DOI: 10.1002/cbic.200500235

Protein-Structure Prediction by Recombination of Fragments

Janusz M. Bujnicki*[a]

The field of protein-structure prediction has been revolutionized by the application of "mix-and-match" methods both in template-based homology modeling and in template-free de novo folding. Consensus analysis and recombination of fragments copied from known protein structures is currently the only approach that allows the building of models that are closer to the native structure of the target protein than the structure of its closest homologue. It is also the most successful approach in cases in which the target protein exhibits a novel three-dimensional fold. This review 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. A convergence between the "protein folding" and "protein evolution" schools of thought is postulated.

Introduction

The high-resolution three-dimensional structure of a protein is the key to the understanding and manipulation of its biochemical and cellular functions. However, the rate of protein structure determination by X-ray crystallography lags behind the rate of determination of new protein sequences. As of May 2005, the National Center for Biotechnology Information's Non-Redundant (nr) GenPept database[1] contained 2456090 sequences, while the Protein Data Bank[2] contained only 21177 proteins structures with non-redundant sequences (30672 structures in total). The size of the sequence database doubles approximately every 18 months, while that of the structure database doubles every three years, so the gap between the number of known structures and the number of known sequences will continue to widen in the foreseeable future and it is unlikely that it will ever be closed, that is, structures will never be solved experimentally for all proteins.

More than 40 years ago, Anfinsen demonstrated that all of the information necessary for RNase A to fold into its native structure is contained in its amino acid sequence.[3] This finding has been generalized to most globular proteins, suggesting that a protein's structure could be calculated (modeled) from knowledge of its sequence and our understanding of the sequence–structure relationships. The current structural genomics initiative aims to solve experimentally the structures of only the most important or most representative proteins, while it is hoped that the others may be modeled computationally.[4] The theoretical prediction of the native structure of a protein from its amino acid sequence, however, remains one of the most challenging problems in contemporary life sciences.

Protein-structure prediction Methods—Classification and Critical Evaluation

Efforts to solve the protein folding problem have traditionally been rooted in two schools of thought (Figure 1). One is based on the principles of physics: that is, 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.[5] 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.[6] 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), to model the overall fold of the target based on that of the template, and to infer how the target structure will be changed, relative to the template, as a result of substitutions, insertions, and deletions (for reviews, see refs. [7,8]).

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. The de novo approach can be further subdivided into ab initio methods (that is, those based exclusively on

[a] *Dr. J. M. Bujnicki*
Laboratory of Bioinformatics and Protein Engineering
International Institute of Molecular and Cell Biology
Trojdena 4, 02-109 Warsaw (Poland)
Fax: (+48) 22-668-5288
E-mail: iamb@genesilico.pl

© 2006 Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim

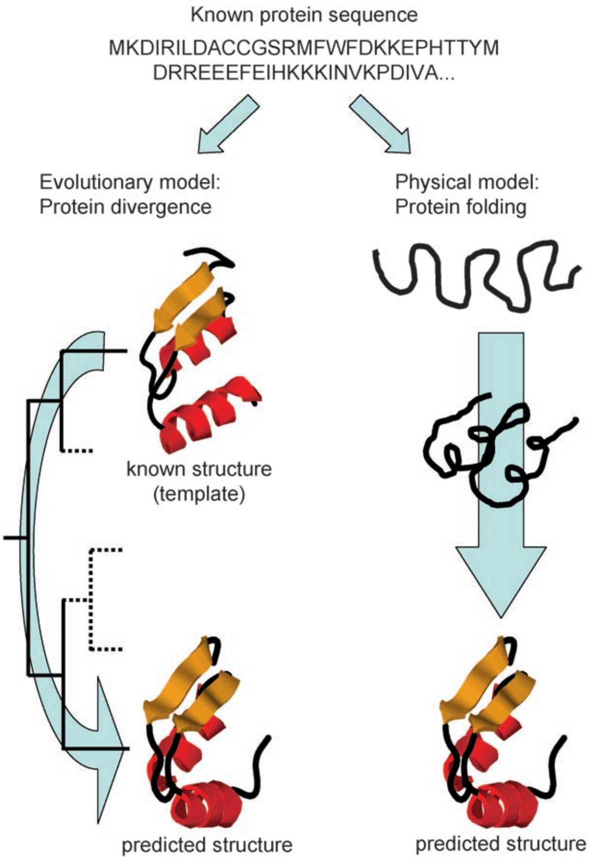

Figure 1. The evolutionary and physical approaches to protein-structure prediction. Given the amino acid sequence, a simulation of either protein evolution or protein folding is carried out, according to quantitative models of either divergence of sequences and structures or physical interactions within the molecule and between the molecule and the solvent.

the physics of the interactions within the polypeptide chain and between the polypeptide and the solvent[9]), and "knowledge-based" methods that utilize statistical potentials based on the analysis of recurrent patterns in known protein structures and sequences.[10]

