Protein Science (1998), 7:1431–1440. Cambridge University Press. Printed in the USA.

Fold prediction by a hierarchy of sequence, threading, and modeling methods

ŁUKASZ JAROSZEWSKI,[1] LESZEK RYCHLEWSKI,[2] BAOHONG ZHANG,[2] AND ADAM GODZIK[2]

[1]Department of Chemistry, University of Warsaw, Warszawa, Poland
[2]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037

(RECEIVED December 10, 1997; ACCEPTED March 18, 1998)

Abstract

Several fold recognition algorithms are compared to each other in terms of prediction accuracy and significance. It is shown that on standard benchmarks, hybrid methods, which combine scoring based on sequence-sequence and sequence-structure matching, surpass both sequence and threading methods in the number of accurate predictions. However, the sequence similarity contributes most to the prediction accuracy. This strongly argues that most examples of apparently nonhomologous proteins with similar folds are actually related by evolution. While disappointing from the perspective of the fundamental understanding of protein folding, this adds a new significance to fold recognition methods as a possible first step in function prediction.

Despite hybrid methods being more accurate at fold prediction than either the sequence or threading methods, each of the methods is correct in some cases where others have failed. This partly reflects a different perspective on sequence/structure relationship embedded in various methods. To combine predictions from different methods, estimates of significance of predictions are made for all methods. With the help of such estimates, it is possible to develop a "jury" method, which has accuracy higher than any of the single methods. Finally, building full three-dimensional models for all top predictions helps to eliminate possible false positives where alignments, which are optimal in the one-dimensional sequences, lead to unsolvable sterical conflicts for the full three-dimensional models.

Keywords: distant homology; fold prediction; structure similarity

The protein-folding problem, i.e., the question of predicting the structure of a protein from its amino acid sequence, is one of the most important unsolved problems of molecular biology. Over the last 40 years, the efforts of X-ray crystallographers and, more recently, NMR spectroscopists, yielded thousands of atomic-resolution protein structures. These structures, for the most part available in public databases, form a rich source of knowledge that can be analyzed in search of empirical rules of protein folding. The most powerful rule discovered so far is that proteins with similar sequences have similar structures and often their functions are related. Another interesting observation was made with the increasing number of experimentally known protein structures. Numerous examples of protein pairs or groups without any recognizable sequence similarity but with remarkable structural similarity have been found. The existence of such groups was, at first, treated as mere curiosity, but as their number grew quickly (Pascarella & Argos, 1992; Orengo et al., 1993), it was soon accepted as a new paradigm in protein structure analysis. Is this really a new observation or merely an extension of the "similar sequence–similar structure" rule, but with a new definition of sequence similarity, which goes beyond a traditional letter-by-letter comparison of sequences?

This question touches on an important problem: are proteins such as these related by evolution (i.e., homologous) or not? Are our sequence-based similarity searches simply not sensitive enough to detect very distant homologies? For many such protein groups, there are hints of distant evolutionary relationships, such as some analogy between their functions or limited sequence similarity in the important regions of the protein (Babbitt et al., 1995). For other groups of proteins with the same fold, there are no obvious relations between their function or any other observations that could be used to argue for their homology. The example of a globin-like fold of bacterial toxin colicin comes to mind (Holm & Sander, 1993). The existence of such protein groups can be used to advance a theory that the universe of protein structures, in fact, may be limited (Finkelstein & Ptitsyn, 1987; Chothia, 1992) and proteins end up having similar folds simply by having to choose from a limited set of possibilities.

The difference between these two possibilities is very important for practical reason, as it determines the optimal choices for im-

Reprint requests to: Adam Godzik, Department of Molecular Biology, The Scripps Research Institute, 10550 N. Torrey Pines Road, La Jolla, California 92037; e-mail: adam@scripps.edu.

proving fold prediction strategies. Different tools might be needed to recognize proteins from extended homologous families or from nonhomologous but structurally converging protein groups. The first choice would call for the enhancement of tools of standard sequence analysis. For instance, multiple alignments could be used to create "profiles" where invariant positions are weighted more strongly than the positions with a large variation within the family of related proteins (Gribskov et al., 1987).

On the other side, methods where a compatibility between a sequence and a structure is calculated using energy potentials (Finkelstein & Reva, 1990; Godzik et al., 1992; Jones et al., 1992; Maiorov & Crippen, 1992; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993; Ouzounis et al., 1993) disregard the question of the evolutionary relation between proteins, focusing instead on the fact that two different sequences might have the global energy minimum in the same area of the conformational space. The main idea behind such methods can be compared to a grid search, where a free energy surface for a new protein sequence is tested at a number of points in anticipation that one of these points would fall close to the actual minimum. The goal is to predict a structure that is likely to be adopted by the sequence being studied, avoiding pitfalls of *ab initio* folding simulations such as long simulation times or the necessity to explore conformations that are unlikely to be seen in folded proteins. To allow for scanning of large structural databases within a reasonable length of time, algorithms are optimized for speed, paying the price by using an extremely simplified description of a protein structure. Various groups developed several different fitness functions and algorithms.

The important point here is that one group of methods seeks to enhance the underlying sequence similarity and, thus, it directly strives to search for very distant homologues. The second group of methods searches for compatibility between the structure and the sequence and, thus, disregards a possible evolutionary relationship between the proteins. Of course, the algorithms from two groups often converge and end up being formally equivalent, despite using vastly different nomenclature and ideology. For instance, three-dimensional profiles of Bowie et al. (Bowie et al., 1991) are formally equivalent to the "frozen approximation" of the topology fingerprint method of Godzik et al. (Godzik et al., 1992). In each case, a position dependent mutation matrix is created and used in the dynamic programming alignment. For three-dimensional profiles, it is done based on the classification of environments of each position into several classes (Bowie et al., 1991). In the latter method, it is done by calculating the energy of each possible mutation by summing up interactions at each position (Godzik et al., 1992). In addition, some potential energy parameters used in sequence-structure recognition methods contain a strong sequence-sequence similarity component by being based on the same amino acid features that dominate mutation matrices. For instance, hydrophobicity is a main component in both mutation matrices (Tomii & Kanehisa, 1996) and some interaction parameter sets (Godzik et al., 1995). Finally, some similarities between methods from the two groups might happen despite the authors' intentions, for instance, when potential energy parameters contain a strong "sequence memory" by including contributions from amino acid composition or size (Godzik et al., 1998). There are also methods that explicitly combine elements of both approaches, such as enhancing sequence similarity by residue burial status (Bowie et al., 1990), secondary structure (Luethy et al., 1991), or a generalized "interaction environment" (Bowie et al., 1991). The new generation of algorithms that follow these ideas are still being developed (Yi & Lander, 1994; Fischer & Eisenberg, 1996; Rice & Eisenberg, 1997).

