Full text provided by www.sciencedirect.com

ELSEVIER

Sequence comparison and protein structure prediction

Roland L Dunbrack Jr

Sequence comparison is a major step in the prediction of protein structure from existing templates in the Protein Data Bank. The identification of potentially remote homologues to be used as templates for modeling target sequences of unknown structure and their accurate alignment remain challenges, despite many years of study. The most recent advances have been in combining as many sources of information as possible — including amino acid variation in the form of profiles or hidden Markov models for both the target and template families, known and predicted secondary structures of the template and target, respectively, the combination of structure alignment for distant homologues and sequence alignment for close homologues to build better profiles, and the anchoring of certain regions of the alignment based on existing biological data. Newer technologies have been applied to the problem, including the use of support vector machines to tackle the fold classification problem for a target sequence and the alignment of hidden Markov models. Finally, using the consensus of many fold recognition methods, whether based on profile-profile alignments, threading or other approaches, continues to be one of the most successful strategies for both recognition and alignment of remote homologues. Although there is still room for improvement in identification and alignment methods, additional progress may come from model building and refinement methods that can compensate for large structural changes between remotely related targets and templates, as well as for regions of misalignment.

Addresses
Institute for Cancer Research, Fox Chase Cancer Center, 333 Cottman Avenue, Philadelphia, PA 19111, USA

Corresponding author: Dunbrack, Roland L (roland.dunbrack@fccc.edu)

Current Opinion in Structural Biology 2006, **16**:374–384

This review comes from a themed issue on
Sequences and topology
Edited by Nick V Grishin and Sarah A Teichmann

Available online 19th May 2006

0959-440X/$ – see front matter
© 2006 Elsevier Ltd. All rights reserved.

DOI 10.1016/j.sbi.2006.05.006

Introduction

Even in the 1980s, it was clear that multiple sequence alignments could improve sequence-structure alignment and hence structure prediction, but we rarely had enough sequences for this effort to have a big impact on prediction accuracy. Because of the rapid increase in available sequence and structure data, the linkage of sequence comparison and analysis with protein structure prediction has become even stronger in recent years. Both these areas have broad definitions and comprise many aspects each. It is not possible to review all aspects of these areas and their applications. I have instead chosen to focus on certain areas of sequence comparison related to structure prediction in which there has been important progress — the recognition of remote homologues and the determination of accurate alignments. The review covers primarily the time period from January 2004 to January 2006.

Structure prediction by comparative modeling can be divided into a number of steps:

1. Identification or recognition of a (potentially remote) homologue or homologues of known structure to be used as a template for modeling the target sequence of interest.
2. Improving the alignment of the target sequence with the template structures using alternative alignment methods or manual adjustment.
3. Building coordinates of the three-dimensional model based on the alignment, including the building of loops and sidechains, and the refinement of the entire model away from the template structure toward the target.
4. Assessing the potential accuracy of the model from the alignment or the model.
5. Using the model for biological inference from existing experimental data or to generate ideas for new experiments.

Of course, identification (step 1) generally involves aligning the target sequence with a set of available template sequences and structures, but this step also includes some kind of ranking and assessment of the statistical significance of the hits identified. Even with easy identification, the second step may involve using a number of methods to produce more accurate alignments and manual adjustment of alignments [1]. In this review, I cover methods used in the first two steps and their assessment. The methods discussed in this review that are publicly available via the Internet are listed in Table 1.

Assessing identification of template structures and alignment accuracy

Before discussing the progress made in the many methods for detecting remote homologues and producing alignments, it is important to review some of the approaches used to assess the abilities of these methods, in regard to both detection and alignment accuracy. Many calculated

Table 1

Availability of programs discussed in the text.

Program	Type	Download	Web	Address	References
HHsearch	HMM-HMM[a]	Yes	Yes	http://www.protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred http://protevo.eb.tuebingen.mpg.de/download/	[66••]
PRC	HMM-HMM	Yes		http://supfam.org/PRC/	Unpublished
QCComp	HMM-HMM		Yes	http://liao.cis.udel.edu/website/servers/modmod	[65]
PSI-BLAST-ISS	ISS[b]	Yes		http://www.ibt.lt/bioinformatics/iss/	[92]
3D-Jury	Meta-server		Yes	http://bioinfo.pl/meta/	[88]
GeneSilico	Meta-server		Yes	http://genesilico.pl/meta	[89]
Pcons5	Meta-server	Yes	Yes	http://www.sbc.su.se/(bjornw/Pcons5/ http://www.bioinfo.se/pcons/	[90]
MUSCLE	MSA[c]	Yes	Yes	http://www.drive5.com/muscle/ http://phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py	[32••,33]
3D-Coffee	MSA		Yes	http://www.igs.cnrs-mrs.fr/Tcoffee	[39••]
Kalign	MSA	Yes	Yes	http://msa.cgb.ki.se	[34]
Dialign	MSA	Yes	Yes	http://bibiserv.techfak.uni-bielefeld.de/dialign/	[36]
MuSiC	MSA		Yes	http://genome.life.nctu.edu.tw/MUSIC	[37,38]
Palign	Profile-profile	Yes		http://www.bioinfo.se/palign/	[55,62•]
SP^3	Profile-profile		Yes	http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sp3.html	[45••]
SALIGN (Modeller)	Profile-profile	Yes		http://salilab.org/modeller/	[56]
FORTE	Profile-profile		Yes	http://www.cbrc.jp/forte/	[63]
COACH (Lobster)	Profile-profile	Yes		http://www.drive5.com/lobster/	[64]
HMAP	Profile-profile		Yes	http://trantor.bioc.columbia.edu/hmap/main.html	[41,42]
FFAS03	Profile-profile		Yes	http://ffas.ljcrf.edu	[99]
COMPASS	Profile-profile	Yes		ftp://iole.swmed.edu/pub/compass/	[51,67]
Gist	SVM[d]	Yes		http://microarray.cpmc.columbia.edu/gist/	[76,83•]
SVM-BALSA	SVM		Yes	http://www.bioinfo.rpi.edu/applications/bayesian/balsa/manual/balsa.html	[78]
Saigo *et al.*	SVM		Yes	http://sunflower.kuicr.kyoto-u.ac.jp/~hiroto/svm/index.html	[81]
Prospector	Threading	Yes	Yes	http://www.gatech.edu/directories.php?entry=jskolnick3	[58•]

[a] Alignment of two HMMs.
[b] Intermediate sequence search.
[c] Multiple sequence alignment.
[d] Support vector machine.

parameters and test sets are used for these assessments. It is generally very difficult to compare the abilities of different programs from papers describing new methods, because the test sets and the parameters used to judge accuracy are usually different. Often the new method is compared only to PSI-BLAST [2]. This is quite inadequate, as many methods have been shown to be superior to using PSI-BLAST alone. However, the difficulty of using many distributed standalone programs discourages developers from making direct comparisons with other available programs. Such difficulties may include awkward input/output formats, poor program design, compilation difficulties and other problems. Some programs are available only as web servers and it is difficult to run a large test set on other people's web servers. Nevertheless, most new methods are compared with existing programs in some way; here, I discuss some of the common procedures.

