Protein Science (1999), *8*:750–759. Cambridge University Press. Printed in the USA.

Factors limiting the performance of prediction-based fold recognition methods

XAVIER DE LA CRUZ AND JANET M. THORNTON

Department of Biochemistry and Molecular Biology, University College, Gower Street, London WC1E 6BT, United Kingdom

(RECEIVED July 29, 1998; ACCEPTED December 4, 1998)

Abstract

In the past few years, a new generation of fold recognition methods has been developed, in which the classical sequence information is combined with information obtained from secondary structure and, sometimes, accessibility predictions. The results are promising, indicating that this approach may compete with potential-based methods (Rost B et al., 1997, *J Mol Biol 270*:471–480). Here we present a systematic study of the different factors contributing to the performance of these methods, in particular when applied to the problem of fold recognition of remote homologues.

Our results indicate that secondary structure and accessibility prediction methods have reached an accuracy level where they are not the major factor limiting the accuracy of fold recognition. The pattern degeneracy problem is confirmed as the major source of error of these methods. On the basis of these results, we study three different options to overcome these limitations: normalization schemes, mapping of the coil state into the different zones of the Ramachandran plot, and post-threading graphical analysis.

Keywords: fold recognition; protein function identification; protein structure prediction; remote homologues; secondary structure and accessibility predictions; sequence annotation; threading

In light of the vast amount of information generated by the different genome projects, improvement in the performance of fold recognition methods has become an important challenge for those researchers working in the genomics field (Lander, 1996). These methods, also known as threading methods, provide a very promising approach to the problem of protein structure prediction and function identification, as can be seen by the results of the last CASP2 prediction experiment (Marchler-Bauer & Bryant, 1997). The first fold recognition methods generally used either distance-based (Jones et al., 1992; Sippl & Weitckus, 1992; Bryant & Lawrence, 1993) or profile-based (Bowie et al., 1991; Ouzounis et al., 1993) scoring functions. However, since the original work by Sheridan et al. (1985), different researchers (Fischel-Ghodsian et al., 1990; Fischer & Eisenberg, 1996; Russell et al., 1996, 1998; Rice & Eisenberg, 1997; Rost et al., 1997; Aurora & Rose, 1998) have developed a series of related methods that combine sequence information with secondary structure (SS) and accessibility (AC) predictions. In these prediction-based methods (PBM), all the structural information is encoded into a one-dimensional (1D) string of symbols, thus allowing matching in 1D and the use of classical dynamic programming algorithms (Needleman & Wunsch, 1970; Smith & Waterman, 1981). When tested in different cases (Rost et al., 1997; Rice et al., 1997), these methods have given promising results, comparable to the more complex distance-based methods (DBM) (Rost et al., 1997). Furthermore, matching in one dimension is about 10 times faster (Rost, 1996), which is important for the vast sequence searches related to genome projects. Despite these promising results, there is no clear understanding of the factors affecting/limiting the recognition ability of the PBM. In the case of the DBM, researchers from several laboratories have studied different aspects of their performance. In a thorough study, Kocher et al. (1994) analyzed the ability of different potential terms and side-chain models to recognize the native-fold of the query sequence within a set of candidates. Later, researchers from the same group (Lemer et al., 1995), studying the results from the first CASP prediction experiment, suggested that fold recognition may be achieved, despite poor alignment quality, by a generally unspecific maximization of the hydrophobic interactions, and a reasonably good prediction of the local secondary structure. Westhead et al. (1995) compared the behavior of two different threading algorithms when used together with the distance-based potentials. Bryant (1996), using his DBM (Bryant & Lawrence, 1993), shows that there is a clear relationship between the percentage of residues of the query sequence aligned with its remote homologue and the recognition specificity of the method. This result has been supported by the analysis of the results of the CASP2 prediction experiment (Marchler-Bauer et al., 1997). For the PBM the different authors have provided serious descriptions on the behavior

Reprint requests to: Janet M. Thornton, Department of Biochemistry, University College, Gower Street, London WC1E 6BT, United Kingdom; e-mail: thornton@biochem.ucl.ac.uk.

Abbreviations: AC, accessibility; DBM, distance-based methods; PBM, prediction-based methods; SS, secondary structure.

of their respective methods. However, to the best of our knowledge no systematic studies have yet been published similar to the ones described for the DBM.

In this paper we utilize a scoring function that shares the main characteristics of those used in the PBM and a representative dynamic programming algorithm to study different effects contributing to the performance of these methods. In particular, we concentrated in the study of their limitations when applied to the fold recognition of remote homologues. Remote homologues (Russell et al., 1997) are proteins, which, despite their low sequence homology, are evolutionarily related and generally have a similar function. The interest in this problem lies in the identification of function by recognition of a relationship between the query sequence and a protein of known function (Russell et al., 1998). Our results indicate that inaccuracies in the predictions are not the main limiting factor preventing improved specificity. Secondary structure and sequence pattern degeneracies have a more important effect. We explore how the use of functional annotations may help to improve discrimination.

Results and discussion

Performance in the fold recognition of remote homologues

To test the performance of the threading method, we used a set of 73 pairs of proteins that are remote homologues (see Methods). The sequence of the first protein in each pair (the query sequence) was used to search the structural database for the second protein in the pair (the target structure). Using the optimized parameters (see Methods), the fold recognition program ranked the target structure in the first position in 29 queries, while in the remaining 44 examples a nonhomologous structure was ranked first. Therefore, the success rate (29/73) was 39.7% higher than the 29% success rate reported by Rost et al. (1997). (Note: This calculation increases dramatically when homologues for the query sequence are not specifically excluded from the database; see below.) Several factors may account for this difference in performance:

(1) The parameter optimization was done on the same set of proteins used to evaluate the performance of the method.

(2) The test sets are not the same. In our case we were only interested in the fold recognition of remote homologues, while Rost et al. (1997) include analogous proteins and consider the general threading problem.

(3) We are using a smaller database of structures (627 proteins) than Rost et al. (1997) (701 proteins), who have shown that the smaller the database, the better the performance.