It is very likely that examples of both types of structural similarity are present in nature. Therefore, both types of methods might be useful, each for a different type of target/template relationship. In the present contribution, we will compare structure prediction methods based solely on sequence similarity, structure/sequence compatibility and hybrid methods, mixing these two types of contributions. At the same time, we address the question of the prediction significance. This problem was extensively studied in the context of the sequence similarity between two proteins (Karlin & Altschul, 1990; Waterman, 1995), but it is still not completely understood, and the calculations of significance of threading predictions are still in their infancy (Bryant & Altschul, 1995). The problem of prediction reliability has a very practical aspect, because, as discussed above, we can expect at least two types of structural similarity and it is very likely that different algorithms have to be used in different cases. The reliable significance estimate would allow combining predictions done with different algorithms.

The question of homology between proteins with similar structures is also very important from the viewpoint of the possible applications of fold recognition algorithms. In the "distant homology" paradigm, establishing structural similarity opens the way to functional predictions, since even the most distant homologues share some level of functional similarity. On the other hand, in the "random structural similarity" paradigm, such predictions could be much more difficult.

This paper is organized as follows. In the first part of the Results section, prediction accuracy for different fold recognition algorithms is compared for several extensive benchmarks, including the set of Critical Assignment of Structural Predictions Meeting, Asilomar 1996, targets. Different methods are compared not only on the overall prediction accuracy, but also on a case-by-case basis. Prediction reliability estimates are used to create a "jury" method, which achieves an accuracy higher than any of the individual methods. Modeling of the best scoring templates is used as a final validation tool in the last part of this section. Possible applications as well as insights into the homology vs. random similarity question are discussed later in the paper. All methods, databases, and benchmarks are described in Methods at the end of the paper.

Results

The ultimate test of fold recognition methods is the prediction of the folds of new proteins when only their sequences are known and before any structural information is available. There are hundreds of thousands of proteins with known sequences, but without any information about structures that are potential targets of fold recognition. Unfortunately, at any given time, only some are of interest to a wider group of researchers. For this reason, it is difficult to test fold prediction methods in a unbiased fashion. Two recent CASP (Critical Assessment of Techniques for Proteins Structure Predictions) meetings addressed this issue by soliciting information about structures, which are already solved but still not publicly known, and then inviting groups developing prediction algorithms to test them on these examples. However, such meetings happen too infrequently to be used as the only means of testing and validation of new algorithms. Therefore, the fold recognition methods presented in this paper are tested on four benchmarks: two created

at the UCLA-DOE Laboratory of Structural Biology, a set of CASP2 targets, and our own "in-house" benchmark. Predictions using the topology fingerprint threading algorithm (Godzik et al., 1992) were presented at both CASP meetings and also will be discussed here. Other methods presented here were developed after the CASP2 meeting and, therefore, the results presented in Table 1 do not represent a genuine prediction.

Each benchmark consists of a set of proteins whose structure is to be predicted. We call them prediction targets. For each prediction target, its sequence is matched against a large number of proteins with known structures, representing all currently known protein folds. These proteins are potential modeling templates, the structures of which could be used to make a detailed prediction of the target structure. The goal of the fold recognition algorithm is to identify the most appropriate template protein. For each of the methods presented here, the identification is made on the basis of the alignment score as compared to the distribution of scores for the entire template database. In a benchmark, the real structure of prediction targets is known. Therefore, the quality of a given prediction method can be measured by a number of targets for which the template chosen by the algorithm was indeed similar to its real structure. For a comparison, a set of newly developed methods is compared to the topology fingerprint-based threading method (Godzik et al., 1992). The fold prediction WEB site maintained at UCLA-DOE (UCLA, 1996) provides information about the prediction accuracy of several algorithms developed at UCLA, and a number of current and forthcoming publications discusses the performance of several methods on the set of CASP2 prediction targets.

The four benchmarks discussed above are used to compare the performance of several fold recognition algorithms. The algorithms can be divided into three broad groups, depending on the type of information used to assign a score to matches between positions in the template and the target protein:

- sequence information for both the target and the template,
- sequence information for the target and structural information for the template, and
- structural information for both the target and the template. Such information is either known from the experiment (for the template) or predicted (for the target).

Specific energy terms are discussed in Methods as well as in footnote a to Table 1 and in the following discussion.

The results presented in Table 1 were calculated with parameters of various methods optimized for the number of correct predictions on the UCLA#1 benchmark or for the Scripps benchmark (threading). Results for the benchmarks used to optimize parameters are identified in bold to stress the fact that these numbers could not be treated as an indication of the given method's accuracy. As discussed later, this type of memorization effect is quite strong and if the parameters are optimized and tested on the same benchmark, it can give misleading indications about the prediction accuracy of the fold prediction algorithms.