SCOP [3] is almost universally used for the assessment of recognition of remote homologies. SCOP is quite convenient for this purpose, as nearly all superfamily pairs (i.e. two proteins from different families, but the same superfamily) consist of remotely homologous proteins. It is also possible to select a diverse set of structure pairs by picking one per fold or superfamily pair. When SCOP is used, the receiver-operator characteristic curve (or ROC curve) is most often determined to assess fold recognition. When used correctly, this curve is a plot of the parametric function of the fraction of correct identifications (the number of true positives divided by the number of correct pairs) versus the fraction of incorrect identifications (the number of false positives divided by the number of incorrect pairs). The integral under the curve (the 'ROC score') is a reasonable measure of the tradeoff between sensitivity and specificity for remote homologue detection methods.

Using SCOP can lead to problems, however. For instance, many superfamilies within a fold classification are very likely to be related to each other and thus superfamily identification is not a good standard for accurate recognition. Even at the fold level, some folds are, in fact, likely to be related to each other. We and others [4] have pointed out that there are several folds in SCOP with the topology of the Rossmann fold, but each with additional unique features. Some of these are quite likely to be homologous. Even related proteins sometimes diverge enough in structure that it becomes difficult to recognize the similarity in structure, no less the homology between them.

Sequence alignment accuracy can be measured in different ways. The first involves structure-based alignment of the target-template pair, deriving a sequence alignment from this structure alignment and then comparing the predicted sequence alignment with the structure-based alignment. This is generally the most common way and several scores have been developed based on this method. A number of groups have used three scores that I and others [4,5] developed, now often referred to as $Q_{modeler}$, $Q_{developer}$ and $Q_{combined}$. The first is the number of correct pairs divided by the number of pairs in the predicted sequence alignment, and is effectively a measure of the specificity of the alignment. The second is the number of correct pairs divided by the number of pairs in the structure-based alignment, and is effectively a measure of sensitivity.

$Q_{modeler}$ does not penalize under prediction and $Q_{developer}$ does not penalize over prediction. Consequently, Yona and Levitt [5] defined a measure, $Q_{combined}$, that penalizes both under and over prediction of aligned pairs. It is defined as the number of correct pairs divided by the number of unique pairs in the structure alignment and predicted alignment. That is, pairs that exist in both alignments ('correct' pairs) are counted only once in the denominator. In assessing profile-profile alignment methods and parameters, we found that it was much better to optimize for higher $Q_{combined}$ than for the other scores [6•]. Cline *et al.* [7] developed Q_{cline}, which measures alignments in terms of shifts of the predicted alignment from the structure alignment, as well as penalizing over and under prediction. The curated structure alignments of BaliBase [8] and HOMSTRAD [9] are often used for comparison with predicted alignments. Otherwise, SCOP superfamily pairs may be aligned with any of a number of structure alignment programs, although the results unfortunately vary significantly with the structure alignment program used. Using multiple structure alignment methods and looking for a consensus of aligned pairs is one way of dealing with this problem [6•,10].

The second method for assessing alignments involves building a model of the target by numbering and renaming the amino acids of the template with the target sequence numbering and residue types, according to the predicted target-template sequence alignment. The model of the target structure and the experimental target structure can be compared using a sequence-dependent alignment method. That is, the model and target have the same sequence and numbering scheme, so the structures can be aligned with a predefined sequence alignment in mind. This avoids the difficulties associated with sequence-independent alignments that are used when aligning targets with templates. In the CASP6 (Critical Assessment of Structure Prediction) fold recognition assessment [10], we found that, when models differed substantially from the target structure, sequence-independent alignment methods often failed to align any residues correctly between model and target. In these cases, sequence-dependent alignments sometimes managed to identify at least a substructure of the prediction that was reasonably close to the target experimental structure.

The CASP meetings use the GDT-TS score, which is based on such a sequence-dependent alignment and is roughly defined as the average of the percentage of the target that can be aligned to within 1, 2, 4 and 8 Å using four independent sequence-dependent alignments to maximize the number of aligned residues within these distances [11]. If the model is refined, perhaps with loops added or other backbone coordinates refined, then this score no longer reflects the sequence alignment alone and therefore should be used cautiously in evaluating purely alignment accuracy. This score has been criticized, as small changes in the structure can change the score significantly and overly compacted structures can result in higher scores than expected. Zhang and Skolnick [12••] have developed the TM-score, which overcomes some of the deficiencies of GDT-TS. Karplus *et al.* [13] have also developed a smoothed GDT for similar reasons, which was used in analysis of their CASP6 predictions.

Identification and alignment of sequences and structures can be done in many ways. The alignment accuracy of course depends on the alignment method, whereas the identification method depends on the scoring system and the estimated statistical significance of the alignment. Next, I review recent developments in the various methods in order of increasing complexity.

Pairwise alignments: substitution matrices and gap penalties

When determining relationships between remote homologues, emphasis is usually placed on multiple sequence alignments and profiles (see below). However, these methods often depend on pairwise sequence alignments, such as an initial BLAST search in a multiple-round PSI-BLAST search [2] or initial pairwise alignments before multiple sequence alignments. The determination of substitution matrices and gap penalties is often dependent on pairwise sequence and structure alignments, which are often used without modification for profile alignments and multiple sequence alignments. Also, for sequence families with very few members, pairwise alignment accuracy is even more important.