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 sequences from a database of previously observed loops, followed by replacement of the mutated side chains with rotamers that satisfy the stereochemical criteria, together with (optionally) limited energy optimization, as implemented in Swiss-Model.[11] 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 the idealization of the geometry and packing of the entire chain through satisfaction of stereochemical constraints derived from the database of protein structures, as implemented in Modeller.[12] The CM approach has been also extended to "fold-recognition" (FR), in which one attempts to identify a template with a similar fold that does not need to exhibit significant sequence similarity to the target (i.e., the target and the template may or may not be homologous, but they need to share the common fold).[13,14] While the early FR methods relied mostly on the "threading" approach (that is, 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 homologues rather than unrelated structural analogues (for reviews, see refs. [15–17]).

In order to make an objective assessment of the abilities (and inabilities) of different methods for protein-structure prediction, Moult and co-workers organized the biennial Community-Wide Experiment on the Critical Assessment of Techniques for Protein-structure prediction (CASP). 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.[18] Since then, the progress in the field of protein-structure prediction has been significant, especially in the template/knowledge-based category (i.e., 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 of numerous new sequences that can serve as convenient evolutionary intermediates in homology searches.[19]

"Meta" Approaches to Template-Based Prediction

The series of CASP experiments has 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, Daniel Fischer, Leszek Rychlewski, Arne Elofsson, and the author of this article, pioneered the idea of meta-prediction in CASP4, by comparing the models generated by FR servers participating in the satellite experiment CAFASP-2 (CAFASP = critical assessment of techniques for fully automated protein-structure prediction) 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.[20] It was thus demonstrated that the recurrence of a particular protein fold within the sets of top 10 models returned by different servers (and not necessarily in first position in their ranking) increases the likelihood of a correct prediction and that, on average, no single FR method is better than the combination of a few top methods. Since then, meta-prediction based on FR (Figure 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.

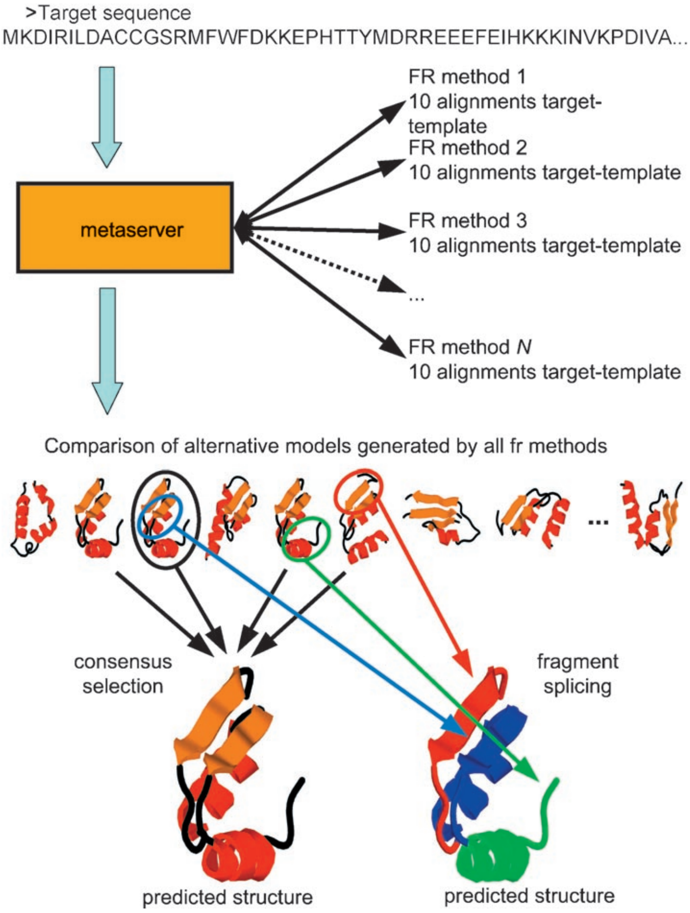
>Target sequence
MKDIRILDACCGSRMFWFDKKEPHTTYMDRREEEFEIHKKKINVKPDIVA...
FR method 1
10 alignments target-template
FR method 2
10 alignments target-template
metaserver
FR method 3
10 alignments target-template
...
FR method N
10 alignments target-template
Comparison of alternative models generated by all fr methods
...
consensus selection
fragment splicing
predicted structure
predicted structure

Figure 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, to collect the results (target–template alignments), to build the corresponding models, to compare them with each other, and either to select the most representative structure or to construct a hybrid model from the most frequently represented fragments.

