1874–1891 *Nucleic Acids Research, 2005, Vol. 33, No. 6*
doi:10.1093/nar/gki327

SURVEY AND SUMMARY

Practical lessons from protein structure prediction

Krzysztof Ginalski[1,2,3], Nick V. Grishin[3,4], Adam Godzik[5] and Leszek Rychlewski[1,*]

[1]BioInfoBank Institute, ul. Limanowskiego 24A, 60-744 Poznań, Poland, [2]Interdisciplinary Centre for Mathematical and Computational Modelling, Warsaw University, Pawińskiego 5a, 02-106 Warsaw, Poland, [3]Department of Biochemistry, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9038, USA, [4]Howard Hughes Medical Institute, University of Texas, Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, TX 75390-9050, USA and [5]The Burnham Institute, 10901 N. Torrey Pines Road, La Jolla, CA 92037, USA

Received December 20, 2004; Revised and Accepted March 8, 2005

ABSTRACT

Despite recent efforts to develop automated protein structure determination protocols, structural genomics projects are slow in generating fold assignments for complete proteomes, and spatial structures remain unknown for many protein families. Alternative cheap and fast methods to assign folds using prediction algorithms continue to provide valuable structural information for many proteins. The development of high-quality prediction methods has been boosted in the last years by objective community-wide assessment experiments. This paper gives an overview of the currently available practical approaches to protein structure prediction capable of generating accurate fold assignment. Recent advances in assessment of the prediction quality are also discussed.

INTRODUCTION

Methodological advances in DNA sequencing resulted in an outbreak of sequence information (1–3). Compared to about 3 million publicly available different protein sequences, the number of experimentally determined spatial structures lags two orders of magnitude behind. Despite the structural genomics initiatives and biochemical efforts to characterize protein families, the fastest way to gain information about the structural and functional properties of a protein is through computational inference from an experimentally studied homolog (4–9). Structure prediction, even in the absence of homology, is an important first step in the sequence-to-structure-to-function paradigm (10) (Figure 1). Recognizing numerous pitfalls in the naïve application of this paradigm, we agree that knowledge of the spatial structure assists functional prediction, because proteins function as 3D objects. Theoretically, it should be possible to deduce structure from sequence by accurate simulation of physical processes (11). We are very far from achieving this goal, and the methods of practical importance were traditionally based on the observation that proteins with similar sequences are structurally similar as well (12). Simple sequence similarity-based approaches, such as BLAST, were potent in making structure–functional predictions (13), since statistically significant sequence similarity usually signifies homology, similar fold and related function. However, standard sequence methods leave a prohibitively large gap of more than 30% of the proteome while assigning protein function (14), whereas at the fold prediction level the effectiveness of these methods is even much worse. Frequent examples of proteins with similar function and structure, but undetectable sequence similarity have prompted development of more sensitive structure prediction algorithms (15,16).

In the 1990s threading (inverse folding, fold recognition) methods (17–20) emerged. These methods, matching a sequence to a structure, gave a promise to extend the power of sequence-to-sequence alignments. Initially, fold recognition was developed with a hope to detect analogous proteins with no evolutionary relationships, but with common fold. However, in many cases the homology was confirmed later by new, more sophisticated sequence comparison methods, such as PSI-BLAST (21). The fold recognition research left a significant impact on the protein classification field not only because of the development of new sensitive prediction approaches but also because it established the prediction accuracy assessment standards (22–34). Since structure diverges slower than sequence, fold similarity and structure-based alignments can be used for benchmarking prediction success for both fold recognition and distant sequence similarity detection methods. This led to the direct competition of both types of methods and inspired further development. Threading was routinely weighed against PSI-BLAST, but with time evolved into hybrid approaches, which in many aspects profited from the fast and sensitive sequence alignment algorithms (horizontal

*To whom correspondence should be addressed. Tel: +48 604 628805; Fax: +48 61 8643350; Email: leszek@bioinfo.pl

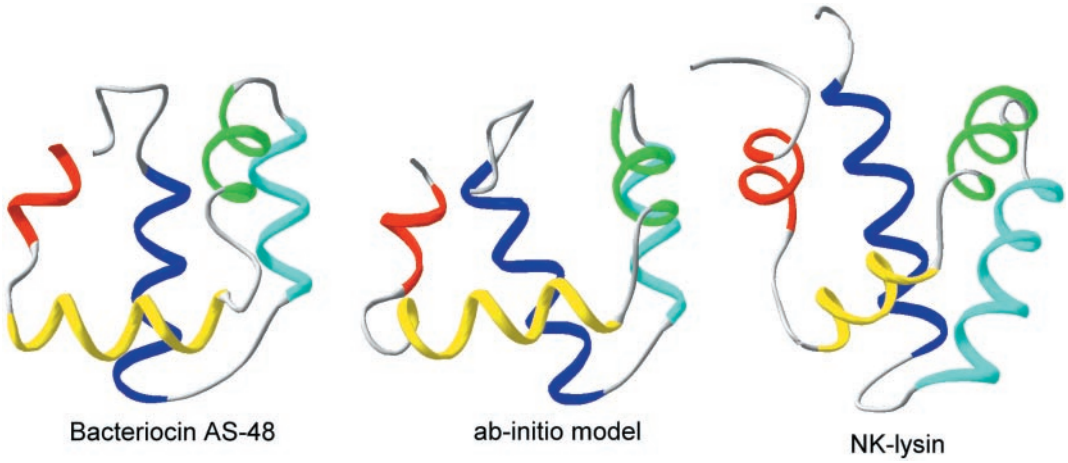

Figure 1. Sequence-to-structure-to-function paradigm. The leftmost picture shows the structure of Bacteriocin AS-48 (1e68, left) from *Enterococcus faecalis*, a 70 residues long cyclic bacterial lysin (100). This protein is structurally and functionally related to mammalian NK-lysin (101) (1nkl, right) despite undetectable sequence similarity, as only 4% of residues are identical after structural superposition. The Bacteriocin sequence was a target T0102 in the CASP-4 experiment. An excellent model (middle) was obtained by the Baker (59) group using the *ab initio* method Rosetta with an RMSD of 3.5 Å over all 70 residues. No other method was able to predict this fold with similar accuracy. A search of the protein structure database with this model yielded NK-lysin as the first structural match of comparable length. This illustrates that the *ab initio* approach was able to predict the structure that could be used to predict the function of the protein.

software transfer). The secondary structure prediction components were often based on PSI-BLAST profiles, and sequence profile scoring was soon added to the fold recognition scoring functions. At the time when pure threading was replaced by hybrid approaches, gapped alignment of two sequence profiles provided an accuracy boost to methods, which completely ignore the structural information (35). Since the first community-wide benchmarking of servers (27) in 1998, such 'sequence-only' algorithms have proven to be competitive in structure prediction tests. Until today, the advantage of using the structural information available for one partner in comparing two protein families has not been clearly demonstrated in benchmarks. Nevertheless, generation of structural models in the course of prediction has proven to be very useful, i.e. when clustering structural models in meta predictions.

In a meta prediction, the query sequence is subjected to a variety of different prediction approaches, the results are collectively analyzed for consistency to generate consensus predictions and to estimate their reliability. With the appearance of the first protein structure prediction, meta server (36) developers obtained convenient access to many different 3D models produced with various prediction methods, but standardized in terms of their format. It has become possible to cluster a large set of models by structural comparison. Such clusters could contain similar models based on evolutionary diverse templates additionally supporting the putative structural predictions. This idea was soon exploited by experts (28) and first automated meta predictors (37) that selected representatives of large clusters of models rather than models with the highest score. This strategy turned out to be very successful and soon after the first versions of meta predictors became available, they took over individual methods as shown by evaluation of blind predictions, including the community-wide assessment of protein structure prediction [CASP-5 (26)]. Multiple tests confirm that consensus methods are more powerful than individual prediction servers in sensitivity and specificity even if some meta predictors use as little as three component servers as sources of models.

The biannual CASP (Critical Assessment of techniques for protein Structure Prediction) experiments objectively evaluate the prediction protocols used by experts. It is increasingly evident that good results are obtained by groups, which utilize as many different sources of information as feasible given the short time and available human resources. The most common successful strategies include the initial analysis of many models collected with meta servers and processed by various meta predictors, which sometimes include *ab initio* components, followed by manual selection and tuning of final models supported by extensive literature analysis. To fully benefit from the growing number of resources, it is important to know the differences between available prediction methods and how to come to correct consensus decisions and improve structural model for a target of interest. This paper is aimed at providing an overview of current techniques that have proven to be most successful judging by the results of the CASP-5 experiment. We summarize the lessons from many years of testing the structure prediction methods and suggest possible applications for 3D protein models in biological research.

PROTEIN STRUCTURE PREDICTION METHODS

Although we are still far from the precise computational solution of the folding problem, a variety of different approaches to protein structure prediction are available after more than 50 years of research. They range from those based solely on physical principles to purely statistical methods and methods that rely on utilization of evolutionary information. The methods rooted in physics are still in their infancy and are not yet capable of large-scale generation of meaningful protein models. We focus on practical solutions that, despite the absence of theoretical rigor in them, can be and are successfully used by biologists in their research. Figure 2 provides an overview of different classes of algorithms described in more detail below. Table 1 lists servers that offer structure prediction service for the community of researches.

