J. Mol. Biol. (1997) **270**, 471–480

JMB

Protein Fold Recognition by Prediction-based Threading

Burkhard Rost[1,2]*, Reinhard Schneider[1] and Chris Sander[1,2]

[1]*EMBL, 69012, Heidelberg Germany*

[2]*EBI, Cambridge, CB10 1RQ UK*

**Corresponding author*

In fold recognition by threading one takes the amino acid sequence of a protein and evaluates how well it fits into one of the known three-dimensional (3D) protein structures. The quality of sequence-structure fit is typically evaluated using inter-residue potentials of mean force or other statistical parameters. Here, we present an alternative approach to evaluating sequence-structure fitness. Starting from the amino acid sequence we first predict secondary structure and solvent accessibility for each residue. We then thread the resulting one-dimensional (1D) profile of predicted structure assignments into each of the known 3D structures. The optimal threading for each sequence-structure pair is obtained using dynamic programming. The overall best sequence-structure pair constitutes the predicted 3D structure for the input sequence. The method is fine-tuned by adding information from direct sequence-sequence comparison and applying a series of empirical filters. Although the method relies on reduction of 3D information into 1D structure profiles, its accuracy is, surprisingly, not clearly inferior to methods based on evaluation of residue interactions in 3D. We therefore hypothesise that existing 1D-3D threading methods essentially do not capture more than the fitness of an amino acid sequence for a particular 1D succession of secondary structure segments and residue solvent accessibility. The prediction-based threading method on average finds any structurally homologous region at first rank in 29% of the cases (including sequence information). For the 22% first hits detected at highest scores, the expected accuracy rose to 75%. However, the task of detecting entire folds rather than homologous fragments was managed much better; 45 to 75% of the first hits correctly recognised the fold.

Keywords: protein structure prediction; threading; remote homology detection; fold recognition; secondary structure

Introduction

Reducing the sequence-structure gap by homology modelling

Large scale gene-sequencing projects accumulate gene data, and consequently protein sequences, at a breathtaking pace (Oliver *et al.*, 1992; Fleischmann *et al.*, 1995; Dujon, 1996; Johnston, 1996) . However, information about three dimensional (3D) structure is available for only a small fraction of known proteins (Bernstein *et al.*, 1977). Thus, although experimental structure determination has improved (Lattman, 1994), the sequence-structure gap continues to increase. One of the main tasks of theoretical biology is to reduce this gap by predictions. However, the only somewhat reliable way to predict 3D structure is homology modelling (Greer,

Abbreviations used: 3D, three-dimensional; 1D, one-dimensional; rmsd, root-mean-square deviation; *U*, protein sequence of unknown 3D structure (e.g. search sequence); PDB, Protein Data Bank of experimentally determined 3D structures of proteins; SWISS-PROT, database of protein sequences; DSSP, database containing the secondary structure and solvent accessibility for proteins of known 3D structure; FSSP, database of remote homologues of known 3D structure; MaxHom, dynamic programming algorithm for conservation weight based multiple sequence alignment; PHD, profile-based neural network prediction of secondary structure (PHDsec) and solvent accessibility (PHDacc).

0022–2836/97/280471–10 $25.00/0/mb971101

1991; Lesk & Boswell, 1992; May & Blundell, 1994): the structure of a protein of unknown structure (dubbed *U*) can be modelled by homology if a protein of known 3D structure is found which has more than 25 to 30% pairwise sequence identity with *U* (Chothia & Lesk, 1986; Doolittle, 1986; Sander & Schneider, 1991).

Possible scope of remote homology modelling

Two naturally evolved proteins can have rather different sequences and still fold into homologous structures. Currently there are thousands of remote homologues, i.e. homologues with less than 25% pairwise sequence identity, stored in a database of structurally aligned remote homologues (Holm *et al*., 1993; Holm & Sander, 1994). To illustrate the possible scope of remote homology modelling by numbers: homology modelling is currently applicable to over 11,000 SWISS-PROT (Bairoch & Apweiler, 1996) sequences with more than 25% pairwise sequence identity to a known structure. However, the majority of all homologous proteins has supposedly less than 25% pairwise sequence identity (unpublished results). Thus, for a significant fraction of the currently known sequences remote homology modelling could yield 3D predictions.

Long way from fold recognition to remote homology modelling

The problem of detecting remote homologues is of the ''needle in the haystack'' type: aligning the unique folds (150) against the entire PDB (3000) would yield 450,000 pairs, of which about 1500 are remote homologues (Holm & Sander, 1994), i.e. the goal is to find the one true homologue among 100 to 300 decoys. A test of threading methods at the first meeting to evaluate structure prediction accuracy (Moult *et al*., 1995) suggested levels of 10 to 40% accuracy in correctly detecting the homologous fold (Lemer *et al*., 1995; Shortle, 1995). However, detection of the homologue is the simpler part of a successful remote homology modelling. More problematic is to correctly align the homologous proteins and to correctly build the model (Bryant & Altschul, 1995; Lemer *et al*., 1995; Sippl, 1995). Only for a few cases has threading been shown to yield correct 3D models (Flöckner *et al*., 1995).

Here, we extend our previously proposed novel method for threading predictions of one-dimensional (1D) structure into 3D structures (Rost, 1995a,b). First, 1D structure profiles were predicted from multiple sequence alignments. Then, the 1D predictions were aligned to 1D projections of known structures. The novel aspect reported here was the combination of information from 1D predictions and sequences. We had to focus on the main aspects of the method, a detailed description of the algorithm is electronically available (Rost, 1996c). The accuracy of the method in detecting remote homologues was evaluated on a data set of 89 unique protein folds. The ability to correctly build remote homologous models was investigated for all correctly detected remote homologues. Finally, we compared the performance of the method to other tools based on three different data sets.