Recently, a number of groups have developed new sets of substitution matrices based on structure. These substitution matrices are often meant to be used either when one knows the structure of one protein (the potential template, for instance) or when one has a predicted secondary structure (for the target), or both. That is,

unlike BLOSUM matrices, the matrix used depends on the known and/or the predicted secondary structures of one or both of the sequences. These kinds of matrices were suggested by Eisenberg [14], among others. Blundell's group [15] has recently updated them with both functional and structural environment-dependent matrices to improve alignments. Huang and Bystroff [16] used their I-sites method to assign possible structural motifs to sequences and then produced 281 20×20 matrices for each of the 281 I-sites contexts in HMMSTR [17]. They produced improved pairwise alignments compared with BLOSUM-type matrices and structure-based matrices. Gelly *et al.* [18] recently provided a decision-tree method for deriving such structure-based substitution matrices, and Yu and Altschul [19] constructed matrices for proteins of unusual amino acid composition. Another approach is to derive properties of the 20 amino acids from physical data or from co-occurrence in multiple sequence alignments. These properties are often divided into principal components to explain the observed variation and the relative values on these scales can be used to score alignments simply by comparing the single amino acid values for each property. Wrabl and Grishin [20] recently performed this kind of analysis using variance maximization on BLOCKS alignments, and Atchley *et al.* [21] used a large number of physical measurements and principal component analysis.

Most scoring methods for alignments produce a higher score for more similar proteins and thus cannot be used as a distance metric, for which smaller scores mean more similar structures and the triangle inequality requirement must be met. Xu and Miranker [22], and Sonnhammer and Hollich [23•] have produced 'metric' substitution matrices, similar to PAM matrices, which provide higher scores for more dissimilar amino acids and lower scores (all greater than 0) for similar amino acids. These are true metrics that satisfy the axioms of a metric and can be used to judge the evolutionary distance between two sequences, in a way that sums of BLOSUM scores with gap penalties cannot.

In the past two years, comparative analysis of homologous protein structures has suggested new approaches to the determination of gap penalties in protein sequence alignment, following earlier work by Qian and Goldstein [24], and Benner *et al.* [25]. Goonesekere and Lee [26] derived a two-component gap penalty that fit the data from a set of 3992 structure alignments, with a steeper penalty for gaps up to length three and a flatter penalty for longer gaps. Also, Chang and Benner [27] found an inverse power law relationship between probability and gap length, with an exponent of 1.8, using a data set of alignments with only one indel event, but not using structure alignments. Generalized affine gap penalties have also been proposed, in which some regions of both proteins are unaligned with the other. This penalty is motivated by protein structure analysis showing that loops of very different length and sequence simply cannot be meaningfully aligned with a gap corresponding to the difference between their lengths. Rather, they should be unaligned and the gap penalty should reflect this. Zachariah *et al.* [28] found that a generalized affine gap penalty resulted in higher per residue accuracy ($Q_{modeler}$), even though it resulted in shorter alignments and hence lower $Q_{developer}$ scores. The use of residue composition statistically observed next to indel sites may also improve the accuracy of pairwise and multiple sequence alignments [27,29].

Multiple sequence alignments

Multiple sequence alignment remains an important area of research, as biological inferences can be made from the conservation or variation within aligned positions, especially with reference to the structure of at least one of the aligned sequences. For remote homologue detection using profiles or generalized profiles in the form of hidden Markov models (HMMs), more accurate multiple sequence alignments produce better models, and hence better detection and more accurate sequence-structure alignments. It is often quite obvious that a multiple sequence alignment produced by PSI-BLAST has many inaccuracies due to the nature of the program. PSI-BLAST aligns each database sequence with the single query sequence, and hence insertions and deletions are placed at slightly varying positions in the database sequences, resulting in a multiple sequence alignment that is inaccurate and aesthetically unpleasing, with more gaps than aligned positions.

Multiple sequence alignment is a large field of research, and I will only discuss a few recent papers and programs. ClustalW [30] and T-Coffee [31] remain popular choices for multiple protein sequence alignment. These programs employ progressive alignment using a guide tree first to align the most closely related sequences and then to align the profiles of each alignment. Often amino acids are obviously misaligned by such methods because of small misalignments early in the process. The MUSCLE program, recently developed by Edgar [32••,33], has shown improved performance over other methods. MUSCLE uses first the similarity of segments of length k between all protein pairs to determine distances and then performs a progressive alignment on this guide tree. Calculation of this tree is very fast, because it does not require pairwise alignments. This is followed by using the Kimura distance, which does require an existing alignment, to produce another tree and a second multiple sequence alignment. This tree is then broken at certain edges and multiple sequence alignments are produced for the subtrees, which are then recombined with a profile-profile alignment. If the sum-of-pairs score is increased, the new alignment is kept. This process is repeated for a fixed number of iterations or until the score converges. Although other methods have been developed or

improved recently [34,35], MUSCLE is more accurate than these when subjected to benchmarks such as BaliBase [8] and generally faster.

Two methods have been developed by a number of groups to improve multiple sequence alignments: the use of constraints, and the combination of structure and sequence alignments. Morgenstern *et al.* [36] have included the ability to use constraints or 'anchor points' in their DIALIGN segment alignment program, which is available as a stand-alone program and as a web server. Lu *et al.* [37,38] have also developed a multiple sequence alignment method that uses constraints. Such constraints are usually few in number and biologically meaningful. More generally, using structure alignment in combination with sequence alignment methods (for those related sequences without known structure) is more powerful, effectively resulting in a large number of constraints based on the structure alignment. For this purpose, the T-Coffee method of Notredame *et al.* has been improved by the use of pairwise structure alignments in 3D-Coffee [39••]. This method combines pure sequence alignment, pure structure alignment, and sequence-structure alignment or threading. Although not applied to proteins of unknown structure, Shatsky *et al.* [40] have combined multiple structure alignment with sequence alignment methods to produce multiple sequence alignments from multiple structure alignment that are more biologically meaningful than those produced by the multiple structure alignment alone, because of the inherent ambiguities of deriving a sequence alignment from structure superposition.

Sequence-profile alignments and improvements in HMMs

Although multiple sequence alignment is useful for biological analysis, for structure prediction we need first to identify a homologue of known structure for a target sequence of unknown structure, and then to align them accurately. The goals are therefore different from those of multiple sequence alignment, in which sequences known to be related are aligned and any one pairwise alignment is not more important than any other. Beyond simple pairwise alignment and multiple sequence alignment, most remote homologue detection and sequence-structure alignment methods use profiles of the target and/or the template sequence families. Since the development of PSI-BLAST, profiles have continued to be the basis of many remote homologue searching methods. I first discuss the asymmetric case, whereby, for instance, a target profile is used to search sequences from the Protein Data Bank (PDB) or a target sequence is compared to a database of template profiles. Profiles can be variations on the $20 \times L$ matrix used in PSI-BLAST, where L is the length of the generating sequence. These variations might include a gap character, for example. Profiles are built from multiple sequence alignments, usually obtained from a database search using pairwise alignments. Profile HMMs are also built from multiple sequence alignments of the target or template family, but include more information than the standard profile, including the positions of common insertions and deletions, and transition probabilities to and from match states at each position.