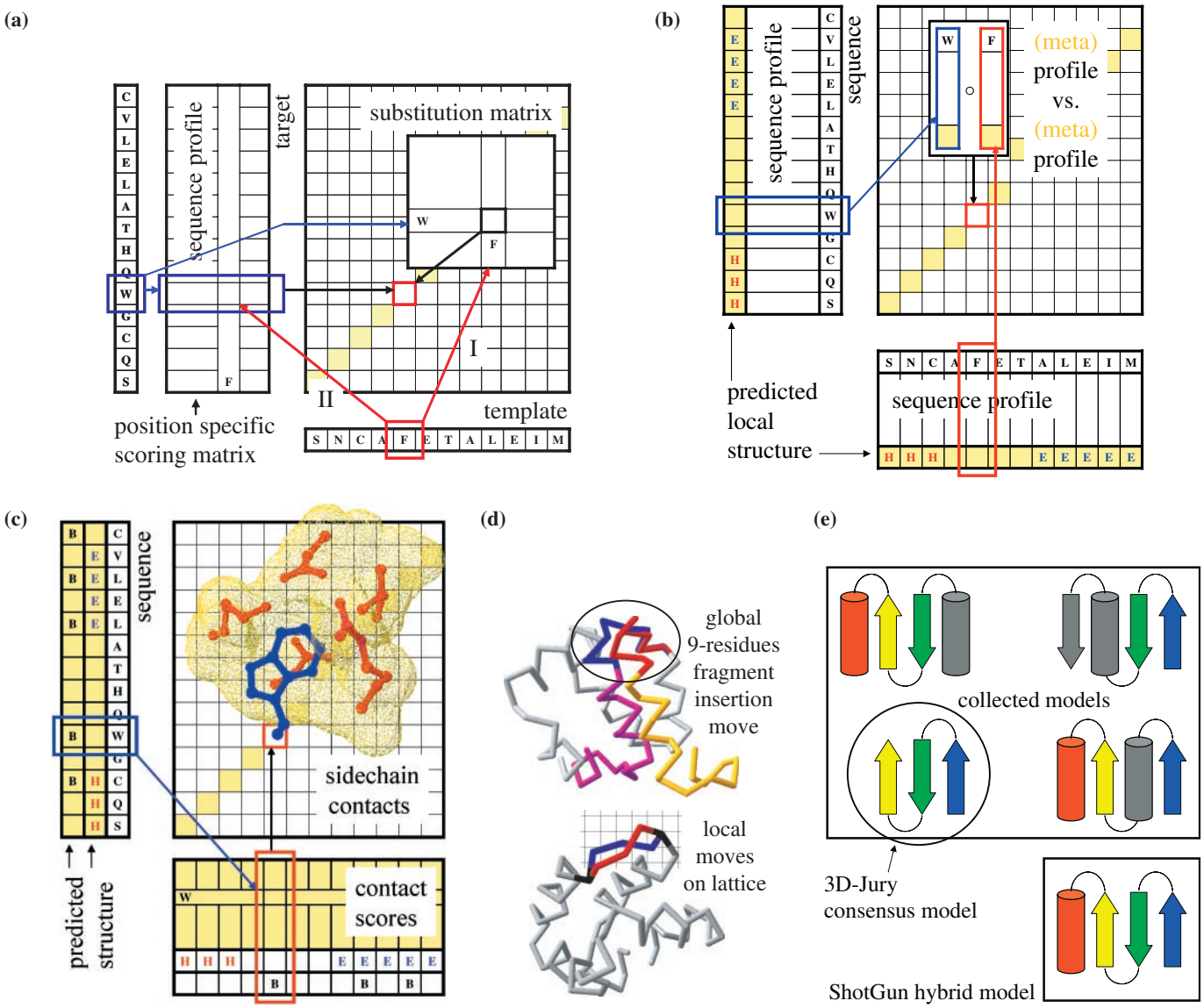

Figure 2. Protein structure prediction methods. (**a**) Sequence–sequence, profile–sequence, sequence–profile comparison methods represent a traditional evolutionary-based approach to predict structures of proteins. The simplest method (I) aligns the sequence of the target with the sequence of the template using a substitution matrix. More sensitive methods (II) define scores for aligning different amino acids separately for each position of the target sequence (PSI-BLAST) or the template sequence (RPS-BLAST). The scores are taken from the analysis of sequence variability in multiple alignments of the corresponding sequence families. Such position-specific scores are also called profiles. They are similar in format to the representation of sequence families used by prediction methods based on HMMs. (**b**) Profile–profile comparison methods utilize the profiles generated by the above mentioned sequence alignment methods. Instead of a lookup of a substitution score, they compare two vectors with each other when building the dynamic programming matrix used to draw the alignment. The comparison is usually conducted by calculating a dot product of the two positional vectors (as shown in the figure) or by multiplying one vector times a substitution matrix time the other vector. Depending on the choice of the comparison function the vectors are often rescaled before the operation. The sequence variability vectors are sometimes also augmented with meta information, such as predicted secondary structure as indicated in the figure. (**c**) Threading or hybrid methods utilize the structure of the template protein in the comparison function. The position-specific alignment scores are computed for the template protein by replacing the side-chain of a residue with side-chains of all possible amino acids and by calculating the resulting substitution scores using statistically derived contact potentials. In addition, factors such as matching of predicted and observed secondary structure or burial preferences are also taken into account when aligning two positions. Most threading methods use frozen approximation where the sequence is threaded through the template structure and contacts are calculated between the target side-chain and side-chains of the residues of the template. In the much slower, defrosted threading template side-chains are replaced with side-chains of the target according to the alignment before calculating the contact scores. (**d**) *Ab initio* methods represent a physical approach to predict the structure of the target protein. The methods are based on an energy function, which estimates the conformational energy of the chain of the modeled protein. The energy can be calculated in a similar fashion as in the threading methods, i.e. utilizing contact potentials. The advantage of *ab initio* is that the database of folds does not constrain the set of possible results and theoretically any conformation can be generated and tested. *Ab initio* methods differ in employed energy functions and in the way conformational modifications are generated. Most common methods employ fragment insertion techniques or constrain the move set by placing the molecule on a lattice. (**e**) Meta predictors represent statistical approaches to improve the accuracy of protein structure predictions. Simple meta predictors collect models from prediction servers, compare the models and select the one, which is most similar to other models. The consensus model corresponds to a model selected from the collected set and represents the final prediction. More advanced meta predictors are able to modify the set of collected models either by filing missing parts with *ab initio* or loop modeling or by creating hybrid models from segments of structures collected from prediction servers. Hybrid models have a higher chance to provide a more complete model but are sometimes unphysical in terms of chain connectivity.

Table 1. Publicly available fold recognition servers

Code	
	Sequence-only methods (no structural information required)
PDBb	PDB-BLAST is based on the PSI-BLAST (21) program. PSI-BLAST is iterated five times on the non-redundant protein sequence database clustered at 70% identity and masked with low complexity filters. Before the fifth iteration, the sequence profile is saved and used as query against sequences of proteins with known structures (from PDB). This server is a default reference most fold recognition servers are compared with.
FFAS FFA3	FFAS (45) is a profile–profile comparison method. Profiles are generated for protein families in a different way than in PSI-BLAST, but PSI-BLAST is used to collect the sequences of the families. The old version FFAS is now obsolete and replaced with the new version FFAS-03, which uses vector times matrix times vector multiplication when aligning two positions and improved transformation of raw alignment scores into *Z*-scores. FFAS is one of the first profile–profile comparison servers.
ORFs ORF2	ORFeus (51) is a meta-profile with meta-profile comparison method (meta profiles include sequence profiles and predicted secondary structure). It uses vector times vector multiplication. The old version returns the raw alignment score, while the new version ORFeus-2 translates the score into a *Z*-score.
mBAS BasD BasP	Meta-BASIC (mBAS) (69) is a local meta predictor, which uses six different versions of meta-profile alignment methods, including two versions of ORFeus. Distal-BASIC (BasD) uses two versions of low stringency meta profiles (five PSI-BLAST iterations) aligned with vector times vector and vector times matrix times vector multiplication. Proximal-BASIC (BasP) uses high stringency meta profiles (only three PSI-BLAST iterations). The strongest asset of these algorithms is their high specificity.
ST99	Sam-T99 (44) builds a multiple alignment (the SAM-T99 alignment) by iterated search using HMMs. It uses the alignment to predict secondary structure (with various methods) and to build an HMM for searching PDB for similar proteins. Also, a library of HMMs built by similar methods from PDB sequences is used to score the target sequence. This server has a long tradition and was one of the best servers in CAFASP-1.
SFAM SFPP	SUPERFAMILY (102) is a library of HMMs based on SCOP. The server uses HMMs and the SAM methodology as does Sam-T99. SUPFAM_PP is the next generation of SUPERFAMILY. Both servers are capable of generating hybrid models using partial alignments to various templates. The top 10 generated models are sometimes quite redundant.
FRT1	FORTE-1 (103) is a profile–profile comparison method. The correlation coefficient is used as similarity measure of two aligned profile positions. The profiles are generated using PSI-BLAST.
	Hybrid methods (use structural information of the template)
ST02	SAM-T2K (104) iterated search procedure is used to create a multiple alignment of homologs. Templates are aligned with three different target HMMs (using different secondary structure predictions and also no secondary structure prediction at all) and the target is aligned with template HMMs. Many alignments are made and the top five distinctly different ones are reported. This server has a higher accuracy than Sam-T99.
3DPS	3D-PSSM (105) is based on a hybrid threading approach using 1D and 3D sequence profiles coupled with secondary structure prediction and solvation potential. 3D-PSSM is one of the first fold recognition servers. It was rated as very sensitive in LiveBench-2.
GETH MGTH	GenTHREADER (GETH) (106) uses a combination of various methods, including sequence alignment with structure-based scoring functions as well as a neural network-based jury system to calculate the final score for the alignment. mGenTHREADER (MGTH) is an enhanced version of GenTHREADER. It takes as input a PSI-BLAST profile calculated for the target sequence. Both versions took part in the first CAFASP evaluation and have a long history. mGenTHREADER was rated as very specific in LiveBench-2.
FUG2 FUG3	In FUGUE (107), environment-specific substitution tables were derived from the structure-based alignments in the HOMSTRAD database. Each alignment in HOMSTRAD was converted into a scoring template (profile) using the environment-specific substitution tables with environment-dependent gap penalties and enhanced by homologous sequences. FUGUE takes a sequence or sequence alignment and searches against the library of profiles. FUGUE is a relatively new server.
RAPT	RAPTOR (108) uses a threading technique for fold recognition. It minimizes an energy function consisting of mutation, singleton, pair-wise and secondary structure terms. The method is formulated as a large-scale integer programming problem. Support Vector Machine technique is used to assess the alignment reliability. RAPTOR is quite new and was very successful in CASP-5.
SPKS	SPARKS (Sequence, secondary structure Profiles And Residue-level Knowledge-based Score for fold recognition) (109) uses single-body residue-level knowledge-based energy score combined with sequence profile and secondary structure information for fold recognition.
PRO2	PROSPECT (PROtein Structure Prediction and Evaluation Computer Toolkit) (110) is a threading-based protein structure prediction system. The system uses the following terms: mutation energy (including position-specific score matrix derived from multiple sequence alignments), singleton energy (including matching scores to the predicted secondary structures), pairwise contact potential (distance dependent or independent) and alignment gap penalties.
INBG	INBGU (111) is a combination of five methods, which exploit sequence and structure information in different ways and produces one consensus prediction of the five. It uses predicted versus observed secondary structure and sequence profiles for both the target and for the folds in the library. The precursor of INBGU (frsvr) was one of the best performing servers in CAFASP-1.
SHGU	ShotGun-INBGU (68) uses the ShotGun consensus layer to create alternative consensus models from the INBGU components. The server is much more accurate than INBGU. It uses only in-house components, but it is almost as accurate as structure meta predictors.
	Structure meta predictors (build consensus form other servers)
PCO2 PCO3 PCO4 PCO5 PMOD PMO3 PMO4	Pcons (37) comes in various versions and uses various sets of component servers to generate consensus predictions. The largest set includes SUPFAM_PP, FFAS-03, FFAS, SAM-T2K, FUGUE-3, PROSPECT, mGenTHREADER, INBGU, 3D-PSSM, ORFeus, FORTE-1 and PDB-BLAST. Pcons (PCO . . .) returns one of the models obtained from component servers while Pmodeller (PMO . . .) runs Modeller (65) using the alignments collected from the servers. Pcons is the first automated structure meta predictor and receives very good scores since LiveBench-2.
3DS3 3DS5	ShotGun (68) is a consensus predictor which utilizes the results of FFAS-03, 3D-PSSM and INBGU (3DS3) and in the larger version also FUGUE and mGenTHREADER (3DS5). It compiles a hybrid model from the models produced by the component servers by combining partial structures. The generated structures are sometimes unphysical but the server has very high sensitivity and specificity (reliability estimation).
3JAa 3JBa 3JCa 3JA1 3JB1 3JC1	3D-Jury (64) is an interactive meta predictor. The user can select the set of servers used for consensus building. 3D-Jury can also include other meta predictors making this server a 'meta–meta predictor' (the 3JB and 3JC version). It can operate in single model (one model per server, suffix '1') or multiple model (suffix 'a') modes. The default 3JA1 version uses 8 component servers and the single model mode.

Table I. *Continued*

Code	
	Ab initio *meta predictors (use meta predictors and ab initio modules)*
RBTA	Robetta (66) produces full chain models with the Rosetta *de novo* and comparative modeling methods. *De novo* models are built by fragment insertion simulated annealing. Comparative models are built by detecting a parental PDB structure with PSI-BLAST or Pcons2, aligning the query to the template with the K*SYNC alignment method, and modeling variable regions with a modified version of the *de novo* protocol. Robetta is one of the best performing servers as evaluated in CASP-5.
PRCM	PROTINFO-CM (67) uses 3D-Jury as initial alignment provider. Initial models are then built for each alignment and scored. Loops and side-chains are built on the best scoring models using a frozen approximation. The relatively slow method (not available publicly at present) does a sophisticated graph–theory search to mix and match between various main-chain and side-chain conformations. Good results have been obtained in LiveBench-7 but due to limited computational resources it was withdrawn from LiveBench-8.

The table provides a short description of selected, publicly available servers that took part in LiveBench-7 or LiveBench-8 and gives a short description of the underlying algorithm. The new *ab initio* meta predictors are quite slow and because of this not yet available for public use. Their additional weakness is the lack of a confidence score. Sequence-only methods neglect by definition any information about the structure of the template and in contrast to hybrid methods can be used as general homology inference methods between any protein families. Structure meta predictors offer currently the highest utility by producing accurate models with reliable confidence assessment.

Sequence similarity-based methods

Sequence–sequence comparison. The most popular and simple method to assign the fold of a protein is by finding a close homolog with a known structure. Simple sequence–sequence comparison methods, such a BLAST (13) or FASTA (38), can assign the fold for ∼30% genes in microbial genomes (39). First versions of BLAST provided ungapped local alignment and required only a substitution matrix, which defines a score of aligning residues of two types. Currently, all common programs provide gapped alignment requiring parameters defining the penalty for allowing gaps, usually a gap initiation and a gap extension penalty. Both alignment parameters and substitution matrices have been extensively evaluated to obtain best alignment and highest discriminative power between significant scores for homologs and expected random scores (40). The main advantage of methods like BLAST is their extreme speed achieved through tailored initial screening of sequences in the database, which have a chance to obtain a sufficiently high alignment score. The disadvantage of such methods is that conserved and variable positions are treated with the same weight and have the same effect on the final alignment score. In contrast to newer approaches, the ability to detect distant homologs with simple sequence alignment is limited.