Methods

Brief outline of the algorithm

The algorithm started from a protein sequence which was aligned by MaxHom (Sander & Schneider, 1991) against SWISS-PROT (Bairoch & Boeckmann, 1994) (Figure 1). The resulting multiple sequence alignment was used as the input to neural network systems predicting secondary structure (PHDsec; Rost & Sander, 1994a) and solvent accessibility (PHDacc; Rost & Sander, 1994b). The predictions were converted into 1D structural profiles. Up to this point the method was constrained to a straight prediction in 1D, i.e. without any reference to 3D structure or the final goal of threading. Effectively, the amino acid sequence had now been translated into a 1D string of structure symbols (''predicted structure profile''), with some cooperativity taken into account. The idea was now to find the 3D fold that had the most similar structure profile (in terms of secondary structure and accessibility). The next step was to represent each of the known folds in the database as an observed structure profile. Finally, predicted and observed 1D structure profiles were optimally aligned by a dynamic programming algorithm (MaxHom). The best hit of the alignment procedure was recorded, and the final best hit was taken as the predicted fold. The predicted 3D structure was modelled based on the alignment of the input sequence into the predicted fold.

Alignment of 1D structure

Three alternatives for the aligned strings

For a practical application of the method, predicted 1D structure profiles were aligned to observed 1D structure profiles (PHD *versus* PDB). To investigate the influence of the accuracy of 1D structure prediction, we performed the following calibration experiment: observed 1D structure profiles were aligned against observed 1D structure profiles (PDB *versus* PDB). Another possible extension of the concept was the alignment of predicted against predicted 1D structure profiles (PHD *versus* PHD). Such a search could yield a prediction of a fold identity between two proteins both of unknown structure. Alignments of 1D structure strings can reveal structural homologues as 1D structure is conserved between remote homologues (Rost, 1996b).

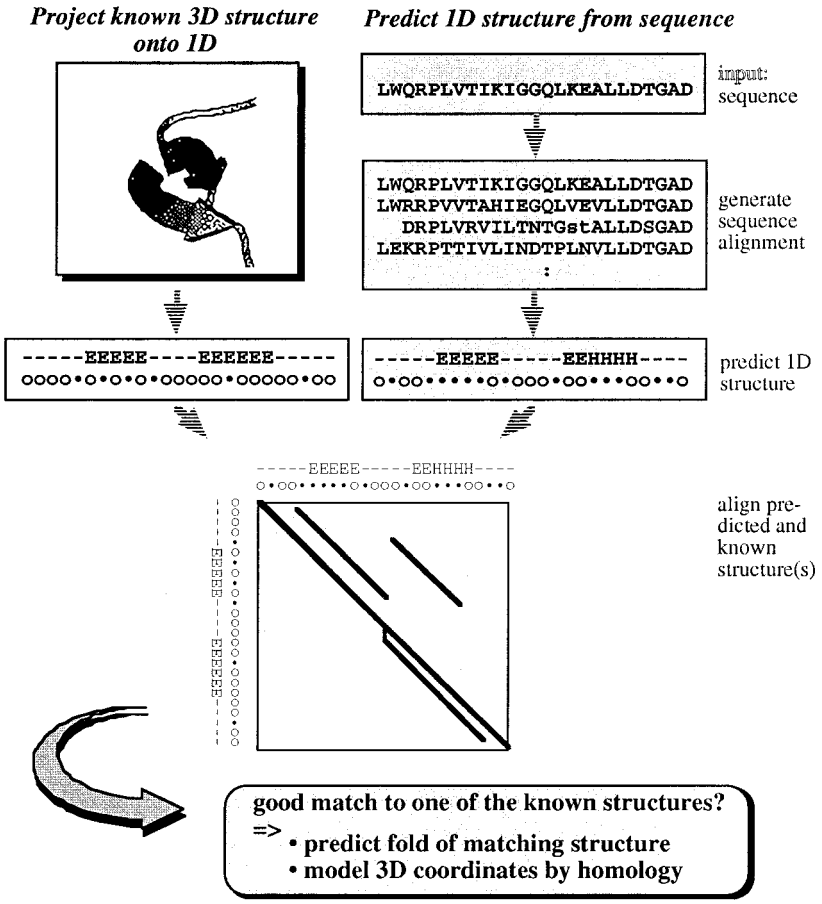

Figure 1. Threading predicted 1D structure profiles into known 3D structures. (1) A multiple sequence alignment is generated for a given sequence of unknown structure (*U*). (2) The alignment profile of *U* is used as the input to a neural network system (PHD) that predicts secondary structure and relative solvent accessibility. (3) The resulting predicted 1D structure profile for *U* is aligned by dynamic programming (program MaxHom; Sander & Schneider, 1991) to 1D structure strings assigned from known structures by the program DSSP (Kabsch & Sander, 1983). Abbreviations: H, helix; E, strand; L, rest; ●, buried (<15% solvent accessible); ○, exposed (⩾15% solvent accessible).

Free parameters for dynamic programming

The predicted strings were aligned based on a Smith-Waterman type dynamic programming algorithm (Smith & Waterman, 1981). This algorithm was implemented in the program MaxHom (Sander & Schneider, 1991; Schneider, 1994). The following free parameters had to be adjusted: (1) the similarity matrix, and (2) the penalties associated with the introduction of gaps in the alignment.

