. A group of four human predictors including Daniel Fischer, Leszek Rychlewski, Arne Elofsson, and Janusz M. Bujnicki, pioneered the idea of "meta-prediction" in CASP4, by comparing the models generated by FR servers participating in the satellite experiment CAFASP-21 and submitting manually selected "consensus" predictions as the "CAFASP-CONSENSUS" group. This group performed better than any of the

individual servers and ranked seventh among all predictors of CASP4 (Fischer et al. 2001). Thus, it was demonstrated that the recurrence of a particular protein fold within the sets of top ten models returned by different servers (and not necessarily at the first position of their ranking) increases the likelihood of a correct prediction and that on the average, no single FR method is better than the combination of a few top methods. Since then, meta-prediction based on FR (Fig. 10.2) has become the most successful approach for template-based modeling and has been applied by a large number of human predictors, including the best performers in CASP5 and CASP6.

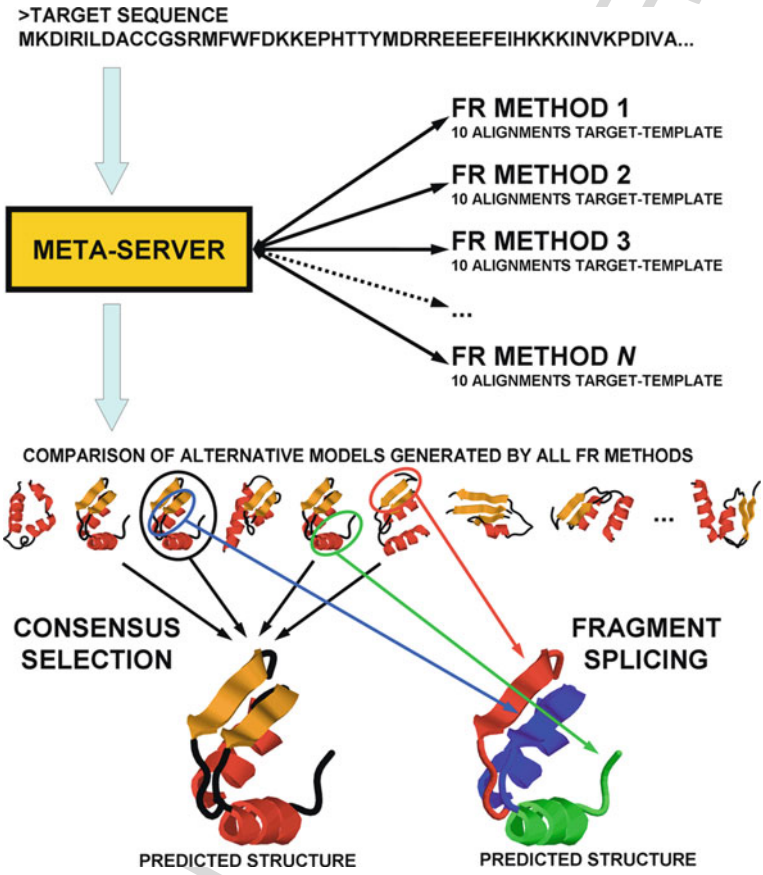

Fig. 10.2 The meta-server approach for protein structure prediction. The meta-server is used as a gateway to send the target sequence to various "primary" fold-recognition servers, collect the results (target-template alignments), build the corresponding models, compare them with each other, and either select the most representative structure or construct a hybrid model from the most frequently represented fragments (Bujnicki, 2006, Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission)

Following the proven success of manual "meta-predictors," several groups have implemented fully automated "meta-servers" (Bujnicki et al. 2001b; Douguet and Labesse 2001; Kurowski and Bujnicki 2003; Wu and Zhang 2007a). One of the earliest meta-predictors was the neural network PCONS developed by Elofsson and coworkers (Lundstrom et al. 2001), which collects a set of top models generated by different FR servers and selects the models that were most similar to other models in the set. The second edition of an independent assessment experiment LiveBench (Bujnicki et al. 2001a) organized shortly after CASP4 revealed that PCONS2 (version trained for a few specific "primary" FR servers) exhibited the sensitivity comparable to the most sensitive primary method and the specificity higher than any primary method. The newest version of PCONS is reinforced by methods for protein model evaluation (Wallner and Elofsson 2003), exhibits even higher specificity and is able to use as an input model generated by any set of methods. PCONS is available as a part of various meta-servers, as well as a standalone server Pcons.net (http://pcons.net/) (Wallner et al. 2007).

3D-JURY developed by Rychlewski and coworkers (Ginalski et al. 2003) is another automated meta-predictor that simply selects models from those produced by other servers. It takes as input any set of models, compares all against all, and selects one that appears to contain the largest subset of commonly superimposable coordinates. The most important feature contributing to the success of 3D-JURY and its popularity among the users is its scoring system, which allows confidently identifying the models with correctly predicted fold, even though it does not necessarily recognize the absolutely best model among similar "top solutions." 3D-JURY is available as an integral part of the Bioinfo.pl meta-server (http://meta.bioinfo.pl/). Based on this algorithm, a very similar method 3D-Consens has been implemented as an optional post-processing tool in the GeneSilico meta-server (https://genesilico.pl/meta2/).

10.4 From Multiple Template-Based Models to Hybrids

Apart from meta-predictors that simply select models from the input set, another breed of meta-predictors have been developed that use the unrefined models generated by primary servers as a "structural scrap-yard" from which to obtain "spare parts" to generate new models. 3D-SHOTGUN developed by Fischer (Fischer 2003) was the first fully automated meta-predictor designed to assemble hybrid models from fragments of models obtained from independent FR methods (e.g., from different components of the BIOINBGU server (Fischer 2000)). In the first step, regions of structural similarity are identified for all initial models by pairwise superposition. Subsequently, for each residue in each model, the number of its occurrences in the superimposed regions of other models is counted and a hybrid model is assembled by taking the coordinates of each residue from a model having the highest count. Thus, for each initial model a hybrid model is constructed that contains the most common structural features of all models, often including more residues than any

of the initial models. In the second step, the assembled models are assigned scores based on the combination of the original scores of their parent models (normalized to a similar scale) and the scores describing the structural similarity of the assembled model to other models, as determined by MAXSUB (Siew et al. 2000). For each cluster of highly similar assembled models only one representative model with the highest score is reported.

