'Meta' Approaches to Protein Structure Prediction

J.M. Bujnicki, D. Fischer

1 Introduction

The computational assignment of three-dimensional structures to newly determined protein sequences is becoming an increasingly important element in experimental structure determination and in structural genomics (Fischer et al. 2001a). In particular, fold-recognition methods aim to predict approximate three-dimensional (3D) models for proteins bearing no evident sequence similarity to any protein of known structure (see the review by Cymerman et al., this Vol.). The assignment is carried out by searching a library of known structures (usually obtained from the Protein Data Bank) for a compatible fold. A variety of fold-recognition methods has been published, both *structure-dependent* (i.e.*threading*) (Sippl and Weitckus 1992; Godzik et al. 1992; Jones et al. 1992; Ouzounis et al. 1993; Bryant and Lawrence 1993; Rost 1995; Alexandrov et al. 1996; Di Francesco et al. 1997; Fischer 2000; Kelley et al. 2000; Shi et al. 2001) and *sequence-only dependent* (Karplus et al. 1998; Rychlewski et al. 2000). The state-of-the-art in the field of fold recognition is currently to combine the evolutionary information available from multiple sequence alignments for the target and the template (to detect remote homology between protein families) and the structural information from the template (to detect similarities of folds of compared proteins regardless of their evolutionary relationship, i.e. analogs and homologues as well).

J.M. Bujnicki
Bioinformatics Laboratory, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland
D. Fischer
Bioinformatics, Dept. Computer Science, Ben Gurion University, Beer-Sheva 84015, Israel

2 The Utility of Servers as Standard Tools for Protein Structure Prediction

Automatic structure prediction has witnessed significant progress during the last few years. A large number of fully automated servers, covering various aspects of structure prediction, are currently available to the scientific community. In addition to the biannual Critical Assessment of Structure Prediction (CASP) experiment, which evaluates the state-of-the-art in the methodology and the skills of modeling teams and individual modelers (Moult et al. 1995, 1997, 1999, 2001), ta number of evaluation experiments exist that are aimed at assessing the capabilities and limitations of the servers. These experiments assess the reliability of the programs when applied to specific prediction targets and provide predictors with valuable information that can help them in choosing which programs to use and thereby make best use of the automated tools. One of these experiments is CAFASP (Fischer et al. 1999, 2001b), where the evaluation is carried out over the set of the CASP prediction targets by fully automatic web servers that submit the predictions without any human-expert intervention. CAFASP servers cover various aspects of protein structure prediction, such as secondary structure, inter-residue contacts, and tertiary structure. Another experiment is LiveBench, which differs from CAFASP in that it is run continuously and on a much larger set of targets. The targets are selected from protein structures newly submitted to the Protein Data Bank, if their sequences show no trivial similarity to any of the previously available structures (Bujnicki et al. 2001a, b).

However, despite significant progress, protein structure prediction methods still have a number of limitations . Fully automated fold-recognition methods can currently produce reliable sequence-structure assignments for only a fraction of target sequences with no significant sequence similarity to proteins of known structure (Bujnicki et al. 2001b). In the case of remote structural similarities, the sequence alignments between the target and the template reported by fold recognition often contain large errors (shifts). Needless to say, fold-recognition methods perform poorly when the target protein exhibits only partial structural similarity (i.e. not the same, but a related fold) to proteins in the database or when the sought fold is completely novel and cannot be recognized among the known structures. Another limitation of fold-recognition methods is the uncertainty as to the identity of the best model among the *top candidates*. Quite often, the correct fold is reported within the best ten predictions, but with a non-significant confidence score, buried among *false positives*.

2.1 Consensus 'Meta-Predictors': Is the Whole Greater Than the Sum of the Parts?

The use of a number of models and methods to produce better predictions has already proven useful in a number of areas, including artificial intelligence and computer vision (Marr 1982). Not surprisingly, this approach works well also in protein structure prediction. It has been observed in protein secondary structure prediction (consensus of various methods (Cuff et al. 1998; Selbig et al. 1999; Cuff and Barton 2000)), in homology modeling (multiple-parent structures; (Marti-Renom et al. 2000)) and in ab initio protein folding methods (clustering models and deriving recurring constraints from various models (Simons et al. 1999; Kihara et al. 2001; Kolinski et al. 2001).

The most vigorous development of *meta* approaches has been recently in the field of protein fold recognition. From the series of CASP experiments, it has become clear that often a correct protein fold prediction can be obtained by one server but not by the others. It has also been observed that no server can reliably distinguish between weak hits (predictions with below-threshold scores) and wrong hits, and that often a correct model is found among the top hits of the server, but scoring below a number of incorrect models. From such, and other, observations, many human expert predictors have realized that in order to produce better predictions, the results from a number of independent methods need to be analyzed.