Similarity matrix for six states

Various strategies were explored to find the optimal matrix for weighting matches between 1D structure pairs (Rost, 1995a,b). Here we used a matrix refined and starting from database counts (Rost, 1996c). Finally, we simplified the resulting matrix by making it symmetrical and slightly more balanced.

Similarity for 120 states

The combination of information from 1D structure and sequence was accomplished by combining the 1D structure similarity matrix (Figure 3 of Rost, 1996c) with a McLachlan (McLachlan *et al.*, 1984) or a Blosum62 (Henikoff & Henikoff, 1992) exchange matrix:

$$M_{ij} = \alpha \times M_{ij}^{\text{1D structure}} + (100 - \mu) \times M_{ij}^{\text{sequence}} \quad (1)$$

where M_{ij} determined the score for a match at a given position between state i in the first string and state j in the second string, and $\mu = 0$ to 100 tuned the percentage of 1D structure contribution to the final alignment score E (note that $\mu = 0$ corresponded to a simple sequence alignment; $\mu = 100$ marked an alignment based on 1D structure only).

Gap open and gap elongation penalty

The optimal choice of gap penalties depends on the context, i.e. the particular alignment pair (Vingron & Waterman, 1994). For an alignment of a search sequence against a database, there is a trade-off between coverage (correct hits found *versus* all possible correct hits) and accuracy (correct hits *versus* all hits found) of detection for the choice of the gap parameters go (penalty for opening a gap) and ge (penalty for continuing an open gap). We compiled results for various gap open penalties. The relative values of the two were found to be of marginal importance; we used: ge = 0.1× go.

Evaluation of prediction accuracy

Cross validation and parameter optimisation

To ascertain that knowledge about structure was not used for the 1D prediction we used prediction networks that had been trained on proteins with less than 25% pairwise sequence identity with the predicted protein (cross validation). Furthermore, free parameters for the dynamic programming algorithm were optimised before the final results were compiled. This was achieved by varying free parameters based on a data set of 46 non-unique protein structures (listed by Rost, 1995b).

Measuring the accuracy of fold recognition

Prediction accuracy was defined as the cumulative percentage of correct predictions up to rank R, $Q(R)$ (defined in equation 6 of Rost, 1996c). To measure the accuracy obtained on subsets selected according to a fixed z-score (equation 5 of Rost, 1996c) we defined $\mathrm{Cor}^R(\theta)$ as the cumulative percentage of correct hits up to rank R for a given threshold $z > \theta$ (equation 7 of Rost, 1996c). The corresponding coverage was $\mathrm{Cov}^R(\theta)$, defined as the percentage of hits found at R for θ (equation 8 of Rost, 1996c). $\mathrm{Cor}(\theta)$ and $\mathrm{Cov}(\theta)$ determined the trade-off between accuracy and coverage. Results will be given for first ranks ($R = 1$), only. The definitions for coverage and accuracy *versus* a given cut-off address the following questions. What is the expected accuracy to find correct homologues if the hit list is cut at rank R and at a z-score $> \theta$? And for what proportion of the proteins are predictions made at the given cut-off?

Measuring accuracy of remote homology modelling

We measured alignment quality by (1) the percentage of pairwise sequence identity between the predicted and the structural alignment; (2) the average number of residues shifted between the predicted and the structural alignment; and (3) an alignment shift score (equation 9 of Rost, 1996c). For the quality of the model we simply determined backbone root mean square deviations (rmsd; Sippl, 1982). The superposition was based on the sequence alignment obtained from the threading without any further optimisation (loop regions were included when compiling the rmsd values). We regarded the structural alignments taken from the FSSP database (Holm & Sander, 1994) as the correct ''standard-of-truth''. However, alignments between two structures are not always unique (Zu-Kang & Sippl, 1996) . In some cases the alternative correct structural alignment might have fitted the prediction better.

Data sets used for validation

Set of 89 unique folds

As of early 1996, there were more than 200 unique protein folds in the PDB (Holm & Sander, 1994). These were used as a starting point to compile a set of 89 proteins used to evaluate the accuracy in detecting remote homologues (Rost, 1996a). The resulting list of remote homologues comprised a rather difficult test set, as it included many cases for which the structural alignment covered only fragments of the two aligned proteins rather than extending over the entire ''folds''. Consequently, the results provided conservative estimates for the accuracy of 1D structure threading. The ''correct''

Table 1. Accuracy of fold detection

Method	Seq. matrix	smax (3)	go	μ (2)	$Q(1)$ (6)	$\mathrm{Cor}^1(z > x)$ (7) with $x =$ 4.5	3.5	$\mathrm{Cov}^1(z > x)$ (8) with $x =$ 4.5	3.5
PDB *versus* PDB	–	1	4	100	35 ± 5	100	85	7	22
PHD *versus* PDB	–	1	4	100	23 ± 4	100	33	2	23
PHD *versus* PDB	McLachlan	1	2	50	28 ± 5	80	63	5	12
PHD *versus* PDB	Blosum62	2	2	50	29 ± 5	88	75	10	22
PHD *versus* PHD	Blosum62	2	2	50	27 ± 4	100	85	10	15
Sequence only	McLachlan	3	2	0	16 ± 4	30	29	25	26
Random pick	–	–	–	–	2				