The rationale of the 3D-SHOTGUN strategy is the same as in the consensus methods (selectors) acting on complete models, namely, that recurring structural features observed in models obtained from different FR methods are more likely to be correct. The initial version of 3D-SHOTGUN generated models that comprised only Cα atoms and commonly exhibited stereochemical problems, such as implausible distances and angles and steric clashes between fragments taken from different parent models. In terms of coverage and root mean square deviation (RMSD) between the model and the native structure, however, the approach of hybrid model construction is superior to selection of one of the stereochemically more acceptable input models, as the hybrids are on the average more complete and superimpose better on the native structure than the initial models. The method is sensitive to initial alignment errors – if none of the initial alignments is correct for a given region, it is unlikely that this region will be modeled correctly in the final structure. A new automated version SHGUM includes a crude refinement step, using MODELLER (Sali and Blundell 1993), to generate full-atom models with idealized stereochemistry and without gaps and collisions and even with a slight improvement in the overall RMSD (Sasson and Fischer 2003). The method is available via the INUB server at http://inub.cse.buffalo.edu/query.html.

FRankenstein's Monster is another approach to meta-prediction by consensus and recombination of fragments, developed in the authors' laboratory (Kosinski et al. 2003, 2005). It is similar to 3D-SHOTGUN, but goes beyond the identification of geometrical consensus by including evaluation of the models by statistical potentials and features an additional step of local realignment of uncertain regions. This helps not only to overcome the problem of selection of the optimal template, but also to correct initial alignment errors. Briefly, the GeneSilico meta-server (Kurowski and Bujnicki 2003) is used as a gateway to run diverse FR methods and to generate preliminary full-atom models from initial pairwise target-template alignments. The local quality of sequence–structure fit in these models is evaluated by a fitness function of local MetaMQAP score (Pawlowski et al. 2008). The most probable folds are identified by clustering. For each fold, a hybrid model is assembled from fragments that are structurally similar in >40% of all preliminary models, while the remaining non-consensus fragments are selected based on the MetaMQAP local score. The initial hybrid model (the "FRankenstein's monster") typically exhibits stereochemical problems similar to those found in models generated by the 3D-SHOTGUN method. However, in the FRankenstein strategy the hybrid model is not directly refined, but instead it is superimposed onto the structures of the templates used, yielding a new target-multiple template sequence alignment, which is used to generate a new, stereochemically acceptable model by an "orthodox" CM procedure. The sequence–structure fit in the new model is re-evaluated with

MetaMQAP, and regions of low local score are selected for further refinement. For each poorly scored non-consensus region, a set of new alignments is generated by progressively shifting the target sequence with a step of one residue in the direction of either terminus, within the region of overlap between the secondary structure elements found in the template structure and those predicted for the target. All resulting alignments are used to generate a new family of intermediate models, which are again evaluated and recombined to produce a hybrid model. The procedure is iterated until all regions in the protein core obtain acceptable score or if the score cannot be further improved.

The FRANKENSTEIN3D method generates models that retain the fragments confidently predicted by consensus (regardless of their fitness according to statistical potentials) and attempts to refine the alignment in the uncertain regions to maximize the fitness score. As demonstrated in CASP5 (Kosinski et al. 2003) and CASP6, where the groups from the author's laboratory ranked very high in the CM and FR categories (Kolinski and Bujnicki 2005; Kosinski et al. 2005), the application of this approach leads to very accurate target-template alignments, often more accurate than any of the initial alignments, provided that a template with a correct fold is identified by at least one of the FR servers used. The method automatically clusters all templates available from FR servers and allows to automatically build multiple alternative models before alignment optimization. The current version of the method is available as a FRANKENSTEIN3D server at http://genesilico.pl/frankenstein/.

Another approach to overcome the problem of template selection and correction of alignment errors by recombination of alternative models was developed by Bates and coworkers (Contreras-Moreira et al. 2003a,b). The in silico Protein Recombination method starts with a population of models built from alternative alignments to one or more templates sharing the same fold and uses a genetic algorithm with two mutually exclusive genetic operators – "recombination" of parent models with crossover points outside the regions of secondary structure and "mutation" by averaging the coordinates of two parent models. The fitness function acting as a selection agent is a free-energy estimate based on protein contact pair-potentials and side-chain solvation energies, estimated from their solvent accessible area. The method was shown to be able to improve alignments by recombining well-aligned regions from the initial models and to produce recombinant models that are comparable to the best initial model. However, the quality of the initial models is the upper limit for the quality of the final model (e.g., unlike the FRANKENSTEIN method, it does not produce new, potentially better alignments). It is also critically dependent on the confident identification of a correct fold. The in silico Protein Recombination method is available as a web server at http://www.bmm.icnet.uk/servers/3djigsaw/recomb/.

Another method that implements a genetic algorithm for comparative modeling was developed by Sali and coworkers (John and Sali 2003). It is similar to the FRANKENSTEIN3D approach in that it continuously refers to the target-template alignments, modifies them locally, and assesses the result of these changes by evaluation of the corresponding models, generated by MODELLER (Sali and Blundell 1993). The genetic operators include recombination of the parent alignments (one

and two-point crossovers) and gap insertions/deletions/shifts that actually generate local changes in the parent alignment. The fitness function is based on a score that combines the evaluation of the model by a statistical potential (Melo et al. 2002), target-template sequence identity, and a measure of structural compactness. The method was shown to increase the average quality of the target-template alignment and the corresponding models, but is dependent on the initial choice of the templates; in addition, the inaccurate statistical potential is generally unable to choose the best model (John and Sali 2003).