Following the proven success of manual meta-predictors, several groups have implemented fully automated meta-servers[21–24] (Table 1; for a review, see ref. [25]). One of the earliest meta-predictors was the neural network PCONS developed by Elofsson and co-workers,[26] which collects a set of top models generated by different FR servers and selects the models most similar to other models in the set. The second edition of an independent assessment experiment, LiveBench,[27] organized shortly after CASP4, revealed that PCONS2 (version trained for a few specific primary FR servers) exhibited a sensitivity comparable to that of the most sensitive primary method and a higher specificity than any primary method. (Unlike CASP, which employs human assessors to evaluate models submitted by human predictors and computer programs, Livebench and CAFASP experiments are focused on the fully automated assessment of methods (servers) that exclude human intervention in the process of protein-structure prediction.) The newest version of PCONS5, reinforced by the PROQ method for protein model evaluation,[28] exhibits even higher specificity and is able to use models generated by any set of methods as input.

3D-Jury, developed by Rychlewski and co-workers,[29] 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 the 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 users is its scoring system, which allows confident identification of the models with correctly predicted folds, even though it does not necessarily recognize the absolutely best model among similar top solutions.

From Multiple Template-Based Models to Hybrids

As well as those meta-predictors that simply select models from the input set, another breed of meta-predictors that use the unrefined models generated by primary servers as a structural scrap-yard from which to obtain spare parts to generate new models has also been developed. 3D-Shotgun, developed by Fischer,[30] was the first fully automated meta-predictor designed to assemble hybrid models from fragments of models obtained from independent FR methods (i.e., from different components of the Bioinbgu server[31]). 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 with the highest count. Thus, for each initial model, a hybrid model is constructed, containing the most common structural features of all models, and often including more residues than any of the initial models. In the second step, the assembled models are assigned scores based on a combination of the original scores of their parent models (normalized to a similar scale) and scores describing the structural similarity of the assembled model to other models, as determined by Maxsub.[32] 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 containing 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 hybrid model construction approach is superior to selection of one of the stereochemically more acceptable input models, as the hybrids are on average more complete and superimpose better than the initial models on the native structure. The method is sensitive to initial alignment error—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[12] to generate full-atom models with idealized stereochemistry and without gaps and collisions,

Table 1. Summary of key features of methods analyzed in this article.

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	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 potential (VERIFY3D)	FR 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	3 and 9 aa fragments from PDB
Simfold	de novo fragment splicer	Multicanonical Monte Carlo ensemble	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
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	FR models
Frank/Cabs	FR/CM/de novo fragment splicer	replica exchange Monte Carlo (on a lattice)	statistical potential	FR models
structural descriptors	FR	NA	NA	descriptors (groups of >2 fragments >3 aa long)

and even with a slight improvement in the overall RMSD.[33] The method is available via the INUB server at http://fischerlab.bioinformatics.buffalo.edu/inub/.

Frankenstein's Monster is another approach to meta-prediction by consensus and recombination of fragments, developed in the author's laboratory.[34] 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 in overcoming the problem of selection of the optimal template, but also in correcting initial alignment errors. Briefly, the GeneSilico meta-server[23] 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 based on statistical potentials, such as implemented in Verify3D[35] or other similar methods (for a review, see ref. [36]). 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 by the Verify3D 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, the hybrid model in the Frankenstein strategy is not directly refined, but is instead 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 reevaluated with Verify3D, 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 1 aa 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 an acceptable score or the score cannot be further improved.

The Frankenstein's Monster 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[34] and CASP6, in which the groups from the author's laboratory ranked very high in the CM and FR categories,[37,38] application of this approach gives 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 current version of the method is available as a FRANKENSTEIN3D server at http://genesilico.pl/frankenstein3d/

Another approach to overcoming the problem of template selection and correction of alignment errors by recombination of alternative models was developed by Bates and co-workers.[39] 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 areas. 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.[40] However, the quality of the initial models is the upper limit for the quality of the final model (i.e., 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 co-workers.[41] It is similar to the FRANKENSTEIN'S MONSTER 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.[12] 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,[42] 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.[41]

Fragment Assembly: A New Trend in de novo Protein-Structure Prediction

Modeling a 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 simulation of up to about 1 microsecond of folding of full-atom representations of only small proteins (<100 aa), while most proteins are larger and fold over timescales of milliseconds or even seconds. The solvent is therefore usually treated implicitly and various simplified models that have fewer degrees of freedom and exploit the repetitive natures of protein structures are used. These simplified models typically retain only certain atoms, such as Cα or Cβ or united atoms in which several atoms, such as the centers of mass of the side-chains, are grouped together.[43,44] The protein structures may be represented by a number of simplified schemes such as lattices or bond angles with discrete values.[10,45] Despite the considerable reduction in 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.[46]

Significant progress in the field of de novo protein-structure prediction has been stimulated by the observation that the structure of a protein backbone can be represented quite accurately through the use of short fragments taken from other proteins.[47,48] Fragments up to 10 aa long provided an efficient method for interpreting electron density maps in protein crystallography and in building protein models from nuclear magnetic resonance (NMR) data.[49] Classification of protein loops has proven useful in comparative modeling, in which the incomplete framework of a protein core has to be amended by de novo insertion of polypeptide segments[50–52] (see also above). Several groups have classified peptide backbone units with fixed or variable lengths into collections of fragments.[53–58] Analysis of such recurring fragments has identified local sequence–structure correlations in proteins[59,60] and suggested a new method for de novo protein-structure prediction.