CASP has shown that the combined use of human expertise and automated methods can often result in successful predictions. This, however, requires extensive human intervention, because a human predictor has to improve the model manually, has to determine whether the rank-1 model obtained is correct, whether there is a lower ranking model that corresponds to a correct prediction, or whether the results of the method indicate that no prediction at all can be obtained. To this end, human expert predictors have developed a number of semi-automated strategies. One such strategy has been the application of a number of independent methods to extract a prediction from the top ranking predictions. This has proven useful because for some prediction targets, one method may succeed in producing a correct prediction while others fail, yet for other targets, this same method may fail while the others succeed. Because it is impossible to determine a priori for which targets a given method will succeed, human expert predictors attempt to extract any useful information from results obtained with different methods.

To study whether it was possible to obtain a better prediction using a very simple consensus method that utilized the information from several servers, in CASP4, a group of four human predictors, Leszek Rychlewski, Arne Elofsson and both authors of this chapter, pioneered the consensus idea by submitting to CASP manually selected *consensus* predictions under the group-name CAFASP-CONSENSUS. The consensus predictions were obtained by

analyzing the predictions of the fold-recognition servers that participated in the parallel CAFASP2 experiment. This group performed better than any of the CAFASP servers and ranked seventh among all other human predictors of CASP (Fischer et al. 2001b). This finding illustrated the utility of the servers' results when taken as a whole. Since then, meta-prediction has become the most successful approach, and has been applied by a large number of human predictors, including some of the best CASP5 performers.

For example, in the comparative modeling section of CASP5, three groups excelled (Tramontano 2003), including the GeneSilico group (Janusz Bujnicki and colleagues). This group applied a new semi-automated multi-step meta-protocol named *Frankenstein's Monster*, which uses the results of diverse fold-recognition methods to generate initial target-template alignments (Kosinski et al. 2003; Kurowski and Bujnicki 2003). Full-atom models were built by a series of steps aimed at assembling hybrid models using the most conserved and most reliable fragments from the various models. Because this procedure required extensive human intervention (over 24 h/model), it is clear that *human-meta-predicting* is a difficult task requiring extensive expertise, and that automated procedures are sorely needed.

2.2 Automated Meta-Predictors

Following the proven success of manual meta-predictors, several groups have already implemented fully-automated versions of the meta-approach (Table 1). Automated meta-predictors can be divided into two types: (1) selectors, which simply select models from the input and (2) added-value meta-predictors, which use the input models to generate new models.

One of the earliest meta-predictors was developed by Arne Elofsson by implementing the CAFASP-CONSENSUS ideas from CASP4 into the automated program Pcons (Lundstrom et al. 2001). Pcons receives, as input; the top models produced by different fold-recognition servers and selects the models that are evaluated to be more likely to be correct, based on the structural similarities among the input models. That is, it does not produce any new models, only re-ranks the existing ones, based on their mutual similarity and the original scores assigned by the individual servers. Pcons corroborated the strength of the consensus idea in the subsequent LiveBench experiments (Bujnicki et al. 2001b). It was demonstrated that PCONS2 (version trained specifically for a few *original*, i.e. non-meta servers) combined the sensitivity of the most sensitive original method (3D-PSSM; Kelley et al. 2000) with a very high specificity (higher than any individual server). The most important feature contributing to the improved performance of an early version of PCONS was its scoring system, which allowed to confidently identify the correct models, although it was not always able to identify the absolutely best model among similar *top solutions*. The newest version of PCONS, reinforced

Table 1. Meta-servers for protein structure prediction. Most fold-recognition (FR) servers utilize PSI-BLAST and SS methods and therefore are considered "Meta[0]". Here, only those that explicitly utilize more than one FR method are included