Accuracy scores (equation numbers in italics refer to Rost, 1996c): $Q(1)$, percentage of correct first hits (± values estimate one standard deviation; $\mathrm{Cor}^1(z > x)$, percentage of correct first hits with a z-score $> x$, data given for $x = 4.5$ and 3.5 (values refer to small subsets of 1 to 23 proteins, thus the standard deviations are >10); $\mathrm{Cov}^1(z > x)$, percentage of proteins for which the first hit reached a z-score $> x$; go, gap open penalty; smax, maximal entry of the normalised combined matrix; Seq. matrix, matrix used for the sequence part (equation (1)); and μ, influence of 1D structure relative to sequence (equation (1)), i.e. $\mu = 100$, only 1D structure aligned; $\mu = 50$, 1D structure : sequence = 50 : 50, $\mu = 0$, only sequence alignment. The abbreviations for the methods refer to the alternatives of aligning known structures against a database of known structures (i.e. optimal prediction scenario, PDB *versus* PDB), predicted structures against a database of known structures (i.e. the fold-detection scenario, PHD *versus* PDB) and predicted structures against a database of predicted structures (i.e. detection of remote homology between pairs of unknown structure, PHD *versus* PHD); Random pick refers to the likelihood of hitting the correct remote homologue in a list of 723 proteins by chance (significantly higher than 1/723 as for some of the 89 proteins there was more than one remote homologue in the list of 723).

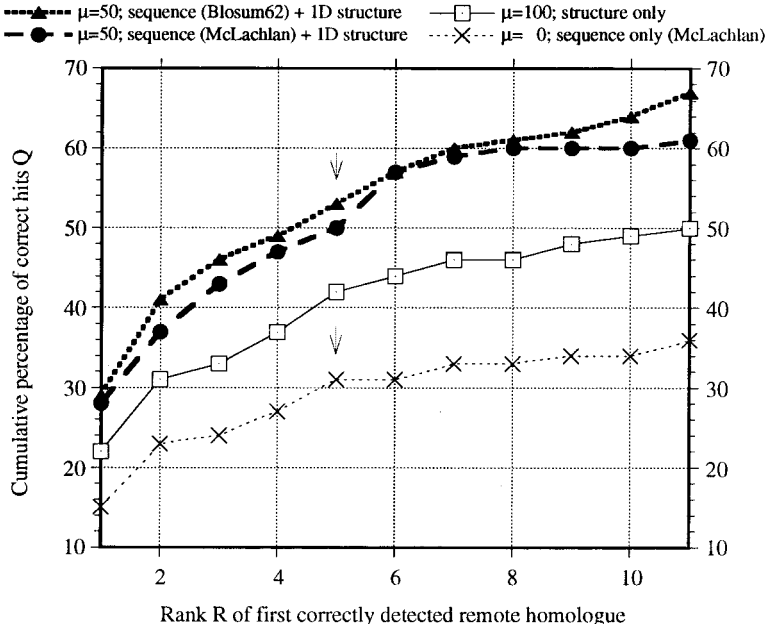

Figure 2. Cumulative accuracy of detection *versus* rank of hit. How many of the homologues were detected up to a certain rank R of the alignment list? For ranks $R = 1$ to 11, the cumulative percentages of correctly detected folds is shown ($Q(R)$, for $R = 1, ..., 11$; see equation 6 of Rost, 1996c). Thick lines: alignments mixing 1D structure and sequence 50:50 ($\mu = 50$, equation (1); filled circles, McLachlan exchange matrix; filled triangles, Blosum62 matrix. Thin lines: alignments based on sequence (McLachlan matrix, crosses) and on 1D structure information only (open squares). For all results, 1D structure information was obtained by cross-validated prediction (PHD), i.e. the knowledge about the 3D structure of the threaded sequence had been removed from the experiment and was used only to evaluate the results; the gap open penalty was chosen as 2. For example (arrow), the correct hit was found among the first five hits in more than 50% of the cases for an alignment including 1D structure, and in less than 15% of the cases for a simple sequence alignment. Or: 40% of the remote homologues were identified among the first two hits when combining 1D structure and sequence; among the first five when using only 1D structure and among the first 15 (not shown) when using sequence alignment only.

remote homologues for the 89 search proteins and the 723 sequence-unique (<25% pairwise sequence identity) proteins used to search remote homologues are listed on the World Wide Web (Rost, 1996a).

Data sets for comparison with other methods

Finally, we compiled the results of our method based on three tiny sets of proteins for which results were published in the literature: (1) a set of 11 proteins used by Jones *et al.* (1992) to evaluate the performance of the program THREADER (Table 4 of Rost, 1996c); (2) a set of 11 representative protein families used by Russell *et al.* (1996) to evaluate the performance of the programs THREADER and MAP (Table 5 of Rost, 1996c); and (3) a set of 11 proteins used for the Asilomar 1994 prediction contest (Lemer *et al.*, 1995; Table 6 of Rost 1996c).

Results

Fold recognition

Loss of information by projection onto 1D limiting factor

When threading 1D structure profiles taken from the DSSP (Kabsch & Sander, 1983) assignments based on coordinates of known 3D structures (in other words completely correct ''predictions''), the first hit was correct in 35% of all test cases (PDB *versus* PDB, $\mu = 100$; Table 1). When using real predictions from PHD (at an average accuracy of about 70%), the first hit was correct in 23% of the cases (PHD *versus* PDB, $\mu = 100$, Table 1). Thus, the limited prediction accuracy of PHD (70%) reduced detection accuracy by ''only'' 12 percentage points; whereas the loss of information by projecting 3D structure onto 1D accounted for 75 percentage points in reducing detection accuracy.