ROSETTA, developed by Baker and co-workers,[61] 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 a particular conformation being assumed is based on the similarity of the local sequence and predicted secondary structure of the target to sequences and structures from the template library, as in the traditional template-based methods for protein modeling. The fragments (of 3 and 9 aa residues) are assembled by a Monte Carlo (MC) simulated annealing (SA) search strategy, in which fragments are randomly inserted into the protein chain by replacement of 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 with full-atom representation, special conformation modification operators, and a refined (more physical) energy function. During the simulation, a large set of decoys (1000–100 000) are generated, and these 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 in ref. [62]).

The difference between Rosetta and most 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). Rosetta can thus generate a de novo model by allowing the full-length polypeptide chain to explore conformational space by 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.[63] As demonstrated in CASP5,[64] 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 a template that matches only a part of the modeled protein.

Rosetta is 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 Robetta meta-server developed by the Baker group[65] (http://robetta.bakerlab.org/) 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.[66]

Simfold is another fragment assembly method, recently developed by Shoji Takada and co-workers.[67] The original method, which performed quite well in CASP6 as ROKKO and ROKKY,[68] is similar to Rosetta. It uses 4- to 9-residue fragments, and its energy function consists of various interactions that are based on physical considerations.[69] Simfold exhibits an important difference, namely, it introduces reversible fragment insertion. 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 reinserted. This 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 that the conventional SA method[67] implemented in the ROKKY server or in many other methods described in this article. Simfold has been made available as a web server (http://www.proteinsilico.org/), albeit with a heavily limited functionality in relation to the method used by the ROKKO team in CASP6 (SA instead of MEMC, limited length of the simulation, etc.).

Profesy, developed by Lee and co-workers,[70] 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.[71] The fragment library is constructed by use of the secondary structure prediction method Predict and comprises a collection of 15 aa-long backbone structures. This method is not yet publicly available.

Fragfold, developed by Jones, uses supersecondary structural fragments (made up of two or three sequential secondary structures) from a library of high-resolution protein structures as well as small (3, 4, and 5 aa) fragments.[72,73] Possible supersecondary fragments are assigned to the target sequence by a threading procedure similar to that used in the GenThreader FR algorithm.[74] The global structure is assembled by a genetic algorithm or a simulated annealing method, in which half the random moves correspond to the insertion of a preselected 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 yet publicly available.

Undertaker is a method developed by Karplus and co-workers that assembles the target structure by use of fragments of known structures obtained from three sources: a generic library of very short segments (1–4 aa) that must exactly match the target sequence, medium-length segments (9–12 aa) that are assigned by the Fragfinder program from the SAM suite, and variable-length segments assigned by FR analysis.[75] 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 recombination of different conformations. The cost function used to assess the decoys includes many tunable parameters, the most important of which—as the name of the method implies—is the burial. Undertaker is not yet publicly available.

Able, developed by Shimizu and co-workers,[76] 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. The Able method has two interesting features that help in avoiding problems if the initial distribution of decoys is too broad and no clusters can be identified from the RMSD as a measure of the distance between the conformations. Firstly, it uses the unit-vector root mean square distance (URMS)[77] as a measure of structural similarity. Secondly, 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. Able is not yet publicly available.

Hybrid Methods Involving Fragment Assembly and Lattice-Based Folding Simulations

An alternative approach to fragment assembly, and one with a long history, is that of lattice representation, in which residues are restricted to points in a regular three-dimensional lattice.[78,79] 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. Very recently, following the success of Rosetta and other fragment-assembly methods, several hybrid methods combining the strengths of both approaches have arisen.

Tasser, developed by Skolnick and co-workers,[80] starts with FR analysis based on the Prospector threading method,[81] identifying either a single consensus fold or a set of templates with globally distinct folds. From the FR alignments, the protein chain is divided into contiguous aligned regions of at least 5 aa (20.7 aa on average, according to the authors' own benchmark), and gapped unaligned regions. The conformation of the 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 co-workers.[82,83] A series of initial models is generated and subjected for assembly and refinement to the parallel hyperbolic Replica Exchange Monte Carlo (REMC) sampling method. Structures generated in the lowest-temperature replicas are subjected to iterative clustering with Spicker[84] to identify the final models based on the cluster density. Tasser performed very well in the recent CASP-6 evaluation, comparably with Rosetta and the Frankenstein's Monster/CABS hybrid method described below. Tasser is not yet publicly available.