As with multiple sequence alignments, better profiles and HMMs can be built by using structure alignments of remote homologues and by adding sequences of unknown structure that can be easily aligned with each structure. These profiles are usually able to find more remote homologues than profiles built by PSI-BLAST from iterated searches of the sequence databases. This kind of technique has been used by Honig [41,42], Russell [43], Orengo [44] and colleagues. Zhou and Zhou [45••] pointed out that using structure alignments to build better profiles has often resulted in modest improvements in remote homologue detection or alignment accuracy. This is probably due to the non-uniqueness of sequence alignments generated from structure alignments, especially in the vicinity of large insertions or deletions or significant structural changes. These may result in misalignments between sets of sequences related to each structure. Their solution was to generate fragments of proteins and use these to build profiles in their SP^3 method.

HMMs have been improved in a number of ways. Wistrand and Sonnhammer [46] used a heuristic technique to adjust the transition parameters from match states to deletion and insertion states by adding sequences from outside the protein family of the model, so that positive sequences were scored more highly than the unrelated sequences. In another study, these authors compared the performance of HMMER and SAM, and found that, although SAM models were better because of better sequence or prior weighting, HMMER model scoring was better [47]. Combining SAM model building with HMMER scoring resulted in better remote homologue detection. Karplus *et al.* [48] recalibrated the E-values provided by the SAM program by optimizing parameters to the observed distribution using multiple functions, instead of a single Gumbel distribution.

One often-neglected issue is how to determine the sequences that should be placed in the multiple alignment that generates the profile or profile HMM. I refer to this issue as 'sequence choice'. It is often assumed that retaining as many remote homologues as possible in the multiple sequence alignment is the best choice. However, one can imagine that aligning sequences more distant from the target than the existing template(s) might in fact degrade the alignment accuracy or detection ability. Johnston and Shields [49] explored this issue by using a random sampling strategy to generate many

subsets of a larger set of sequences. They found that combining HMMs built from multiple sequence alignments of these subsets performed better than using the single HMM built from an alignment of all of the sequences. Some subsets were chosen on the basis of shared features; using a set of such HMMs enabled more sensitive and specific identification of remote homologues. Mihalek *et al.* [50•] recently examined sequence choice in the context of identifying functional sites in protein structures, rather than for comparative modeling, using a Monte Carlo method to choose sequences for the multiple sequence alignments. Sadreyev and Grishin [51] found, in general, that using more sequences in their COMPASS profile-profile alignment program always produced better profile alignments (see below). Rosenberg [52] recently showed that two sequences, A and B, could be aligned better if a third sequence, C, was added to make a multiple sequence alignment, if the AC distance was half that of the BC distance. That is, the perfect intermediate sequence was not half way between A and B, but closer to one of the sequences than the other. They did not examine the effect of adding a sequence that was further from both A and B than A and B were from each other. In general, the sequence choice for a particular target-template pair has not been examined in detail. Such a choice might occur after identification, during the stage of improving the alignment once the template has been identified.

Profile-profile and HMM-HMM alignments

As a generalization of sequence-profile alignments or sequence-HMM comparisons, profile-profile and HMM-HMM alignments have gained popularity in recent years. That is, instead of using profiles (or HMMs) for only the target or template, they are used for both and are compared to one another. It is assumed that, even when the target and template are known to be homologous, the target profile will be built from sequences closer to the target and the template profile will be built from sequences closer to the template. These profiles can be either traditional position-specific scoring matrices (PSSMs) or HMMs. In the past two years, there has been a focus on improving the scoring functions and gap penalties of profile-profile alignment methods. HMMs are a generalization of profiles and methods have recently been developed to align two HMMs with one another, hence producing a generalization of standard profile-profile alignment methods.

Whereas profile-profile alignment methods have been around for some time [53,54], several papers were published in 2004 comparing the various column-column scoring methods that had been proposed, along with optimization of gap penalties and the addition of predicted and actual structural information [6•,55–57]. It was commonly found that, although the different column-column scoring methods performed similarly once gap penalties were optimized for each scoring method, the methods differed in their remote homologue detection ability [6•].

The inclusion of structure, either predicted or actual, also improves profile-profile alignments [6•]. This may be only at the level of secondary structure or, in fact, may be combined with threading — that is, the evaluation of pairwise contacts when a target sequence is threaded through a template structure [58•,59]. New publicly available methods (downloadable programs or web servers) have been derived based on some of these results [56,60,61,62•,63], each usually presenting self-assessments indicating greater detection ability and sequence alignment accuracy than other methods.

As HMMs are a generalization of profiles with more detailed information on the positions of insertion-deletion states and a more sophisticated statistical framework, it became clear to a number of groups that HMMs might be aligned with each other and therefore be an improvement over standard profile-profile alignment methods [64,65,66••]. Kahsay *et al.* [65] used a reduced representation of a HMM, in terms of a quasi-consensus sequence derived from one HMM, scored this sequence versus the other HMM and repeated the process in reverse. This is not strictly HMM-HMM comparison, but it performed comparably to the sequence-profile alignment program COMPASS for a large test set and was considerably faster. COACH is a hybrid method that also compares a multiple sequence alignment with an HMM [64]. On the other hand, Söding [66••] used a dynamic programming alignment method to produce an alignment through the match and insertion-deletion states of two HMMs; the alignments are superior to those produced by COMPASS [67]. This is a more 'true' HMM-HMM alignment method than that of Kahsay *et al.* [65] and COACH. Madera and Chothia have also produced a true HMM-HMM alignment method, although it remains unpublished (see http://supfam.org/PRC/).