Significant improvement by including sequence information

When 1D structure and sequence information was combined (equation (1)) detection accuracy increased markedly: for a 50:50 mixture of 1D-structure-to-sequence ($\mu = 50$ in equation (1)), 29% of the first hits were correct (Table 1); and in half of the test cases, the correct homologue was detected among the first five alignment hits (Figure 2). The choice of a particular sequence matrix (McLachlan *versus* Blosum62) yielded different alignments (and most often different first hits). However, the overall accuracy for the entire test set was similar (Figure 2, Table 1). For a random prediction, the first hit would be correct in 2% of the cases. For a sequence alignment method (MaxHom with McLachlan matrix), the first hit was correct in about 15% of all cases (Table 1).

Stronger hits more likely to be correct

When the alignment list was cut off at a z-score > 4.5 (equation 5 of Rost, 1996c), the first hit was correct in 88% of the cases (Table 1). At this higher level of accuracy only ten out of the 89 test

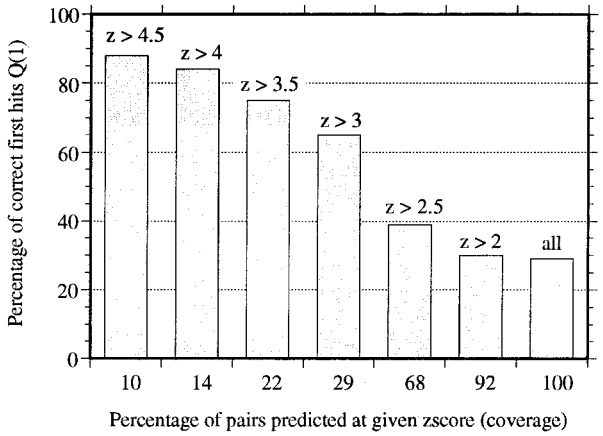

Figure 3. Focusing on stronger predictions. The percentage of correct first hits can be increased by focusing on hits detected with higher *z*-scores. However, the increase of accuracy was at the expense of coverage. For example, at $z > 3.5$ 75% of all first hits were correct, but only for 22% of all test proteins did the first hit reach a *z*-score > 3.5. In other words, a fifth of the test cases predicted most strongly reached an accuracy of 75% (Q(1)).

proteins were detected (Figure 3). The correlation between *z*-score and prediction accuracy illustrated, in particular, the strength of prediction-based threading, as opposed to simple sequence alignment. The sequence alignment used as reference resulted in relatively many correct first hits (15%), but it was very difficult to separate the chaff from the wheat: for 25% of the first hits the *z*-score was above 4.5, and of these only 30% were predicted correctly (Table 1). In other words, sequence alignments reached a similar level of accuracy as prediction-based threading for every fourth protein.

Successful detection of remote homology in the absence of 3D information

One of the features of prediction-based threading is that the detection of remote homology is not restricted to knowing the structure of the target. Instead, a sequence of unknown structure can be threaded through a library of predicted 1D structure assignments. The result was surprisingly not much inferior to the case of using known 3D structures: 27% of the hits were correctly detected at first rank (PHD *versus* PHD; Table 1).

Better recognition of entire folds than of shorter fragments

The test set of 89 proteins was deliberately chosen to answer the following question: how accurately can the method detect any remote homologous fragment in a library of protein structures (remote homology detection)? An easier task is to detect similarities between entire folds (fold detection). We generated subsets of our full test set by excluding all cases for which the structural alignments covered only a small fraction of the aligned pair. For example, if the goal is to detect similarities that cover at least 70% of the lengths of both proteins, the expected accuracy (correct first hit) rose to 50% (Figure 4 of Rost, 1996c). Thus, prediction-based threading was clearly more successful in capturing homologies between entire folds than in detecting homologies between local regions.

Remote homology modelling

Few correct predictions of 3D structure

Given a correctly detected remote homologue, how accurate was the alignment? This question was addressed in two ways. First, the predicted alignments were compared to the structural alignments. For the hits correctly detected at ranks 1 and 2, the average shift score (equation 9 of Rost, 1996c) was 38%, the average identity of the residues between predicted and structural alignments was 33%, and the average shift 11 (Table 3 of Rost, 1996c). More than half of the hits correctly detected at first rank reached an alignment shift score above 50% (15 out of 25); and one half (13 out of 25) had more than 50% of the residues identical with the structural alignment (Table 3 of Rost, 1996c) ; three representative alignments are given in Figure 7 of Rost, (1996c). For the second way to evaluate the alignment, we simply superimposed the backbone model resulting from the predicted alignment with the known structure of the search protein. For only six of the test cases correctly detected at first rank (total of 25) the final model for the 3D structure of the threaded sequence deviated less than 2 Å rmsd from the optimal superposition of the two structures (Table 3 of Rost, 1996c).

Comparison to other threading methods

A favourable set of 11 proteins

Russell *et al*. (1996) recently evaluated their prediction-based threading method (MAP) and the THREADER program of Jones *et al*. (1995) based on a small set of 11 proteins. For the first hit they reported an accuracy of 37 to 45% (depending on the threshold used for defining homologue structures) for MAP and of only 9 to 19% for THREADER (Jones *et al*., 1992). With the same 11 families, our prediction-based threading resulted in 78% correct first hits (Table 5 of Rost, 1996c). The reported quality of the alignments (percentage identical residues between predicted and structural alignment) was 15% for MAP and 11% for THREADER (Russell *et al*., 1996). For our prediction-based threading the average number of correctly aligned residues was 27% (Table 5 of Rost, 1996c). Thus, although the set used by Russell *et al*. (1996) was

much more conservative than the one used initially by Jones *et al*. (1992) (both THREADER and our method yielded 100% correct first hit on that set; Table 4 of Rost, 1996c), it still yielded very optimistic estimates for prediction accuracy when compared to the performance on our set of 89 proteins. Did we select a set that yielded too pessimistic estimates of performance accuracy?