Profile–sequence and sequence–profile comparison. The assumption that the alignment of conserved motifs is more important than the alignment of variable regions led to the development of position-specific substitution matrices (41). The generic 20 times 20 substitution matrix is replaced by an N times 20 substitution matrix, called profile, which defines the score for aligning any of the 20 amino acids to each of the N residues of the protein for which the profile is built. Such profiles are generated based on the variability of amino acids found in multiple sequence alignment of the target with its close homologs. Thus, the profile describes a family of homologs rather than a single sequence. The calculation of the profile requires an initial multiple alignment as input. This calculation is the only additional computational requirement relative to the simple sequence alignment tools described earlier. The speed of aligning a profile to a sequence is approximately the same as aligning two sequences, because the score of aligning two positions is calculated through a lookup in a profile or a matrix, respectively.

The most popular profile–sequence comparison method is PSI-BLAST (21). It enables the iterative generation of multiple alignments and profiles for the query protein. Its popularity is partially attributed to its high speed, which comes from the same initial screening technique for potential high scoring hits as implemented in BLAST. RPS-BLAST, a recent addition to the BLAST-based tools, offers the possibility of searching a database of profiles calculated, for example, for conserved domains, with a query sequence (42). It enables a much faster analysis of the query protein because no iterative generation of multiple alignments and profiles for the target protein is needed. Nevertheless, the search is limited to a relatively small set of several thousands of protein families in contrast to currently about 3 million non-redundant sequences that can be aligned to the query profile with PSI-BLAST. The database of profiles used by RPS-BLAST requires much more space because at least 20 values must be stored for each residue. This database has to be read from the disk each time a new prediction is conducted if it exceeds the RAM limits of the computer system.

Other closely related methods are based on the application of hidden Markov models (HMMs) (43,44). The models describe the sequence variability of the protein family and contain the probability of occurrence of each of the 20 amino acids at each position of the query protein. This is essentially identical to the information stored in position-specific substitution matrices, but instead of gap penalties HMMs operate with position-specific deletion and insertion probabilities. The HMMs are also calculated based on multiple alignment of the family of homologous proteins. However, HMM search tools do not use any initial filtering of a database. This greatly reduces the speed of the comparison but has a positive effect on the sensitivity of the method. In most benchmarks, HMMs outperform PSI-BLAST based procedures.

Profile–profile comparison. Using profile with sequence (profile–sequence and sequence–profile) alignment methods, it is possible to compare both query profile with template protein and query sequence with template profile, which results in two potentially different alignments. The attempt to make this comparison more symmetric resulted in the direct alignment of two profiles (45–50). Instead of scoring an amino acid in a sequence with a position of the profile, the direct

profile–profile alignment methods compare two profiles with each other. There are many ways to convert two positional vectors into one similarity score. The first and simple approach is based on the calculation of a dot product, which equals to the sum of products of all 20 pairs of amino acid substitution scores. Scores equal to the first vector times a substitution matrix times the second vector are also used, and the vectors correspond to observed amino acid frequencies at a given position. The main problem when developing such methods was the definition of the alignment parameters that include the gap initiation and gap extension penalties and a constant value, which is subtracted from the vector comparison score. The constant value ensures that the expected score of aligning two positions remains below zero. Otherwise, the expected optimal alignment would be global and will typically span the entire sequences of both proteins.

The reason for developing profile–profile comparison methods was to obtain more sensitive tools applicable even if none of the alignment partners has a known structure. It was shown that such methods can detect similarity between two families, undetectable when using profile with sequence comparison for any member of the query or the template families (45). Recent advance in profile comparison tools is based on the application of meta profiles, which add predicted secondary structure preferences as additional three values to the position-specific substitution scores, as implemented in ORFeus (51). The secondary structure prediction is based solely on the sequence profiles themselves, thus no additional source of information is required. Nevertheless, this approach of presenting the information to the alignment program seems to result in a more sensitive detection of similarity between protein families.

Threading methods: sequence-to-structure scoring

Two observations inspired the development of threading methods. First, proteins sharing similar structure while showing negligible sequence similarity were discovered. This led to the conclusion that sequence similarity is not necessary for structural similarity, suggesting that convergent evolution can drive completely unrelated proteins to adopt the same fold. Second, the analysis of spatial arrangements of amino acids in protein structures resulted in the identification of interaction preferences and development of residue contact potentials. It is easy to explain the observation that hydrophobic residues have a higher chance to be in contact with other hydrophobic residues just owing to the fact that those residues are expected to be packed in the interior of the protein forming the hydrophobic core. While these interactions played an undoubtedly dominant role in the calculation of potential, the developers claimed that the contact-based scoring matrices contain information about other essential and specific interactions shaping the native structures of proteins. The success of this concept is partly owing to the simplicity of representing the interactions between various amino acids in form of a matrix, which has the format of a substitution matrix.

The threading methods take their name from the conceptual threading of the sequence of the query protein through the structure of the template. The structural environment around a residue could be translated into substitution preferences by summing the contact preferences of surrounding amino acids. This means that, knowing the structure of a template, the preferences for 20 amino acids in each position can be calculated and expressed in the form of a profile of N times 20 values. This profile has the same format as the position-specific scoring profile used by sequence alignment methods, such as PSI-BLAST, and can be used to evaluate the fitness of a sequence to a structure. The fact that the sequence of the protein is usually not the sequence with the highest fitness score to this protein's structure did not discourage developers. This phenomenon can be easily explained by the concept that native proteins may fold through a process of elimination of other unfavorable conformations. Thus, the native sequence adopts the native structure because the fitness score to the native structure is much higher than that to other possible conformations. Such an energy gap was not necessarily observed for sequences designed artificially to optimize the fitness score to one particular structure.

However, one problem remains unsolved. If threading methods are designed to find similarity between evolutionary distant or even unrelated proteins, which share much <30% sequence identity, then the actual structural environments should also change dramatically. On average, one could expect that for each considered (defined as a central) residue <30% of surrounding amino acids are identical in both structures. The handling of this problem divides the threading methods in those using the 'frozen approximation' and those using the 'defrosted approximation'. In the first approach, while analyzing the fitness of the query sequence to the template structure, the surrounding structural environments for each residue of the query are kept identical to those observed in the template structure. This procedure is as fast as aligning a profile with a sequence but has important disadvantage that calculated in such a way local environments have little in common with those that might be observed in the native structure of the query protein. Most of them are essentially wrong, as majority of surrounding residues in template structure are replaced by different amino acids in the query protein. In contrast, the defrosted approach updates the surrounding amino acids of the template with the aligned amino acids of the query protein when calculating the fitness of the central residue (52). This has a dramatic negative effect on the speed of the optimal alignment, which now cannot be evaluated on a local basis because the alignment score depends on the alignment of other parts of the protein. Stochastic methods are used to update the alignment, and the fitness of the sequence to the structure is evaluated each time the alignment changes. The calculation of the fitness of a sequence to a structure can take hours and the scanning of databases of folds requires large computer resources. Nevertheless, methods using the defrosted approach are much more accurate in predicting the fold of a protein than threading with frozen approximation. Fold recognition in CASP-3 (24) in 1998 was dominated by such methods (53). However, with the growing number of known structures, the computational requirements become prohibitive. While running a dozen of CASP targets is feasible, genome annotations cannot be conducted without massive computer resources.

Hybrid methods: combining sequence similarity with threading

The impractical speed of orthodox threading programs motivated the quest for other sensitive fold recognition methods,

which could be applied on the rapidly increasing number of targets resulting from genome sequencing projects. The fast growth of the sequence databases contributed to a critical review of the threading concept, which was based on the assumption that the local structural environment has an effect on the amino acid substitution pattern of each considered residue. If this effect would be the main determinant of the mutational behavior of a residue, it should manifest itself in the mutations observed in homologous proteins. This reasoning led to the conclusion that profiles generated with sequence alignment methods, such as PSI-BLAST, already include the mutation preferences imposed by the native conformation. Threading algorithms would then be required only if insufficient information exists about the sequences of proteins homologous to the template protein. However, the majority of protein families with known structure do have sufficient homologs to calculate local substitution preferences from multiple alignments.

This observation gave rise to hybrid methods, which were designed to utilize sequence information from multiple sequence alignments if available, but also added terms such as residue-based secondary structure preferences or preferences to be buried in the core of the protein. The hybrid methods can use the frozen approximation when aligning a query protein to a template structure, because the secondary structure or the pattern of exposed and buried residues in structurally similar protein shows smaller variation than the amino acids and the local structural environments. The dynamic programming matrix of scores between all residues of two aligned proteins can be calculated using the employed terms before the alignment is drawn. This alignment defines the initial model of the structure of the query protein, which in some cases is additionally evaluated with previously mentioned contact potentials. Such methods have been successfully applied in genome-wide structure prediction experiments and claimed a higher fold assignment rate than that obtained with PSI-BLAST, which is routinely used as performance reference. However, direct comparison with profile–profile alignment methods turned out to be surprisingly favorable for the latter ones, which became serious competitors in protein structure prediction. Presently, the advantage of including the structural information in the fitness function cannot be clearly proven in benchmarks.

Practical *ab initio* methods

According to various benchmarks, fold recognition methods fail to select the correct fold from the database for ~50% of the cases when no significant sequence similarity exists. Fold recognition has also the limitation that no novel folds can be proposed since all predictions are based on already known structures. On the other hand, it is largely accepted that the structure and function of the protein is determined by its sequence (54). *Ab initio* methods are aimed at finding the native structure of the protein by simulating the biological process of protein folding. These methods perform iterative conformational changes and estimate the corresponding changes in energy. The main problems here are the inaccurate energy functions and the vast number of possible conformations a protein chain can adopt. The second problem is approached by reduced representation of conformations and coarse search strategies. The most successful approaches include lattice-based simulations of simplified protein models (55,56) and method building structures from fragments of proteins (57,58). The recent progress in the field is mainly attributed to the clustering of final conformations obtained after a large number of simulations. Representatives of large clusters are preferred as final models, which decrease the emphasis on calculated energy values.

Ab initio methods demand substantial computational resources. Nevertheless, they have been used successfully in last two CASP experiments on targets where fold recognition methods failed. The quality of the models remains quite low and it is difficult to say, which parts of which model are correct. Nevertheless, the course model can be used to query structural similarity searches against the database of known folds to detect distant functional similarity (59) (Figure 1). *Ab initio* methods are increasingly applied in large-scale annotation projects, including fold assignments for small genomes. *Ab initio* methods are also the only methods that can be used to design new proteins (60,61). However, only few sites provide *ab initio* structure prediction service for the community. Biologists have to rely on pre-calculated results available, for example, for selected Pfam (62) families. The *ab initio* methods are also quite difficult to use and expert knowledge is needed to translate the results into biologically meaningful predictions. Nevertheless, these methods are expected to have a huge impact on the future of structural biology.

Meta predictors: consensus from multiple methods

One of the main lessons from the last CASP experiments is that experts that utilize diverse sources of information are more successful than groups relying on a single structure prediction method. Hints influencing the selection of final models may come, for example, from biological expertise or literature searches (63). Such procedures are difficult to implement in an automated and reproducible fashion. However, the large diversity can also be obtained by utilizing the growing number of diverse prediction algorithms. A framework to profit from this diversity was created by Meta Servers (36) collecting and analyzing models from many prediction services spread around the globe.

The first successful attempt to benefit from the diversity of models was based on the simple approach of selecting the most abundant fold represented in the set of high scoring models, a procedure reminiscent of clustering simulated structures by *ab initio* prediction protocols. This procedure was easy to automate and resulted in the first fully automated meta predictor, Pcons (37). Several others followed soon. All benchmarking results obtained in the last 2 years indicate that meta predictors are more accurate than the independent fold recognition methods. Their strength is mainly attributed to the structural clustering of initial models. Even if many of them are wrong, it can be expected that structures of incorrectly predicted fragments of the models have random conformations; and only structures of fragments corresponding to preferred conformations occur with higher than expected frequency.