Support vector machines

Remote homology detection is a classification problem. It can be framed as a series of questions concerning the known structures or folds to determine whether the target protein is a member of family F or not. Given known features of the target of interest, $\mathbf{x}$, such as its sequence and sequence relatives, secondary structure prediction, amino acid content and so on, we wish to determine whether $y = 1$ (the target protein is a member of family F) or $y = 0$ (the target protein is not a member of family F). Such classifiers can be broadly classed as generative or discriminative. Generative classifiers model the data distribution, $p(\mathbf{x}|y)$, for each value of y, so that $p(y|\mathbf{x}) = p(\mathbf{x}|y)p(y)/(p(\mathbf{x}|1)p(1) + p(\mathbf{x}|0)p(0))$ can be calculated with Bayes' rule. Discriminative classifiers, such as the logistic regression model, attempt to represent

$p(y|\mathbf{x})$ directly. Neural networks and support vector machines (SVMs) are also discriminative classifiers, and have been used widely in computational biology [68–71]. SVMs determine a hyperplane that separates the data points based on the value of y (taking values of 1 or −1), given $\mathbf{x}$. However, the data may not be linearly separable. SVMs rectify this problem through the kernel function, $k(\mathbf{x}_i,\mathbf{x}_j)$, which measures the similarity of data points. The kernel function takes the data into a feature space in which the data points are linearly separable. The separating hyperplane is dependent on only a subset of the data, in the region that separates the data into positives and negatives. These are the support vectors. The SVM is easy to calculate once the kernel and its parameters are chosen. Ordinarily, a grid search over the parameters is used to determine the best values given the training set data. There are a number of commonly used kernels, including the polynomial and Gaussian kernels. SVMs have been used for remote homology detection, including SVM-Fisher [72], SVM-k-spectrum [73], SVM-pairwise [74], SVM-I-sites [75] and SVM-mismatch [76].

In the past two years, a number of new SVM-based methods for remote homology detection have been presented, usually differing in both the feature vector and kernel functions used. Hou *et al.* [77] have developed SVM-HMMSTR, as a successor to SVM-I-sites, which derives hidden Markov states that represent local folding patterns (from HMMSTR) in both a sequence-order independent and sequence-order dependent way. These states act as inputs for an SVM using a radial basis kernel. Webb-Robertson *et al.* combined probabilistic Bayesian alignment scores (BALSA) [78] as input vectors for an SVM with a quadratic kernel function to derive SVM-BALSA [79]. Han *et al.* [80] used profile-profile alignment scores of fragments to train an SVM with linear and radial basis kernel functions. Saigo *et al.* [81] have developed 'local alignment' kernels that mimic the Smith–Waterman alignment scores, but remain valid kernels ($k(\mathbf{x}_i,\mathbf{x}_j)$ must be a positive semi-definite matrix). Rangwala and Karypis [82•] have also developed new kernel functions based on the profile-profile alignment scores themselves, using either windows of fixed length or Smith–Waterman alignments. Noble and co-workers [83•] have recently used cluster kernels to improve remote homology detection.

Consensus methods for template recognition and sequence-structure alignment

One of the outcomes of the recent CASP experiments [10,84,85] was the dominance of consensus methods that combine the results of a number of fold recognition servers into a single prediction. These 'meta-servers' clearly outperform many of the individual methods they are built from, some of which are described above: sequence-profile alignments, HMMs, profile-profile alignments and threading. Some of the earliest meta-servers include the Pcons series [86], 3D-SHOTGUN [87] and 3D-JURY [88,89]. These meta-servers usually compare all the hits and/or just the top hits from the servers using structure alignment methods, and then predict the structure using the structure that is most similar to all of the other predictions.

Recently, Wallner and Eloffson [90] added to the consensus structure calculation an assessment of model quality using their ProQ software. This program evaluates surface accessibility, contacts, agreement of predicted and model secondary structure, and other values. They also added a simple method for using the scores provided by the servers as an indication of the predicted quality of the model by benchmarking each of the servers on previous LiveBench experiments. Another development is consensus prediction of several profile-profile alignment methods using different scoring systems. This is the basis of the meta-BASIC method [91••] and the method used by Tomii *et al.* [63] in CASP6. Finally, the PSI-BLAST-ISS method of Venclovas [92] is, in some sense, a consensus alignment method, as the final alignment is a consensus of many target-intermediate-template sequence alignments with different intermediates.

Model quality assessment

An important aspect of structure prediction is assessment of the likely quality of the model, even when the structure of the target is not known. In this review, we have not considered the generation of coordinates, loop and side-chain modeling, or the refinement of structures. Therefore, in this context, model quality assessment can be performed either by estimating the probability of each residue pair being correctly aligned in the target-template sequence alignment or by analyzing the structural model that is produced merely by copying the backbone coordinates of the template according to the alignment without further refinement. An example of the first method is provided by Tress *et al.* [93,94], who use the template profile, the target-template sequence alignment, a smoothing algorithm and secondary structure to produce a reliability estimate for each position in the target-template alignment. Although not designed for sequence-structure alignment evaluation, both Sadreyev and Grishin [95], and Lassmann and Sonnhammer [96] have developed methods for assessing the confidence of aligned positions in multiple sequence alignments.

It would be useful to have a model quality assessment program (referred to as a MQAP) that could determine which of several models is likely to be the best. Pettitt *et al.* [97•] recently compared four such programs [98], including their own MODCHECK based on a threading potential that includes a pairwise energy term and a solvation energy potential. They found that, for LiveBench9 targets, MODCHECK was able to rank models

better than several of the participating servers themselves, especially FFAS [54], FFAS03 [99] and Pcon4 [86], and better than three other competing MQAPs. Thus, even if the model alignment is determined by profile-profile alignment methods, threading information (pairwise interactions, solvation) still has a role in choosing the best model and, potentially, in improving initial alignments.

Future perspectives

As stated in the introduction, authors use many different test sets and evaluation criteria to judge the fold identification ability and alignment accuracy of their new methods. The role of community-wide experiments, such as CASP, EVA (http://cubic.bioc.columbia.edu/eva) and LiveBench, in comparing methods under identical conditions is very important. The most recent CASP experiment papers were published in December 2005 [10,84,85,100–102], and the most recent LiveBench results in January 2005 [103].

It is unfortunate that remote homology detection is assessed more frequently by developers of new methods than sequence alignment accuracy. As the size of the structural database increases, finding a template may become easier, as judged by statistical significance or functional relationships. However, improvements in sequence alignment accuracy may be harder to come by. Certainly, there remains room for improvement, as existing methods still do not always reach the level of the best alignment possible, according to a structure alignment.