The 11 Asilomar 1994 targets

A final test of our method on 11 proteins that were used as threading targets at the first Asilomar meeting for the evaluation of prediction methods (Lemer *et al*., 1995; Moult *et al*., 1995) suggested that the estimates derived on our initial set of 89 proteins might be closer to the ''reality'' of using automated threading than those derived on favourable test sets. For the Asilomar 11 we correctly detected the remote homologues at first rank in four cases (i.e. 36%; Table 6 of Rost, 1996c). The average percentage of correctly aligned residues was 21%; the average shift nine residues; and the alignment shift score on average AS = 26% (equation 9 of Rost, 1996c). Thus, the alignments were mostly wrong. How did the results compare to the blind predictions made for the meeting? The best methods performed better than our method: (1) the expert-driven usage of THREADER by David Jones and colleagues (Jones *et al*., 1995) detected five out of nine proteins correctly at first rank; and (2) the best alignments of the potential-based threading method perfected by Manfred Sippl and colleagues (Flöckner *et al*., 1995) were clearly better than our best ones.

Remote homology modelling

Correctness of the alignment and consequently the 3D model obtained by threading has hardly been evaluated in the literature. One common example is the homology between the heat shock protein 70 (PDB code: 2hsc) and the A chain of the muscle protein actin (PDB code: 2atnA). Searching with 2hsc, the 1D-profile threading brought up 2atnA at first rank. The predicted alignment agrees for 44% of the residues with the structural alignment taken from FSSP (Holm & Sander, 1994; Figure 8 of Rost, 1996c). For a threading method based on energy calculations, Abagyan *et al*. (1994) published the predicted alignment for the last 232 residues of the same pair. They report that the alignment was wrong for the C-terminal part of the molecules, for the 232 aligned residues their alignment is for 14% of the residues identical with the structural alignment. Interestingly, for the same region the prediction-based threading has 22% of the residues identical with the structural alignment, i.e. clearly worse than the average for the entire protein.

Conclusion

Successful fold recognition by threading predicted 1D structure profiles

Fold motifs could be detected automatically by aligning predicted and known 1D structure profiles (secondary structure and solvent accessibility). However, even for an (in practice unrealistic) optimal prediction of 1D structure (assignment from known coordinates), the first hit was correct in only 35% of all test cases ($Q(1)$, Table 1). A realistic prediction of 1D structure (obtained by cross-validated PHD predictions) yielded 23% detection accuracy. This result suggested two conclusions. (1) The loss of information by projecting 3D information onto 1D structure profiles was the bottleneck of the method. To illustrate this problem: at least 16 unrelated structures contain the secondary structure motif ''H-E-E-H-E-E'' (data not shown). An additional incorporation of information about inter-residue distances may open that bottle-neck. (2) Further improvements of 1D structure predictions could improve the accuracy of prediction-based threading significantly.

Better fold recognition by combining 1D structure profiles and sequence information

The novel step introduced here (combining 1D structure profiles with sequence information, equation (1)) increased detection accuracy significantly: 29% of all first hits were correct (Table 1), and in about 53% of the test cases the correct homologues was found among the first five hits (Figure 2). Thus, the prediction-based threading was clearly superior to sequence alignments (15% correct first hits, Figure 2). Furthermore, accuracy could be increased by focusing on the subset of those hits which were predicted with higher z-scores. For example, for the 10% of all proteins predicted at $z > 4.5$ (equation 5 of Rost, 1996c) the expected accuracy of correctly detecting the fold at first rank rose to 88% (Table 1, Figure 3). Homologous folds were detected more accurately than homologous fragments. For example, for a test set with true homologues for which the alignment covered 70% of both aligned sequences, one half of the first hits were correct (Figure 4 of Rost, 1996c). A feature of prediction-based threading that may become particularly interesting for applications in practice is that remote homology can successfully be detected between protein pairs without knowledge of 3D structure: when using 1D structure predictions as a fold library, we correctly detected the remote homologue in 27% of the test cases at first rank (Table 1).

Prediction-based threading competitive with other threading techniques

A recent analysis based on a small set of 11 structure families (Russell *et al*., 1996), suggested a

Table 2. Performance on a test set of 11 proteins used by Russell *et al.* (1996)

	Homologous pairs				PHDthreader			Others	
idSeq	id^{RCB}	idStr	pide	zDALI	*R*	*Z*	ali(%)	R^{RCB}	R^{JTT}
1bfg	4fgf	1irp	16	9.0	1	3.5	12	10	6
2fal	1mba	1hlb	13	15.4	1	5.0	58	1	1
1hnf		3cd4	15	12.7	2	2.3	0	1	>10
1lkkA	1shaA	2pna	24	9.5	1	4.1	60	>10	>10
2mhr	2hmqA	1ilk	16	2.8	1	2.8	0	2	>10
2pgd		1fps	11	2.8	2	2.4	0	1	2
1plc		1aac	21	11.0	1	4.2	75	2	1
1rcb		1fps	7	3.3	4	2.3	0	2	>10
1thx	2trxA	1trw	21	16.3	1	4.7	60	1	1
1ubi	1ubq	1frr	13	4.0	1	3.1	34	1	4
1ubsA	1wsyA	1nal	14	12.7	1	3.1	0	3	3