The first observation that can be made from the data presented in Table 1 is that even though the benchmarks were constructed specifically to include only proteins without obvious sequence

Table 1. *The comparison of several types of fold prediction methods on four different benchmarks* [a]

	Benchmark number of targets	UCLA#1 68	UCLA#2 28	CASP2 7	Scripps 25
Methods using only sequence	BLAST	27/30/33	8/10/10	1/2/2	5/6/7
	Sequence	**40/50/52**	**9/13/16**	**2/5/7**	**12/16/18**
	Burial	**11/23/29**	5/14/15	0/0/2	5/15/18
	r14	**16/26/30**	4/6/10	0/2/2	9/17/17
Methods using only target structure	2b interactions	**4/9/17**	1/1/5	0/0/0	1/3/5
	Burial+r14+2b	**33/41/52**	8/19/19	0/3/3	9/17/21
	Secondary str.	**21/34/42**	9/16/20	0/0/4	13/20/20
	ss+burial+r14+2b	**36/47/51**	**12/19/21**	**1/3/4**	**11/21/24**
	s+r14	**43/50/57**	12/16/18	3/4/4	9/19/19
Hybrid methods, using both sequence and structure contributions	s+burial	**46/50/57**	11/14/18	3/4/6	13/19/20
	s+int	**43/50/53**	9/12/16	1/2/4	14/19/20
	s+burial+r14	48/53/56	15/19/21	4/4/4	12/17/20
	s+burial+r14+2b	**50/54/55**	13/19/19	4/4/5	14/20/20
	s+ss	**49/55/56**	16/19/20	3/4/4	15/22/22
	s+ss+burial+r14	**54/58/60**	**14/20/21**	**4/6/6**	**16/21/22**
Topology fingerprint threading		22/30/34	8/12/17	2/3/3	**10/14/15**

[a]Methods are identified by the type of scoring function used to evaluate the similarity between a target sequence and a template sequence/profile, see text and Methods for details. Prediction accuracy is described by three numbers: the number of correct templates at the first position, within top 5 or within top 10, respectively. The abbreviations: s, sequence; ss, secondary structure; 2b, two-body interaction preferences; r14, local structure preferences. See Methods for more detailed discussion of all energy terms. Threading was done with the topology fingerprint method (Godzik et al., 1992). The results for methods used later in Figures 1 and 4 are highlighted. The predictions were made based on the *p*-value as compared to the distribution of scores in the entire structural database (in the database of known structures). Numbers in bold identify results that were obtained with parameters optimized on this particular set. Note that for the sequence based scoring, independently optimized parameters were used. Threading predictions for CASP2 targets were submitted to the CASP2 conference All other results were obtained when the correct answer was known; therefore, they do not represent genuine predictions.

similarities, methods using only sequence information perform very well. Even the BLAST (Altschul et al., 1990) search correctly recognizes 27 targets from the UCLA#1 benchmark. The "sequence only" alignment using the dynamic programming algorithm correctly recognizes 40 prediction targets in this benchmark. This is more than the best recognition method that does not use the sequence information (ss+burial+r14+2b in Table 1). There is only one benchmark (UCLA#2) where the no-sequence scoring method is better than the sequence-only one. It is also interesting to note that many of these predictions, while being correct, would be judged as statistically insignificant by most of the standard sequence comparison algorithms that use distribution of randomized scores to assess score significance (Waterman & Vingron, 1994).

Methods using structural information about the template, but not its sequence, are not very powerful fold predictors on their own, but the information carried by these contributions is often independent from that of sequence comparison. As shown in Figure 1, out of the 40 and 36 targets correctly predicted by the sequence-based or structure-only-based scoring method, respectively, only 23 are recognized simultaneously by both methods. This observation explains why hybrid methods, combining both types of information, tend to perform better than either sequence only or no-sequence methods alone.

From the local structure-based scoring contributions, following energy terms contribute to the fold recognition in the descending order of importance:

- Predicted secondary structure (ss in Table 1).
- Preference for the template local backbone conformation as described by distances between Ca atoms separated by three bonds (r14 in Table 1).
- Burial preferences (burial in Table 1).
- Two-body side-chain interaction preferences (2b in Table 1).

The last contribution is by far the weakest, and in a hybrid scoring system that adds together all possible contributions, two-body interaction preferences do not add any new information. This might be explained by the use of "frozen approximation" where, in order to use dynamic programming alignment algorithm, this nonlocal contribution was made local by "freezing" interaction partners to that in the template protein (Godzik et al., 1992). It was shown recently that the average energy contribution from two-body interactions calculated this way is essentially random (Zhang et al., 1997). Consistent with these observations, threading predictions based solely on template structural information are much worse than the sequence-based predictions, and even worse than the newly optimized scoring systems based on local structural information (burial+r14+2body and ss+burial+r14+2body in Table 1). This last result is probably due to the poor optimization of the parameters of the threading algorithm, which was done using a very small number of examples known at that time.

Combining sequential and structural information significantly improves the prediction accuracy. Two types of additional scoring terms, namely sequence-structure one based on energy potentials and structure-structure one based on predicted structural information about the target, add a comparable number of correctly predicted targets. However, detailed analysis of the significance of various contributions must be done with some caution. As discussed earlier, all parameters were optimized on one of the benchmarks (UCLA#1) and subsequently used for other benchmarks. Since correct answers are known during optimization, this offers an indirect way of memorization of the results. This memorization allows some types of scoring to be "overoptimized" for a given benchmark, thus giving a misleadingly high prediction accuracy and obscuring general trends. Although the number of parameters is relatively small as compared to the number of proteins in each benchmark, the memorization effect is quite significant, as illustrated in Table 2.

The memorization effect is smaller for simpler methods with a smaller number of parameters, but it is still surprisingly large. Even for methods with only three parameters (gap opening, gap extension, and the zero of the scoring function), the number of correct predictions can change substantially from one parameter set to another.

Figure 1 illustrates another interesting observation in that despite the "hybrid method" being much better than others, each method has some targets for which only that particular method recognized the right template, while other methods fail. For instance, there are three protein pairs that can be identified at the first position by a method that does not use sequence information, but are not seen in the first ten scores when this information is added.

Subsequently, the number of targets predicted correctly by at least one method is larger than such a number for the best method. For instance, for the UCLA#1 benchmark, the best method recognizes 54 targets on the first position, 57 in the first 5 and 59 in the first 10. Taking the best results from all methods, this could be improved to 58/61/64. The same observation applies to other benchmarks, where for UCLA#2, CASP2, and SCRIPPS benchmarks, respectively 20, 5, and 21 targets were predicted at the first position by at least one method. The interesting question that arises at this point is how to create a "super" method, which would combine the best prediction from different algorithms.

We will explore two possible solutions to this problem. The first involves calculating the significance measure of a prediction and combining predictions with higher significance. The second approach studies alignments in the context of a template structure, trying to determine if alignments could lead to realistic models.