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 combination of the Frankenstein's Monster method[34] (see also above), for generation of initial models, with the reduced lattice model CABS.[10] Briefly, preliminary hybrid models are generated with the template-based recombination method and are scored with Verify3D to identify well-folded fragments, as described earlier. These fragments are not used directly, but are used as a source of spatial restraints to guide the REMC folding simulation by the CABS model. The resulting decoys are clustered by use of the HCPM method[85] to identify the final models. This method performed very well in the recent CASP-6 evaluation,[37] with the second-best average score among all teams according to the unofficial evaluation available at http://bioinformatics.buffalo.edu/casp6/). The whole method is not yet available as an integrated package, but the individual components are available as the Frankenstein3D server (see above) and the set of parameters of the CABS model (http://www.biocomp.chem.uw.edu.pl/files/papers/CABS/cabs.html).

Other Methods Based on Fragment Prediction

All the methods for protein-structure prediction by recombination described above use contiguous fragments of protein backbone. Another, novel approach to protein-structure prediction is based on the concept of three-dimensional structural descriptors developed by Kryshtafovych and Fidelis (unpublished analysis cited in ref. [86]): substructures that encompass, for example, a set of noncontiguous protein backbone fragments residing within a spatial neighborhood of a specific residue. Calculation of descriptors for all known protein structures, followed by clustering of similar descriptors into groups, revealed certain sequence preferences that can be interpreted as propensities of particular residues to be accommodated within particular substructures,[86] similarly to the observation made for single contiguous fragments in, for example, the I-SITES library.[87] From these correlations it is possible to identify descriptors matching the target sequence and to predict a three-dimensional fold most compatible with these descriptors, without building an explicit three-dimensional model of the target structure.[86] In principle, it may be possible to assemble the tertiary structure of the target from descriptors containing multiple backbone fragments but to the best of the author's knowledge no such method has yet been developed. The structural descriptor approach may also be useful for protein-structure prediction in combination with other methods that allow reconstruction of complete atomic structures (as a source of restraints for methods like CABS, for instance).

Why Are the Fragment 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, such as for construction of models for proteins encoded in whole genomes. However, all template-based methods suffer from the fundamental limitation of being able to recognize only folds that have already been observed. The results of structural genomics initiatives reveal that the majority of proteins belong to previously characterized fold classes, 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 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 co-workers[58] has 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 co-workers[88,89] have shown that most of the proteins in the PDB have significant structure alignments to other proteins in different secondary structure and fold classes. Modeling of new folds can therefore be greatly facilitated by assembling them from fragments of known structures identified by local fold recognition, rather than by attempting to model the whole process of protein folding from first principles.

The success of methods based on fragment assembly lies not only in the restriction of the conformational space, which can be also achieved by other reduced models (e.g. pure lattice models) that are less successful. As emphasized by Takada and co-workers,[67] one of the problems of the contemporary energy functions—those based on physics and statistics alike—is the limited ability to capture the subtleties of interactions between neighboring residues (side-chain/main-chain hydrogen bonding, side-chain configurational entropy loss etc.), which govern the local torsional propensities. Ab initio computation of the local interaction energies may give rise 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. A single fragment substitution thus 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 the focus to be shifted 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, Ponting, and Russell[90] proposed a scenario in which modern proteins have evolved from ancient short-peptide ancestors, called antecedent domain segments (ADSs). They suggested that the ancestors of contemporary (sub)domains arose by spontaneous noncovalent 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 have revealed that successfully recombined fragments called "schemas" often correspond to known supersecondary structural elements.[91] This hypothetical mix-and-join scenario convincingly explains the structures of repetitive proteins such as propellers, TIM-barrels, helical bundles, etc., but may also be invoked to explain the origins of more complicated and asymmetrical domains.[92]

Author's Perspective

It is intriguing that those features of fragment assembly methods that make them so successful are in fact common to the method of homology modeling. The process of fragment assignment may be seen not as mere identification of sequence–structure compatibility, but as a true search for remote homology to ancient short-peptide ancestors represented by families of their descendants observed in a variety of different folds. Thus, it is tantalizing to observe the competition (e.g. in CASP) between the recently developed methods for fragment assembly that use different libraries of fragments, as this may to some extent resemble the evolutionary competition between different peptides that might have existed in the primordial preprotein world. The "winner" library would possibly correspond to the set of short peptides that were also most successful in the evolution. While the field of protein-structure prediction clearly benefits from studies on protein evolution, the field of evolutionary biology could also be inspired by a particular model of protein folding by recombination of fragments.