For some examples we used the representative of the family used by Russell *et al.* (1996) that we found in the current FSSP release (Holm & Sander, 1994). Abbreviations: idSeq, PDB + chain identifier for the search sequence; id^{RCB}, family member used by Russell *et al.* (1996); idStr, identifier for the aligned homologue; *R*, rank of first correct hit for our method; *Z*, z-score for our method; ali(%), percentage of residues identical between predicted and structural alignment; R^{RCB}, rank of first correct hit for MAP method (Russell *et al.*, 1996); R^{JTT}, rank of first correct for THREADER method (Jones *et al.*, 1992) (result taken from Russell *et al.* (1996)). Alignment averages: $\langle AS \rangle = 32\%$, $\langle ali\% \rangle = 27\%$, $\ll S \gg = 16$; zDALI, structural similarity (Holm & Sander, 1994).

significant detection accuracy (correct first hits) below 20% for the potential-based threading program THREADER (Jones *et al.*, 1992) . The prediction-based threading method of Russell *et al.* reached 37 to 45% accuracy. For the same 11 families our method had 75% correct first hits (one standard deviation > 15%; Table 2). When (in retrospect) evaluating our method on the 11 threading targets used for the Asilomar 1994 prediction contest (Moult *et al.*, 1995) we had four first hits correct (36%). However, other methods performed better (Lemer *et al.*, 1995): an expert-driven usage of THREADER had more correct first hits (Jones *et al.*, 1995), and the potential-based threading by Sippl and colleagues obtained the best alignments more accurately (Flöckner *et al.*, 1995). Fischer and Eisenberg (Fischer & Eisenberg, 1996; Fischer *et al.*, 1996) have recently developed a method for prediction-based threading that is very similar to the one presented here They evaluated their own and previous potential-based threading methods based on a large set of 64 remote homologues and reported 31% correct hits for potential-based threading (Bowie *et al.*, 1990, 1991; Lüthy *et al.*, 1991, 1992) and 48% correct hits for prediction-based threading (Fischer & Eisenberg, 1996). This confirms the conclusions suggested by the results presented here and previously (Rost, 1995a,b): in correctly identifying the first hit, prediction-based threading is, at least, as accurate as potential-based threading.

Correct prediction of 3D structure by remote homology modelling for single cases

The correct detection of remote homology is the precondition for remote homology modelling. However, correct detection does not imply correct alignments. On the contrary, for most correctly detected remote homologues the alignment was, at least, partially wrong (for one half of the hits correctly predicted at first rank the identity between predicted and structural alignment was above 50%; Table 3 of Rost, 1996c). The same is true for most other threading techniques (Flöckner *et al.*, 1995; Lemer *et al.*, 1995; Shortle, 1995; Fischer & Eisenberg, 1996; Russell *et al.*, 1996). How can a false alignment result in the detection of the true remote homologue among a huge set of decoys? The answer remains open.

Method available by automatic prediction service

The prediction-based threading of 1D structure profiles (PHDthreader) is available *via* an automatic prediction service (send the word help to the internet address PredictProtein@EMBL-Heidelberg.DE, or use the World Wide Web (WWW) site http://www.embl-heidelberg.de/predictprotein/). By default, input strings (1D structure profile) are generated by a PHD prediction; however, users can also opt to provide their own predictions of secondary structure and solvent accessibility.

Will threading replace structure determination?

The number of different protein folds is probably limited (Chothia, 1992). Thus, will threading eventually close the sequence-structure gap by remote homology modelling? Three reasons make this appear an over-optimistic science fiction. (1) Correct alignments are still the exception rather than the rule. (2) Even when the alignments are correct, remote homology modelling at levels of less than 30% pairwise sequence identity is yet another unsolved problem (even for close homologues, modelling is not always successful). (3) The more unique folds are contained in the database, the more difficult the detection will become. This was illustrated by the following experiment. We aligned our 89 test proteins against three different ''fold libraries'': (1) the largest set of sequence-unique proteins as of spring 1996 (723 chains; Rost, 1996a), (2) the largest set of 1995 (449 chains), and (3) a set of unique folds (plus the detectable homol-

ogues, 403 chains). The percentage of correctly detected first hits was inversely proportional to the size of the data set: 29% (1), 31% (2) and 33% (3). This result, probably, stems from the fact that the selection procedure is non-linear. Thus, the likelihood of random errors is increased by increasing the fold library. In other words, we doubt that threading is likely to close the sequence-structure gap in the future, but it can contribute to bridging it today.

Acknowledgements

First of all, thanks to Manfred Sippl (University of Salzburg) for discussions and help. Furthermore, thanks to Michael Braxenthaler (CARB, Washington, DC), Séan O'Donoghue (EMBL, Heidelberg), and Daniel Fischer and David Eisenberg (both UCLA, Los Angeles) for helpful dialogues; and to Rob Hooft (EMBL, Heidelberg) for software assistance. Last, not least, thanks to all those who deposit protein structures and protein sequences in public databases and those maintaining high quality databases (we mention, in particular, Amos Bairoch and his group (Basel)) thereby enabling the design of prediction methods.

References

Abagyan, R., Frishman, D. & Argos, P. (1994). Recognition of distantly related proteins through energy calculations. *Proteins: Struct. Funct. Genet.* **19**, 132–140.

Bairoch, A. & Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucl. Acids Res.* **24**, 21–25.

Bairoch, A. & Boeckmann, B. (1994). The SWISS-PROT protein sequence data bank: current status. *Nucl. Acids Res.* **22**, 3578–3580.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. & Brice, M. D., *et al*. (1977). The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U., Clarke, N. D., Pabo, C. O. & Sauer, R. T. (1990). Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins: Struct. Funct. Genet.* **7**, 257–264.

Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164–169.

Bryant, S. H. & Altschul, S. F. (1995). Statistics of sequence-structure threading. *Curr. Opin. Struct. Biol.* **5**, 236–244.

Chothia, C. (1992). One thousand protein families for the molecular biologist. *Nature,* **357**, 543–544.

Chothia, C. & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**, 823–826.

Doolittle, R. F. (1986). *Of URFs and ORFs: a Primer on How to Analyze Derived Amino Acid Sequences.* University Science Books, Mill Valley, CA.

Dujon, B. (1996). The yeast genome project: what did we learn? *Trends Genet.* **12**, 263–270.

Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.

Fischer, D., Elofsson, A., Rice, D. & Eisenberg, D. (1996). Assessing the performance of fold recognition methods by means of a comprehensive benchmark. In *Pacific Symposium on Biocomputing, Hawaii, 1996* (Hunter, L. & Klein, T., eds), pp. 300–318, World Scientific Publishing, Singapore.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A. & Kirkness, E. F., *et al*. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science,* **269**, 496–512.

Flöckner, H., Braxenthaler, M., Lackner, P., Jaritz, M. & Ortner, M., *et al*. (1995). Progress in fold recognition. *Proteins: Struct. Funct. Genet.* **23**, 376–386.

Greer, J. (1991). Comparative modeling of homologous proteins. *Methods Enzymol.* **202**, 239–252.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA,* **89**, 10915–10919.

Holm, L. & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* **22**, 3600–3609.

Holm, L., Ouzounis, C., Sander, C., Tuparev, G. & Vriend, G. (1993). A database of protein structure families with common folding motifs. *Protein Sci.* **1**, 1691–1698.

Johnston, M. (1996). Towards a complete understanding of how a simple eukaryotic cell works. *Trends Genet.* **12**, 242–243.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86–89.

Jones, D. T., Miller, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387–397.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen bonded and geometrical features. *Biopolymers,* **22**, 2577–2637.

Lattman, E. E. (1994). Protein crystallography for all. *Proteins: Struct. Funct. Genet.* **18**, 103–106.

Lemer, C. M.-R., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337–355.

Lesk, A. M. & Boswell, R. D. (1992). Homology modelling: inferences from tables of aligned sequences. *Curr. Opin. Struct. Biol.* **2**, 242–247.

Lüthy, R., McLachlan, A. D. & Eisenberg, D. (1991). Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins: Struct. Funct. Genet.* **10**, 229–239.

Lüthy, R., Bowie, J. U. & Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. *Nature,* **356**, 83–85.

May, A. C. W. & Blundell, T. L. (1994). Automated comparative modelling of protein structures. *Curr. Opin. Biotech.* **5**, 355–360.

McLachlan, A. D., Staden, R. & Boswell, D. R. (1984). A method for measuring the non-random bias of a codon usage table. *Nucl. Acids Res.* **12**, 9567–9575.

Moult, J., Pedersen, J. T., Judson, R. & Fidelis, K. (1995). A large-scale experiment to assess protein structure

prediction methods. *Proteins: Struct. Funct. Genet.* **23**, 2–4.
Oliver, S., van der Aart, Q. J. M., Agostioni-Carbone, M. L., Aigle, M. & Alberghina, L., *et al*. (1992). The complete DNA sequence of yeast chromosome III. *Nature,* **357**, 38–46.
Rost, B. (1995a). Fitting 1-D predictions into 3-D structures. In *Protein Folds: a Distance Based Approach* (Bohr, H. & Brunak, S., eds), pp. 132–151, CRC Press, Boca Raton, FL.
Rost, B. (1995b). TOPITS: threading one-dimensional predictions into three-dimensional structures. In *Third International Conference on Intelligent Systems for Molecular Biology* (Rawlings, C., Clark, D., Altman, R., Hunter, L. & Lengauer, T., *et al*., eds), pp. 314–321, Menlo Park, CA: AAAI Press, Cambridge, England.
Rost, B. (1996a). Appendix to ''Protein fold recognition by prediction-based threading''. EMBL Heidelberg, Germany WWW document (http://www.embl-heidelberg.de/ ~ rost/Papers/JMB96-threading.html).
Rost, B. (1996b). Average conservation of 1D structure between remote homologues. EMBL Heidelberg, Germany, WWW document (http://www.embl-heidelberg.de/ ~ rost/Res/96E-ConservationOf1D.html).
Rost, B. (1996c). 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. (1994a). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.
Rost, B. & Sander, C. (1994b). Conservation and prediction of solvent accessibility in protein families. *Proteins: Struct. Funct. Genet.* **20**, 216–226.
Russell, R. B., Copley, R. R. & Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349–365.
Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.
Schneider, R. (1994). Sequenz und sequenz-struktur vergleiche und deren anwendung für die struktur- und funktionsvorhersage von proteinen. PhD thesis, University of Heidelberg.
Shortle, D. (1995). Protein fold recognition. *Nature Struct. Biol.* **2**, 91–92.
Sippl, M. J. (1982). On the problem of comparing protein structures. *J. Mol. Biol.* **156**, 359–388.
Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229–235.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197.
Vingron, M. & Waterman, M. S. (1994). Sequence alignment and penalty choice. *J. Mol. Biol.* **235**, 1–12.
Zu-Kang, F. & Sippl, M. J. (1996). Optimum superimposition of protein structures: ambiguities and implications. *Folding Design,* **1**, 123–132.

Edited by F. E. Cohen

(Received 5 December 1995; received in revised form 19 September 1996; accepted 19 September 1996)