(4) Our probabilities are dependent on the reliability indexes, and this may improve the performance of our method (Fischer & Eisenberg, 1996).

(5) These results may also indicate that, on the average, it is slightly easier to identify remote homologues than analogues due to their higher conservation of sequence and structural properties. This accords with the results of Russell et al. (1998), who achieved higher accuracies for homologues (64%) than analogues (56%), using a smaller test set including multiple targets.

Fischer and Eisenberg (1996) present higher success rates (74%) on their WWW server, but this reflects the sequence similarity between many of the sequences in their database. A simple sequence search using the PAM250 matrix gives a success rate on their dataset of 51%, while on ours the same search identifies only 9% of the targets.

The quality of the alignments obtained here is comparable to previous PBM methods. Comparing our predicted alignments against the structure-based alignments derived using the SSAP program (Taylor & Orengo, 1989), we find that 62% of correct hits show more than half of their residues correctly aligned. This 62% is close to the value derived by Rost et al. (1997). If we use the STAMP program (Russell & Barton, 1992) structural alignments, this value drops to 38%. This illustrates one of the problems when trying to assess the accuracy of the threading alignments: different structure comparison methods provide different alignments (Godzik, 1996; Zu-Kang & Sippl, 1996).

Thus the method used here shows a success rate comparable to previous methods. As it is based on essentially the same principles as these other PBMs, we can reasonably assume that the results obtained from the performance analysis of our method will be generally applicable.

Effect of the number of fold representatives in the searched database

To compute the performance of our method for each query sequence, all its remote homologues present in the original database (Rost, 1996) were included in the target database (see Methods). The average number of target structures per query sequence increased from 1 to 3.8, on average, giving a much improved performance, with an average success rate of 67.4%. This result generalizes the initial observation by Lemer et al. (1995), and shows that searched databases should include the maximum available number of fold representatives.

Comparison between the structural and the threading alignments

One of the goals of the threading methods is to provide an alignment of the query sequence against the target sequence to allow an accurate model to be built by using homology modeling techniques (Rost et al., 1997). However, the outcome of the modeling critically depends on the accuracy of the alignment, as shown by the results of the CASP prediction experiment (Sali et al., 1995; Samudrala et al., 1995; Martin et al., 1997). When comparing the observed and the predicted alignments, we found that there was no case for which the threading alignment coincided entirely with the structural alignment. To explore the reasons for the nonrecognition of the structural alignments, we compared the score decomposition (see Methods) for the predicted and the observed alignments (Table 1). To take into account the structural alignment ambiguities (Godzik, 1996; Zu-Kang & Sippl, 1996), two sets of structural alignments were used, derived from SSAP (Taylor & Orengo, 1989) and STAMP (Russell & Barton, 1992). The results for both sets were similar.

For the observed structural alignments, we see that both the sequence and the gap terms make unfavorable negative contributions to the alignment scores (-0.100 and -0.258, respectively, for the SSAP alignments). Only the structure term has a positive contribution (0.049), derived solely from the SS prediction term. This reflects the fact that SS is both better predicted and more conserved between homologues than AC. The very negative con-

Table 1. *Average values for the different scoring terms in the observed structural and predicted threading alignments*

	Observed alignments		Predicted alignments
	SSAP[a]	STAMP[b]	Threading
Sequence	−0.100 (0.055)	−0.090 (0.063)	0.083 (0.073)
Structure[c] (SS + AC)	0.049 (0.058)	0.057 (0.050)	0.188 (0.045)
Gap	−0.258 (0.224)	−0.349 (0.155)	−0.103 (0.045)
Structural decomposition[d]			
SS[e]	0.049 (0.047)	0.053 (0.040)	0.121 (0.029)
AC[f]	0.000 (0.022)	0.004 (0.020)	0.067 (0.026)

[a]Structure contribution obtained for the SSAP alignments (Taylor & Orengo, 1989).
[b]Structure contribution obtained for the STAMP alignments (Russell & Barton, 1992).
[c]Total contribution of the structure term (SS + AC).
[d]Decomposition of the structural term.
[e]Contribution due to the SS term.
[f]Contribution due to the AC term.

tributions of the gap penalties can be attributed to the high number of gaps observed in the structural alignments of remote homologues (Russell et al., 1997).

In the predicted threading alignments, both the structure and sequence terms have positive values of 0.188 and 0.083, respectively. In addition, decomposition of the structure term shows that both the secondary structure and the accessibility terms make positive contributions to the alignment scores. It is not surprising that the predicted alignments show better scores, since they were derived to optimize these values. However, this highlights the inadequacies of the scoring function (an incorrect alignment scores better than the correct observed alignment).

This conclusion is essentially independent of the ambiguities in the structural alignments (Godzik, 1996; Zu-Kang & Sippl, 1996), as similar results are reached when using the STAMP (Russell & Barton, 1992) structural alignments (Table 1).

Factors affecting the fold recognition specificity of remote homologues

Above we used a simple score decomposition corresponding to the three terms of the scoring function: structure, sequence, and gap. Unless otherwise stated, we will use score differences to describe the contribution of the different terms to the successes or failures of the method (see Methods).

When looking at the successes of the method, we see that on average the sequence, structure, and gap terms of the scoring function (0.051, 0.041, and 0.013) all favor the recognition of the correct fold. In particular, the sequence and the structure terms are the most discriminating. If we plot the sequence identity distribution for both the successes and failures of the method (Fig. 1), we observe that sequence identity is in general higher for the successful than for the failed queries, emphasizing that homologues sharing higher sequence similarities are easier to recognize. However, there is a substantial overlap between both distributions, and for 60% of the successes, the contribution of the structure term is higher than that of the sequence term. These results confirm that at this level of sequence identity the structure term is needed for a fruitful fold recognition.