A recent addition to the repertoire of "recombinators" that act on the level of target-template alignments is the MULTICOM-cluster method (which is also a Model Quality Assessment protocol). It performs greedy merging of the top-scoring FR alignment with other alternative alignments that can fill gaps in the structure, generates models with MODELLER, and clusters the resulting models. In the first iteration, the clustering is used to find references among all single-template models, to obtain a global quality score for each model. In the second iteration, models are compared to references to obtain local quality scores, based on average superposition accuracy to the best reference. The models with best global quality are identified as solutions (Cheng et al. 2009).

10.5 Fragment Assembly: A New Trend in De Novo Protein Structure Prediction

Modeling protein structure "de novo" without the template is very difficult, because the conformational space to be searched is so vast that it is practically impossible to simulate the folding of a model that includes all atoms of the polypeptide chain and the surrounding solvent molecules. Methods and resources currently available allow to simulate up to about 1 ms of folding of full-atom representations of only small proteins (<100 residues), while most proteins are larger and fold in the timescale of milliseconds or even seconds. Therefore, the solvent is usually treated implicitly and various simplified models are used that have fewer degrees of freedom and exploit the repetitive nature of protein structure (see the chapter by Liwo and coworkers in this volume). These simplified models typically retain only certain atoms, such as Cα or Cβ or "united atoms" in which several atoms are grouped together, such as the centers of mass of the side chains (Liwo et al. 2005; Sun 1995). The protein structure may be represented using a number of simplified schemes such as lattices or bond angles with discrete values (Geetha and Munson 1996; Kolinski 2004). Despite the considerable reduction of dimensionality of the structure space in simplified models, the polypeptide main chain remains highly flexible and requires many variables per residue to model the protein conformation accurately (Hunter and Subramaniam 2002).

Significant progress in the field of "de novo" protein structure prediction has been prompted by the observation that the structure of protein backbone can be represented quite accurately by using short fragments taken from other proteins (Claessens et al. 1989; Jones and Thirup 1986). Fragments up to ten residues long

provided an efficient method for interpreting electron density maps in protein crystallography and in building protein models from nuclear magnetic resonance (NMR) data (Kraulis and Jones 1987). Classification of protein loops has proven useful in comparative modeling, where the incomplete framework of a protein core has to be amended by "de novo" insertion of polypeptide segments (Donate et al. 1996; Oliva et al. 1997; Tramontano et al. 1989) (see also above). Several groups classified peptide backbone units with fixed or variable lengths into collections of fragments (Bystroff and Baker 1997; Camproux et al. 1999; Kolodny et al. 2002; Micheletti et al. 2000; Unger et al. 1989). Analysis of such recurring fragments identified local sequence–structure correlations in proteins (Bystroff et al. 1996; Han and Baker 1996) and suggested a new method for "de novo" protein structure prediction.

10.5.1 De Novo Modeling by Fragment Assembly (and Subsequent Refinement)

ROSETTA developed by Baker and coworkers (Simons et al. 1997) implements a model of folding in which short fragments of the protein chain alternate between different local conformations copied from segments of known, not necessarily homologous, protein structures. The probability of assuming a particular conformation is based on the similarity of the local sequence and predicted secondary structure of the target to sequence and structure from the template library, as in the "traditional" template-based methods for protein modeling. The early version of ROSETTA used the I-SITES library of fragments 7–19 residues in length that corresponded to one of 82 patterns of sequence/structure motifs commonly present in all known structures (Bystroff and Baker 1998). The current version uses a library of fragments 3–9 residues in length extracted from known structures, which are assembled by using a Monte Carlo (MC)-simulated annealing (SA) search strategy, in which fragments are randomly inserted into the protein chain by replacing the backbone torsion angles with those in the fragment. The resulting "decoy" conformation is then evaluated according to a database-derived pseudoenergy function that rewards native protein-like properties. Additionally, a number of heuristic filters can be used to discriminate "protein-like" decoys by virtue of contact order, topology of β-sheets, etc. In the standard protocol, ROSETTA uses a reduced representation with backbone heavy atoms and Cβ atoms explicitly included and the side chains represented by single centroids. ROSETTA is also capable of refinement of models using full-atom representation, special conformation modification operators, and a refined (more "physical") energy function. During the simulation, a large set of decoys are generated (1,000 to many thousands or even millions, depending on the protein length and computing power used), which are then clustered to identify the largest populations of similar global conformations which correspond to the broadest free-energy minima. Full-atom models with explicit side-chain rotamers can be rebuilt before or after clustering (see the recent review of ROSETTA by Das and Baker (2008)).

The difference between ROSETTA and most of template-based methods for fragment recombination lies in the stochastic and iterative character of this process and in the utilization of multiple small fragments of different, unrelated proteins (template-based methods use the whole structure of one protein or a few-related templates). Thus, ROSETTA can generate a "de novo" model by allowing the full-length polypeptide chain to explore conformational space via the fragment insertion search method, even if no homologous or analogous template structure is available. Nonetheless, if a template structure is available, the conserved parts of the target can be built as in traditional CM, while the variable parts are allowed to explore the conformational space with fragments in fashion similar to the "de novo" protocol, but in the context of the template (Rohl 2005). As demonstrated already in CASP5 (Bradley et al. 2003), ROSETTA is capable of generating native-like protein models either de novo (i.e., without any template structure) or by adding long insertions and N- and C-terminal extensions to the template that matches only a part of the modeled protein.