The positive evaluation results boosted further development of meta predictors. Currently available versions differ in the way the initial models are compared and the final model is generated, and in the use of the initial scores assigned to the

models by individual servers. Structural comparison of models must be fast. If 10 servers each providing 10 models are used to create a consensus prediction, almost 5000 structural comparison are required. Fast, sequence-dependent methods (see below), such as MaxSub or LGscore, are used to accomplish this task. When models do not exhibit high structural similarity, the initial scores assigned to each model by the original prediction method can be consulted to improve the selection procedure. However, this is not simple, because different fold libraries and scoring schemes were employed by different prediction servers. Some meta predictors develop server-specific neural networks to translate the initial values into uniform scores (Pcons). Others ignore the scores altogether and base their consensus evaluation only on the abundance of folds or structural motifs [3D-Jury (64)]. The final consensus model is either identical to one of the original models (3D-Jury, Pcons) or additional modifications are performed. Pmodeller runs the Modeller program (65) on the selected initial model. The ShotGun server combines the final model from fragments taken from several initial models. Servers such as Robetta (66) or ProtInfo (67) conduct *ab initio* calculations on parts of the models guided by the initial structures obtained from other servers, usually other simpler meta predictors. This procedure leads to the creation of meta–meta predictors theoretically unlimited in the complexity of combining various components. Technical aspects, such as the delay in waiting for the response of servers, limit the number of components to be included. As a consequence, some meta predictors are created from components available at a single site, such as ShotGun-INBGU (68), which uses the ShotGun fragment assembly technology in combination with the traditional local components of the INBGU hybrid structure prediction method.

The main advantage of developing all components in-house is that the scaling of scores and the libraries of templates can be standardized. This concept is applied in Meta-BASIC (Bilaterally Amplified Sequence Information Comparison) (69), which uses two versions of algorithms conducting gapped alignment of meta profiles, such as those used by ORFeus (see above). The scores computed by Meta-BASIC have a very high specificity, but the main advantage of this approach is that it does not require any structural information of the two aligned proteins. This makes Meta-BASIC (non-structural meta predictor) applicable to comparison of any sequence families, such as those from Pfam (62) or COG (70) databases.

EVALUATION OF PREDICTION METHODS

With considerable growth of protein structure prediction online services, an objective evaluation of available methods became essential. The launch of the CASP program in 1994 represents a crucial milestone in the protein structure prediction field. The first experiment started with the collection of sequences of proteins from the crystallographic community, for which the structure was about to be solved within the next few months. The sequences were made publicly available and the structure prediction community was challenged to respond with predictions for the released set of targets. At the time of the release, the native structure remained unknown to the predictors, the organizers and the crystallographers. Several months after releasing the targets, the crystallographers were asked to report on the progress of their structure determination efforts. Solved structures obtained so far have been collected and used to assess the predictions collected earlier. This procedure made a completely blind and relatively objective evaluation of the different structure prediction approaches and is being followed until today with ever-increasing interest and response from the scientific community.

There are few minor shortcomings of the CASP experiments. For example, evaluators (assessors) selected by the CASP organizing committee establish the exact assessment protocol after the release of the structural information by the experimentalists. Another aspect of CASP makes it difficult for biologists to use best-performing methods. The most accurate CASP predictions represent the work of groups of experts rather than the outcome of particular algorithms. The same prediction program can show different results in the hands of different experts owing to a different choice of parameters and databases. As a response to this, the CAFASP (Critical Assessment of Fully Automated Structure Prediction) experiment was launched in 1998 after the CASP-3 session. The first CAFASP experiment was conducted on the CASP-3 targets with a handful of automated programs. The goal was to provide assessments of methods with which any user would obtain the same results. The growth of the CASP community is accompanied by a fast growth of the number of methods taking part in the subsequent sessions of CAFASP. Almost 50 prediction servers took part in the last CAFASP-3 round in 2002. The experiments are now conducted in parallel with CASP in a completely blind fashion. The assessment procedures are essentially the same in all rounds and are published before each experiment.

A shortcoming of the CASP and CAFASP experiments is the relatively long delay between the development of a new method, a process which has been substantially expedited in the last years, and the availability of objective assessment of accuracy. In addition, it takes about a year from the beginning of the experiment to publication of the results. The LiveBench program is a response to this problem. Launched in 1998, it has followed a more instant assessment protocol. Protein structures released weekly in the PDB (Protein Data Bank) (71) are immediately submitted to the prediction servers, with the hope that the fold databases used by the prediction methods are updated in a slower fashion. After some time, this assumption was proven wrong but the procedure was modified to allow only models built using templates, which are at least 1 month older than the target. The main advantage of this program over the blind prediction experiments is that the evaluation is available almost instantly after the release of the target and delayed only by the time the servers need to compute the models. The disadvantage is obviously that the predictions are not blind any more and results obtained with the LiveBench program have to be compared with the outcome of blind prediction tests. To approach this problem other experiments, such as EVA or PDB-CAFASP, use now sequences of 'on-hold' entries available from PDB. The structures of 'on-hold' entries are released usually several months after releasing the sequences, but the delay varies significantly between cases.

The two programs (EVA and PDB-CAFASP) have the same problem with the delay of the evaluation as the traditional experiments CASP and CAFASP. Despite the transparent

Table 2. Selected evaluation measures used to assess the quality of 3D models

GDT TS (Global Distance Test) (112) measure performs sequence-independent superposition of the model and the native structure and calculates the number of structurally equivalent pairs of C-alpha atoms that are within specified distance *d*. The GDT TS score is the average of four scores obtained with $d = 1, 2, 4$ and 8 Å divided by the number of residues of the target. Despite being slow, GDT TS is the standard measure used in CASP, but it is not part of LiveBench evaluation.

LG-score (113) superimposes the model with the native structure to maximize the Levitt–Gerstein score (114), as in MaxSub (below). The final score is translated into a *P*-value, which estimates the chance of obtaining this score given the length of the model. LG-score can operate in sequence-dependent and sequence-independent modes. The second is much slower. Because of limited computational resources, it has been removed from standard LiveBench evaluations.

Mammoth (115) computes the optimal similarity of the local backbone chains to establish residue correspondences between residues in both structures in the first step. In the second step, the largest subset of residues found within a given distance threshold is calculated with MaxSub (below). This sequence-independent structural similarity is translated into *P*-values.

MaxSub (116) identifies the largest subset of C-alpha atoms of a model that superimpose well (below 3.5 Å) over the experimental structure. MaxSub calculates a variant of the Levitt–Gerstein score (114), which equals to $\sum \{1/[1 + (d/3.5\ \text{Å})^2]\}$, summed over all superimposed pairs of C-alpha atoms and divides it by the number of residues in the target. MaxSub is the official CAFASP evaluation method.

3D-score (32) optimizes the sum of $\exp[-\ln(2)*(d/3\ \text{Å})^2]$, where *d* is the distance between the superimposed C-alpha atoms. This sum behaves very similar to the score used in MaxSub or LGscore, but it has no cutoff value and it decays faster with higher distance. The final score is not divided by the length of the target.

CA-atoms<3 Å (32) returns the maximum number of atoms within 3 Å after superposition generated by optimization of the 3D-score. This very simple measure shows good performance in distinguishing biologically relevant predictions and is very intuitive and easy to understand.

Q(CA-atoms<3 Å) is aimed at evaluating the specificity of the alignment and penalizes wrong sections of the models. It is equal to the square of (CA-atoms<3 Å) divided by the number of residues in the model. This is the only measure used in LiveBench, which penalizes overpredictions (too long alignments). Servers that return coordinates always for all residues of the target perform worse than if evaluated with other measures.

Contact(A&B) (32) calculate the distance map overlap between the model and the native structure. The calculation is performed in sequence dependent manner and no rigid body superposition is required. Two ways to normalize the overlap are used resulting in two scores Contact(A) and Contact(B). These two are the only contact measures used in LiveBench.

Methods performing sequence-independent superposition (first three) are relatively slow and are not used in current LiveBench experiments. Only one measure [Q(CA-atoms<3 Å)] penalizes for wrong parts of models. All methods, except the contact measure [Contact (A&B)], conduct rigid body superposition. The contact measure can handle the evaluation of multiple domains. GDT TS and MaxSub divide the score by the size of the target. Mammoth and LG-score estimate the probability of non-random structural similarity expressed as *E*-value. The scores of the others are proportional to the size of the model.

character of targets, LiveBench is a good approximation of accuracy of prediction methods; it requires much less maintenance and can operate on a much larger set of targets than CASP or CAFASP (both use the same targets). In addition, a permission from the authors to use each target structure is not required as it is in CASP/CAFASP. The number of targets released each week is limited mostly by the throughput of participating prediction servers or by the number of available new structures without close homologs of known structure if only predictions for non-trivial targets are evaluated. Currently, approximately five new non-trivial targets enter the LiveBench process each week.

Assessment of 3D models

The concept of using different methods in fold recognition to confirm the predictions can be applied also to the evaluation process. There is a large set of available programs that can be used to compare a model with the native structure. Table 2 provides a short description of frequently used model assessment methods. Most of the methods use rigid body superposition algorithms to find the best structural alignment, dependent or independent of the assignment of residue identities in the model. The so-called sequence-independent methods ignore the identities of the residues in the model, thus ignoring possible alignment errors and focusing only on the general shape of the model. This is somewhat equivalent to verifying if the target and the template used to build the model have a similar architecture, topology and fold. Sequence-independent methods are computationally intensive. It is much faster to require correct alignment and to evaluate the spatial proximity of equivalent residues in the model and native structure in sequence-dependent superposition as most methods do.

Alternative to rigid body superposition, local contacts or distances between corresponding residue pairs in two structures can be compared. This procedure is also very fast if conducted in sequence-dependent fashion. Sequence-independent procedures require finding the best superposition of contact or distance maps, which is very time-consuming. Such methods have not found application in the evaluation of fold recognition results. In contrast, sequence-dependent contact or distance scores are used and offer the advantage of much higher tolerance to relative movements between domains in multi-domain models. If there are differences in relative domain placement between the model and the structure, rigid body superposition methods can properly evaluate only one domain. Thus, the models are frequently divided into regions corresponding to domains and assessed independently. This is the case in CASP where the evaluation is conduced under strong human supervision. Fully automated evaluation programs, such as LiveBench, would require robust automated domain detection methods. Additionally, most prediction servers return as the first model the structure of the domain, which is easiest to predict, unless a suitable multi-domain template is found. For these two reasons, division of targets into domains is not used in LiveBench.

The differences between model assessment methods contribute to variations in the ranking of servers produced by different assessment procedures. Obviously, all model evaluation methods give higher scores to models that are closer to the native structure. However, different assessment methods use different criteria of closeness to the native structure. For example, a hypothetical model built using the native structure but with substantial alignment errors will be preferred by sequence-independent methods over a model with completely correct alignment built using a slightly distorted template. The opposite ranking of these two models will be given by sequence-dependent methods.

The structure prediction community has failed so far to define a standard model assessment algorithm. The main reason for this is the lack of an exact definition of similarity between the native structures of two proteins. Different

structural classifications of proteins, such as SCOP (72), CATH (73) or FFSP (74), disagree in many cases when assessing a weak similarity between two native structures, which is at the level of similarity between models for difficult to predict targets and the correct structure. Evolutionary relations between proteins cannot be used to replace the structural classification as they often remain hypothetical. As a result of the ambiguity of structural classification, the annotation of a model as 'correct fold' or 'incorrect fold' remains rather arbitrary. As a consequence, a single ranking of prediction methods should be viewed with the necessary caution. Consulting various rankings is always recommended, as is the application of various prediction methods in structural annotation projects.

Evaluation protocols

There is a number of ways the performance of a prediction server on a set of test targets can be presented, even if using just one model assessment program. Prediction servers can be tuned to obtain the highest alignment score for selecting the best template or to generate the most accurate alignment ignoring the score. Both optimizations result in different sets of alignment parameters. Traditionally this has been addressed in most evaluation experiments by dividing the targets into several categories depending on the level of difficulty of finding the correct fold. CASP and CAFASP divide targets in 'homology modeling', 'fold recognition' and 'novel folds' subdividing the categories further in easy and difficult cases. LiveBench ignores the easy 'homology modeling' targets referring to them as 'trivial targets' and divides the remaining targets into 'easy' and 'hard'. The boundaries between categories are defined rather arbitrary. In LiveBench, fold assignment for trivial targets must be possible with BLAST and for easy targets with five iterations of PSI-BLAST, in both cases using the 0.001 *E*-value as confidence threshold. The relative performance of servers is not constant across different target difficulty categories. Because of this, moving the barriers of the categories has an effect on the ranking. The definition of boundaries based on PSI-BLAST has also the effect that methods, which are similar to PSI-BLAST, may have improved performance on easy targets and reduced performance on difficult targets.