Prediction significance

Table 1 summarizes the information about the number of correct predictions, but does not tell us about the reliability of the predic-

```
s+ss+burial+r14 54 11111111111111111111111111111111*11311191114*11*11113*11181*1**16*
   sequence     40 11111111111111111111121111111*111111156*141**4*1**11*17*12**********
 ss+burial+r14  36 11151131111*521*1151111***2111**11*31*1*11111**4181****121161*3*112*
```

Fig. 1. Comparison between case-by-case prediction accuracy of different fold prediction algorithms for the UCLA#1 benchmark. The first number gives the overall accuracy of a given method, equal to the first number in Table 1. Next, the line of single digit numbers from 1–9 gives the position of the correct answer, * means that the correct answer was not in the top 9 scores.

Table 2. *The comparison of several types of fold prediction methods with different gap penalties* [a]

Benchmark	UCLA#1	UCLA#2	Scripps
Number of targets	68	28	25
	40/50/52	9/13/16	12/16/18
Sequence	**42/50/50**	9/13/16	12/15/16
	38/47/50	7/14/15	**14/18/20**
	40/49/50	**11/14/15**	11/19/19
	36/47/51	12/19/21	11/21/24
Local	35/45/48	13/16/18	**15/23/24**
	34/46/48	**15/19/20**	13/20/22
	54/58/60	14/20/21	16/21/22
s+ss+burial+r14+int	45/55/57	15/21/22	**21/23/23**
	50/53/56	**18/21/23**	15/22/22

[a] As before, results obtained with parameters optimized on a given benchmark are identified in bold. Prediction accuracy is described by three numbers, the number of correct templates at the first position, within top 5 or within top 10, respectively.

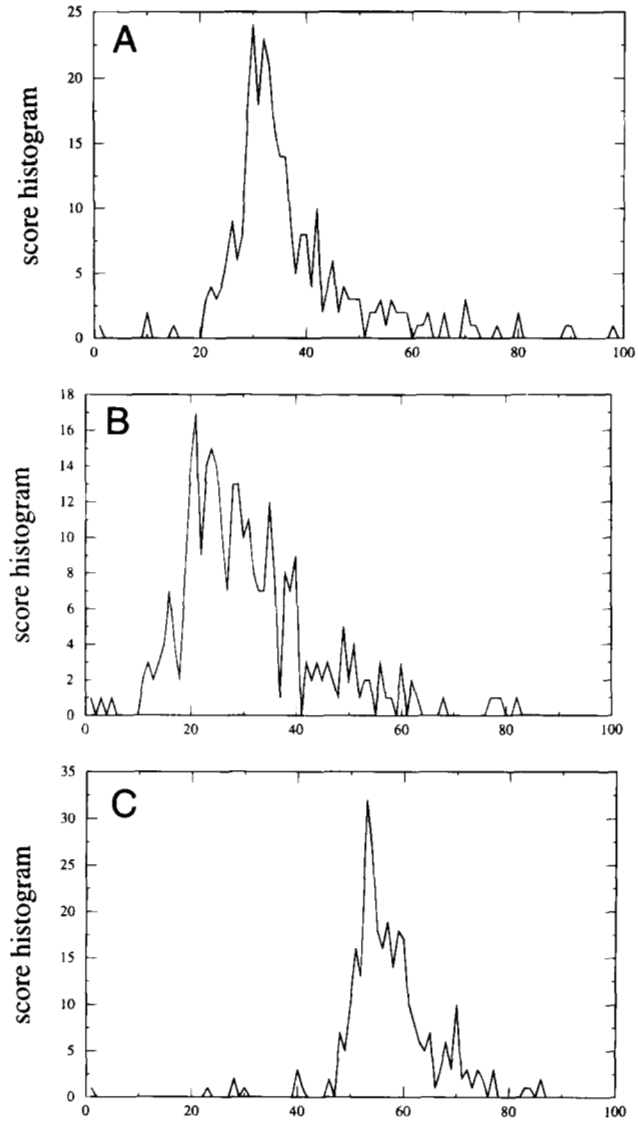

Fig. 2. The example of the distribution of scores for the local (i.e., ss+burial+r14+2b) sequence and the hybrid scoring methods. The distribution of scores for the 2fbjL target is shown for sequence (**A**), local (**B**), and the hybrid scoring (**C**), respectively.

tion. Following tradition in sequence homology searches (and common sense), prediction reliability could be defined as related to the probability that a given sequence/template score would be obtained by chance by comparing unrelated sequences. In particular, in subsequent figures and tables we define a significance for a given target-template match as the inverse of the probability that the score for that template is a part of the distribution of scores between unrelated proteins.

The distribution of sequence similarity scores between random sequences follows the extreme value distribution (Waterman, 1995) described by Equation 1 in Methods. As illustrated in Figure 2, for an example from the UCLA#1 benchmark, the score distributions for the sequence, local, and combined scores could be well described by the extreme value distribution, with the possible exception of the unexpectedly long tail of unfavorable scores. To avoid the bias from these high scoring proteins, only central 80% of scores are used to fit the parameters of the distribution. There are some anecdotal observations that for scorings based solely on local structure information the score distribution might have two maxima, describing scores of different structural classes (see Fig. 2). Unfortunately, because of the random fluctuations caused by the small number of templates in the protein structure database (ca. 400), it is not easy to determine exactly what distribution is followed by scores obtained in the calculations described here. For instance, it is possible to fit both the Gaussian and the extreme value distributions to the actual score distributions. With such a small number of samples, the difference between these two distributions is not statistically significant. However, the significance calculated with the extreme value distribution is much more reliable and, in particular, the significance calculated from the Gaussian distribution tends to create a lot of false positives—apparently high reliability scores that turn out to be false (results not shown). An example of this can be seen in Figure 2A, where the score for the best scoring (flase) template has a Z-score of 3.5, but a p-value of only 0.25.