Among the "winners" of CASP, ROSETTA continues to be the only "de novo" method that has been made available to the academic community both as a source code of the standalone program and as a web server. It is available in two versions: as a part of the GINZU/ROBETTA meta-server developed by the Baker group (Kim et al. 2004) (http://robetta.bakerlab.org/) that uses a hybrid strategy of template-based, if template is available, and template-free modeling if not, and in conjunction with the alternative fragment library I-SITES and the fragment assignment method HMMSTER (http://www.bioinfo.rpi.edu/~bystrc/hmmstr/), developed and maintained by the Bystroff group (Bystroff and Shao 2002).

SIMFOLD is another fragment-assembly method developed by Shoji Takada and coworkers (Chikenji et al. 2003). The original method, which performed quite well in CASP6 as "ROKKO" and "ROKKY", is similar to ROSETTA. It uses 4–9 residue fragments and its energy function consists of various interactions that are based on physical considerations (Fujitsuka et al. 2004). Yet, SIMFOLD exhibits an important difference: it introduces reversible fragment insertion, e.g., when a new fragment is inserted at a junction between two fragments, the replaced "old" conformation comprising elements of two different fragments is added to the library of fragments, so it can be re-inserted. The latter operation is not possible in ROSETTA, which uses only fragments from the original database. This modification satisfies the detailed balance condition, providing the basis for the application of a multicanonical ensemble Monte Carlo method (MEMC), as used by the human team ROKKO in CASP6. MEMC is more effective in finding low-energy conformations than the conventional SA method (Chikenji et al. 2003). SIMFOLD has been available as a server, but unfortunately ceased to be available publicly.

PROFESY developed by Lee and coworkers (Lee et al. 2004) is similar to SIMFOLD in that it attempts to improve the poor sampling efficiency of the traditional SA method and uses a physics-based energy function rather than a statistical potential. The global minimization of the energy function is thus performed by the conformational space annealing (CSA) method (Liwo et al. 1999). The fragment library is constructed using the secondary structure prediction method

PREDICT and comprises a collection of 15-residue long backbone structures. To our knowledge, this method is not yet publicly available.

FRAGFOLD developed by Jones uses supersecondary structural fragments (comprising 2 or 3 sequential secondary structures) from a library of high-resolution protein structures as well as small (3, 4, and 5 residues) fragments (Jones and McGuffin 2003). Possible supersecondary fragments are assigned to the target sequence by a threading procedure similar to that in the GenTHREADER FR algorithm (Jones 1999). The global structure is assembled by a genetic algorithm or a simulated annealing method, in which half of random moves correspond to the insertion of a pre-selected supersecondary structure fragment and the other half involve a completely free choice of one of the small fragments. Conformations that lack steric clashes and pass the checks for protein-like compactness and hydrogen bonding are clustered to identify representatives of the most probable folds. FRAGFOLD is not available as a server, but a standalone version that runs on GPU exists (http://bioinfadmin.cs.ucl.ac.uk/downloads/gpufragfold/).

UNDERTAKER is a method developed by Karplus and coworkers that assembles the target structure using fragments of known structures obtained from three sources: a generic library of very short segments (1–4 residues) that must exactly match the target sequence, medium-length segments (9–12 residues) that are assigned by the FRAGFINDER program from the SAM suite, and variable-length segments assigned by FR analysis (Karchin et al. 2003). In addition to fragment replacement, UNDERTAKER implements an alignment replacement operation in which a complete FR match is imported into the model, allowing the replacement of several segments at once in the same orientation as they occur in the template structure. UNDERTAKER uses a genetic algorithm for the stochastic search and includes a crossover operation that allows recombining different conformations. The cost function used to assess the decoys includes many tunable parameters, among which the most important one, as the name of the method implies, is the burial. To our knowledge, UNDERTAKER is not yet publicly available.

ABLE developed by Shimizu and coworkers (Ishida et al. 2003) is also based on fragment assembly, but it assigns main-chain dihedral angles individually to each residue. The energy function is similar to that used in ROSETTA. ABLE method has two interesting features that help to avoid problems if the initial distribution of decoys is too broad and no clusters can be identified based on the RMSD as a measure of the distance between the conformations. First, it uses the unit-vector root mean square distance (URMS) (Kedem et al. 1999) as a measure of structural similarity. Second, if not enough clusters with sufficient size and density are obtained, the fragment assembly search is reiterated, but with additional spatial restraints obtained from the consensus substructures in the models generated by the previous minimization procedure. To our knowledge, ABLE is not yet publicly available.

The SALAMI method developed by Andrew Torda and coworkers uses a small alphabet of 300 fragment templates derived from clusters of 6-residue fragments in a non-redundant version of the Protein Data Bank (PDB) database (Margraf et al. 2009). Each fragment template includes information about average geometry of residues and distribution of deviations from average values. These distributions

are used to derive energies that along with steric clash potentials are used to find a lowest-energy conformation. This conformation is a result of simulated annealing in the torsion angle space for each of randomly chosen sequence of fragments. Results for different sequences of fragments are treated as set of decoys and clustered to obtain the final solution. This approach gives more complete coverage of fragment space than many other methods, because it generalizes single, average geometries found in ROSETTA (for example) to fragment template distributions, defined as histograms of torsion angles. SALAMI is available as a web server at http://www.zbh.uni-hamburg.de/salami.

10.5.2 Hybrid Methods Involving Fragment Assembly and Folding Simulations

ZAM is an interesting method developed by the group of Kenneth Dill, which uses a unique combination of assembly and ab initio-folding simulation (Ozkan et al. 2007). The target-protein sequence is divided into small pieces, which are folded by a physics-based ab initio method and only then assembled. It will be interesting to see if computational methods such as this one that explore conformation of fragments by ab initio calculations will match the performance of de novo methods that simply derive the fragments from a database of known structures. To our knowledge, the ZAM method is unfortunately not available publicly.