The easiest way to compare prediction methods in each category is by counting the number of correct predictions that were generated. This requires to set cutoffs for correct and false models. This has been performed, although rather arbitrary, for various assessment methods. At a first glance, it seems pointless to count the number of correct hits in the trivial or easy target category because servers are expected to do much better than BLAST or PSI-BLAST and should obtain the maximum score there. Practice shows that this is not the case. In LiveBench, for some targets PSI-BLAST hits with *E*-value below 0.001 are false positives (do not find a correct template) and those targets contaminate the 'easy' category. The LiveBench community has decided to leave such targets in the set because correctness of a model depends on the model assessment program. Despite the problem of apparently misclassified targets, there is a variation between the number of correct models generated by different prediction methods. *Ab initio* methods can produce incorrect folds for any target, if information about the structure of homologous proteins is disregarded. Fold recognition methods can miss a trivial prediction if the template is missing in their fold libraries. This problem is easy to detect if the generated prediction turns out to be wrong. However, the less severe problem of not having the best possible template, which results in building the model using less suitable template, is almost undetectable owing to complete lack of synchronization between the fold libraries. As a consequence, the number of correct models generates a ranking of servers in the easy categories, which reflects mainly technical aspects of the implementation of the prediction servers. Unfortunately, the much finer evaluation, which takes into account the quality of each model, is strongly affected by these technical problems. This problem is not easy to solve for an independent server, because a robust routine to update the fold databases represents a rather unscientific (or at least biologically not attractive) and non-trivial task. On the other hand, this problem is negligible for meta predictors, which can assume that an easy hit will be missed only by a minority of servers and the consensus procedure will not be severely affected.

Finer procedures to assess the performance of servers usually add the scores obtained for each model as assigned by the chosen model assessment method. Depending on the procedure of assessing models, the evaluation can weight each target equally, or dependent on its length. Manual homology modeling requires an effort, which is proportional to the size of the protein, i.e. the number of loops or segments that have to be corrected increases with the size of the protein. Accordingly, most methods used in LiveBench score the quality of the model proportional to the number of correctly placed residues. In contrast to this, the MaxSub program used in CAFASP or the transformation of the score into a *Z*-score as performed often in CASP result in assigning a length-independent weight to each target. The main difference between CASP and both LiveBench and CAFASP is that in CASP all models contribute to the evaluation to some extent, independent of their correctness. In LiveBench and CAFASP models, which are assessed as false do not contribute to the sum. As a result, the ranking in the 'fold recognition' category in CASP (equivalent to the hard category in LiveBench) is affected stronger by the ability to create models *de novo* by *ab initio* methods, since for the many of these difficult targets fold recognition servers give wrong answers. Having an assessment score for a set of models allows the application of statistical methods, which can measure the significance of the difference between two servers. This is often conducted in CASP, despite the fact that only few participants return models for all targets. The experience of LiveBench shows that the completeness of the fold library and the ability to select correct templates has a much stronger effect on the performance than the fine differences in alignments and models. Because of this, a simple table is provided showing the number of times each method has generated predictions which were missed by other methods ('added value plots').

Other options to generate ranking of servers include: giving points only to the methods, which provided the best model for each target or running the evaluation using the best out of 5 or 10 top models instead of looking only at the first one. Each of the evaluation schemes can be used to answer a different question. Number of correct predictions reflects the completeness of the fold library in the easy target category and the

ability to detect the correct fold for more difficult targets. The alignment quality can be assessed using the sum of model scores, however, keeping in mind that servers, which can recognize the folds of more targets will have an advantage in this evaluation scheme as well. This is also true in the ranking based on the number of best models with the addition that in this category weaker performing servers can demonstrate their contribution to the community of servers, if obtaining a count higher than zero. The ranking using the best of the top models reflects the potential to improve the performance of a server by reshuffling the top scoring predictions using additional selection criteria.

However, one very important question remains unanswered, namely when can the user trust the model. With the exception of some servers, which rely heavily on *ab initio* components, most servers assign a confidence score to each model, with the goal to provide a way to distinguish probably correct models from likely wrong results. The quality of this score is irrelevant in CASP, which stopped to assess it, but is very important for biologists when conducting large-scale structural annotation projects. Only this score enables a user to define a set of targets, e.g. genes in a genome, for which a fold can be assigned using computational approaches. Because of this, some developers tune their methods to obtain highest correlation between the confidence score and the quality of returned models. This often results in a different choice of parameters than those obtained in optimization for the best alignment or the highest number of correct fold assignments. The evaluation of the specificity of the score is usually provided using the receiver operator characteristic curve (75). The curve represents a plot of correct predictions on the *y*-axis versus false predictions on the *x*-axis. The higher the area below the curve, the more correct predictions are obtained before making an error. In LiveBench, the specificity score is proportional to the area under the curve, but the curve is followed only until 10 errors are made. This corresponds to a reasonable cutoff below a 20% error rate. Models for all targets are taken into account in this calculation. The results of LiveBench-7 and LiveBench-8 are summarized in Tables 3 and 4.

Results of evaluation experiments

Years of experience with benchmarking prediction methods taught the community to treat the results with appropriate caution. The tests are affected by many technical problems, which distort the evaluation of the method performance. The most severe problem is missing predictions. In CASP, which evaluates groups of experts, it is quite common that some participants send only models for selected targets, e.g. when they feel confident in a prediction or have special expertise. Such groups have a chance to obtain a higher average model accuracy than groups or methods, which predicted all models, but the sum of scores will be probably higher in the later case. The time factor is also very important in CASP. Currently, CASP offers on average one target domain per working day in the prediction period of ~3 months. Models can be improved, if more time can be spent on their preparation, but the time limits are very strict and the groups have to run a well-planned schedule to obey the deadlines. CAFASP has allowed the servers only a 2 days delay between releasing the target and collecting the results. A server, which is down over this period, will not be able to respond in time. If the target is trivial, missing it means almost automatically that the server has lost the chance to rank number 1 in the most common classification based on the sum of scores. In LiveBench, these deadlines are flexible and servers are allowed to file late predictions or even replace predictions, if severe technical errors were found. However, this causes additional problems in the assessment of results. The later a prediction is computed, the easier on average is the fold recognition problem. This is owing to the constant growth of the number of determined protein structures resulting in larger fold libraries and better coverage of the sequence space with homologs with known structure. It is also owing to the growth of sequence databases providing more information about sequence variations in protein families. This improves the quality of profiles used in fold recognition. LiveBench tolerates these distortions to avoid the more severe problem of missing predictions.

The evaluation of the reliability of the confidence score assigned to models, which is very important to the users, has a few shortcomings as well. The problem of significant similarity in short segments of proteins is most profound. Such short segments can harbor important functional features and are biologically clearly related. Nevertheless, the small size of the segments makes it impossible for the model assessment methods to detect significant structural similarity and such models are deemed to be false. In LiveBench, models that are shorter than 50 residues are ignored. On the other hand, CASP defines sometimes targets that are below 50 residues. One of the assessment methods used in LiveBench requires that at least 40 residues are correctly positioned in space to assess the model as correct. For short targets, such models are sometimes difficult to obtain even if the prediction is highly confident. High scoring wrong predictions have obviously a dramatic effect on the evaluation of specificity. Such errors are partially owing to short motifs, but surprisingly may be owing to errors in the experimental protein structures used as the standard of truth. In LiveBench-4, several targets were removed after the predictors realized that confident consistent predictions produced by fold recognition servers are clearly different than the 'native' structures deposited in PDB. In one case, the target structure was affected (76), while in another case, the closest homolog in the template database (1fznD) was incompatible with the target (77). Authors of the protein structures confirmed afterwards the concerns and removed or replaced the structures in PDB. This problem is detectable, if the errors in protein structures are substantial. Relative minor errors, such as tracing shifts in small segments of proteins, are more frequent (78,79), but essentially undetectable with current methods. This is one of the reasons why details of models, such as side-chain placement, are neglected in LiveBench, despite clear biochemical relevance.

Because of many various factors providing distortion to the ranking of methods, a single ranking produced in one experiment is not very meaningful. The size of the test set should be increased from the current number of ~100 targets evaluated in each round to at least 1000, but this is currently not feasible. This would require a very long data collection period and would surpass the upgrade and improvement cycles of methods, which according to our experience takes ~6 months, the period of a LiveBench round. Despite these difficulties, practical conclusions for automated and manual structure

Table 3. Comparison of selected servers participating in LiveBench-8

LiveBench-7 (115 targets)									LiveBench-8 (172 targets)							
Code	Sum	FR	Code	ROC%	All	Score	3	Lost	Code	Sum	FR	Code	ROC%	All	Score	3	Lost
3JCa	2300	32	3DS5	55.7	70	5.658	62	0	3JCa	2920	39	3JA1	59.6	110	45.33	100	0
3DS5	2133	29	PMOD	54.3	69	1.511	64	2	3DS3	2757	38	3JCa	58.8	107	16.76	100	0
3JC1	2042	30	FFA3	54.2	66	−9.3	62	0	3JA1	2749	42	3JC1	57.9	108	63.45	100	0
3JAa	1927	30	PCO2	53.8	66	1.12	62	0	3JC1	2720	39	3DS3	56.0	106	51.81	93	0
3DS3	1908	27	3DS3	53.7	67	31.41	62	0	PMO4	2639	38	BasD	53.5	101	14.11	91	0
PMO3	1830	27	3JA1	53.2	67	40.56	61	0	3JAa	2622	37	ORFs	53.3	100	7.55	93	1
PMOD	1793	28	3JC1	52.5	71	62.62	61	0	PCO5	2585	37	PCO5	53.1	104	1.688	91	3
3JA1	1786	27	PMO3	52.2	66	1.76	60	1	3DS5	2531	36	PMO4	53.0	106	1.847	90	2
SHGU	1708	24	PMO4	51.9	66	1.546	60	2	RBTA	2407	33	3JAa	52.7	105	15.13	88	0
PCO2	1688	26	PCO4	51.1	64	1.173	60	2	PCO4	2336	35	mBAS	52.4	102	16.38	91	3
FFA3	1687	26	FUG3	49.9	60	4.8	56	0	mBAS	2306	34	PCO4	51.7	103	1.489	87	2
PMO4	1679	26	PCO3	49.7	64	1.858	53	1	BasD	2261	33	BasP	50.9	98	13.36	86	0
PCO3	1655	26	FUG2	49.7	61	4.54	58	0	SHGU	2241	31	ORF2	50.6	98	27.47	87	2
PCO4	1614	26	3JCa	49.0	73	27.54	54	0	ORFs	2196	32	SFST	49.5	97	6E-06	83	5
RBTA	1597	24	SHGU	48.8	62	38.33	52	0	BasP	2157	31	FUG3	48.8	93	6.59	84	0
RAPT	1496	24	3JAa	48.2	70	12.32	55	0	ORF2	2124	31	SHGU	48.8	98	25.19	85	0
ORFs	1455	23	3DPS	46.8	64	0.242	53	0	FFA3	2118	31	FFA3	48.6	98	−14.8	83	0
3DPS	1426	23	PRO2	46.4	59	4.545	49	2	SFST	1943	29	STMP	47.7	96	1E-05	79	5
FUG2	1415	22	INBG	44.9	59	21.1	51	0	STMP	1860	28	FUG2	47.6	90	6.26	81	0
INBG	1397	22	RAPT	44.3	61	8.08	49	0	INBG	1792	25	3DS5	47.4	103	7.242	65	1
FUG3	1332	22	ORFs	44.3	60	9.3	48	0	FUG3	1756	27	INBG	45.6	91	20.6	75	0
PRO2	1318	20	MGTH	43.2	57	0.564	53	0	MGTH	1743	26	SPKS	45.6	88	−2.75	79	0
MGTH	1307	22	SFAM	41.7	51	0.009	48	0	FUG2	1703	25	PRO2	43.5	86	4.091	72	3
FRT1	1121	19	ST99	41.0	55	13.09	52	6	PRO2	1700	26	SFAM	43.5	84	5E-11	64	0
SFPP	1107	17	FRT1	40.3	56	12.07	44	0	3DPS	1685	25	RAPT	42.2	81	6.66	74	0
SFAM	1067	16	SFPP	39.9	52	1E-12	44	0	SPKS	1646	23	ST99	41.8	82	22.73	66	16
ST99	1020	17	PDBb	36.3	46	0.009	42	1	RAPT	1573	24	SFPP	41.3	85	3E-24	67	0
GETH	980	16	GETH	34.6	48	0.596	38	0	SFPP	1534	23	PDBb	38.3	73	2E-04	64	8
FFAS	414	7	FFAS	23.3	34	7.63	27	0	FRT1	1294	20	MGTH	37.8	91	0.678	51	0
PDBb	344	6							SFAM	1163	18	3DPS	36.4	91	0.06	62	0
									GETH	1140	17	GETH	31.9	78	0.619	50	0
									ST99	1081	18	FRT1	30.1	82	22.72	30	0
									PDBb	386	7	FFAS	16.5	49	12.17	22	0
									FFAS	130	3						