It is tempting to speculate that in the near future we will see more integration of the most successful approaches—meta-prediction and assembly of fragments, for example—and further convergence of the evolutionary and physical schools. The currently available hybrid methods generate low-resolution models that already seem to be sufficiently accurate and confident to be widely used by biologists to make structural and functional predictions. In the author's opinion, in the nearest future both the quality and the confidence of the theoretical models will improve significantly. It remains to be seen, however, if the combination of meta-predicting with fragment assembly will finally result in the solution of the protein folding problem or if some radically new approach will be required.

Acknowledgements

I would like to thank Phil Bradley, Daniel Fischer, and Shoji Takada for discussions and for useful comments on the manuscript (special thanks to Phil for extensive proof-reading). My research in the field of protein bioinformatics is funded by the Polish Ministry of Scientific Research and Information Technology (grant PBZ-KBN-088/P04/2003), by the NIH (Fogarty International Center grant R03 TW007163-01), and by the European Commission (grant LSHG-CT-2003-503238). I am also supported by the EMBO/HHMI Young Investigator Programme.

Keywords: molecular evolution · molecular modeling · protein folding · protein models · protein structures

[1] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, D. L. Wheeler, *Nucleic Acids Res.* **2005**, *33*, D34–38.
[2] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, *Nucleic Acids Res.* **2000**, *28*, 235–242.
[3] C. B. Anfinsen, E. Haber, M. Sela, F. H. White Jr., *Proc. Natl. Acad. Sci. USA* **1961**, *47*, 1309–1314.
[4] D. Baker, A. Sali, *Science* **2001**, *294*, 93–96.
[5] C. B. Anfinsen, *Science* **1973**, *181*, 223–230.
[6] C. Chothia, A. M. Lesk, *EMBO J.* **1986**, *5*, 823–826.
[7] E. Krieger, S. B. Nabuurs, G. Vriend, *Methods Biochem. Anal.* **2003**, *44*, 509–523.
[8] "From Molecular Modeling to Drug Design", M. Cohen-Gonsaud, V. Catherinot, G. Labesse, D. Douguet in *Practical Bioinformatics, Vol.* 15*: Nucleic Acids and Molecular Biology* (Ed.: J. M. Bujnicki), Springer, Berlin, **2004**, pp. 35–71.
[9] H. A. Scheraga, *Biophys. Chem.* **1996**, *59*, 329–339.
[10] A. Kolinski, *Acta Biochim. Pol.* **2004**, *51*, 349–371.
[11] M. C. Peitsch, *Bio/Technology* **1995**, *13*, 658–660.
[12] A. Sali, T. L. Blundell, *J. Mol. Biol.* **1993**, *234*, 779–815.

[13] D. T. Jones, W. R. Taylor, J. M. Thornton, *Nature* **1992**, *358*, 86 – 89.

[14] A. Godzik, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 12 098 – 12 102.

[15] K. Ginalski, N. V. Grishin, A. Godzik, L. Rychlewski, *Nucleic Acids Res.* **2005**, *33*, 1874 – 1891.

[16] "Computational Methods for Protein-structure prediction and Fold Recognition", I. A. Cymerman, M. Feder, M. Pawlowski, M. A. Kurowski, J. M. Bujnicki in *Practical Bioinformatics, Vol.* 15*: Nucleic Acids and Molecular Biology* (Ed.: J. M. Bujnicki), Springer, Berlin, **2004**, pp. 1 – 21.

[17] A. Godzik, *Methods Biochem. Anal.* **2003**, *44*, 525 – 546.

[18] *Proteins* **1995**, *23* (whole issue).

[19] C. Venclovas, A. Zemla, K. Fidelis, J. Moult, *Proteins* **2003**, *53 Suppl 6*, 585 – 595.

[20] D. Fischer, A. Elofsson, L. Rychlewski, F. Pazos, A. Valencia, B. Rost, A. R. Ortiz, R. L. Dunbrack, Jr., *Proteins* **2001**, *Suppl 5*, 171 – 183.

[21] J. M. Bujnicki, A. Elofsson, D. Fischer, L. Rychlewski, *Bioinformatics* **2001**, *17*, 750 – 751.

[22] D. Douguet, G. Labesse, *Bioinformatics* **2001**, *17*, 752 – 753.

[23] M. A. Kurowski, J. M. Bujnicki, *Nucleic Acids Res.* **2003**, *31*, 3305 – 3307.

[24] D. Juan, O. Grana, F. Pazos, P. Fariselli, R. Casadio, A. Valencia, *Proteins* **2003**, *50*, 600 – 608.

[25] "'Meta' Approaches to Protein-Structure Prediction", J. M. Bujnicki, D. Fischer in *Practical Bioinformatics, Vol.* 15*: Nucleic Acids and Molecular Biology* (Ed.: J. M. Bujnicki), Springer, Berlin, **2004**, pp. 23 – 34.