An alternative approach to fragment assembly, and one with a long history, is that of the lattice representation, in which residues are restricted to points on a regular three-dimensional lattice (Hinds and Levitt 1992; Skolnick and Kolinski 1991). These methods allow very fast sampling of the conformational space, but their ability to represent the atomic details and to use physics-based energy function is limited. Following the success of fragment-assembly methods, several hybrid methods arose, which combine the strengths of both approaches.

TASSER developed by Skolnick and coworkers (Zhang and Skolnick 2004a) starts with an FR analysis based on the PROSPECTOR threading method (Skolnick et al. 2004) that identifies either a single consensus fold or a set of templates with globally distinct folds. Based on the FR alignments, the protein chain is divided into contiguous aligned regions of at least 5 residues (20.7 residues on average, according to the authors' own benchmark) and gapped unaligned regions. The conformation of aligned regions is copied from the templates and remains unchanged during the assembly, while the unaligned regions are represented on an underlying cubic lattice as in the earlier models developed by Skolnick, Kolinski, and coworkers (Kihara et al. 2001). A series of initial models is generated and submitted for assembly and refinement to parallel hyperbolic Replica Exchange Monte Carlo (REMC) sampling method. Structures generated in the lowest-temperature replicas are subjected to iterative clustering using SPICKER (Zhang and Skolnick 2004b) to identify the final models based on the cluster density. I-TASSER is a newer variant of this method, which substitutes fragments modeled de novo for

analogous fragments found in known structures in the PDB database and uses them to provide restraints to guide the folding simulation during the following iteration. TASSER and its variants have been very successful in CASP, since CASP6. I-TASSER is available as a web server at http://zhanglab.ccmb.med.umich.edu/I-TASSER/ (Wu et al. 2007). TASSERLite, a version that allows for refinement of comparative models but not full de novo modeling (Pandit et al. 2006), is available from the Skolnick group website (http://cssb.biology.gatech.edu/skolnick/webservice/tasserlite/index.html).

Another hybrid method, involving the recombination of fragments and lattice-based modeling, was developed by the author of this article in cooperation with Andrzej Kolinski, by combining the FRankenstein's Monster method (Kosinski et al. 2003) (see also above) for generation of initial models, with the reduced lattice model CABS (Kolinski 2004). Briefly, preliminary hybrid models are generated with the template-based recombination method and scored with an MQAP software to identify well-folded fragments, as described earlier. These fragments are not considered directly, but are used as a source of spatial restraints to guide the REMC-folding simulation using the CABS model. The resulting decoys are clustered using the HCPM method (Gront and Kolinski 2005) to identify the final models. This method performed very well in CASP6 evaluation (Kolinski and Bujnicki 2005), but has been never implemented as fully integrated and automated software.

One of the authors of this chapter (M.P.) has recently developed a new fragment-based method, which has participated as "GS-KudlatyPred" in CASP8 (Pawlowski 2009). This method is to some extent related to the "FRankenstein's Monster approach", but operates essentially only on the three-dimensional fragments, without toggling between the three-dimensional and alignment representation. As an input this method takes a set of models built with any methods (comparative or de novo) and extracts fragments comprising 1–4 sequentially occurring secondary structures. Each of the initial models is scored by MetaMQAPcons (a derivative of method described in the article by Pawlowski et al. (2008)). Five models with the best global score are selected as the reference models. All possible combinations of supersecondary structure fragments are generated and ranked based on the sum of three components: (1) MetaMQAPcons score of all fragments in the combination; (2) a GDT_TS score describing the fit of each fragment onto the closest "root model"; and (3) degree of structural similarity in the area of overlap between neighboring fragments. In the last step, hybrid models are built for 200 top-scored combinations of fragments, using MODELLER (Fiser and Sali 2003) in a multi-template mode and ranked by the MetaMQAPcons method.

10.5.3 Other Methods Based on Fragment Prediction

All the aforementioned methods for protein structure prediction by recombination use contiguous fragments of protein backbone. Another approach to protein

structure prediction is based on the concept of three-dimensional structural descriptors developed by Kryshtafovych and Fidelis (unpublished analysis cited in Hvidsten et al. (2003)), e.g., substructures that encompass a set of non-contiguous protein backbone fragments residing within a spatial neighborhood of a specific residue. The calculation of descriptors for all known protein structures followed by clustering similar descriptors into groups revealed certain sequence preferences that can be interpreted as propensity of particular residues to be accommodated within particular substructures (Hvidsten et al. 2003), similar to the observation made for single contiguous fragments, e.g., in the I-SITES library (Bystroff et al. 2000). Based on these correlations it is possible to identify descriptors matching the target sequence and to predict a three-dimensional fold that is most compatible with these descriptors, without building an explicit three-dimensional model of the target structure (Hvidsten et al. 2003). In principle, it may be possible to assemble the tertiary structure of the target from the descriptors that contain multiple backbone fragments but to the author's knowledge such a method has not yet been developed. It remains to be seen if the three-dimensional descriptor approach will ever lead to practically a useful method for tertiary structure prediction that would become available publicly.

10.6 Why Are the Fragments-Assembly Methods So Successful?

Template-based methods, especially FR meta-servers, have been found to produce exceptionally good predictions and are now widely used for protein structure prediction. In particular, their relatively low-computational cost makes them very useful for large-scale analyses, e.g., for building models for proteins encoded in whole genomes. However, all template-based methods suffer from the fundamental limitation of being able to recognize only the folds that have been already observed. The results of structural genomics initiatives reveal that the majority of proteins belong to the previously characterized folds, but the percentage of structures with "new" folds or variations of "old" folds that cannot be accurately predicted by FR methods remains significant. On the other hand, physics-based methods for "ab initio" folding are extremely costly in terms of the computing power even if they use reduced representation and do not yet successfully fold large proteins. However, even when a novel fold is discovered, it usually turns out to be composed of common structural motifs, often at the level of supersecondary or even larger structures. Levitt and coworkers (Kolodny et al. 2002) have demonstrated that all proteins in the PDB can be modeled accurately from rigid fragments of unrelated proteins that are concatenated without any degrees of freedom. Skolnick and coworkers (Kihara and Skolnick 2003; Zhang et al. 2005) have shown that most of proteins in the PDB have significant structural alignments to other proteins in a different secondary structure and fold class. Thus, modeling of new folds can be greatly facilitated by assembling them from fragments of known structures identified by "local fold recognition," rather than attempting to model the whole process of protein folding based on first principles.