Only publicly available servers that provide a description of the underlying algorithm are listed. Results obtained in LiveBench-7 are also displayed if available. Results for the PROTINFO-CM server are not presented because of late predictions. Servers are colored blue (sequence only methods), red (hybrid methods) and black (structure meta predictors). The 'Code' column shows the code of the method as provided in Table 1. Results obtained using the 3D-score assessment measure are shown (see Table 2). The 'Sum' column prints the sum of scores obtained for correct models of difficult targets (no PSI-BLAST assignment with *E*-value below 0.001). The 'FR' column shows the number of correct models generated for difficult targets. The 'All' column shows the number of correct models generated for all targets including the easy ones. The 'ROC' (Receiver Operator Characteristic) value describes the specificity of the confidence scores reported by the methods. It corresponds to the average number of correct models that have a higher confidence score than the first, second ... tenth false prediction. The 'ROC%' column prints the 'ROC' value divided by the total number of targets and multiplied by 100. Robetta is not listed here since it does not provide confidence scores. The 'Score' column reports the score of the third false positive prediction and the '3' column presents the number of correct predictions with higher score than the third false one. The score can be used as an approximate value for the confidence threshold, below which false positive predictions become frequent. The 'Lost' column shows the number of missing predictions for each server. Servers that have more than a few missing predictions cannot be properly evaluated. Some servers that entered LiveBench in the eighth round have missing scores in LiveBench-7. The new sequence-only methods exhibit high specificity and can compete with structure meta predictors in this ranking. The structure meta predictors rank much higher in the sensitivity (FR) and model quality (Sum) based evaluation. In LiveBench-8 the meta–meta predictors, such as 3JC1 and 3JCa, which use results of other meta predictors, profit greatly from its parasitic nature. Servers, which are not maintained over long period of time become obsolete due to outdated fold libraries. As an example, FFAS now seems to perform similarly to PDBb, while it used to perform much better in first rounds of LiveBench. High-quality servers are able to generate ∼50% more correct models than PDBb ('All' column).

prediction projects can be drown from the experiments and include:

(1) Individual threading methods and hybrid methods, utilizing structural information in the scoring function, are probably not more accurate than well-tuned sequence profile comparison methods. In fact, the latter methods seem to lead the sensitivity and specificity rankings of the latest LiveBench rounds. This has to be confirmed in blind tests in future CAFASP experiments.

(2) Meta predictors are clearly superior to simple individual methods. For quite some time, meta predictors are heading the rankings in sensitivity and specificity. These results have also been obtained in last CASP-5 and CAFASP-3 rounds. In sensitivity, meta predictors using structural information of the models for building the consensus predictions are leading the field. The top ranks in the specificity evaluation are also occupied by meta predictors disregarding any 3D structural information, such as Meta-BASIC. Meta-BASIC uses predicted secondary structure to compliment sequence profiles. We conclude that structural meta predictors build superior models, while the calculated score of the models is not yet estimated that well.

(3) In contrast to common sequence alignment methods, structure prediction servers are generally not prepared to deal with multi-domain targets. Division of a sequence into

Table 4. Comparison of rankings obtained using different evaluation measures

LiveBench-7 (115 targets)								LiveBench-8 (172 targets)							
Code	3D	MS	CA	Q	CA	CB	Avg	Code	3D	MS	CA	Q	CA	CB	Avg
3JC1	3	3	3	2	2	2	2.5	3JC1	4	1	3	3	1	1	2.2
3JCa	1	2	1	1	1	9	2.5	3JCa	1	2	1	2	3	8	2.8
3JAa	4	8	4	4	4	5	4.8	3JA1	3	5	4	7	4	2	4.2
PMO3	6	4	7	7	3	4	5.2	PMO4	5	6	5	5	2	5	4.7
3DS5	2	1	2	3	8	17	5.5	PCO5	7	7	6	1	5	3	4.8
3JA1	8	7	6	6	5	3	5.8	3JAa	6	10	7	4	6	4	6.2
FFA3	11	12	8	5	10	6	8.7	PCO4	10	9	9	8	7	7	8.3
PMO4	12	11	12	8	6	7	9.3	3DS3	2	3	2	11	12	22	8.7
3DS3	5	5	5	10	13	19	9.5	3DS5	8	4	8	17	8	15	10.0
PMOD	7	10	9	13	9	10	9.7	RBTA	9	8	10	19	9	6	10.2
RBTA	15	6	14	23	7	1	11.0	mBAS	11	12	12	10	11	11	11.2
PCO2	10	14	13	9	12	12	11.7	BasD	12	11	11	12	13	9	11.3
PCO3	13	9	10	14	17	15	13.0	ORFs	14	19	13	13	10	10	13.2
PCO4	14	17	15	12	11	13	13.7	BasP	15	13	16	14	14	14	14.3
ORFs	17	16	17	11	15	8	14.0	FFA3	17	17	15	16	15	13	15.5
SHGU	9	13	11	15	16	22	14.3	ORF2	16	18	17	15	16	12	15.7
RAPT	16	15	16	16	14	11	14.7	SFST	18	15	18	6	18	19	15.7
INBG	20	19	18	17	18	16	18.0	STMP	19	16	19	9	20	21	17.3
3DPS	18	21	20	24	19	14	19.3	SHGU	13	14	14	20	17	27	17.5
FUG2	19	20	19	20	22	20	20.0	INBG	20	24	20	18	19	18	19.8
PRO2	22	22	21	22	20	18	20.8	PRO2	24	21	23	24	22	17	21.8
FUG3	21	18	22	25	21	21	21.3	FUG3	21	22	21	21	28	26	23.2
MGTH	23	24	23	26	24	23	23.8	MGTH	22	23	22	25	24	23	23.2
SFAM	26	27	26	19	23	25	24.3	3DPS	25	25	25	26	23	20	24.0
FRT1	24	23	24	27	27	24	24.8	RAPT	27	27	28	28	21	16	24.5
SFPP	25	26	25	21	25	27	24.8	FUG2	23	26	24	23	26	28	25.0
ST99	27	25	27	18	26	26	24.8	SPKS	26	20	26	29	25	24	25.0
GETH	28	28	28	28	28	28	28.0	SFPP	28	28	27	22	27	25	26.2
PDBb	30	29	29	29	30	29	29.3	FRT1	29	29	29	31	29	29	29.3
FFAS	29	30	30	30	29	30	29.7	SFAM	30	32	31	30	30	30	30.5
								ST99	32	30	32	27	31	31	30.5
								GETH	31	31	30	32	32	32	31.3
								PDBb	33	33	33	33	33	33	33.0
								FFAS	34	34	34	34	34	34	34.0

Only publicly available servers participating in LiveBench-8 that provide a description of the underlying algorithm are listed. Results obtained in LiveBench-7 are also displayed if available. Servers are colored blue (sequence only methods), red (hybrid methods) and black (structure meta predictors). The 'Code' column shows the code of the method as provided in Table 1. Rankings of servers obtained using five different assessment measures (see Table 2): 3D-score, MaxSub, CA-atoms<3 Å, Q(CA-atoms<3 Å), Contact(A) and Contact(B) are shown in columns '3D', 'MS', 'CA', 'Q', 'CA' and 'CB', respectively. The 'Avg' column prints the average ranking of the server.

domains and iterative submission of corresponding domain sequences to prediction servers is strongly advised. This is especially important for eukaryotic proteins, since many of them contain several structured domains and additionally possess a few disordered regions. The disordered regions in proteins can be predicted with specially tailored methods, such as GlobPlot (80) or PONDR (81). Despite high biological importance of unstructured regions (82), attempts to predict their structure as globular are destined to fail. Disordered regions are abundant in eukaryotic proteins making clear that not all protein segments have to fold into a single low-energy native conformation in order to perform a function.

(4) The score for hits reported by some meta predictors (e.g. 3D-Jury) is sometimes artificially increased, if one of the component servers generate many alignments to very similar proteins. Thus, a high score is only significant, if several independent servers confirm the fold assignment.

(5) For the majority of difficult cases, such as the targets in the 'fold recognition' category in CASP, the confidence scores reported by the servers are below the reliability threshold and the correct models are not always the top ranking ones. Expert users are sometimes able to select the correct predictions using additional knowledge, such as the similarity of function between the target and the template or conservation of essential amino acids or short sequence patterns. In many cases, the experts have to conduct extensive literature analysis and sequence searches to guess the right fold. This time-consuming but frequently very fruitful exercise is advised to all predictors, provided that there is enough time to analyze the target of interest.

(6) In very difficult cases, results of *ab initio* methods or servers using *ab initio* components can be consulted. Difficult targets can be distinguished from others based on low scores of meta predictions. In general, such models have very low quality independent of their source, but some biological hints can be gained. Nevertheless, *ab initio* methods improved significantly over the past years. The ultimate goal of the community is to be able to predict the structure of the protein independent of structural information available for its homologs. Significant efforts are devoted to this goal. A solution to the folding

problem would obviously have an immense impact on structural biology. Thus, one should keep an eye on the progress in this field.

(7) Models can be improved manually by experts, as shown in the CASP experiments. Detailed analysis of the target and the template families, including extensive literature searches, is mandatory in such cases. This helps to identify functionally crucial residues often misaligned by servers, if both families exhibit very week similarity. A model can be improved significantly, if the expert detects a substantial error in the alignment resulting, for example, from the insertion of an entire domain. In many cases, expert improvements, however, remain marginal.

(8) It is possible to estimate, which parts of the models are likely to be correct and which parts are more or less random. The most reliable approach (consensus approach) is by evaluating the set of alignments obtained from different servers and searching for structurally well-conserved regions, i.e. where the alignments by different servers are consistently the same or very similar. Regions where different methods report different alignment to similar templates are more likely to be misaligned, or structurally diverged. The quality of models can also be evaluated with programs, such as Verify3D (83). Unfortunately, the quality of difficult fold recognition models is below the standards of the benchmarks used to tune the majority of quality assessment methods, making application of such methods problematic in difficult cases.

(9) Most online protein structure meta predictors are too slow to be used in high-throughput annotation projects. For such purpose, it is better to construct in-house meta predictors using several simple, but diverse and independent components. A simple function, which tells by how many components the prediction was confirmed, can be used as reliability score. This is a general suggestion not only for structural annotation but also for sequence homology annotation, which is routinely conducted with only one method, PSI-BLAST, e.g. when annotating genomes. An alternative to designing an in-house meta predictor is to use fast online meta predictors, such as Meta-BASIC, which do not utilize 3D structural information.

UTILITY OF PROTEIN STRUCTURE PREDICTIONS

Drug design is the one of the major financial driving forces behind biomedical research. Unfortunately, the protein structure prediction field is currently unsuccessful in keeping its promise of making the drug development process much more efficient. Predicted protein structures can be used if very close homologs with known structure are available, but in most cases rational drug design requires iterative co-crystallization of the protein–ligand complexes. In the majority of cases, predicted models are of insufficient quality to offer the atomic details necessary for lead optimization. Currently available structure prediction methods do not allow for high-quality predictions of the quaternary structure of protein complexes and for the prediction of interactions between proteins. Current benchmarks indicate that methods predicting interactions can be successful mainly in cases when structures exhibit minimal conformation changes upon complex formation. Substantial errors observed in predicted models go beyond the limits tolerated by such methods (84).