Ultimately, however, at very low sequence identity, structures diverge significantly enough that alignment of some parts of the target and template structures is not even meaningful, and so there is a limit to how accurate a sequence alignment can be [104]. It may therefore be the case that using reliable parts of the alignment for well-conserved core regions, and better refinement and *ab initio* modeling of highly divergent regions may be a much more promising approach than trying to eke out small improvements in alignment quality. This was shown by the Baker [105], Skolnick [59] and other groups at the CASP6 meeting in December 2004. It was clear at the CASP6 meeting that many groups used one of a very small number of coordinate-generating programs, such as Modeller [106], once they had identified a template and fixed a target-template alignment. Improvements and more diverse approaches in these model-building programs — the refinement problem in homology modeling — are probably more important tasks than making small improvements in alignments. Refinement at both low and high sequence identity has become a funding priority of the National Institutes of Health, and we can expect to see progress in this area in the coming years [107], along with perhaps even better alignment quality yet to be realized.

Acknowledgements

Funding from the National Institutes of Health and the Pennsylvania Tobacco Settlement, and an appropriation from the Commonwealth of Pennsylvania are gratefully acknowledged.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Canutescu AA, Dunbrack RL Jr: **MolIDE: a homology modeling framework you can click with**. *Bioinformatics* 2005, **21**:2914-2916.

2. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of database programs**. *Nucleic Acids Res* 1997, **25**:3389-3402.

3. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J Mol Biol* 1995, **247**:536-540.

4. Sauder JM, Arthur JW, Dunbrack RL Jr: **Large-scale comparison of protein sequence alignment algorithms with structure alignments**. *Proteins* 2000, **40**:6-22.

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

6. • Wang G, Dunbrack RL Jr: **Scoring profile-to-profile sequence alignments**. *Protein Sci* 2004, **13**:1612-1626.
The authors present a comparison of several algorithmic choices in the alignment of profiles, including the scoring function, sequence weighting, sequence choice, determination of gap penalties and inclusion of structural information.

7. Cline M, Hughey R, Karplus K: **Predicting reliable regions in protein sequence alignments**. *Bioinformatics* 2002, **18**:306-314.

8. Thompson JD, Koehl P, Ripp R, Poch O: **BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark**. *Proteins* 2005, **61**:127-136.

9. Stebbings LA, Mizuguchi K: **HOMSTRAD: recent developments of the homologous protein structure alignment database**. *Nucleic Acids Res* 2004, **32**:D203-D207.

10. Wang G, Jin Y, Dunbrack RL Jr: **Assessment of fold recognition predictions in CASP6**. *Proteins* 2005, **61**:46-66.

11. Zemla A: **LGA: a method for finding 3D similarities in protein structures**. *Nucleic Acids Res* 2003, **31**:3370-3374.

12. •• Zhang Y, Skolnick J: **Scoring function for automated assessment of protein structure template quality**. *Proteins* 2004, **57**:702-710.
The authors present a scoring function for comparing models of proteins with their experimental structures that overcomes some of the deficiencies of the GDT-TS measure commonly used at the CASP meetings.

13. Karplus K, Katzman S, Shackleford G, Koeva M, Draper J, Barnes B, Soriano M, Hughey R: **SAM-T04: what is new in protein-structure prediction for CASP6**. *Proteins* 2005, **61(suppl 7)**:135-142.

14. Rice DW, Eisenberg D: **A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence**. *J Mol Biol* 1997, **267**:1026-1038.

15. Chelliah V, Blundell T, Mizuguchi K: **Functional restraints on the patterns of amino acid substitutions: application to sequence-structure homology recognition**. *Proteins* 2005, **61**:722-731.

16. Huang YM, Bystroff C: **Improved pairwise alignments of proteins in the twilight zone using local structure predictions**. *Bioinformatics* 2006, **22**:413-422.

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

18. Gelly JC, Chiche L, Gracy J: **EvDTree: structure-dependent substitution profiles based on decision tree classification of 3D environments**. *BMC Bioinformatics* 2005, **6**:4.

19. Yu YK, Altschul SF: **The construction of amino acid substitution matrices for the comparison of proteins with non-standard compositions**. *Bioinformatics* 2005, **21**:902-911.

20. Wrabl JO, Grishin NV: **Grouping of amino acid types and extraction of amino acid properties from multiple sequence alignments using variance maximization**. *Proteins* 2005, **61**:523-534.

21. Atchley WR, Zhao J, Fernandes AD, Druke T: **Solving the protein sequence metric problem**. *Proc Natl Acad Sci USA* 2005, **102**:6395-6400.

22. Xu W, Miranker DP: **A metric model of amino acid substitution**. *Bioinformatics* 2004, **20**:1214-1221.

23. • Sonnhammer EL, Hollich V: **Scoredist: a simple and robust protein sequence distance estimator**. *BMC Bioinformatics* 2005, **6**:108.
The authors propose a correction-based protein sequence distance estimator that uses a logarithmic correction of observed divergence based on the alignment score according to the BLOSUM62 score matrix. Scoredist is more robust than other methods for measuring the evolutionary distance of two sequences.

24. Qian B, Goldstein RA: **Distribution of Indel lengths**. *Proteins* 2001, **45**:102-104.

25. Benner SA, Cohen MA, Gonnet GH: **Empirical and structural models for insertions and deletions in the divergent evolution of proteins**. *J Mol Biol* 1993, **229**:1065-1082.

26. Goonesekere NC, Lee B: **Frequency of gaps observed in a structurally aligned protein pair database suggests a simple gap penalty function**. *Nucleic Acids Res* 2004, **32**:2838-2843.

27. Chang MS, Benner SA: **Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments**. *J Mol Biol* 2004, **341**:617-631.

28. Zachariah MA, Crooks GE, Holbrook SR, Brenner SE: **A generalized affine gap model significantly improves protein sequence alignment accuracy**. *Proteins* 2005, **58**:329-338.

29. Wrabl JO, Grishin NV: **Gaps in structurally similar proteins: towards improvement of multiple sequence alignment**. *Proteins* 2004, **54**:71-87.

30. Thompson JD, Higgins DG, Gibson TJ: **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**:4673-4680.

31. Notredame C, Higgins DG, Heringa J: **T-Coffee: a novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**:205-217.

32. •• Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**:1792-1797.
MUSCLE is a robust, accurate and fast multiple sequence alignment program that is becoming a standard in the field.

33. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.

34. Lassmann T, Sonnhammer EL: **Kalign-an accurate and fast multiple sequence alignment algorithm**. *BMC Bioinformatics* 2005, **6**:298.

35. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B: **DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment**. *BMC Bioinformatics* 2005, **6**:66.