The success of methods based on fragment assembly lies not only in the restriction of the conformational space, but which can be also achieved by other reduced models (e.g., "pure" lattice models) that are less successful. As emphasized by Takada and coworkers (Chikenji et al. 2003), one of the problems of the contemporary energy functions, those based on physics and statistics alike, it the limited ability to capture the subtleties of interactions between the neighboring residues (side-chain/main-chain hydrogen bonding, side-chain configurational entropy loss, etc.), which govern the local torsional propensities. Computing the local interaction energies "ab initio" may lead to accumulation of inaccuracies and greatly decrease the chances of obtaining a globally correct model. Methods that utilize fragments avoid this problem by sampling local conformations that exist in native protein structures, which provides implicit, yet accurate representation of local interactions. Thus, a single-fragment substitution corresponds to instantly transporting the modeled protein from one local energy minimum to another, without the necessity of overcoming local energetic barriers. This enormously speeds up the search for the global energy minimum and allows shifting the focus to the generation of non-local conformational changes and identification of globally native-like structures.

The conservation of local structure may have not only physical, but also evolutionary sense. Lupas et al. (2001) proposed a scenario, in which modern proteins evolved from ancient short-peptide ancestors, called antecedent domain segments (ADSs). They suggested that the ancestors of contemporary (sub)domains arose by spontaneous non-covalent association of peptides with native-like and/or tertiary-like structural features, and since such assemblies provided functional advantage (e.g., due to improved stability of the individual fragments or their increased efficient concentration), the fusion of primitive genes encoding these fragments was preferentially selected by evolution. It is noteworthy that attempts to form folded and functional proteins by recombination revealed that successfully recombined fragments called "schemas" often correspond to known supersecondary structural elements (Voigt et al. 2002). This hypothetical "mix-and-join" scenario convincingly explains not only the structure of repetitive proteins, such as propellers, TIM-barrels, and helical bundles, but may also be invoked to explain the origin of more complicated and asymmetrical domains (Soding and Lupas 2003).

10.7 Conclusions and Outlook

In the earlier version of this chapter the main author (J.M.B.) predicted that in the "near future" with respect to the year of 2006 (i.e., near past with respect to the year of 2010) we should have seen more integration of the most successful approaches, that is, meta-prediction and assembly of fragments, and further convergence of the evolutionary and physical schools. In opinion of the authors of this chapter this indeed has happened. Essentially all methods that currently score best in CASP rely on some sort of meta-prediction, either with the use of external servers or ones constructed in house. In parallel, emphasis increases on the use of physics-based methods for the refinement of models that are close to the native structure, but could

be even closer. So what is our prediction of the nearest future, e.g., a few years following the publication of this volume? The results of recent CASP demonstrate that the progress in protein structure prediction is negligible and comes more from the area of "information technology" than "science." In our opinion this indicates that the field of protein structure prediction has grown old and the only alternative to retirement is a major breakthrough rather than just recombination of what is already available. While we keep our fingers crossed for such a breakthrough on any line of the stalled protein front, we predict that an increased number of researchers will turn away from the field of protein structure prediction to move to other fields. One of the interesting directions is the testing of the applicability of techniques developed for protein three-dimensional structure prediction in the emerging field of RNA three-dimensional structure prediction, which on the other hand offers many interesting solutions that could be used to refresh the aging field of protein bioinformatics. The Baker group has already developed a version of "ROSETTA for RNA", dubbed FARNA (Fragment Assembly of RNA) (Das and Baker 2007). So is the RNA structure the New World for conquistadors from the Protein Continent? It remains to be seen. See you there!

Acknowledgments Our recent research in the field of structural bioinformatics has been funded by the Polish Ministry of Scientific Research and Higher Education (grant numbers: POIG.02.03.00-00-003/09, 188/N-DFG/2008/0, N301 106 32/3600, and PBZ-MNiI-2/1/2005), by the NIH (grant numbers R01GM081680 and R03TW007163-01), by the European Commission (grant numbers LSHG-CT-2003-503238, LSHG-CT-2005-518238, MRTN-CT-2005-019566, 229676, and RIDS 011934), and by HFSP program RGP 55/2006.

References

Anfinsen CB (1973) Principles that govern the folding of protein chains. Science 181:223–230

Anfinsen CB, Haber E, Sela M, White FH Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. Proc Natl Acad Sci USA 47:1309–1314

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

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL (2005) GenBank. Nucleic Acids Res 33:D34–38

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. Nucleic Acids Res 28:235–242

Bradley P, Chivian D, Meiler J, Misura KM, Rohl CA, Schief WR, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss CE, Baker D (2003) ROSETTA predictions in CASP5: successes, failures, and prospects for complete automation. Proteins 53 Suppl 6: 457–468

Bujnicki JM (2006 Jan 9) Protein structure prediction by recombination of fragments. Chem Bio Chem 7(1):19–27

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001a) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. Proteins Suppl 5:184–191

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001b) Structure prediction meta server. Bioinformatics 17:750–751

Bystroff C, Baker D (1997) Blind predictions of local protein structure in CASP2 targets using the I-sites library. Proteins Suppl 1:167–171

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

Bystroff C, Shao Y (2002) Fully automated ab initio protein structure prediction using I-SITES, HMMSTR and ROSETTA. Bioinformatics 18(Suppl 1):S54–S61