Method	Input	Reference (http:// URL)	Comments
Meta[1]			
BIOINBGU selector	GONP, GONPM, SEQPPRF, SEQPMPRF, PRFSEQ	www.cs.bgu.ac.il/~bioinbgu/	Consensus prediction obtained from the results of five different methods run locally
LIBELULLA selector	SAM-T99, 3DPSSM	www.pdg.cnb.uam.es/servers/libellula.html	
Meta[2], utilize Meta[1]			
PCONS/PROQ selector		Available only via 3D-JURY and GeneSilico servers	Predicts the quality of models generated by various FR methods by all-against-all comparison and evaluation with the ProQ method
@TOME selector	PDB-BLAST, 3D-PSSM GENTHR., FUGUE, SAM-T99	bioserv.cbs.cnrs.fr	Predicts the quality of models generated by FR methods using the TITO method. Top models ranked by Verify3D and PROSAII
SHGU/SHGUM added-value	BIOINBGU	www.cs.bgu.ac.il/~bioinbgu/	Utilizes the 3D-SHOTGUN approach to create hybrid models. SHGU produces raw C-α models, SHGUM produces full-atom models using MODELLER
3DS3/3DS5* added-value	BIOINBGU, FFAS, 3DPSSM, FUGUE*, GENTHR.*	www.cs.bgu.ac.il/~bioinbgu/	Utilize the 3D-SHOTGUN approach to create hybrid models (C-α). Assemble hybrid models from fragments of models generated by original and consensus methods

Table 1. (*Continued*)

Method	Input	Reference (http:// URL)	Comments
Meta[3], utilize Meta[2] ROBETTA added-value	PDB-BLAST PCONS, ROSETTA	Not available publicly	Takes the Pcons results to initiate the folding simulation using the Rosetta method
GENESILICO selector	A variety of primary FR methods + PCONS	genesilico.pl/meta/	Common input and output. Predicts SS with several methods, runs several FR methods and re-ranks the results using Pcons. All models additionally ranked by Verify3D
Meta[4], utilize Meta[3] 3D-JURY selector	A variety of primary FR methods + SHOTGUN, PCONS	bioinfo.pl/meta/	Common input and output. Predicts SS with several methods, runs a variety of FR methods and consensus methods. All results (including the original models and consensus models) are re-ranked by the 3D-JURY system
Meta[5], utilize Meta[4] PRCM added-value	3D-JURY	protinfo.compbio.washington.edu	Builds, minimizes and ranks the full-atom models starting from the crude FR models selected by 3D-JURY
Meta[6], utilize Meta[5] ALEPH0-JURY selector	ROBETTA, PRCM, SHGUM	Fisher (unpubl.)	Selects the best full-atom model

by the PROQ method for protein model evaluation (Wallner and Elofsson 2003), exhibits even higher specificity; moreover, it is able to use the set of any external (original or meta) methods as model generators.

LIBELULLA (Juan et al. 2003) is a system of neural networks trained to select correct folds from among the results of two *primary* fold-recognition methods implemented as web servers, SAM-T99 (Karplus et al. 1999) and 3DPSSM (Kelley et al. 2000). It uses a set of associated characteristics such as the quality of the sequence-structure alignment, distribution of sequence features (sequence-conserved positions and apolar residues), and compactness of the resulting models.

Another fully-automated meta-predictor that simply selects models from those produced by other servers is 3D-JURY (Ginalski et al. 2003). It takes as input any set of models, structurally compares all against all using MaxSub (Siew et al. 2000), and selects one that appears to contain the largest recurrent subset of common coordinates. It does not use any special characteristics of the models or of the servers. 3D-JURY is coupled to the BioInfo.PL Meta-server and, thus, can use any model including selection of the *most common* model from a user-defined subset.

2.3 Hybrid Methods: Going Beyond the "Simple Selection" of Models

Some automated meta-predictors go beyond the simple selection. PMOD uses MODELLER (Sali and Blundell 1993) to generate full-atom models based on the selection of fold-recognition results reported by PCONS, amended by secondary structure predicted by PSI-PRED (Jones 1999). These models are evaluated using the PROQ method (Wallner and Elofsson 2003). ROBETTA (D. Baker, unpubl.) builds full-atom models using the ROSETTA fragment insertion method (Simons et al. 1997), starting from structures detected by PDB-BLAST or PCONS and aligned by the K*SYNC alignment method. PRCM takes as input the top models selected by 3D-JURY and builds full-atom models, which are minimized and evaluated using energy functions. ALEPH0-JURY (D. Fischer, unpubl.) selects a model from those of ROBETTA, PRCM, and SHGUM using a combination of the 3D-SHOTGUN technology (see below) and evaluation using knowledge-based potentials.