For each of the targets, Equation 1 was used to calculate the distribution of scores in the entire database and the corresponding significance of the best score. Figure 3 shows the significance (calculated as the inverse of the probability that the best score could be obtained by chance) against the relative position of the correct template. As seen in this figure, the significance of the best score could be used as a very good indicator of the prediction accuracy. For all three scoring systems, there are no false positives with probabilities below 0.05. Thus, 33 (out of 54), 22 (out of 40), and 16 (out of 36) targets can be predicted with high confidence. Using the same reliability criteria, there are no false positives in all other benchmarks. However, the number of reliable predictions is much smaller than the number of correct predictions and, in this sense, the high prediction accuracy reported in Table 1 and Figure 1 is rather misleading. At the same time, reliability calculations explain the results shown in Table 2, because the differences in results obtained with different parameters happens exclusively for marginally significant predictions, where the correct answer closely

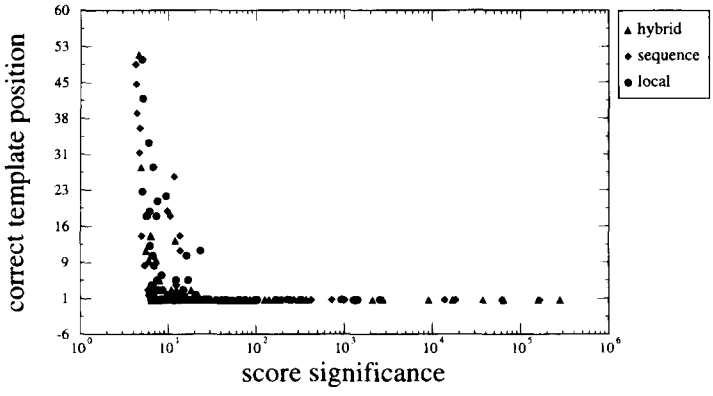

Fig. 3. For each of the 68 prediction targets in the UCLA#1 benchmark, the relative position of the correct template is shown as the function of the inverse of the probability that the score of the best template could be obtained by chance.

competes with wrong answers. The more accurate picture of accuracy of different prediction methods is shown in Figure 4.

In comparing Figures 1 and 4, it is interesting to note that in a number of cases, although none of the methods produces a reliable prediction, different methods give consistently correct answers. This enables us to build a "jury" prediction method where the prediction is made when two methods agree or if one of the three different predictions has high reliability. Such a jury method is correct in 55 cases, wrong in 3, and unable to make a prediction in 10 cases. It is encouraging to note that in all three cases where the jury method is wrong, there are quite significant similarities between target structures and a wrong template chosen by the prediction algorithm. For instance, 2rhe is chosen as a template for 1ten—both structures share the same β-barrel Greek key topology. In this case, the 2rhe template should probably be included as a correct prediction in the benchmark. Two other errors are more severe when spectrin is chosen as a prediction for a five-helical bundle structure of apo-lipophorin (1aep) and actinoxantin is chosen as a prediction for a viral protease 2snv. In both cases despite a different overall topology, there is a high similarity in the number, type, and location of secondary structure elements.

Alignment analysis

Sequences or property profiles are treated as a one-dimensional object by the alignment algorithms when, in fact, they represent three-dimensional protein structures. Therefore, it is possible for the alignment procedure to introduce gaps or deletions resulting in structure fragments that cannot form a continuous model in three-dimensional space. For instance, a gap in the alignment might result in two consecutive (in the target sequence) residues being separated by a long distance in the template structure. On the other hand, an insertion in the target sequence that happens in the core of the template will be difficult to accommodate, leading to severe steric overlaps. Unfortunately, such constraints could not be directly incorporated into the dynamic programming, but there were many attempts to introduce them indirectly into the alignment procedure. A variable gap penalty, large in the middle of secondary structure elements and smaller in the loops, is often introduced to mimic insertions/deletions patterns in real structures. When compared for prediction accuracy, such as presented in Table 1, variable gap alignments did not improve the fold prediction; on average, they were actually less accurate (results not shown).

For the UCLA#1 benchmark, alignments between the best scoring pairs for each of the three alignment methods shown in Figure 1 were prepared for the ten cases where the jury method could not make a prediction (see Fig. 4). At first, alignments were analyzed for breaks and overlaps in the model, as discussed above (columns "b" and "o" in Table 3, see Methods for details). In a second step, the alignments were used to produce three-dimensional models of the target by an automated modeling procedure (Sali, 1994; Pawlowski et al., 1997). As implemented, the modeling step mostly addressed the question of whether the models were possible to build at all. In each of the ten cases, the goal was to distinguish

```
s+ss+local 33 111111111117111117?1111??1?111?11???1?111?1????????11?????????????????

sequence   22 11111?111111????11?????1?1???11?????1?1????1?1?????????????????????????????

ss+local   16 111?1??1111?????????11?????111?????????????1?1?????????????????????????1???

-------------------------------------------------------------------------------------------

jury       55 1111111111111111111111111111111111111w111?111?w11?1111??111?1w??11??
```

Fig. 4. Comparison between case-by-case prediction accuracy of different fold prediction algorithms for the UCLA#1 benchmark with predictions weighted by their significance. The first number gives the overall accuracy of a given method. A jury prediction is a prediction where two methods agree or one of the predictions has high (p-value less than 0.05) significance. A question mark means that no prediction could be made—highest score had low significance or all methods disagreed. w means a wrong prediction.

Table 3. *Modeling of the ten cases where the jury method was not able to make the prediction*[a]

	Sequence scoring				Local scoring				Hybrid scoring			
		b	o	en		b	o	en		B	o	en
1bgeB	1gp1A	8	20	−17.7	**1huw**	1	2	−28.8	1mbc	2	24	−48.0
1cewI	2wrpR	2	22	−1.5	**2sarA**	—	44	−13.8	1molA	4	—	−1.8
1cid	2ms2A	1	40	0.9	1tie	2	40	1.4	**1ubq**	–	2	−0.1
1dsbA	2mtaC	5	40	−2.4	**1gp1A**	1	5	−40.1	5p21	4	4	−16.9
1gp1A	1ltz1	4	22	0.2	**2gcr**	1	5	−9.6	3chy	2	40	−0.8
1ltsD	3b5c	2	4	−0.4	**1bovA**	—	25	−6.0	1ubq	2	—	−10.7
1sacA	2mtaC	3	40	1.2	**2ayh**	1	—	−44.3	8fabA	7	40	−92.5
1tie	2pcy	3	23	—	**1tnfA**	—	40	−6.3	2bb2	4	20	−38.8
2azaA	3b5c	2	24	0.7	**1hfh**	—	—	−15.5	**3hlaB**	—	12	−20.3
3chy	1ifc	3	5	−4.1	2rslA	12	40	0.2	1bsgE	2	4	−18.6

[a]Column "b" gives the number of breaks in the chain, column "o" the number of steric overlaps in the model, and en the threading energy. Note that number of breaks is calculated directly from the alignment before the modeling step. Number of overlaps is calculated before the minimization step in the modeling procedure. Correct templates (if present) are highlighted. Templates chosen based on small number of breaks in the chain are denoted by bold.

between three different template choices. In three cases, the correct template (according to the benchmark) was among the three possibilities, while in seven others, it was not. Table 3 summarizes the results of the analysis of the alignments and the resulting models.