Bystroff C, Simons KT, Han KF, Baker D (1996) Local sequence-structure correlations in proteins. Curr Opin Biotechnol 7:417–421

Bystroff C, Thorsson V, Baker D (2000) HMMSTR: a hidden Markov model for local sequence-structure correlations in proteins. J Mol Biol 301:173–190

Camproux AC, Tuffery P, Chevrolat JP, Boisvieux JF, Hazout S (1999) Hidden Markov model approach for identifying the modular framework of the protein backbone. Protein Eng 12: 1063–1073

Cheng J, Wang Z, Tegge AN, Eickholt J (2009) Prediction of global and local quality of CASP8 models by MULTICOM series. Proteins 77 Suppl 9:181–184

Chikenji G, Fujitsuka Y, Takada S (2003) A reversible fragment assembly method for de novo protein structure prediction. J Chem Phys 119:6895–6903

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

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

Cohen-Gonsaud M, Catherinot V, Labesse G, Douguet D (2004) From molecular modeling to drug design. In: Bujnicki JM (ed) Practical bioinformatics, vol. 15. Springer, Berlin, pp 35–71

Contreras-Moreira B, Fitzjohn PW, Bates PA (2003a) In silico protein recombination: enhancing template and sequence alignment selection for comparative protein modelling. J Mol Biol 328:593–608

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

Cymerman IA, Feder M, Pawlowski M, Kurowski MA, Bujnicki JM (2004) Computational methods for protein structure prediction and fold recognition. In: Bujnicki JM (ed) Practical bioinformatics, vol 15. Springer Berlin, pp 1–21

Das R, Baker D (2007) Automated de novo prediction of native-like RNA tertiary structures. Proc Natl Acad Sci USA 104:14664–14669

Das R, Baker D (2008) Macromolecular modeling with ROSETTA. Annu Rev Biochem 77: 363–382

Donate LE, Rufino SD, Canard LH, Blundell TL (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

Douguet D, Labesse G (2001) Easier threading through web-based comparisons and cross-validations. Bioinformatics 17:752–753

Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pacific Symp Biocomp 5:119–130

Fischer D (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins 51:434–441

Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. Proteins Suppl 5:171–183

Fiser A, Sali A (2003) Modeller: generation and refinement of homology-based protein structure models. Methods Enzymol 374:461–491

Fujitsuka Y, Takada S, Luthey-Schulten ZA, Wolynes PG (2004) Optimizing physical energy functions for protein folding. Proteins 54:88–103

Geetha V, Munson PJ (1996) Simplified representation of proteins. J Biomol Struct Dyn 13: 781–793

946 Ginalski K, Elofsson A, Fischer D, Rychlewski L (2003) 3D-jury: a simple approach to improve
947 protein structure predictions. Bioinformatics 19:1015–1018
948 Ginalski K, Grishin NV, Godzik A, Rychlewski L (2005) Practical lessons from protein structure
949 prediction. Nucleic Acids Res 33:1874–1891
950 Godzik A, Kolinski A, Skolnick J (1992) Topology fingerprint approach to the inverse protein
951 folding problem. J Mol Biol 227:227–238
952 Gront D, Kolinski A (2005) HCPM – program for hierarchical clustering of protein models.
953 Bioinformatics 21:3179–3180
954 Han KF, Baker D (1996) Global properties of the mapping between local amino acid sequence and
955 local structure in proteins. Proc Natl Acad Sci USA 93:5814–5818
956 Hinds DA, Levitt M (1992) A lattice model for protein structure prediction at low resolution. Proc
957 Natl Acad Sci USA 89:2536–2540
958 Hunter CG, Subramaniam S (2002) Natural coordinate representation for the protein backbone
959 structure. Proteins 49:206–215
960 Hvidsten TR, Kryshtafovych A, Komorowski J, Fidelis K (2003) A novel approach to fold recog-
961 nition using sequence-derived properties from sets of structurally similar local fragments of
962 proteins. Bioinformatics 19(Suppl 2):II81–II91
963 Ishida T, Nishimura T, Nozaki M, Inoue T, Terada T, Nakamura S, Shimizu K (2003) Development
964 of an ab initio protein structure prediction system ABLE. Genome Inform Ser Workshop
965 Genome Inform 14:228–237
966 John B, Sali A (2003) Comparative protein structure modeling by iterative alignment, model
967 building and model assessment. Nucleic Acids Res 31:3982–3992
968 Jones DT (1999) GenTHREADER: an efficient and reliable protein fold recognition method for
969 genomic sequences. J Mol Biol 287:797–815
970 Jones DT, McGuffin LJ (2003) Assembling novel protein folds from super-secondary structural
971 fragments. Proteins 53(Suppl 6):480–485
972 Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition.
973 Nature 358:86–89
974 Jones TA, Thirup S (1986) Using known substructures in protein model building and crystallogra-
975 phy. Embo J 5:819–822
976 Karchin R, Cline M, Mandel-Gutfreund Y, Karplus K (2003) Hidden Markov models that use
977 predicted local structure for fold recognition: alphabets of backbone geometry. Proteins 51:
978 504–514
979 Kedem K, Chew LP, Elber R (1999) Unit-vector RMS (URMS) as a tool to analyze molecular
980 dynamics trajectories. Proteins 37:554–564
981 Kihara D, Lu H, Kolinski A, Skolnick J (2001) TOUCHSTONE: an ab initio protein struc-
982 ture prediction method that uses threading-based tertiary restraints. Proc Natl Acad Sci USA
983 98:10125–10130
984 Kihara D, Skolnick J (2003) The PDB is a covering set of small protein structures.
985 J Mol Biol 334:793–802
986 Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the ROBETTA
987 server. Nucleic Acids Res 32:W526–531
988 Kolinski A (2004) Protein modeling and structure prediction with a reduced representation. Acta
989 Biochim Pol 51:349–371
990 Kolinski A, Bujnicki JM (2005) Generalized protein structure prediction based on combination
991 of fold-recognition with de novo folding and evaluation of models. Proteins 61(Suppl 7):
992 84–90
993 Kolodny R, Koehl P, Guibas L, Levitt M (2002) Small libraries of protein fragments model native
994 protein structures accurately. J Mol Biol 323:297–307
995 Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM (2003) A
996 "FRankenstein's monster" approach to comparative modeling: merging the finest fragments
997 of fold-recognition models and iterative model refinement aided by 3D structure evaluation.
998 Proteins 53(Suppl 6):369–379