Nevertheless, low-resolution models obtained using structure predictions methods can find other applications. It is much easier to handle a protein in experiments if at least a crude model of its structure is available. An example from our work is the beta-ketoacyl synthase domain of mycocerosic acid synthase from *Mycobacterium bovis* BCG. According to Fernandes and Kolattukudy (85), the domain spanning amino acids 1–341 exhibits selectivity for methylmalonyl-CoA over malonyl-CoA. However, results of structural modeling of the domain show that this construct lacks large section of the core and parts of the catalytic pocket, which should be essential for the protein to perform its function. According to the model, the correct boundaries of the domain span residues 1–437. Surprisingly, experimental results revealed that both versions of the domain provide the ability to incorporate methylmalonyl-CoA into fatty acids *in vivo*. Nevertheless, the incomplete domain is probably not folded quite right or distorted and has biochemical properties, which make a crystallization of the protein impossible. A simple modeling step enabled to find proper domain boundaries and resulted in a better-characterized protein for future experimental work.

Another application of low-resolution models can be the analysis of active sites in proteins and detection of compensating mutations. As an example, models of a PD-(D/E)XK-like domain enabled us to reveal a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs (86). These proteins have an essential glutamate three residues away from its equivalent position in other related enzymes. Despite the shift in sequence, the main-chain alpha carbon atom is only one α-helical turn away and the side-chain of glutamate points to the same direction as in the canonical configuration. Such active site residue compensatory replacements may be difficult to detect if only a sequence multiple alignment is available, which can only suggest a lack of conservation of crucial catalytic residues. Structural analysis can help to rationalize the mutations and confirm the general enzymatic classification of the protein.

Low-resolution models can be used to guess the function through a sequence-to-structure-to-function paradigm. Significant success of this approach has been reported by the group of David Baker (87). The *ab initio* method Rosetta was used to predict the fold of 130 proteins. For each family, five models were generated and each model was classified by structural comparison with a fold library. In 35% of the cases, one of the five models could be used successfully to identify the correct fold (SCOP superfamily) of the native protein. This procedure has been applied to annotate 510 protein families collected from Pfam database. Only families with no links to proteins with known structure and average length of the protein below 150 residues were selected. Many interesting assignments have been generated, of which some were confirmed by functional similarity between the target family and the families with the same fold as the predicted one. This procedure is unable to assign novel function to a protein family, but the ability to guess the general function in a sizable fraction of the targets can be regarded as a success. Such functional hints may be valuable for future biochemical characterization. This example also shows that fold recognition results can be validated using independent observations, such as the similarity

of function. Fold predictions can also support function predictions obtained with non-homology-based methods, such as gene fusion (88,89), gene neighborhood (90,91) or phylogenetic profiles (92,93).

A much simpler approach to follow the sequence-to-structure-to-function paradigm is by confirming weak sequence similarity by fold recognition. In many cases, homologous proteins divert beyond the level of detectable sequence similarity while keeping the general fold unchanged. Methods decoding and amplifying the structural preferences buried in sequence can help us to bridge diverged families. Threading was designed to solve this problem but is now increasingly replaced by hybrid or profile–profile comparison methods. Benchmarks indicate that fold recognition methods can generate ~50% more assignments than PSI-BLAST in cases of non-trivial sequence similarity, undetectable with simple BLAST. Confirmation of weak sequence similarity by fold recognition methods is probably the most common application of structure prediction servers, conducted sometimes in genome scale or routinely during screening of structural genomics targets. On the other hand, confirmation of very weak similarities with below threshold fold assignments can be attempted with *ab initio* methods, such as Rosetta as described previously (94).

Low-resolution models can also be used to aid the experimental structure determination process. The previously mentioned example of detection of errors in the PDB by the LiveBench program demonstrates the power of fold recognition methods in this respect. Crystal structures obtained with a resolution of ~3 Å are susceptible to tracing shifts that misassign some side-chains along the backbone. Such errors can be detected if a sufficient evolutionary signal exists and reliable alignment to other more confident structures can be used as evidence supporting the hypothesis of a shift. Structure factors collected during the X-ray experiment can then be used to validate the proposed improved model. A related application represents the determination of protein structures using molecular replacement. In this technique, an initial model, which shows sufficient structural similarity to the native structure, is used to solve the phasing problem. With increasing quality of the models generated by fold recognition methods, more distant homologs can be used to produce suitable initial models (95).

An alternative approach to protein structure determination is represented by NMR experiments, generating a large number of distance restrains used later to build the model of the native protein. In some cases, the information obtained from NMR experiments is too scant to enable direct reconstruction of the structure. Structure prediction methods have been used successfully in combination with sparse restrains obtained from nuclear Overhauser effects, residual dipolar couplings or backbone chemical shifts (96–98). Limited experimental information can greatly reduce the fold space, which is explored by structure prediction program. This has a positive effect on the speed of the calculation and on the final accuracy of the model. In most of the reported cases, the combined methods were able to deliver low-resolution models with essentially correct folds, although some errors were observed for proteins with internal symmetries. However, increasing the number of restrains also resulted in increased quality of the model. Sparse NMR data are relatively easy to obtain and the appearance of public repositories of NMR parameters, such as BioMagResBank (99), is likely to boost the development of methods, which use limited experimental information.

CONCLUSIONS

The goal of structural genomics initiatives is to provide template structures for most protein families. Structure prediction approaches are destined to become limited to comparative modeling, since a close homolog of known structure would be available for most targets. However, the limited success of the structure genomics programs contributes to the booming interest in structure prediction methods as measured by the growing number of servers and groups taking part in community-wide prediction quality assessment experiments. It is clear that prediction methods are not expected to replace experimental determination of protein structures in the nearest future, but are likely to complement such efforts and meanwhile fill the growing gap between the number of sequences and structures. Confident fold prediction models can be viewed as low-resolution structures and will be used by biologists to guide experimental design. The growing number of user-friendly prediction servers will hopefully result in increased awareness of the benefits of a low-resolution 3D protein model for biologists.

ACKNOWLEDGEMENTS

The authors would like to express their cordial thanks to the structure prediction community for supporting the evaluation effort and inspiring the progress in the prediction field. This work has been supported by MNI and the European Commission with grants to L.R. (QLRT-2001-02884, QLRT-2000-00349, LSHG-CT-2003-503265 and LSHG-CT-2004-503567). The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al*. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
2. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
3. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al*. (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
4. Kim,S.H. (1998) Shining a light on structural genomics. *Nature Struct. Biol*., **5** (Suppl.), 643–645.
5. Shapiro,L. and Harris,T. (2000) Finding function through structural genomics. *Curr. Opin. Biotechnol*., **11**, 31–35.
6. Baker,D. and Sali,A. (2001) Protein structure prediction and structural genomics. *Science*, **294**, 93–96.
7. Hood,L. and Galas,D. (2003) The digital code of DNA. *Nature*, **421**, 444–448.
8. Brenner,S.E. (1999) Errors in genome annotation. *Trends Genet*., **15**, 132–133.
9. Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet*., **17**, 429–431.

10. Fetrow,J.S., Godzik,A. and Skolnick,J. (1998) Functional analysis of the *Escherichia coli* genome using the sequence-to-structure-to-function paradigm: identification of proteins exhibiting the glutaredoxin/ thioredoxin disulfide oxidoreductase activity. *J. Mol. Biol.*, **282**, 703–711.
11. Levitt,M. and Warshel,A. (1975) Computer simulation of protein folding. *Nature*, **253**, 694–698.
12. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
13. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
14. Iliopoulos,I., Tsoka,S., Andrade,M.A., Janssen,P., Audit,B., Tramontano,A., Valencia,A., Leroy,C., Sander,C. and Ouzounis,C.A. (2001) Genome sequences and great expectations. *Genome Biol.*, **2** interactions 0001.1–0001.3.
15. Brenner,S.E., Chothia,C. and Hubbard,T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
16. Martin,A.C., Orengo,C.A., Hutchinson,E.G., Jones,S., Karmirantzou,M., Laskowski,R.A., Mitchell,J.B., Taroni,C. and Thornton,J.M. (1998) Protein folds and functions. *Structure*, **6**, 875–884.
17. Narayana,S.V. and Argos,P. (1984) Residue contacts in protein structures and implications for protein folding. *Int. J. Pept. Protein Res.*, **24**, 25–39.
18. Bowie,J.U., Luthy,R. and Eisenberg,D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.
19. Blundell,T.L. (1991) Comparative analysis of protein three-dimensional structures and an approach to the inverse folding problem. *Ciba Found. Symp.*, **161**, 28–36; discussion 37–51.
20. Godzik,A., Kolinski,A. and Skolnick,J. (1992) Topology fingerprint approach to the inverse protein folding problem. *J. Mol. Biol.*, **227**, 227–238.
21. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
22. Moult,J., Pedersen,J.T., Judson,R. and Fidelis,K. (1995) A large-scale experiment to assess protein structure prediction methods. *Proteins*, **23**, ii–v.
23. Moult,J., Hubbard,T., Bryant,S.H., Fidelis,K. and Pedersen,J.T. (1997) Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins*, **29** (Suppl. 1), 2–6.
24. Moult,J., Hubbard,T., Fidelis,K. and Pedersen,J.T. (1999) Critical assessment of methods of protein structure prediction (CASP): round III. *Proteins*, **37** (Suppl. 3), 2–6.
25. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2001) Critical assessment of methods of protein structure prediction (CASP): round IV. *Proteins*, **45** (Suppl. 5), 2–7.
26. Moult,J., Fidelis,K., Zemla,A. and Hubbard,T. (2003) Critical assessment of methods of protein structure prediction (CASP)-round V. *Proteins*, **53** (Suppl. 6), 334–339.
27. Fischer,D., Barret,C., Bryson,K., Elofsson,A., Godzik,A., Jones,D., Karplus,K.J., Kelley,L.A., MacCallum,R.M., Pawowski,K. *et al.* (1999) CAFASP-1: critical assessment of fully automated structure prediction methods. *Proteins*, **37** (Suppl. 3), 209–217.
28. Fischer,D., Elofsson,A., Rychlewski,L., Pazos,F., Valencia,A., Rost,B., Ortiz,A.R. and Dunbrack,R.L.,Jr (2001) CAFASP2: the second critical assessment of fully automated structure prediction methods. *Proteins*, **45** (Suppl. 5), 171–183.
29. Fischer,D., Rychlewski,L., Dunbrack,R.L.,Jr, Ortiz,A.R. and Elofsson,A. (2003) CAFASP3: the third critical assessment of fully automated structure prediction methods. *Proteins*, **53** (Suppl. 6), 503–516.
30. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-1: continuous benchmarking of protein structure prediction servers. *Protein Sci.*, **10**, 352–361.
31. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) LiveBench-2: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **45** (Suppl. 5), 184–191.
32. Rychlewski,L., Fischer,D. and Elofsson,A. (2003) LiveBench-6: large-scale automated evaluation of protein structure prediction servers. *Proteins*, **53** (Suppl. 6), 542–547.
33. Rost,B. and Eyrich,V.A. (2001) EVA: large-scale analysis of secondary structure prediction. *Proteins*, **45** (Suppl. 5), 192–199.
34. Koh,I.Y., Eyrich,V.A., Marti-Renom,M.A., Przybylski,D., Madhusudhan,M.S., Eswar,N., Grana,O., Pazos,F., Valencia,A., Sali,A. *et al.* (2003) EVA: evaluation of protein structure prediction servers. *Nucleic Acids Res.*, **31**, 3311–3315.
35. Rychlewski,L., Zhang,B. and Godzik,A. (1998) Fold and function predictions for *Mycoplasma genitalium* proteins. *Fold Des.*, **3**, 229–238.
36. Bujnicki,J.M., Elofsson,A., Fischer,D. and Rychlewski,L. (2001) Structure prediction meta server. *Bioinformatics*, **17**, 750–751.
37. Lundstrom,J., Rychlewski,L., Bujnicki,J. and Elofsson,A. (2001) Pcons: a neural-network-based consensus predictor that improves fold recognition. *Protein Sci.*, **10**, 2354–2362.
38. Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
39. Wang,Y., Bryant,S., Tatusov,R. and Tatusova,T. (2000) Links from genome proteins to known 3-D structures. *Genome Res.*, **10**, 1643–1647.
40. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
41. Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Methods Enzymol.*, **266**, 162–184.
42. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
43. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
44. Karplus,K., Barrett,C., Cline,M., Diekhans,M., Grate,L. and Hughey,R. (1999) Predicting protein structure using only sequence information. *Proteins*, **37** (Suppl. 3), 121–125.
45. Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.
46. von Ohsen,N., Sommer,I. and Zimmer,R. (2003) Profile–profile alignment: a powerful tool for protein structure prediction. *Pac. Symp. Biocomput.*, 252–263.
47. Sadreyev,R. and Grishin,N. (2003) COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. *J. Mol. Biol.*, **326**, 317–336.
48. Panchenko,A.R. (2003) Finding weak similarities between proteins by sequence profile comparison. *Nucleic Acids Res.*, **31**, 683–689.
49. Yona,G. and Levitt,M. (2002) Within the twilight zone: a sensitive profile–profile comparison tool based on information theory. *J. Mol. Biol.*, **315**, 1257–1275.
50. Heger,A. and Holm,L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.
51. Ginalski,K., Pas,J., Wyrwicz,L.S., von Grotthuss,M., Bujnicki,J.M. and Rychlewski,L. (2003) ORFeus: detection of distant homology using sequence profiles and predicted secondary structure. *Nucleic Acids Res.*, **31**, 3804–3807.
52. Bryant,S.H. and Lawrence,C.E. (1993) An empirical energy function for threading protein sequence through the folding motif. *Proteins*, **16**, 92–112.
53. Panchenko,A., Marchler-Bauer,A. and Bryant,S.H. (1999) Threading with explicit models for evolutionary conservation of structure and sequence. *Proteins*, **37** (Suppl. 3), 133–140.
54. Anfinsen,C.B., Haber,E., Sela,M. and White,F.H.,Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc. Natl Acad. Sci. USA*, **47**, 1309–1314.
55. Ortiz,A.R., Kolinski,A., Rotkiewicz,P., Ilkowski,B. and Skolnick,J. (1999) *Ab initio* folding of proteins using restraints derived from evolutionary information. *Proteins*, **37** (Suppl. 3), 177–185.
56. Skolnick,J., Zhang,Y., Arakaki,A.K., Kolinski,A., Boniecki,M., Szilagyi,A. and Kihara,D. (2003) TOUCHSTONE: a unified approach to protein structure prediction. *Proteins*, **53** (Suppl. 6), 469–479.
57. Simons,K.T., Bonneau,R., Ruczinski,I. and Baker,D. (1999) *Ab initio* protein structure prediction of CASP III targets using ROSETTA. *Proteins*, **37** (Suppl. 3), 171–176.
58. Bradley,P., Chivian,D., Meiler,J., Misura,K.M., Rohl,C.A., Schief,W.R., Wedemeyer,W.J., Schueler-Furman,O., Murphy,P., Schonbrun,J. *et al.* (2003) Rosetta predictions in CASP5: successes, failures, and prospects for complete automation. *Proteins*, **53** (Suppl. 6), 457–468.