Results presented in Table 3 illustrate both the insights that could be gained from this additional step in the prediction hierarchy as well as the problems associated with it. In general, at this level of similarity (or lack thereof) between proteins, most alignments lead to serious problems in the model. There are discontinuities or breaks (column "b" in Table 3) where two consecutive side chains in the target sequence are aligned with residues, which are separated by over 5 Å in the template structure. The existence of such discontinuities could be detected before building the model. There are also the steric overlaps (column "o" in Table 3) where two side chains pack against each other with a hard core repulsion larger than 1 kcal/mol. Such overlaps can be detected early in the modeling process, before the most time-consuming minimization step. In most cases, this latter problem could be solved by more careful repacking of the protein interior; however, large numbers of overlaps clustered in one place in the model could not be resolved easily and might signify deeper problems with the model.

Simply choosing alignments with the smallest number of discontinuities, which could be done without the time-consuming modeling step, allows two correct templates to be recognized (1bovA as a template for 1ltsD and 2ayh as a template for 1sacA). In two additional cases (1huw as a template for 1bgeB and 3hlaB as one of the possible templates for 2azaA), this method identifies templates with topologies very close to the correct target topology, which were not included in the UCLA#1 benchmark by its authors. At the same time, there is only one case where, correctly, none of the templates could be accepted as a correct prediction. Unfortunately, it is difficult to formulate precise criteria for significance of the prediction, as in two other cases where similar criteria led to wrong predictions (1ubq as a template for 1cid and 1tnfA as a template for 1tie). The wrong template choice for 1cewI illustrates the problems with model building as a final step in the prediction hierarchy. Chain A of 2sar is incorrectly chosen over the correct template of 1molA (see Table 3) because the latter leads to four discontinuities in the model, while the former can be used to prepare a well packed and continuous model. It is interesting to note that these two topologies have a lot in common, both having an α-helix packed against an up-down-up-down anti-parallel β-sheet. At the same time, the results presented in Table 3 clearly illustrate the inadequacy of threading energy to evaluate extremely distorted models. Threading energy (see Methods) often favors physically impossible models (such as a model for 1sac built on the 8fabA template with 7 discontinuities of the chain and 40 overlaps). It is in clear contrast to the situation for models close to the correct fold, where threading can detect even subtle structural differences to the experimental structure (Jaroszewski et al., 1998).

Discussion

On several extensive benchmarks, it is shown that fold recognition methods, based on hybrid scoring systems using all available sequence and structure information, achieve a high level of structure prediction accuracy. Benchmarks used in this paper represent most of the known examples of structurally similar proteins with nonobvious homology. For this large group, over 80% of targets could be correctly assigned to the right fold family. Most cases, which are classified as prediction errors, still recognize substantial elements of the global fold and in all cases represent correct structural class.

The results presented here show that the accuracy of fold recognition methods increases with the increasing amount of information about the template used by the program to calculate the comparison score. Amino acid preferences to align with positions buried in the protein interior or exposed to solvent, together with preferences for a particular local backbone conformation, as described by a chiral R14 distance, contribute almost equally to the fold recognition accuracy. Including the predicted structural information about the target protein also adds to the prediction accuracy and this contribution is the single most important nonsequence contribution to the fold prediction. These results mostly parallel the results of the UCLA group, as presented on their fold recognition benchmark site (UCLA, 1996). All three local, nonsequence contributions are almost additive, allowing them to be combined to yield a local-structure based fold prediction method.

In contrast, the contribution from nonlocal two-body interaction preferences is the least successful scoring contribution on its own, and when combined with the sequence and local scoring functions, it does not contribute to the prediction success. For instance, in the most successful prediction method (s+ss+burial+r14) its weight is zero. It is important to note that in the two-body interaction parameters used here, the contribution of one-body effects, such as hydrophobicity, was essentially reduced to zero (Godzik et al., 1992). Most other interaction parameters include substantial contributions from various one-body effects (Godzik et al., 1995); therefore, their contribution might seem important for that reason alone. These results suggest that an ongoing debate about the best way to include nonlocal effects in fold recognition might be mostly irrelevant. Also, a recent analysis of structural alignments between pairs of proteins with similar folds (Zhang et al., 1997) suggests that two-body contributions are very sensitive to alignment details and become attractive only when the alignment is very close to the correct one.

The general trends discussed so far can be observed for all examples studied here; however, specific details of the relative importance of various contributions are blurred by a limited number of examples, as well as by strong memorization effects, such as illustrated in Table 2. The memorization effect is surprisingly strong, despite a relatively small number of parameters, as compared to the number of examples. It allows one to achieve misleadingly good results with some combination of parameters. This is part of the bigger problem of testing algorithms and prediction strategies on examples where the correct answer is known during the development and testing. Despite the relatively large size of available benchmarks, the ultimate test of all fold prediction methods must involve genuine predictions of targets whose structures are unknown when predictions are made.

For most of the pairs of structurally similar proteins with apparently weak sequence similarity, it is still the sequence that carries the most information. This can be measured both by the number of correctly recognized folds in all benchmarks and by the strong significance of the correct predictions. The strong contribution from sequence-based scoring suggests that most of the pairs in all the benchmarks tested here are actually homologous. This marks a significant shift of focus from the first fold recognition algorithms that stressed the structural similarity rather than the evolutionary relationships between structurally similar proteins. The question of whether fold recognition methods should focus on discovering distant homologues or unrelated proteins determines optimal strategies for improving algorithms and projects different applications of such methods. Focusing on distant homologies may seem disappointing from the point of view of understanding the balance of forces responsible for protein folding, but it allows extending fold recognition to function prediction. This dichotomy also provides a rationale for the development of a multi-prong approach where different methods might be more appropriate for different types of targets.