Kosinski J, Gajda MJ, Cymerman IA, Kurowski MA, Pawlowski M, Boniecki M, Obarska A, Papaj G, Sroczynska-Obuchowicz P, Tkaczuk KL, Sniezynska P, Sasin JM, Augustyn A, Bujnicki JM, Feder M (2005) FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. Proteins 61(Suppl 7):106–113

Kraulis PJ, Jones TA (1987) Determination of three-dimensional protein structures from nuclear magnetic resonance data using fragments of known structures. Proteins 2:188–201

Krieger E, Nabuurs SB, Vriend G (2003) Homology modeling. Methods Biochem Anal 44: 509–523

Kryshtafovych A, Fidelis K (2009) Protein structure prediction and model quality assessment. Drug Discov Today 14:386–393

Kurowski MA, Bujnicki JM (2003) GeneSilico protein structure prediction meta-server. Nucleic Acids Res 31:3305–3307

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

Liwo A, Khalili M, Scheraga HA (2005) Ab initio simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. Proc Natl Acad Sci USA 102:2362–2367

Liwo A, Lee J, Ripoll DR, Pillardy J, Scheraga HA (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, Elofsson A (2001) Pcons:a neural-network-based consensus predictor that improves fold recognition. Protein Sci 10:2354–2362

Lupas AN, Ponting CP, Russell RB (2001) On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? J Struct Biol 134:191–203

Margraf T, Schenk G, Torda AE (2009) The SALAMI protein structure search server. Nucleic Acids Res 37:W480–484

Melo F, Sanchez R, Sali A (2002) Statistical potentials for fold assessment. Protein Sci 11:430–448

Micheletti C, Seno F, Maritan A (2000) Recurrent oligomers in proteins: an optimal scheme reconciling accurate and concise backbone representations in automated folding and design studies. Proteins 40:662–674

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

Moult J, Fidelis, K, Rost, B, Hubbard T, Tramontano A (2005) Critical assessment of methods of protein structure prediction (CASP)-round 6. Proteins 61(Suppl 7):3–7

Moult J, Fidelis K, Kryshtafovych A, Rost B, Hubbard T, Tramontano A (2007) Critical assessment of methods of protein structure prediction – Round VII. Proteins 69(Suppl 8):3–9

Moult J, Fidelis K, Kryshtafovych A, Rost B, Tramontano A (2009) Critical assessment of methods of protein structure prediction – Round VIII. Proteins 77(Suppl 9):1–4

Oliva B, Bates PA, Querol E, Avilés FX, Sternberg MJ (1997) An automated classification of the structure of protein loops. J Mol Biol 266:814–830

Ozkan SB, Wu GA, Chodera JD, Dill KA (2007) Protein folding by zipping and assembly. Proc Natl Acad Sci USA 104:11987–11992

Pandit SB, Zhang Y, Skolnick J (2006) TASSER-Lite: an automated tool for protein comparative modeling. Biophys J 91:4180–4190

Pawlowski M (2009) Rozwój metod udokładniania i oceny poprawności teoretycznych modeli struktur białek i zastosowanie ich w technologii Molecular Replacement (MR). Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, p. 111 (in Polish, PhD)

Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics 9:403

Peitsch MC (1995) Protein modelling by e-mail. Bio/Technology 13:658–660

Pruitt KD, Tatusova, T, Maglott DR (2007) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–5

Rohl CA (2005) Protein structure estimation from minimal restraints using ROSETTA. Methods Enzymol 394:244–260

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

Sasson I, Fischer D (2003) Modeling three-dimensional protein structures for CASP5 using the 3D-SHOTGUN meta-predictors. Proteins 53 (Suppl 6):389–394

Scheraga HA (1996) Recent developments in the theory of protein folding: searching for the global energy minimum. Biophys Chem 59:329–339

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

Simons KT, Kooperberg C, Huang E, 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

Skolnick J, Kihara D, Zhang Y (2004) Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. Proteins 56:502–518

Skolnick J, Kolinski A (1991) Dynamic Monte Carlo simulations of a new lattice model of globular protein folding, structure and dynamics. J Mol Biol 221:499–531

Soding J, Lupas AN (2003) More than the sum of their parts: on the evolution of proteins from peptides. Bioessays 25:837–846

Sun S (1995) Reduced representation approach to protein tertiary structure prediction: statistical potential and simulated annealing. J Theor Biol 172:13–32

Tramontano A, Chothia C, Lesk AM (1989) Structural determinants of the conformations of medium-sized loops in proteins. Proteins 6:382–394

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

Voigt CA, Martinez C, Wang ZG, Mayo SL, Arnold FH (2002) Protein building blocks preserved by recombination. Nat Struct Biol 9:553–558

Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086

Wallner B, Larsson P, Elofsson A (2007) Pcons.net: protein structure prediction meta server. Nucleic Acids Res 35:W369–374

Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. BMC Biol 5:17

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

Zhang Y, Arakaki AK, Skolnick J (2005) TASSER: an automated method for the prediction of protein tertiary structures in CASP6. Proteins 61(Suppl 7):91–98

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

Zhang Y, Skolnick J (2004b) SPICKER: a clustering approach to identify near-native protein folds. J Comput Chem 25:865–871