36. Morgenstern B, Werner N, Prohaska SJ, Steinkamp R, Schneider I, Subramanian AR, Stadler PF, Weyer-Menkhoff J: **Multiple sequence alignment with user-defined constraints at GOBICS**. *Bioinformatics* 2005, **21**:1271-1273.

37. Lu CL, Huang YP: **A memory-efficient algorithm for multiple sequence alignment with constraints**. *Bioinformatics* 2005, **21**:20-30.

38. Tsai YT, Huang YP, Yu CT, Lu CL: **MuSiC: a tool for multiple sequence alignment with constraints**. *Bioinformatics* 2004, **20**:2309-2311.

39. •• O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: combining protein sequences and structures within multiple sequence alignments**. *J Mol Biol* 2004, **340**:385-395.
3D-Coffee uses a mixture of pairwise sequence alignments and pairwise structure comparison methods to generate more accurate multiple sequence alignments.

40. Shatsky M, Nussinov R, Wolfson HJ: **Optimization of multiple-sequence alignment based on multiple-structure alignment**. *Proteins* 2006, **62**:209-217.

41. Petrey D, Xiang Z, Tang CL, Xie L, Gimpelev M, Mitros T, Soto CS, Goldsmith-Fischman S, Kernytsky A, Schlessinger A *et al.*: **Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling**. *Proteins* 2003, **53(suppl 6)**:430-435.

42. Tang CL, Xie L, Koh IY, Posy S, Alexov E, Honig B: **On the role of structural information in remote homology detection and sequence alignment: new methods using hybrid sequence profiles**. *J Mol Biol* 2003, **334**:1043-1062.

43. Shah PK, Aloy P, Bork P, Russell RB: **Structural similarity to bridge sequence space: finding new families on the bridges**. *Protein Sci* 2005, **14**:1305-1314.

44. Sillitoe I, Dibley M, Bray J, Addou S, Orengo C: **Assessing strategies for improved superfamily recognition**. *Protein Sci* 2005, **14**:1800-1810.

45. •• Zhou H, Zhou Y: **Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments**. *Proteins* 2005, **58**:321-328.
The SP^3 method combines profile-profile alignment with structural information from protein fragments in a way that is more successful than other similar methods of using structure to produce better profiles.

46. Wistrand M, Sonnhammer EL: **Improving profile HMM discrimination by adapting transition probabilities**. *J Mol Biol* 2004, **338**:847-854.

47. Wistrand M, Sonnhammer EL: **Improved profile HMM performance by assessment of critical algorithmic features in SAM and HMMER**. *BMC Bioinformatics* 2005, **6**:99.

48. Karplus K, Karchin R, Shackelford G, Hughey R: **Calibrating E-values for hidden Markov models using reverse-sequence null models**. *Bioinformatics* 2005, **21**:4107-4115.

49. Johnston CR, Shields DC: **A sequence sub-sampling algorithm increases the power to detect distant homologues**. *Nucleic Acids Res* 2005, **33**:3772-3778.

50. • Mihalek I, Res I, Lichtarge O: **A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins**. *Bioinformatics* 2006, **22**:149-156.
The authors employ a strategy for optimally choosing sequences in multiple sequence alignments for determining probable interaction surfaces of proteins. The concept of optimally choosing sequences for multiple alignments has implications for the determination of profiles for remote homology detection and alignment.

51. Sadreyev RI, Grishin NV: **Quality of alignment comparison by COMPASS improves with inclusion of diverse confident homologs**. *Bioinformatics* 2004, **20**:818-828.

52. Rosenberg MS: **Multiple sequence alignment accuracy and evolutionary distance estimation**. *BMC Bioinformatics* 2005, **6**:278.

53. Pietrokovski S: **Searching databases of conserved sequence regions by aligning protein multiple-alignments**. *Nucleic Acids Res* 1996, **24**:3836-3845.

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

55. Ohlson T, Wallner B, Elofsson A: **Profile-profile methods provide improved fold-recognition: a study of different profile-profile alignment methods**. *Proteins* 2004, **57**:188-197.

56. Marti-Renom MA, Madhusudhan MS, Sali A: **Alignment of protein sequences by their profiles**. *Protein Sci* 2004, **13**:1071-1087.

57. Edgar RC, Sjolander K: **A comparison of scoring functions for protein sequence profile alignment**. *Bioinformatics* 2004, **20**:1301-1308.

58. • Skolnick J, Kihara D, Zhang Y: **Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm**. *Proteins* 2004, **56**:502-518.
A large benchmark of the successful Prospector threading algorithm is presented.

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

60. Simossis VA, Kleinjung J, Heringa J: **Homology-extended sequence alignment**. *Nucleic Acids Res* 2005, **33**:816-824.

61. Chung R, Yona G: **Protein family comparison using statistical models and predicted structural information**. *BMC Bioinformatics* 2004, **5**:183.

62. • Ohlson T, Elofsson A: **ProfNet, a method to derive profile-profile alignment scoring functions that improves the alignments of distantly related proteins**. *BMC Bioinformatics* 2005, **6**:253.
Rather than testing previously proposed scoring functions, the authors derive a scoring function for profile-profile alignments by optimizing the discrimination between correctly and incorrectly aligned positions as a function of distance of aligned structures.

63. Tomii K, Hirokawa T, Motono C: **Protein structure prediction using a variety of profile libraries and 3D verification**. *Proteins* 2005, **61(suppl 7)**:114-121.

64. Edgar RC, Sjolander K: **COACH: profile-profile alignment of protein families using hidden Markov models**. *Bioinformatics* 2004, **20**:1309-1318.

65. Kahsay RY, Wang G, Gao G, Liao L, Dunbrack R: **Quasi-consensus-based comparison of profile hidden Markov models for protein sequences**. *Bioinformatics* 2005, **21**:2287-2293.

66. •• Söding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**:951-960.
An algorithm for remote homology detection by comparison of the match and insertion-deletion states of two profile HMMs is presented. The algorithm is shown to be more sensitive than sequence-HMM and profile-profile comparison methods.

67. Sadreyev R, Grishin N: **COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance**. *J Mol Biol* 2003, **326**:317-336.

68. Nguyen MN, Rajapakse JC: **Multi-class support vector machines for protein secondary structure prediction**. *Genome Inform Ser Workshop Genome Inform* 2003, **14**:218-227.