For the failures of the method, all the terms—sequence, structure, and gap—contribute to recognizing the wrong partner (Table 2), although in this case the more important contributions come from the structure and the gap term, −0.019 and −0.024, respectively. On average the structure term difference is more than twice as large as the sequence term difference (Table 2), which arises from the large contribution of the SS prediction score (−0.020). This probably reflects the highly nonrandom

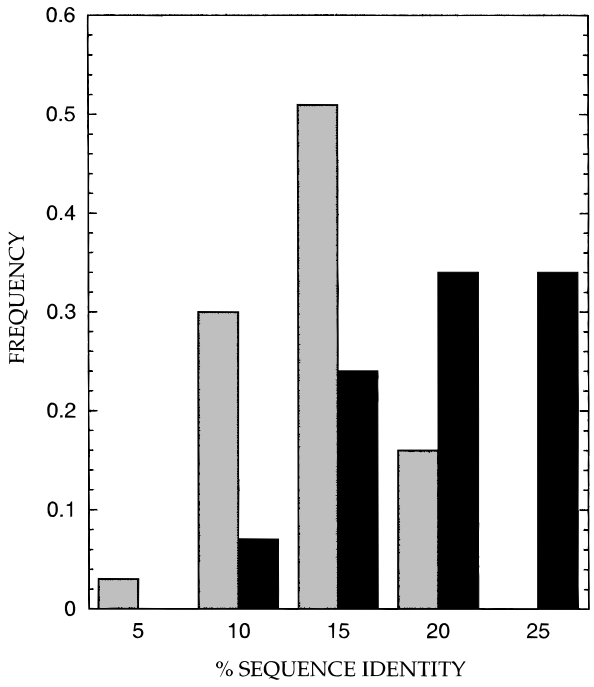

Fig. 1. Distribution of sequence identities from structural alignments for the successes (black) and failures (grey) of the method. The sequence identities for the 73 test cases were derived from the SSAP structural alignments (Taylor & Orengo, 1989). Similar results were obtained using the STAMP alignments (Russell & Barton, 1992) (results not shown).

Table 2. *Summary of the score decompositions for the successes and failures of the fold recognition procedure*[a]

	Successes[b] (28 cases)	Failures[c] (38 cases)
Sequence	0.051 (0.036)	−0.008 (0.055)
Structure[d]	0.041 (0.063)	−0.019 (0.037)
Gap	0.013 (0.053)	−0.024 (0.064)
Structural decomposition[e]		
Correctly predicted[f]	0.066 (0.04)	−0.005 (0.042)
Incorrectly predicted[g]	−0.025 (0.025)	−0.014 (0.028)
SS decomposition[e]		
SS[h]	0.019 (0.028)	−0.020 (0.028)
AC[i]	0.022 (0.025)	0.001 (0.02)
SS decomposition[j]		
Alpha + beta[k]	0.017 (0.016)	−0.007 (0.022)
Coil[l]	0.002 (0.021)	−0.013 (0.018)

[a] Calculated as the differences Δ between the correct match and nonhomologous match with the highest score. Positive values indicate a favorable contribution to the correct candidates while negative scores indicate that the nonhomologous protein scored more highly.

[b] The values are listed for the different contributions computed for each protein using Equation 6 and averaged over the 28 successful fold recognition examples. The standard deviations are given in parentheses.

[c] Same as in footnote b, for the 38 protein pairs that were not successfully matched.

[d] Δ in total contribution of the structure term (SS + AC).

[e] Decomposition of the structural Δ term.

[f] Contribution from those residues for which SS + AC are correctly predicted.

[g] Contribution from those residues for which SS + AC are incorrectly predicted.

[h] Total contribution to Δ of the SS term.

[i] Total contribution to Δ of the AC term.

[j] Decomposition of the total SS term.

[k] Contribution to the Δ SS term from the aligned residue pairs in which the residues belonging to the selected candidates were in helix and sheet.

[l] Same as in footnote k for residues in the coil state.

distribution of the SS states along the sequence (i.e., long stretches of SS can match, even in incorrectly paired sequences).

In the following sections, we discuss in more detail different effects modulating the fold recognition performance of PBM.

Structure prediction accuracy

As expected, the contribution of residues with correctly predicted structure (SS + AC) makes a major contribution (0.066) to the successful recognition of the target structure. In contrast, the average contribution of those residues with incorrectly predicted structure (−0.025) favors recognition of the wrong candidate. This is also true for the failures of the method (Table 2), indicating the relevance of good structure predictions in the recognition of the correct target structure.

To further evaluate the relevance of the prediction accuracy to the recognition process, we computed the secondary structure and accessibility prediction accuracies for the 73 query sequences. The results obtained show that, on average, they are only slightly better when the method succeeds in recognizing the remote homologue (76.0% and 59.5% for SS and AC, respectively) relative to 73.0% and 55.2%, respectively, when the method fails. These differences in prediction accuracy partly explain the failures of the method (see Fischer & Eisenberg, 1996; Rost et al., 1997) but when plotting the histograms for both the SS and AC accuracies (Fig. 2), we can see a clear overlap between the successes and failures of the method.