[26] J. Lundstrom, L. Rychlewski, J. M. Bujnicki, A. Elofsson, *Protein Sci.* **2001**, *10*, 2354 – 2362.

[27] J. M. Bujnicki, A. Elofsson, D. Fischer, L. Rychlewski, *Protein Sci.* **2001**, *10*, 352 – 361.

[28] B. Wallner, A. Elofsson, *Protein Sci.* **2003**, *12*, 1073 – 1086.

[29] K. Ginalski, A. Elofsson, D. Fischer, L. Rychlewski, *Bioinformatics* **2003**, *19*, 1015 – 1018.

[30] D. Fischer, *Proteins* **2003**, *51*, 434 – 441.

[31] D. Fischer, *Pac. Symp. Biocomput.* **2000**, 119 – 130.

[32] N. Siew, A. Elofsson, L. Rychlewski, D. Fischer, *Bioinformatics* **2000**, *16*, 776 – 785.

[33] I. Sasson, D. Fischer, *Proteins* **2003**, *53 Suppl 6*, 389 – 394.

[34] J. Kosinski, I. A. Cymerman, M. Feder, M. A. Kurowski, J. M. Sasin, J. M. Bujnicki, *Proteins* **2003**, *53 Suppl 6*, 369 – 379.

[35] R. Luthy, J. U. Bowie, D. Eisenberg, *Nature* **1992**, *356*, 83 – 85.

[36] J. M. Sasin, J. M. Bujnicki, *Nucleic Acids Res.* **2004**, *32*, W586 – 589.

[37] A. Kolinski, J. M. Bujnicki, *Proteins* **2005**, DOI: 10.1002/prot.20723.

[38] J. Kosinski, M. J. Gajda, I. A. Cymerman, M. A. Kurowski, M. Pawlowski, M. Boniecki, A. Obarska, G. Papaj, P. Sroczynska-Obuchowicz, K. L. Tkaczuk, P. Sniezynska, J. M. Sasin, A. Augustyn, J. M. Bujnicki, M. Feder, *Proteins* **2005**, DOI: 10.1002/prot.20726.

[39] B. Contreras-Moreira, P. W. Fitzjohn, P. A. Bates, *J. Mol. Biol.* **2003**, *328*, 593 – 608.

[40] B. Contreras-Moreira, P. W. Fitzjohn, M. Offman, G. R. Smith, P. A. Bates, *Proteins* **2003**, *53 Suppl 6*, 424 – 429.

[41] B. John, A. Sali, *Nucleic Acids Res.* **2003**, *31*, 3982 – 3992.

[42] F. Melo, R. Sanchez, A. Sali, *Protein Sci.* **2002**, *11*, 430 – 448.

[43] S. Sun, *J. Theor. Biol.* **1995**, *172*, 13 – 32.

[44] A. Liwo, M. Khalili, H. A. Scheraga, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 2362 – 2367.

[45] V. Geetha, P. J. Munson, *J. Biomol. Struct. Dyn.* **1996**, *13*, 781 – 793.

[46] C. G. Hunter, S. Subramaniam, *Proteins* **2002**, *49*, 206 – 215.

[47] T. A. Jones, S. Thirup, *EMBO J.* **1986**, *5*, 819 – 822.

[48] M. Claessens, E. Van Cutsem, I. Lasters, S. Wodak, *Protein Eng.* **1989**, *2*, 335 – 345.

[49] P. J. Kraulis, T. A. Jones, *Proteins* **1987**, *2*, 188 – 201.

[50] A. Tramontano, C. Chothia, A. M. Lesk, *Proteins* **1989**, *6*, 382 – 394.

[51] L. E. Donate, S. D. Rufino, L. H. Canard, T. L. Blundell, *Protein Sci.* **1996**, *5*, 2600 – 2616.

[52] B. Oliva, P. A. Bates, E. Querol, F. X. Aviles, M. J. Sternberg, *J. Mol. Biol.* **1997**, *266*, 814 – 830.

[53] R. Unger, D. Harel, S. Wherland, J. L. Sussman, *Proteins* **1989**, *5*, 355 – 373.

[54] X. Zhang, J. S. Fetrow, W. A. Rennie, D. L. Waltz, G. Berg, *Procedings of the First International Conference of Intelligent Systems for Molecular Biology* (Eds.: L. Hunter, D. Searls, J. Shavik), **1993**, pp. 438 – 446.

[55] C. Bystroff, D. Baker, *Proteins* **1997**, *Suppl 1*, 167 – 171.

[56] A. C. Camproux, P. Tuffery, J. P. Chevrolat, J. F. Boisvieux, S. Hazout, *Protein Eng.* **1999**, *12*, 1063 – 1073.

[57] C. Micheletti, F. Seno, A. Maritan, *Proteins* **2000**, *40*, 662 – 674.