Another successful practice observed in previous CASPs was to build hybrid models from fragments (e.g. Bujnicki and GeneSilico; see above). Automated meta-predictors using this approach have also been developed. Conceptually, the first method to use the *fragment-splicing* approach (which nevertheless should not be considered a meta-server) was David Baker's ROSETTA protein folding simulation algorithm that uses the fragment insertion Monte Carlo approach (Simons et al. 1997). The general premise of this method is that the protein conformation is reasonably well approximated by the distribution of local structures adopted by known, not necessarily homol-
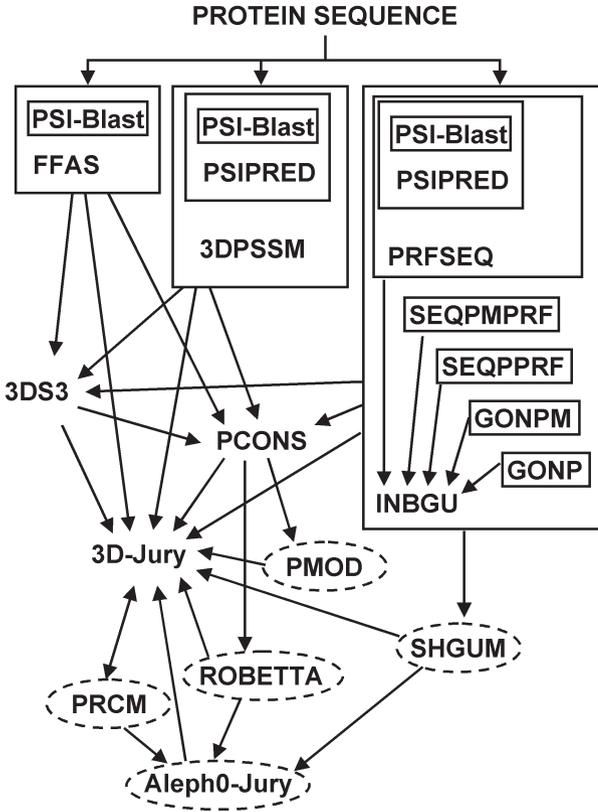
Fig. 1. Mutual interdependencies of meta-servers and their reliance on the original methods. Meta-servers *encircled by broken lines* produce refined (energy-minimized) full-atom models, other meta-servers produce crude C-α models

ogous, protein structures. Protein structure fragments are obtained from the protein structure database (Simons et al. 1997). The original version of ROSETTA utilized the I-sites (invariant or initiation sites) library developed by Chris Bystroff (reviewed elsewhere in this volume), which consists of a set of short motifs, lengths 3 to 19, obtained by a clustering of sequence segments from the Protein Data Bank (Bystroff and Baker 1998). ROSETTA has been notoriously succesful in CASP3, CASP4, and CASP5, demonstrating that protein structure modeling by recombination of fragments derived from experimentally solved structures is a powerful approach.

3D-SHOTGUN (Fischer 2003) is the first fully automated meta-predictor, which assembles hybrid C-α models by combining the structures of individual models, independently obtained from different fold-recognition methods. The 3D-SHOTGUN approach is superior to "pure" selection, as the resulting hybrid models are on the average more complete and more accurate than the input models. There are three versions of 3D-SHOTGUN: (1) an independent version named SHGU using as input models generated by the BIOINBGU server (Fischer 2000); (2) 3DS3 and (3) 3DS5, which uses as input the models

from three or five different independent fold-recognition servers, respectively. A new automated version of the SHOTGUN series, which was very successful in CASP is SHGUM, which generates full-atom, refined models, without the collisions and gaps seen in some of the *raw* spliced models. SHGUM is an independent server using the same input as SHGU (i.e. the results of BIOINBGU).

Figure 1 shows the diagram of mutual interdependencies of "meta^N-servers" and their reliance on the input from original servers and other "meta^{N-1}-servers".

3 Future Prospects

The recent CASP5/CAFASP3 evaluation has clearly shown that the meta-servers, on average, perform much better than these primary servers and the higher the N in the meta^N, the more the meta-server is likely to succeed. This works because no program is suitable for all cases, and each program has its strengths and weaknesses, and with each layer of "meta"- analysis the strengths can be amplified. The idea of meta-servers, or more precisely – the post-CASP5 proliferation of meta-servers – has met, however, with ambiguous reactions of the community of developers of bioinformatic methods. On the one hand, it is much easier to develop meta-servers (especially a relatively simple selector) than to develop a new, original fold-recognition method. On the other hand, because of the *out-sourcing*, the existing meta-servers are very slow: always slower than the slowest of the external servers used to generate primary predictions.

The idea of "meta-prediction" is well known in areas such as artificial intelligence or the stock market, where independent agents are used to obtain a consensus prediction that will be on average more accurate than any of the individual agents. Initially, you consult with various external brokers, but if you have the money, you just hire them to sit at your location. Thus, the richest will be the winner. Likewise, in order to obtain a fast protein fold-recognition meta-server-type method applicable at genomic scales, meta-predictors will run each of the "lower-rank" components locally, without the dependence on other external servers. Many of the existing fold-recognition methods have implemented a local version of PSI-BLAST (Altschul et al. 1997) for database searches and generation of multiple sequence alignment, and PSI-PRED (Jones 1999) (which itself utilizes PSI-BLAST) for prediction of secondary structure. It has been envisaged that the future meta-servers would utilize local implementations of original fold-recognition methods and lower-rank meta-protocols. Hence, some of the criticism attributed to the first generation of meta-servers may not be justified and will certainly fade away in the future, when fast, powerful independent meta-predictors will challenge the best human predictors. Whether this will happen at CASP6 or later, remains to be seen.