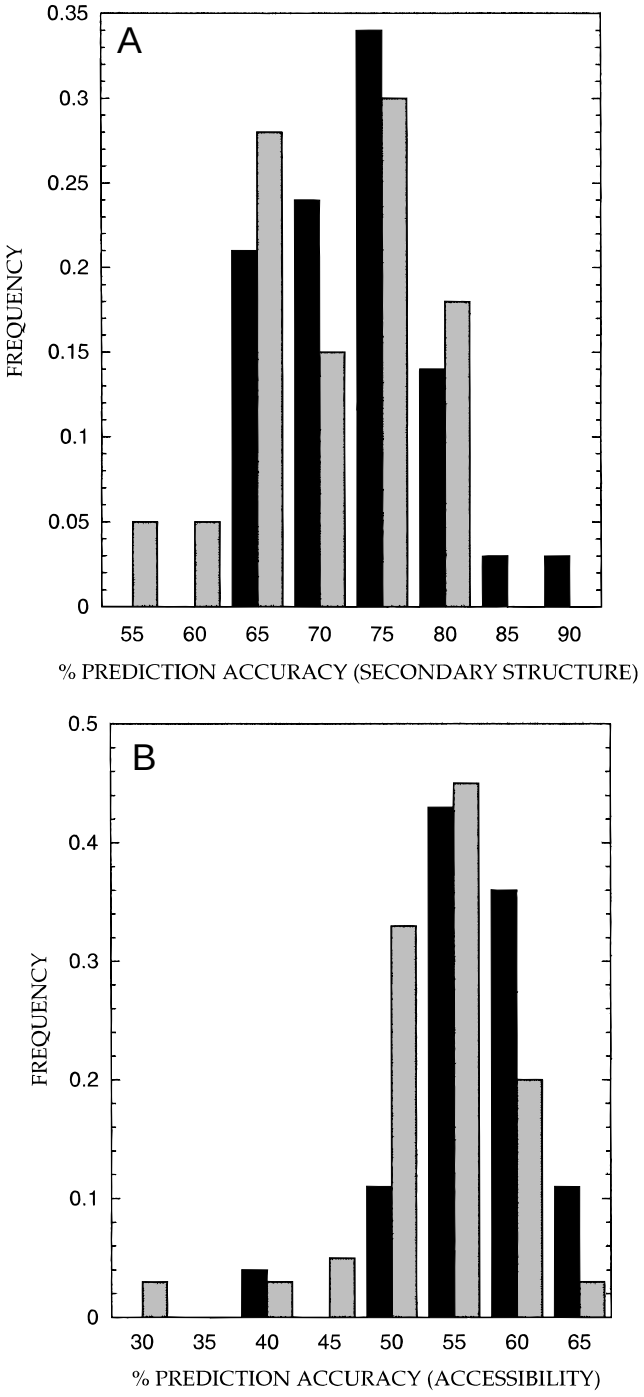

Fig. 2. Prediction accuracy histograms for (**A**) SS and (**B**) AC. The data corresponding to the successes of the method are represented in black. The data corresponding to the failures are represented in grey. In each case the prediction accuracies are computed by comparing the query sequence predictions (from the PHD package, Rost & Sander, 1993) and observed structures (from the DSSP assignments, Kabsch & Sander, 1983).

The previous results indicate that while correct predictions make an important contribution to the success of the method, the present level of structure prediction accuracy is not the main factor discriminating correct from incorrect fold recognition. Two factors have more effect on the correct recognition of the target structure: the degree of similarity between the query and the target structures and the pattern degeneracy problem.

In Figure 3 we display the fraction of secondary structure conserved between the known query and the target structures, for both the successes and the failures of the method. We can see that secondary structure is clearly more conserved for the successes than for the failures of the method. This indicates that even at the present level of secondary structure prediction accuracy, recognition of the target structure will strongly depend on the degree of similarity between the query and the target structures. This is in accordance with the results by Bryant (1996).

Pattern degeneracy in secondary structure and accessibility matching

Usually in PBM the problem of pattern degeneracy refers to the secondary structure pattern degeneracy. It corresponds to the fact that different folds, or parts of them, may have similar 1D structure patterns (Rost et al., 1997). Its origin lies in the information loss due to the projection of the three-dimensional (3D) structure of the protein into a 1D string of symbols (Rost et al., 1997). The information loss is most pronounced in the coil state, because coil residues can map to different zones of the Ramachandran plot. This can be seen in the case of the remote homologue pairs. If we compute the percentage of Ramachandran zone conservation (see Methods), for the aligned residues in the 73 pairs, we can see that it is very high for residues in the alpha or beta states (99.8% and 96.3%, respectively), while it drops clearly (60.7%) when considering residues in the coil state.

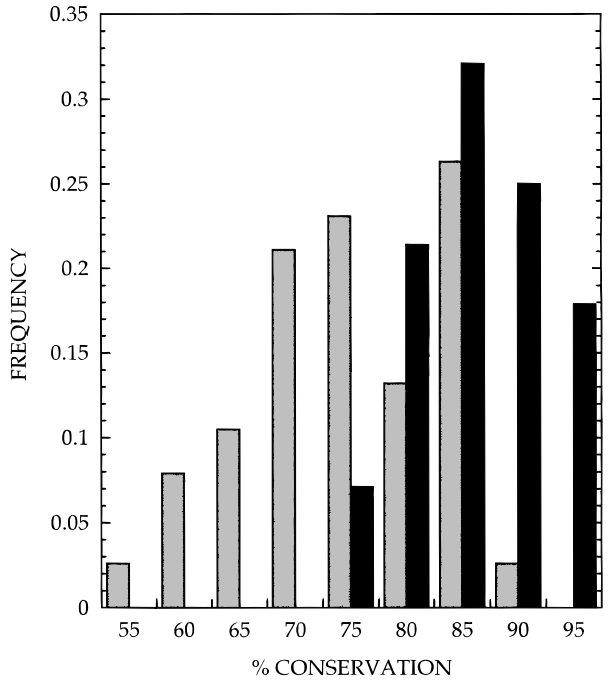

Fig. 3. Histogram of SS conservation between the query and the target structures for the successes (black) and the failures (grey) of the method.

To evaluate, at the score level, the effect of this information loss in the recognition performance of the PBM, we computed the contribution of the coil-based SS terms to the score difference between the correct alignment and the highest ranking incorrect match (Table 2). The results obtained show that the coil term plays an important role in the recognition of the wrong candidate, while the remaining secondary structure elements, alpha and beta, appear to favor more frequently and strongly the recognition of the correct candidate.

Finally, note that secondary structure repetitions, which are very common (e.g., in TIM barrel), exacerbate the pattern degeneracy problem.

The pattern degeneracy problem also affects the accessibility term, as the same accessibility state—buried, exposed, or half-exposed—may be obtained from different combinations of neighbors. However, the average low contribution of the accessibility term to the failures of the method (0.001) indicates that the accessibility cannot distinguish the correct and incorrect matches for the failures at all.

Pattern degeneracy in the sequence term

Sequence effects usually favor the ranking of an incorrect fold in the first position. These unspecific effects are similar to the secondary structure pattern degeneracy, and can be explained by the fact that even for different proteins, hydrophobic residues mainly constitute the core, and hydrophilic residues mainly constitute the surface. Therefore, just by aligning core and surface positions we may obtain positive scores, favoring the alignment between any protein pair, despite underlying structural differences.