Methods

Scoring systems

The score between a position in the target protein and a position in the template protein can be based on different types of information about the target and template proteins.

A sequence-sequence scoring

No structural information is used. This type of comparison can be done even when no template structure is known. The Gonnet mutation matrix was used with optimized gap penalties, as identified by Vogt and Argos (Vogt et al., 1995).

A sequence-structure scoring

Scoring is based on the pseudo-energy of an single amino acid "mounted" in the structural environment of a single position in the template structure. The energy terms depend on the type of structural information used. The possibilities tested in this manuscript include burial status, local secondary structure, and interaction environment

Burial status definition and parameters were adopted from the topology fingerprint threading force field (Godzik et al., 1992).

Local backbone conformation is described by the distance between Cα atoms separated by three positions (R_i^{14}) and modified by the handedness of the torsional angle. The definition of r14 and corresponding energy parameters were adopted from the lattice protein folding force field (Kolinski & Skolnick, 1996).

Interaction preferences. Definitions of all terms, as well as all parameters, were adopted from the topology fingerprint threading method (Godzik et al., 1992). Only the "frozen" approximation was used (Godzik et al., 1992), i.e., no update of the environment was performed and the environment from the template protein was used to define a profile. This approximation is formally equivalent to the three-dimensional profiles (Bowie et al., 1991), albeit with different parameters and environment definitions.

A structure-structure scoring

Scoring is based on the comparison between the predicted structure of the target and the experimentally determined structure of the template. However, similar algorithms can be used to perform a structural alignment (two experimental structures) or enhanced sequence alignment (two predicted structures). An important difference between the sequence/structure and structure/structure scoring is that in the former case, the total score can be factorized into contributions from different target amino acids. For the latter, it is not possible, because the local structure prediction algorithm uses the entire sequence. In all cases, a nearest neighbor algorithm (Rychlewski & Godzik, 1997; Rychlewski & Godzik, 1998) was used. The latest version of this algorithm achieves an average prediction accuracy of 74% in a three-state secondary structure prediction and 73% in a two-state burial prediction. For the results presented here, only the secondary structure prediction was used.

Alignments

Since all tested scoring systems are local, i.e., the score between position i and j does not depend on any other position, standard dynamic programming can be used to find the optimal score for the alignment of two sequences or property profiles. The local-global alignment was used, i.e., end gaps were penalized for the template, but not for the target (Waterman, 1995; Fischer & Eisenberg, 1996). This choice was prompted by the fact that solved structures usually represent entire domains, while sequences might contain extra leading or trailing sequences. This choice assumes that the target sequence is longer than the template and, subsequently, it affects scores of sequences shorter than their desired targets. Effects of

these additional constraints are analyzed in a separate publication (Rychlewski et al., 1998).

In all cases, gap introduction and extension penalties as well as a constant subtracted from the complete, position-by-position scoring matrix, was optimized independently for different combinations of energy terms. For one of the benchmarks, chosen as a "learning set," a search in the parameter space was performed to maximize the number of correct predictions. Results for the "learning benchmark" are identified in bold in Tables 1 and 2. Monte Carlo-based simulated annealing was used as a minimization method. In most cases, multiple solutions, giving the same prediction accuracy, were found in the parameter space. In such cases, parameters were tested for "robustness," i.e., a small perturbation was added to all parameters and the prediction results were tested again. The parameter set that was most stable to perturbations was used on "testing benchmarks."

Benchmarks

Four benchmarks were used in this paper. Two were created at the UCLA-DOE Laboratory of Structural Biology, the first (UCLA#1) consists of 68 targets to be recognized among 301 possible templates; the second (UCLA#2) consist of 29 targets to be recognized among 320 possible targets (UCLA, 1996). In both cases, the benchmarks used in this paper were adopted from the UCLA WEB site. In addition, various methods are compared on a set of CASP2 targets (most targets from CASP meeting in 1994 are incorporated into the UCLA#2 benchmark) and on our own "in-house" benchmark, consisting of 25 targets to be recognized among 380 possible templates. For CASP2 prediction targets, only 7 (targets 2, 4, 14, 20, 22, 31, and 38) were used in a benchmark because no correct templates exists for other targets. A full listing of all benchmarks is available from the original WEB sites, as well as from the authors' group home page. All benchmarks are independent, i.e., there are no identical target-template pair; however, examples from the same families are present in most benchmarks. This situation is difficult to avoid, because the number of protein families displaying significant structural similarity with very weak sequence similarity is limited.

In each benchmark, a number of targets are scanned against a database of templates. The only exception is the BLAST search that used a SWISSPROT 34 (Bairoch, 1994) sequence database. In this case, only relative positions of sequences corresponding to proteins from the template database were compared. Because the structures of targets are known, it is possible to check in advance which of the templates have the largest structure similarity. Recognizing such proteins is registered as a success, failing to recognize them as a failure, and recognition of a protein that, in fact, has a different structure as a false prediction. Various prediction methods can be compared by listing the number of successful predictions vs. the number of failures and false predictions.

Significance analysis

The result of a comparison between a target sequence and a template protein is a single number—a score. To answer the question of whether or not target and structure are similar, we have to know the statistical significance of obtaining such a score by chance while comparing two unrelated proteins. This problem was analyzed extensively in the context of sequence alignments (Waterman, 1995). It can be solved exactly for comparing continuous strings, i.e., in alignment without gaps (Karlin & Altschul, 1990). For alignments with gaps, the distribution cannot be calculated analytically, but numerical experiments have shown that the distribution of scores for alignments with gaps is the same as for the alignments without gaps—it follows the so-called extreme value distribution (Pearson, 1996). The parameters of the distribution must be calculated from a numerical experiment. The distribution depends on two parameters and lengths and compositions of both sequences.

In general, the probability that the alignment score S is larger than X is proportional to

$$P(S > X) = \exp(-\gamma mn\xi^{-X}), \quad (1)$$

where m and n denote the lengths of two sequences being compared and the parameters γ and ξ describe the shape of the distribution curve. Note the difference between Equation 1 and most formulations in the literature, caused by the fact that our score resembles energy in being negative for similar proteins.