References

Alexandrov NN, Nussinov R, Zimmer RM (1996) Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. Pac Symp Biocomput, pp 53–72

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

Bryant SH, Lawrence CE (1993) An empirical energy function for threading protein sequence through the folding motif. Proteins 16:92–112

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001a) LiveBench-1: continuous benchmarking of protein structure prediction servers. Protein Sci 10:352–361

Bujnicki JM, Elofsson A, Fischer D, Rychlewski L (2001b) LiveBench-2: Large-scale automated evaluation of protein structure prediction servers. Proteins 45:184–191

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

Cuff JA, Barton GJ (2000) Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 40:502–511

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ (1998) JPred: a consensus secondary structure prediction server. Bioinformatics 14:892–893

Di Francesco V, Geetha V, Garnier J, Munson PJ (1997) Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. Proteins (Suppl 1):123–128

Fischer D (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. Pac Symp Biocomput, pp 119–130

Fischer D (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. Proteins (in press)

Fischer D, Baker D, Moult J (2001a) We need both computer models and experiments. Nature 409:558

Fischer D, Barrett C, Bryson K, Elofsson A, Godzik A, Jones D, Karplus KJ, Kelley LA, MacCallum RM, Pawlowski K, Rost B, Rychlewski L, Sternberg M (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. Proteins (Suppl 3):209–217

Fischer D, Elofsson A, Rychlewski L, Pazos F, Valencia A, Rost B, Ortiz AR, Dunbrack RL Jr (2001b) CAFASP2: The second critical assessment of fully automated structure prediction methods. Proteins 45 (Suppl 5):171–183

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

Godzik A, Kolinski A, Skolnick J (1992) Topology fingerprint approach to the inverse protein folding problem. J Mol Biol 227:227–238

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292:195–202

Jones DT, Taylor WR, Thornton JM (1992) A new approach to protein fold recognition. Nature 358:86–89

Juan D, Grana O, Pazos F, Fariselli P, Casadio R, Valencia A (2003) A neural network approach to evaluate fold recognition results. Proteins 50:600–608

Karplus K, Barrett C, Cline M, Diekhans M, Grate L, Hughey R (1999) Predicting protein structure using only sequence information. Proteins (Suppl 3):121–125

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

Kelley LA, McCallum CM, Sternberg MJ (2000) Enhanced genome annotation using structural profiles in the program 3D-PSSM. J Mol Biol 299:501–522

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

Kolinski A, Betancourt MR, Kihara D, Rotkiewicz P, Skolnick J (2001) Generalized comparative modeling (GENECOMP): a combination of sequence comparison, threading, and lattice modeling for protein structure prediction and refinement. Proteins 44:133–149

Kosinski J, Cymerman IA, Feder M, Kurowski MA, Sasin JM, Bujnicki JM (2003) A 'Frankenstein's monster' approach to comparative modeling: merging the finest fragments of fold-recognition models and iterative model refinement aided by 3D structure evaluation. Proteins (in press)

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

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

Marr D (1982) Vision. Freeman, New York

Marti-Renom MA, Stuart AC, Fiser A, Sanchez R, Melo F, Sali A (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29:291–325

Moult J, Fidelis K, Zemla A, Hubbard T (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. Proteins (Suppl 5):2–7

Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT (1997) Critical assessment of methods of protein structure prediction (CASP): round II. Proteins (Suppl 1):2–6

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

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

Ouzounis C, Sander C, Scharf M, Schneider R (1993) Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. J Mol Biol 232:805–825

Rost B(1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. ISMB 3:314–321

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

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

Selbig J, Mevissen T, Lengauer T (1999) Decision tree-based formation of consensus protein secondary structure prediction. Bioinformatics 15:1039–1046

Shi J, Blundell TL, Mizuguchi K (2001) Fugue: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. J Mol Biol 310:243–257

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

Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D.(1999) Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 34:82–95

Sippl MJ, Weitckus S (1992) Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. Proteins 13:258–271

Tramontano A (2003) Of men and machines. Nat Struct Biol 10:87–90
Wallner B, Elofsson A (2003) Can correct protein models be identified? Protein Sci 12:1073–1086