The predicted alignments generally comprise correctly and incorrectly aligned stretches of residues (data not shown), a feature obscured by the average alignment shifts. In general, favorable sequence contributions are expected to come only from the correctly aligned residues, e.g., in Figure 4 we see that for the pair (1abe, 1gca), with an average shift of 0.43, the vast majority of the residues are correctly aligned. However, in some cases sequence information may benefit the correct over the incorrect match, even though the alignment is grossly incorrect. For example, for the pair (3blm, 3pte), the correct target is ranked first, favored only by the sequence term, despite an average alignment shift of 48.4. Figure 4 confirms that in this case the number of residue pairs with high shifts is large.

Possible alternatives to improve the performance of prediction-based methods. The previous analysis highlights some of the problems that affect the success of PBM. In this section, we discuss some strategies that could be used to overcome them.

As we have previously seen, the pattern degeneracy problem is the main factor limiting the accuracy of fold recognition using PBM. One approach to eliminate this effect could, in principle, be the use of a normalization scheme to estimate the significance of a match. Bryant and Altschul (1995) suggest a procedure to eliminate unspecific sequence composition effects based on the generation of a reference state by randomly shuffling the aligned residues. In a second step this reference state is used to normalize the score of the alignment. Unfortunately, this procedure cannot be easily applied to scoring functions involving SS terms, due to the high correlation between the SS state of neighboring residues. This problem is likely to affect any correction scheme involving SS

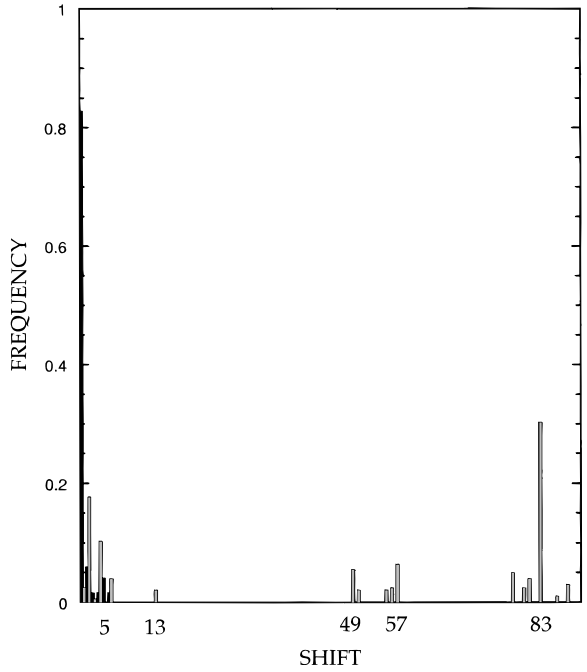

Fig. 4. Frequency distribution of the alignment shifts of the predicted alignments for the pairs 1abe–1gca (black) and 3blm–3pte (grey).

shuffling. For this reason, we decided to explore a different approach to evaluate the reliability of the threading hits and decrease the weight of unspecific effects.

The SS pattern degeneracy problem is due in part to the structural degeneracy of the coil state. In the coil state, residues are characterized by the fact that their ϕ,ψ angles may occupy any allowed zone of the Ramachandran plot. Interestingly, Swindells et al. (1995) have shown that residues in the coil state have well-defined propensities for different zones of the ϕ-ψ map. This suggests that replacing the simple coil state by a set of ϕ-ψ zones may help to improve the recognition ability of PBM.

The feasibility of such an approach relies in the conservation of the ϕ-ψ zones, for the residues in the coil state, across proteins in the same structural families. However, the degree of zone conservation is only 61% for the aligned coil residues in the observed structural alignments and a very similar percentage (57%) was observed in the threading alignments. This suggests a limited usefulness for the proposed approach. A more refined analysis was done in which coil residues were classified according to their accessibility state. The results obtained showed that, for the observed structural alignments, only buried coil residues display a high degree (70%) of ϕ-ψ zone conservation. However, as they only are a small fraction of the coil residues in the protein, it is very unlikely that the use of ϕ-ψ zones for these residues will contribute to increase the success rate of the method. Also a change of ϕ,ψ values for a single residue can critically affect the overall 3D topology of the coil, rendering such predictions very sensitive to errors.

Post-threading analysis

In our threading method, the contribution of all the residues has been given the same weight. However, for remote homologues with similar functions, it is likely that the functionally important residues will be more conserved. Therefore, if we have information to identify these residues (from multiple sequence alignments or structural data), these data should be used to aid recognition. To this end we have utilized a tool already developed in our laboratory, the program SAS (Milburn et al., 1998), to annotate sequence alignments with information extracted from the Protein Data Bank (PDB) file (Bernstein et al., 1977) of the target sequence, e.g., SITE records, contacts with the ligand, etc. Unfortunately, the results obtained are limited by the fact that such information is often not available.

For the 21 pairs in our database for which SITE records were available, only for the pair of protein tyrosine phosphatases (2hnq, 1yts) did the observed coincidence suggest a common function. In this case, 1yts was the top hit in the threading method. The annotated alignment is shown in Figure 5. Protein 1yts shows two segments of active-site residues, residues 350–360 and residues 402–410, according to the PDB file records. It can be seen that for 2hnq several of these residues are conserved. These matches suggest that the query protein is likely to have the same function as the target protein. We have not attempted to quantify the likelihood of these matches, as this is beyond the scope of this paper.

Conclusions

Our results indicate that by enriching the number of remote homologues per fold in the searched database, the average accuracy of PBM may increase substantially. We have also observed that the accuracy level reached by secondary structure and accessibility predictions is often sufficient to allow their use in the recognition of remote homologues by PBM, without introducing any major source of error. Interestingly, at this stage the degree of structural similarity between the query and the target structure becomes a more relevant factor. However, the pattern degeneracy problem, a consequence of using 1D information, is probably the main problem affecting fold recognition by PBM. This suggests two directions for the improvement of prediction-based fold recognition methods: use of more specific function-related sequence information, as shown by our use of structure-derived sequence annotations, or the introduction of 3D information such as that used in distance-based threading methods, in the form of additional terms to the scoring function. We are at present exploring these two alternatives.