The parameters of the Equation 1 were calculated from the empirical distribution function of scores for the database of folds. For every target sequence take number of scores larger than a given value was calculated in 10 score unit intervals. The double logarithm ($-\log$ ($-\log$ empirical distribution)) of the empirical distribution function was fitted to a liner function of score. The parameters of a linear fit will be $\log(\gamma mn) - X\log(\xi)$ were used to calculate the probability of the lowest (the best) score. At the same time, the quality of the fit can be used as a measure that the scores really follow Equation 1.

Acknowledgments

We would like to thank Jacquelyn Fetrow, Andrzej Kolinski, and Jeffrey Skolnick for continuous discussions that helped shape the research described here, as well as Daniel Fischer for providing the list of protein folds. This research was supported by University of Warsaw Grant BST-532/34/96, Howard Hughes Medical Institute Grant 75195-543402 (ŁJ), and NIH Grant No. GM48835 (LR, BZ, and AG).

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Babbitt PC, Mrachko GT, Hasson MS, Hiusman GW, Kolter R, Ringe D, Petsko GA, Kenyon GL, Gerlt JA. 1995. Functionally diverse enzyme superfamily that abstracts the α proton of carboxylic acids. *Science 267*:1159–1161.

Bairoch A. 1994. SwissProt. Protein sequence database. Available at http://expasy.hcuge.ch/sprot/sprot-top.html.

Bowie JU, Clarke ND, Pabo CO, Sauer RT. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins 7*:257–264.

Bowie JU, Luethy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three dimensional structure. *Science 253*:164–170.

Bryant SH, Altschul SF. 1995. Statistics of sequence-structure threading. *Curr Opin Struct Biol 5*:236–244.

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

Chothia C. 1992. Proteins. One thousand families for the molecular biologist. *Nature 357*:543–544.

Finkelstein AV, Ptitsyn OB. 1987. Why do globular proteins fit the limited set of folding patterns? *Prog Biophys Mol Biol 50*:171–190.

Finkelstein AV, Reva BA. 1990. Determination of globular protein chain fold by the method of self-consistent field (in Russian). *Biofizika 35*:402–406.

Fischer D, Eisenberg D. 1996. Fold recognition using sequence derived properties. *Protein Sci 5*:947–955.

Godzik A, Kolinski A, Skolnick J. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci 4*:2107–2117.

Godzik A, Kolinski A, Skolnick J, Jaroszewski Ł. 1998. Dominant effects in-amino acid interactions. Analysis of energy parameter sets. *Proteins.* Forthcoming.

Godzik A, Skolnick J, Kolinski A. 1992. A topology fingerprint approach to the inverse folding problem. *J Mol Biol 227*:227–238.

Gribskov M, McLachlan M, Eisenberg D. 1987. Profile analysis: Detection of distantly related proteins. *Proc Natl Acad Sci USA 84*:4355–4358.

Holm L, Sander C. 1993. Structural alignment of globins, phycocyanins and colicin A. *FEBS Lett 315*:301–306.

Jaroszewski Ł, Pawlowski K, Godzik A. 1998. Multiple model approach: Exploring the limits of comparative modeling. *J Mol Model.* Forthcoming.

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

Karlin S, Altschul SF. 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA 87*:2264–2268.

Kolinski A, Skolnick J. 1996. *Lattice models of protein folding, dynamics and thermodynamics.* Austin, TX: R.G. Landes Company.

Luethy R, McLachlan AD, Eisenberg D. 1991. Secondary structure based profiles: Use of structure conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins 10*:229–239.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 277*:876–888.

Orengo CA, Flores TP, Jones DT, Taylor WR, Thornton JM. 1993. Recurring structural motifs in proteins with different functions. *Curr Biol 3*:131–139.

Ouzounis C, Sander C, Scharf M, Schneider R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3D structures. *J Mol Biol 232*:805–825.

Pascarella S, Argos P. 1992. A data bank merging related protein structures and sequences. *Protein Eng 5*:121–137.

Pawlowski K, Jaroszewski Ł, Bierzynski A, Godzik A. 1997. Multiple model approach: Dealing with alignment ambiguities in comparative protein modeling.In: Altman RB, Dunker AK, Hunter L, Klein TE, eds. *Biocomputing, 97.* Singapore: World Scientific. pp 328–339.

Pearson WR. 1996. Effective protein sequence comparison. In: Doolittle RF, ed. *Methods in enzymology. Computer methods for macromolecular sequence analysis, vol 266.* San Diego, CA: Academic Press. pp 227–258.

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

Rychlewski L, Godzik A. 1997. Secondary structure prediction using segment similarity. *Protein Eng 10*:1143–1153.

Rychlewski L, Godzik A. 1998. Searching for optimal parameters of a sequence structure mapping function. *Protein Eng.* Forthcoming.

Rychlewski L, Jaroszewski Ł, Zhang B, Godzik A. 1998. Fold prediction: Recognition or elimination? *Folding Design.* Forthcoming.

Sali A. 1994. Modeller. A program for protein structure modeling by satisfaction of spatial restraints. Available at http://guitar.rockefeller.edu/modeller/modeller.html.

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

Tomii K, Kanehisa M. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng 9*:27–36.

UCLA. 1996. *The UCLA-DOE benchmark to assess the performance of fold recognition methods.* Los Angeles, CA: University of California Press.

Vogt G, Etzold T, Argos P. 1995. An assessment of amino acid exchange matrices in aligning protein sequences: The twilight zone revisited. *J Mol Biol 249*:816–831.

Waterman MS. 1995. *Introduction to computational biology: Maps, sequences and genomes (Interdisciplinary Statistics).* New York: Chapman & Hall.

Waterman MS, Vingron M. 1994. Rapid and accurate estimates of statistical significance for sequence data base searches. *Proc Natl Acad Sci USA 91*:4625–4628.

Yi TM, Lander ES. 1994. Recognition of related proteins by iterative template refinements. *Protein Sci 3*:1315–1328.

Zhang B, Jaroszewski Ł, Rychlewski L, Godzik A. 1997. Similarities and differences between non-homologous proteins with similar folds. Evaluation of threading strategies. *Folding Design 12*:307–317.