69. Busuttil S, Abela J, Pace GJ: **Support vector machines with profile-based kernels for remote protein homology detection**. *Genome Inform Ser Workshop Genome Inform* 2004, **15**:191-200.

70. Garg A, Bhasin M, Raghava GP: **Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order and similarity search**. *J Biol Chem* 2005, **280**:14427-14432.

71. Bradford JR, Westhead DR: **Improved prediction of protein-protein binding sites using a support vector machines approach**. *Bioinformatics* 2005, **21**:1487-1494.

72. Jaakkola T, Diekhans M, Haussler D: **A discriminative framework for detecting remote protein homologies**. *J Comput Biol* 2000, **7**:95-114.

73. Leslie C, Eskin E, Noble WS: **The spectrum kernel: a string kernel for SVM protein classification**. *Pac Symp Biocomput* 2002:564-575.

74. Liao L, Noble WS: **Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships**. *J Comput Biol* 2003, **10**:857-868.

75. Hou Y, Hsu W, Lee ML, Bystroff C: **Efficient remote homology detection using local structure**. *Bioinformatics* 2003, **19**:2294-2301.

76. Leslie CS, Eskin E, Cohen A, Weston J, Noble WS: **Mismatch string kernels for discriminative protein classification**. *Bioinformatics* 2004, **20**:467-476.

77. Hou Y, Hsu W, Lee ML, Bystroff C: **Remote homolog detection using local sequence-structure correlations**. *Proteins* 2004, **57**:518-530.

78. Webb BJ, Liu JS, Lawrence CE: **BALSA: Bayesian algorithm for local sequence alignment**. *Nucleic Acids Res* 2002, **30**:1268-1277.

79. Webb-Robertson BJ, Oehmen C, Matzke M: **SVM-BALSA: remote homology detection based on Bayesian sequence alignment**. *Comput Biol Chem* 2005, **29**:440-443.

80. Han S, Lee B-c, Yu ST, Jeong C-s, Lee S, Kim D: **Fold recognition by combining profile–profile alignment and support vector machine**. *Bioinformatics* 2005, **21**:2667-2673.

81. Saigo H, Vert JP, Ueda N, Akutsu T: **Protein homology detection using string alignment kernels**. *Bioinformatics* 2004, **20**:1682-1689.

82. • Rangwala H, Karypis G: **Profile-based direct kernels for remote homology detection and fold recognition**. *Bioinformatics* 2005, **21**:4239-4247.
The authors present a remote homology detection method that combines profile-profile comparison with SVMs.

83. • Weston J, Leslie C, Ie E, Zhou D, Elisseeff A, Noble WS: **Semi-supervised protein classification using cluster kernels**. *Bioinformatics* 2005, **21**:3241-3247.
An improvement of this group's string kernels is presented that uses both labeled and unlabeled data to derive more powerful remote homology detection methods by SVMs.

84. Tress M, Ezkurdia I, Grana O, Lopez G, Valencia A: **Assessment of predictions submitted for the CASP6 comparative modeling category**. *Proteins* 2005, **61(suppl 7)**:27-45.

85. Vincent JJ, Tai CH, Sathyanarayana BK, Lee B: **Assessment of CASP6 predictions for new and nearly new fold targets**. *Proteins* 2005, **61(suppl 7)**:67-83.

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

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

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

89. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L: **Structure prediction meta server**. *Bioinformatics* 2001, **17**:750-751.

90. Wallner B, Elofsson A: **Pcons5: combining consensus, structural evaluation and fold recognition scores**. *Bioinformatics* 2005, **21**:4248-4254.

91. •• Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L: **Detecting distant homology with Meta-BASIC**. *Nucleic Acids Res* 2004, **32**:W576-W581.
A method for remote homology detection based on a consensus of profile-profile alignments that use different scoring methods is presented. This method performed well at CASP6.

92. Margelevicius M, Venclovas C: **PSI-BLAST-ISS: an intermediate sequence search tool for estimation of the position-specific alignment reliability**. *BMC Bioinformatics* 2005, **6**:185.

93. Tress ML, Grana O, Valencia A: **SQUARE-determining reliable regions in sequence alignments**. *Bioinformatics* 2004, **20**:974-975.

94. Tress ML, Jones D, Valencia A: **Predicting reliable regions in protein alignments from sequence profiles**. *J Mol Biol* 2003, **330**:705-718.

95. Sadreyev RI, Grishin NV: **Estimates of statistical significance for comparison of individual positions in multiple sequence alignments**. *BMC Bioinformatics* 2004, **5**:106.

96. Lassmann T, Sonnhammer EL: **Automatic assessment of alignment quality**. *Nucleic Acids Res* 2005, **33**:7120-7128.

97. • Pettitt CS, McGuffin LJ, Jones DT: **Improving sequence-based fold recognition by using 3D model quality assessment**. *Bioinformatics* 2005, **21**:3509-3515.
MQAPs based on threading-type potentials are shown to improve the scoring of models produced by other methods, such as profile-profile alignment.

98. Hendlich M, Lackner P, Weitckus S, Flöckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ: **Identification of native protein folds amongst a large number of incorrect models**. *J Mol Biol* 1990, **216**:167-180.

99. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A: **FFAS03: a server for profile-profile sequence alignments**. *Nucleic Acids Res* 2005, **33**:W284-W288.

100. Tress M, Tai CH, Wang G, Ezkurdia I, Lopez G, Valencia A, Lee B, Dunbrack RL Jr: **Domain definition and target classification for CASP6**. *Proteins* 2005, **61**:8-18.

101. Kryshtafovych A, Venclovas C, Fidelis K, Moult J: **Progress over the first decade of CASP experiments**. *Proteins* 2005, **61(suppl 7)**:225-236.

102. Kryshtafovych A, Milostan M, Szajkowski L, Daniluk P, Fidelis K: **CASP6 data processing and automatic evaluation at the protein structure prediction center**. *Proteins* 2005, **61(suppl 7)**:19-23.

103. Rychlewski L, Fischer D: **LiveBench-8: the large-scale, continuous assessment of automated protein structure prediction**. *Protein Sci* 2005, **14**:240-245.

104. Grishin NV: **Fold change in evolution of protein structures**. *J Struct Biol* 2001, **134**:167-185.

105. Bradley P, Malmstrom L, Qian B, Schonbrun J, Chivian D, Kim DE, Meiler J, Misura KM, Baker D: **Free modeling with Rosetta in CASP6**. *Proteins* 2005, **61(suppl 7)**:128-134.

106. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**:779-815.

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