Methods

The set of remote homologue pairs

We used a set of 73 remote homologue pairs derived from the set provided by Russell et al. (1997), after eliminating all the pairs where one of the proteins had missing or unknown residues. In this paper, the first and the second protein in each pair will be considered as the query and the target sequences, respectively.

The searched database

The searched database was a set of 627 proteins, with less than 25% sequence identity between any pair, derived from the set provided by Rost (1996) after eliminating all those proteins having missing or unknown residues.

Fig. 5. Annotated predicted threading alignment between the two tyrosine-phosphatases 2hnq and lyts. 2hnq was used as the query sequence to search the structure database and lyts; the remote homologue of 2hnq was the first hit found by our threading method. The alignment shift errors were 1.7 and 3.8 residues, relative to the SSAP (Taylor & Orengo, 1989) and STAMP (Russell & Barton, 1992) alignments, respectively. Residues colored in red correspond to 1yts active-site residues, according to the PDB file records (Bernstein et al., 1977). The upper case is used for the 1yts sequence to indicate that information on the active-site residues is available. The display was produced using the in-house program SAS (Milburn et al., 1998).

Due to the nature of the database, there are several remote homologues for each of the 73 query sequences. However, when performing a search with a given query sequence, only the remote homologue coinciding with the target sequence was kept in the database. This provided a much more conservative measure of the success of our method (see Results and discussion).

Fold recognition performance and alignment quality

We measured the fold recognition performance of our method by the percentage of correct first hits computed as follows:

$$\frac{100 \cdot nf}{N} \tag{1}$$

where *nf* is the number of pairs for which the query sequence ranked its corresponding target sequence in the first position; *N* is the number of test cases, 73 in our case.

The alignment accuracy was measured using the alignment mean shift error, computed as follows:

$$\frac{100 \cdot \sum_{i=1}^{Nal} |\text{shift}_i|}{Nal} \tag{2}$$

summed over all the query sequence residues present both in the threading and the structural alignments. *Nal* is the total number of these residues. The shift_i is the residue shift between the correct (structural) alignment and the predicted (threading) alignment for a given query sequence residue.

Effect of having multiple target structures in the fold recognition performance

To test the effect of the number of targets/query, the performance of our method was computed using all the homologues of each

query sequence present in the target database, rather than only using one target per query. However, some of the query sequences in our test set belonged to the same SCOP superfamily (Murzin et al., 1995). To avoid any bias due to this fact, we generated a collection of 50 reduced test sets. Each set was constituted by 35 query sequences, randomly chosen from the 73 query sequences of our initial test set, and each belonging to different superfamilies. Finally, the performance of the method was averaged over the values corresponding to the 50 sets.

Structural alignments

Due to the ambiguity in structural alignments (Godzik, 1996; Zu-Kang & Sippl, 1996), two sets of structural alignments were used in this paper: one is available from Russell et al. (1997), who derived them using the STAMP program (Russell & Barton, 1992), and the other set was obtained using the SSAP program (Taylor & Orengo, 1989).

The threading procedure

Our threading procedure is based in the use of a typical dynamic programming algorithm (Needleman & Wunsch, 1970) and a scoring function combining sequence information with SS and AC information derived from predictions. Given a pair (i, j) of aligned residues, i belonging to the query sequence and j to a given protein from the searched database, their contribution, sc_{ij}, to the alignment score was obtained utilizing the following equation:

$$sc_{ij} = w_{seq} \,.\, m_{ij} + w_{stct} \,.\, (ss_{ij} + acc_{ij}) \tag{3}$$

where m_{ij}, ss_{ij}, and acc_{ij} correspond to the sequence, SS, and AC contributions, respectively, and w_{seq} and w_{stct} correspond to their respective weights. Note that the secondary structure and accessibility terms have the same weight.

The sequence contribution to the score, m_{ij}, was obtained using the normalized PAM250 matrix (Dayhoff et al., 1978) from the GCG program manual (GCG, 1994).

The SS and AC terms were derived utilizing database probabilities p. The formula used for the secondary structure case was

$$sc = \ln\left[p\left(\frac{\mathrm{OBS}_{ss}}{\mathrm{PRED}_{ss} \cap rl}\right)\right] - \langle \ln[\ldots] \rangle \tag{4}$$

where sc corresponds to the score for aligning a residue in the target protein, with observed secondary structure (OBS_{ss}), to a residue in the query sequence, with predicted secondary structure (PRED_{ss}), and an associated reliability index (rl). OBS_{ss} and PRED_{ss} may be equal to one of the three secondary structure states: H (helix), E (strand), or C (coil). The normalization term $\langle \ln[\ldots] \rangle$ is equal to

$$\langle \ln[\ldots] \rangle = \frac{1}{3} \sum_{i=1}^{3} \ln\left[p\left(\frac{\mathrm{OBS}_{ss}^{i}}{\mathrm{PRED}_{ss} \cap rl}\right)\right] \tag{5}$$

where OBS_{ss}^{i} goes through all the three possible SS states. The score for the AC was derived in exactly the same way.

The probabilities p in Equation 3 were obtained from the query sequences in our set of 73 remote homologue pairs by aligning, for each residue, its assigned (using the DSSP program, Kabsch & Sander, 1983) and predicted (using the PHD program, Rost & Sander, 1993) SS. That is to say, the probability to observe a residue in the helix state when it is predicted to be in a beta state, with reliability 7, is equal to

$$p\left(\frac{H}{E \cap 7}\right) = \frac{n_{H,E,7}}{n_{E,7}} \tag{6}$$

where $n_{H,E,7}$ is the total number of residues having H as observed secondary structure and E as predicted secondary structure with reliability 7. The denominator $n_{E,7}$ is the total number of residues having E as predicted SS with reliability 7.