[58] R. Kolodny, P. Koehl, L. Guibas, M. Levitt, *J. Mol. Biol.* **2002**, *323*, 297 – 307.

[59] K. F. Han, D. Baker, *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 5814 – 5818.

[60] C. Bystroff, K. T. Simons, K. F. Han, D. Baker, *Curr. Opin. Biotechnol.* **1996**, *7*, 417 – 421.

[61] K. T. Simons, C. Kooperberg, E. Huang, D. Baker, *J. Mol. Biol.* **1997**, *268*, 209 – 225.

[62] C. A. Rohl, C. E. Strauss, K. M. Misura, D. Baker, *Methods Enzymol.* **2004**, *383*, 66 – 93.

[63] C. A. Rohl, C. E. Strauss, D. Chivian, D. Baker, *Proteins* **2004**, *55*, 656 – 677.

[64] P. Bradley, D. Chivian, J. Meiler, K. M. Misura, C. A. Rohl, W. R. Schief, W. J. Wedemeyer, O. Schueler-Furman, P. Murphy, J. Schonbrun, C. E. Strauss, D. Baker, *Proteins* **2003**, *53 Suppl 6*, 457 – 468.

[65] D. E. Kim, D. Chivian, D. Baker, *Nucleic Acids Res.* **2004**, *32*, W526 – 531.

[66] C. Bystroff, Y. Shao, *Bioinformatics* **2002**, *18 Suppl 1*, S54 – 61.

[67] G. Chikenji, Y. Fujitsuka, S. Takada, *J. Chem. Phys.* **2003**, *119*, 6895 – 6903.

[68] See the evaluations at http://predictioncenter.org and the forthcoming special issue of *Proteins*.

[69] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, P. G. Wolynes, *Proteins* **2004**, *54*, 88 – 103.

[70] J. Lee, S. Y. Kim, K. Joo, I. Kim, *Proteins* **2004**, *56*, 704 – 714.

[71] J. Lee, A. Liwo, D. R. Ripoll, J. Pillardy, H. A. Scheraga, *Proteins* **1999**, *Suppl 3*, 204 – 208.

[72] D. T. Jones, *Proteins* **1997**, *Suppl 1*, 185 – 191.

[73] D. T. Jones, L. J. McGuffin, *Proteins* **2003**, *53 Suppl 6*, 480 – 485.

[74] D. T. Jones, *J. Mol. Biol.* **1999**, *287*, 797 – 815.

[75] K. Karplus, R. Karchin, J. Draper, J. Casper, Y. Mandel-Gutfreund, M. Diekhans, R. Hughey, *Proteins* **2003**, *53 Suppl 6*, 491 – 496.

[76] T. Ishida, T. Nishimura, M. Nozaki, T. Inoue, T. Terada, S. Nakamura, K. Shimizu, *Genome Inf. Ser.* **2003**, *14*, 228 – 237.

[77] K. Kedem, L. P. Chew, R. Elber, *Proteins* **1999**, *37*, 554 – 564.

[78] J. Skolnick, A. Kolinski, *J Mol Biol* **1991**, *221*, 499 – 531.

[79] D. A. Hinds, M. Levitt, *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 2536 – 2540.

[80] Y. Zhang, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 7594 – 7599.

[81] J. Skolnick, D. Kihara, Y. Zhang, *Proteins* **2004**, *56*, 502 – 518.

[82] A. Kolinski, M. R. Betancourt, D. Kihara, P. Rotkiewicz, J. Skolnick, *Proteins* **2001**, *44*, 133 – 149.

[83] D. Kihara, H. Lu, A. Kolinski, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10 125 – 10 130.

[84] Y. Zhang, J. Skolnick, *J. Comput. Chem.* **2004**, *25*, 865 – 871.

[85] D. Gront, A. Kolinski, *Bioinformatics* **2005**.

[86] T. R. Hvidsten, A. Kryshtafovych, J. Komorowski, K. Fidelis, *Bioinformatics* **2003**, *19 Suppl 2*, II81 – II91.

[87] C. Bystroff, V. Thorsson, D. Baker, *J. Mol. Biol.* **2000**, *301*, 173 – 190.

[88] D. Kihara, J. Skolnick, *J. Mol. Biol.* **2003**, *334*, 793 – 802.

[89] Y. Zhang, J. Skolnick, *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 1029 – 1034.

[90] A. N. Lupas, C. P. Ponting, R. B. Russell, *J. Struct. Biol.* **2001**, *134*, 191 – 203.

[91] C. A. Voigt, C. Martinez, Z. G. Wang, S. L. Mayo, F. H. Arnold, *Nat. Struct. Biol.* **2002**, *9*, 553 – 558.

[92] J. Soding, A. N. Lupas, *BioEssays* **2003**, *25*, 837 – 846.

Received: June 3, 2005
Published online on November 30, 2005