59. Bonneau,R., Tsai,J., Ruczinski,I. and Baker,D. (2001) Functional inferences from blind *ab initio* protein structure predictions. *J. Struct. Biol.*, **134**, 186–190.
60. Kuhlman,B., Dantas,G., Ireton,G.C., Varani,G., Stoddard,B.L. and Baker,D. (2003) Design of a novel globular protein fold with atomic-level accuracy. *Science*, **302**, 1364–1368.
61. Dantas,G., Kuhlman,B., Callender,D., Wong,M. and Baker,D. (2003) A large scale test of computational protein design: folding and stability of nine completely redesigned globular proteins. *J. Mol. Biol.*, **332**, 449–460.
62. Bateman,A., Birney,E., Cerruti,L., Durbin,R., Etwiller,L., Eddy,S.R., Griffiths-Jones,S., Howe,K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
63. Murzin,A.G. and Bateman,A. (2001) CASP2 knowledge-based approach to distant homology recognition and fold prediction in CASP4. *Proteins*, **45** (Suppl. 5), 76–85.
64. Ginalski,K., Elofsson,A., Fischer,D. and Rychlewski,L. (2003) 3D-Jury: a simple approach to improve protein structure predictions. *Bioinformatics*, **19**, 1015–1018.
65. Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.*, **234**, 779–815.
66. Chivian,D., Kim,D.E., Malmstrom,L., Bradley,P., Robertson,T., Murphy,P., Strauss,C.E., Bonneau,R., Rohl,C.A. and Baker,D. (2003) Automated prediction of CASP-5 structures using the Robetta server. *Proteins*, **53** (Suppl. 6), 524–533.
67. Hung,L.H. and Samudrala,R. (2003) PROTINFO: secondary and tertiary protein structure prediction. *Nucleic Acids Res.*, **31**, 3296–3299.
68. Fischer,D. (2003) 3D-SHOTGUN: a novel, cooperative, fold-recognition meta-predictor. *Proteins*, **51**, 434–441.
69. Ginalski,K., von Grotthuss,M., Grishin,N.V. and Rychlewski,L. (2004) Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.*, **32**, W576–W581.
70. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631–637.
71. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallogr.*, **58**, 899–907.
72. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
73. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
74. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
75. Swets,J.A., Dawes,R.M. and Monahan,J. (2000) Better decisions through science. *Sci. Am.*, **283**, 82–87.
76. Bujnicki,J., Rychlewski,L. and Fischer,D. (2002) Fold-recognition detects an error in the Protein Data Bank. *Bioinformatics*, **18**, 1391–1395.
77. Schumacher,M.A., Hurlburt,B.K. and Brennan,R.G. (2001) Crystal structures of SarA, a pleiotropic regulator of virulence genes in *S.aureus*. *Nature*, **409**, 215–219.
78. Branden,C.I. and Jones,T.A. (1990) Between objectivity and subjectivity. *Nature*, **343**, 687–689.
79. Kleywegt,G.J. and Jones,T.A. (1995) Where freedom is given, liberties are taken. *Structure*, **3**, 535–540.
80. Linding,R., Russell,R.B., Neduva,V. and Gibson,T.J. (2003) GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, **31**, 3701–3708.
81. Romero,P., Obradovic,Z., Li,X., Garner,E.C., Brown,C.J. and Dunker,A.K. (2001) Sequence complexity of disordered protein. *Proteins*, **42**, 38–48.
82. Dunker,A.K., Lawson,J.D., Brown,C.J., Williams,R.M., Romero,P., Oh,J.S., Oldfield,C.J., Campen,A.M., Ratliff,C.M., Hipps,K.W. *et al.* (2001) Intrinsically disordered protein. *J. Mol. Graph Model*, **19**, 26–59.
83. Luthy,R., Bowie,J.U. and Eisenberg,D. (1992) Assessment of protein models with three-dimensional profiles. *Nature*, **356**, 83–85.
84. Mendez,R., Leplae,R., De Maria,L. and Wodak,S.J. (2003) Assessment of blind predictions of protein–protein interactions: current status of docking methods. *Proteins*, **52**, 51–67.
85. Fernandes,N.D. and Kolattukudy,P.E. (1997) Methylmalonyl coenzyme A selectivity of cloned and expressed acyltransferase and beta-ketoacyl synthase domains of mycocerosic acid synthase from *Mycobacterium bovis* BCG. *J. Bacteriol.*, **179**, 7538–7543.
86. Bujnicki,J.M. and Rychlewski,L. (2001) Identification of a PD-(D/E)XK-like domain with a novel configuration of the endonuclease active site in the methyl-directed restriction enzyme Mrr and its homologs. *Gene*, **267**, 183–191.
87. Bonneau,R., Strauss,C.E., Rohl,C.A., Chivian,D., Bradley,P., Malmstrom,L., Robertson,T. and Baker,D. (2002) *De novo* prediction of three-dimensional structures for major protein families. *J. Mol. Biol.*, **322**, 65–78.
88. Marcotte,E.M., Pellegrini,M., Ng,H.L., Rice,D.W., Yeates,T.O. and Eisenberg,D. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science*, **285**, 751–753.
89. Enright,A.J., Iliopoulos,I., Kyrpides,N.C. and Ouzounis,C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.
90. Lathe,W.C.,III, Snel,B. and Bork,P. (2000) Gene context conservation of a higher order than operons. *Trends Biochem. Sci.*, **25**, 474–479.
91. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.
92. Pellegrini,M., Marcotte,E.M., Thompson,M.J., Eisenberg,D. and Yeates,T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.
93. Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.
94. Kinch,L.N., Baker,D. and Grishin,N.V. (2003) Deciphering a novel thioredoxin-like fold family. *Proteins*, **52**, 323–331.
95. Jones,D.T. (2001) Evaluating the potential of using fold-recognition models for molecular replacement. *Acta Crystallogr. D Biol. Crystallogr.*, **57**, 1428–1434.
96. Andrec,M., Harano,Y., Jacobson,M.P., Friesner,R.A. and Levy,R.M. (2002) Complete protein structure determination using backbone residual dipolar couplings and sidechain rotamer prediction. *J. Struct. Funct. Genomics*, **2**, 103–111.
97. Rohl,C.A. and Baker,D. (2002) *De novo* determination of protein backbone structure from residual dipolar couplings using Rosetta. *J. Am. Chem. Soc.*, **124**, 2723–2729.
98. Li,W., Zhang,Y., Kihara,D., Huang,Y.J., Zheng,D., Montelione,G.T., Kolinski,A. and Skolnick,J. (2003) TOUCHSTONEX: protein structure prediction with sparse NMR data. *Proteins*, **53**, 290–306.
99. Doreleijers,J.F., Mading,S., Maziuk,D., Sojourner,K., Yin,L., Zhu,J., Markley,J.L. and Ulrich,E.L. (2003) BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank. *J. Biomol. NMR*, **26**, 139–146.
100. Gonzalez,C., Langdon,G.M., Bruix,M., Galvez,A., Valdivia,E., Maqueda,M. and Rico,M. (2000) Bacteriocin AS-48, a microbial cyclic polypeptide structurally and functionally related to mammalian NK-lysin. *Proc. Natl Acad. Sci. USA*, **97**, 11221–11226.
101. Liepinsh,E., Andersson,M., Ruysschaert,J.M. and Otting,G. (1997) Saposin fold revealed by the NMR structure of NK-lysin. *Nature Struct. Biol.*, **4**, 793–795.
102. Gough,J., Karplus,K., Hughey,R. and Chothia,C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
103. Tomii,K. and Akiyama,Y. (2004) FORTE: a profile–profile comparison tool for protein fold recognition. *Bioinformatics*, **20**, 594–595.
104. Karplus,K., Karchin,R., Barrett,C., Tu,S., Cline,M., Diekhans,M., Grate,L., Casper,J. and Hughey,R. (2001) What is the value added by human intervention in protein structure prediction? *Proteins*, **45** (Suppl. 5), 86–91.
105. Bates,P.A., Kelley,L.A., MacCallum,R.M. and Sternberg,M.J. (2001) Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. *Proteins*, **45** (Suppl. 5), 39–46.

106. McGuffin,L.J., Bryson,K. and Jones,D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, **16**, 404–405.
107. Shi,J., Blundell,T.L. and Mizuguchi,K. (2001) FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties. *J. Mol. Biol.*, **310**, 243–257.
108. Xu,J., Li,M., Lin,G., Kim,D. and Xu,Y. (2003) Protein threading by linear programming. *Pac. Symp. Biocomput.*, 264–275.
109. Zhou,H. and Zhou,Y. (2003) Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins*, **55**, 1005–1013.
110. Kim,D., Xu,D., Guo,J.T., Ellrott,K. and Xu,Y. (2003) PROSPECT II: protein structure prediction program for genome-scale applications. *Protein Eng.*, **16**, 641–650.
111. Fischer,D. (2000) Hybrid fold recognition: combining sequence derived properties with evolutionary information. *Pac. Symp. Biocomput.*, 119–130.
112. Zemla,A., Venclovas,C., Moult,J. and Fidelis,K. (1999) Processing and analysis of CASP3 protein structure predictions. *Proteins*, **37** (Suppl. 3), 22–29.
113. Cristobal,S., Zemla,A., Fischer,D., Rychlewski,L. and Elofsson,A. (2001) A study of quality measures for protein threading models. *BMC Bioinformatics*, **2**, 5.
114. Levitt,M. and Gerstein,M. (1998) A unified statistical framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
115. Ortiz,A.R., Strauss,C.E. and Olmea,O. (2002) MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.*, **11**, 2606–2621.
116. Siew,N., Elofsson,A., Rychlewski,L. and Fischer,D. (2000) MaxSub: an automated measure for the assessment of protein structure prediction quality. *Bioinformatics*, **16**, 776–785.