The final comparison matrix was linearly scaled so that the maximum and minimum sc values were equal to 1 and -1, respectively (Rost, 1996).

The gap function used was a simple linear gap scheme typically used in PBM (Fischer & Eisenberg, 1996; Rice & Eisenberg, 1997; Rost et al., 1997) in which the penalty for opening a gap involving n residues is given by

$$g_o + g_e \cdot n \tag{7}$$

where g_o and g_e are the gap opening and elongation penalties, respectively. No end gap penalties were applied.

The penalty gap values g_o and g_e, as well as those of the two weights w_{seq} and w_{stct}, were obtained following a simple optimization procedure in which our threading method was applied to a set of 73 protein pairs for different sets of values of these parameters. The objective function used for the optimization procedure was the percentage of correct first hits (see above). The different sets were obtained as follows: g_o was systematically varied between -1 and -9 at intervals of -2 U and g_e was set to $0.1 g_o$ (Rost, 1996). The weight w_{seq} was varied between 0 and 1, at intervals of 0.25 U, and w_{stct} was set equal to $1 - w_{seq}$. Finer intervals were not utilized to avoid overfitting problems. During the optimization runs, the values of the probabilities in Equation 3 were computed using a jackknife procedure. Also, for each query sequence, the jackknifed SS and AC predictions were used. In our case we are interested in studying the factors limiting the performance of the PBM, rather than comparing the performance of a newly developed method relative to already developed methods. Therefore, the use of optimal parameters does not affect the conclusions reached. Note that the values of the optimal parameters depend on the distribution of sequence similarities in the test set used. However, due to the careful procedure followed by Russell et al. (1997), to select their set of remote homologue pairs, as well as because of the final number of pairs, 73, the test set used reasonably reflects an average threading scenario. The best results in the optimization runs were then obtained for weights equal to 0.75 and 0.25 for the sequence and structure terms, respectively. Note that this does not mean that the sequence contribution to the final alignment score is higher than that of the structure term. The optimal gap opening and elongation penalties were 3.0 and 0.3, respectively. To test that the results obtained do not depend on some pathological characteristic of the test set, we repeated the optimization runs after eliminating, in turn, each of the 73 pairs. The optimal parameters obtained for the 73 optimization runs were the same, suggesting that the parameters listed are not significantly biased.

The final score *SC* for the alignment between two sequences was given by

$$SC = \Sigma_{(i,j)} sc_{ij} + \Sigma_k (g_o + n_k \cdot g_e) \quad (8)$$

where the indexes (i, j) and k run over all the aligned pairs and the opened gaps, respectively; n_k is the size of k^{th} gap.

Score decompositions

To understand how different factors such as SS prediction accuracy contribute to the performance of the threading method, we decided to use a score decomposition in which each score was divided into three main terms: the sequence, structure, and gap contributions. The structure term was divided in two different ways: contributions from the SS and AC residue states, and contributions from the correctly and incorrectly predicted residue structural states. Finally, the SS contribution was divided into the alpha + beta and the coil terms. An example of the score decomposition is given in Table 3.

These score decompositions can tell us which are the main contributions to the optimal alignment between the query sequence and a database protein. However, given a query sequence, the correct ranking of the corresponding target protein also depends on whether the score for their alignment is better than that of the alignment between the query sequence and the best scoring nonhomologous protein. Therefore, to assess the relevance of the above-mentioned effects, we need to understand their contribution to the difference between the two scores. To that end, for each of the 73 query sequences, we compared the score decomposition of its alignment with the target protein, the remote homologue alignment (RHA), against that of the highest scoring alignment with a nonhomologous protein, the nonhomologue alignment (NHA). Before comparison, the scores and their components were normalized by the number of aligned pairs. The few cases for which the raw and the normalized scores had different relative rankings for the RHA and the NHA were discarded. To avoid any confusion, note that the RHA utilized at this stage is the alignment generated by our threading method, not the structural alignment.

Table 3. *Score decomposition for the remote homologue pair 1abe-1gca*

	Raw score[a]
Sequence	33.00
Structure[b]	64.97
Gap	−29.7
Structural decomposition[c]	
Correctly predicted[d]	67.03
Incorrectly predicted[e]	−2.07
SS decomposition[c]	
SS[f]	41.54
AC[g]	23.43
SS decomposition[h]	
Alpha + beta[i]	26.96
Coil[j]	14.58

[a] In this column are listed the non-normalized contributions of each term cited in the first column.
[b] Total contribution of the structure term (SS + AC).
[c] Decomposition of the structure term.
[d] Contribution due to correct predictions.
[e] Contribution due to incorrect predictions.
[f] Contribution due to the SS term.
[g] Contribution due to the AC term.
[h] Decomposition of the SS term.
[i] Contribution due to residues from the candidate protein (1gca) in helix and sheet.
[j] Same as foonote i for the coil residues.

To quantify the contribution of the different terms of interest (e.g., sequence, structure, etc.) to the successes or failures of the threading method, we divided the 73 test cases in two sets. One set corresponded to the 28 cases where the correct target was identified, and the second set corresponded to the 38 cases where the nontarget protein was ranked in the first position. Then, for a given term X (X = sequence, structure, etc.), the average contribution to the successes or failures of the method is computed as follows:

$$\frac{\sum_i^N [X_i(\text{RHA}) - X_i(\text{NHA})]}{N} \quad (9)$$

where the sum may run over the $N = 28$ successes or $N = 38$ failures of the method. X_i(RHA) and X_i(NHA) are the contributions of the term X to the score of the RHA and NHA alignments, respectively, for the i^{th} case. Notice that a positive sign indicates a favorable contribution of term X to the RHA, while a negative sign indicates a favorable contribution to the NHA.

The ϕ-ψ regions

The ϕ-ψ regions used in this paper are the ones defined by Swindells et al. (1995). They defined four main regions: a (right-handed alpha), b (beta nonaccessible to Pro), p (beta accessible to Pro), and L (left-handed helix). To simplify, we have considered only one beta state, B, as the division between b and p is not relevant for the purposes of this paper.

Acknowledgments

X. de la Cruz acknowledges the support of a "Human Frontier Science Program" fellowship. He also would like to acknowledge Dr. B. Rost for kindly lending us a copy of his SS prediction program PHD, and Drs. B.K. Lee, S.H. Bryant, and A. Marchler-Bauer for useful discussions. This is a publication from the Bloomsbury Center for Structural Biology, funded by the BBSRC Grant no. 31/JRI07365 for computing.

References

Aurora R, Rose G. 1998. Seeking an ancient enzyme in *Methanococcus jannaschii* using ORF, a program based on predicted secondary structure comparisons. *Proc Natl Acad Sci USA 95*:2818–2823.

Bernstein F, Koetzle T, Williams G, Meyer E, Brice M, Rodgers J, Kennard O, Shimanoushi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535–542.

Bowie JU, Luthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known 3-dimensional structure. *Science 253*:164–170.

Bryant SH. 1996. Evaluation of threading specificity and accuracy. *Proteins Struct Funct Genet 26*:172–185.

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 Struct Funct Genet 5*:92–112.

Dayhoff M, Schwartz RM, Orcutt BC. 1978. A model of evolutionary change in proteins. In: Dayhoff M, ed. *Atlas of protein sequence and structure*, vol 5

(suppl 3). Silver Spring, Maryland: National Biomedical Research Foundation. pp 345–352.
Fischel-Ghodsian F, Mathiowitz G, Smith T. 1990. Alignment of protein sequences using secondary structure: A modified dynamic programming method. *Protein Eng 3*:577–581.
Fischer D, Eisenberg D. 1996. Protein fold recognition using sequence-derived predictions. *Protein Sci 5*:947–955.
GCG. 1994. *Program manual for the Wisconsin package*, version 8. Madison, Wisconsin: Genetics Computer Group.
Godzik A. 1996. The structural alignment between two proteins—Is there a unique answer? *Protein Sci 5*:1325–1338.
Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.
Kabsch W, Sander C. 1983. A dictionary of protein secondary structure. *Biopolymers 22*:2577–2637.
Kocher J-P, Rooman MJ, Wodak SJ. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J Mol Biol 235*:1598–1613.
Lander ES. 1996. The new genomics—Global views of biology. *Science 274*:536–539.
Lemer CM-R, Rooman MJ, Wodak SJ. 1995. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins Struct Funct Genet 23*:337–355.
Marchler-Bauer A, Bryant SH. 1997. A measure of success in fold recognition. *TIBS 22*:236–240.
Marchler-Bauer A, Levitt M, Bryant S. 1997. A retrospective analysis of CASP2 threading predictions. *Proteins Struct Funct Genet (Suppl. 1)*:83–91.
Martin ACR, MacArthur MW, Thornton JM. 1997. Assessment of comparative modeling in CASP2. *Proteins Struct Funct Genet (Suppl. 1)*:14–28.
Milburn D, Laskowski R, Thornton JM. 1998. Sequences annotated by structure: A tool to facilitate the use of structural information sequence analysis. *Protein Eng 11*:855–859.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP—A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol 247*:536–540.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J Mol Biol 48*:443–453.
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.
Rice D, Eisenberg D. 1997. A 3D–1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J Mol Biol 267*:1026–1038.
Rice D, Fischer D, Weiss R, Eisenberg D. 1997. Fold assignments for amino acid sequences of the CASP2 experiment. *Proteins Struct Funct Genet (Suppl 1)*:113–122.
Rost B. 1996. Protein fold recognition by merging 1D structure and sequence alignments. EMBL Heidelberg, Germany, WWW document (http://www.embl-heidelberg.de/~rost/Papers/96PreTopits.html).
Rost B, Sander C. 1993. Prediction of secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.
Rost B, Schneider R, Sander C. 1997. Protein fold recognition by prediction-based threading. *J Mol Biol 270*:471–480.
Russell RB, Barton G. 1992. Multiple protein sequence alignment from tertiary structure comparison: Assignment of global and residue confidence levels. *Proteins Struct Funct Genet 14*:309–323.
Russell RB, Copley RR, Barton GJ. 1996. Protein fold recognition by mapping predicted secondary structures. *J Mol Biol 259*:349–365.
Russell RB, Saqi MAS, Bates P, Sayle RA, Sternberg MJE. 1998. Recognition of analogous and homologous protein folds—Assessment of prediction success and associated alignment accuracy using empirical substitution matrices. *Protein Eng 11*:1–9.
Russell RB, Saqi MAS, Sayle RA, Bates PA, Sternberg MJE. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J Mol Biol 269*:423–439.
Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. 1995. Evaluation of comparative protein modeling by MODELER. *Proteins Struct Funct Genet 23*:318–326.
Samudrala R, Pedersen JT, Zhou H-B, Luo R, Fidelis K, Moult J. 1995. Confronting the problem of interconnected structural changes in the comparative modeling of proteins. *Proteins Struct Funct Genet 23*:327–336.
Sheridan RP, Dixon JS, Venkataraghavan R. 1985. Generating plausible protein folds by secondary structure similarity. *Int J Pept Protein Res 25*:132–143.
Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino-acid-sequences of unknown 3-dimensional structure in a data-base of known protein conformations. *Proteins Struct Funct Genet 13*:258–271.
Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. *J Mol Biol 147*:195–197.
Swindells MB, MacArthur MW, Thornton JM. 1995. Intrinsic ϕ,ψ propensities of amino acids, derived from the coil regions of known structures. *Nature Struct Biol 2*:596–603.
Taylor WR, Orengo CA. 1989. Protein structure alignment. *J Mol Biol 208*:1–22.
Westhead DR, Collura VP, Eldridge MD, Firth MA, Li J, Murray CW. 1995. Protein fold recognition by threading—Comparison of algorithms and analysis of results. *Protein Eng 8*:1197–1204.
Zu-Kang F, Sippl MJ. 1996. Optimum superimposition of protein structures: Ambiguities and implications. *Folding Design 1*:123–132.