From: ISMB-95 Proceedings. Copyright © 1995, AAAI (www.aaai.org). All rights reserved.

TOPITS: Threading One-dimensional Predictions Into Three-dimensional Structures

Burkhard Rost

EMBL
69012 Heidelberg, Germany
rost@embl-heidelberg.de

Abstract

Homology modelling, currently, is the only theoretical tool which can successfully predict protein 3D structure. As 3D structure is conserved in sequence families, homology modelling allows to predict 3D structure for 20% of SWISSPROT. 20% of the proteins in PDB are remote homologues to another PDB protein. Threading techniques attempt to predict such remote homologues based on sequence information. Here, a new threading method is presented. First, for a list of PDB proteins, 3D structure was projected onto 1D strings of secondary structure and relative solvent accessibility. Then, secondary structure and accessibility were predicted by neural network systems (PHD). Finally, the predicted and observed 1D strings were aligned by dynamic programming. The resulting alignment was used to detect remote 3D homologues. Four results stand out. Firstly, even for an optimal prediction (assignment based on known structure), only about half the hits that ranked above a given threshold were correctly identified as remote homologues; only about 25% of the first hits were correct. Secondly, real predictions (PHD) were not much worse: about 20% of the first hits were correct. Thirdly, a simple filtering procedure improved prediction performance to about 30% correct first hits. The correct hit ranked among the first three for more than 23 out of 46 cases. Fourthly, the combination of the 1D threading and sequence alignments markedly improved the performance of the threading method TOPITS for some selected cases.

Introduction

Reducing the sequence-structure gap by homology modelling. Large scale gene-sequencing projects produce data of gene, and therefore protein sequences, at a breathtaking pace (Johnston, et al. 1994, Oliver, et al. 1992). However, the three-dimensional (3D [1]) structure is known only for a minority of the known sequences. Although experimental structure determination is becoming a routine procedure (Lattman 1994), the sequence-structure gap is increasing. Currently, the only reliable way to predict 3D structure is homology modelling (Greer 1991, May and Blundell 1994). A structure can be modelled by homology for a sequence of unknown structure (dubbed SOUS) if a protein of known structure (PDB; Bernstein, et al. 1977) has significant sequence identity to SOUS (>25%; Sander and Schneider 1994). Currently, this covers 20% (Sander and Schneider 1994) of the known sequences (SWISSPROT; Bairoch and Boeckmann 1994). What about proteins without sequence homologue in PDB?

Threading techniques may become a second comprehensive tool. There are many remote homologues (Holm and Sander 1994a), i.e., protein pairs which have homologous 3D structures but no significant pairwise sequence similarity. Roughly, 20% of the proteins in PDB are remotely homologous to other PDB proteins (Holm and Sander 1994a). How could we model these remote homologues? One approach is threading, i.e., the attempt to evaluate the fitness of a sequence for a structure (Abagyan, et al. 1994, Bowie, et al. 1990, Bowie, et al. 1991, Bryant and Lawrence 1993, Eisenberg, et al. 1991, Jones, et al. 1992, Nishikawa and Matsuo 1993, Ouzounis, et al. 1993, Sippl 1993b, Sippl and Jaritz 1994, Sippl and Weitckus 1992, Wilmanns and Eisenberg 1993, Wodak and Rooman 1993). The main problem of threading is to distinguish false positives from remote homologues. Furthermore, mean-force potentials (Sippl 1993a) are optimised to detect stresses or instabilities in protein structures caused by subtle changes, e.g., the exchange of a few residues. This permits accurate detection of errors in experimentally determined structures (Sippl 1993b). However, threading implies a search for coarse-grained similarities rather than for fine-grained differences. Is there another way to predict structure when homology modelling is not applicable?

1D predictions are generally applicable and have become significantly better. In general, 3D structure cannot be predicted *ab initio* . One way out is to simplify the prediction task, e.g., by projecting 3D structure onto

[1] **Abbreviations:** 3D, three-dimensional; 1D, one-dimensional; PDB, Protein Data Bank of experimentally determined 3D structures of proteins; SWISSPROT, data base of protein sequences; DSSP, data base containing the secondary structure and solvent accessibility for proteins of known 3D structure; FSSP, data base of remote homologues of known 3D structure; PHD, Profile based neural network prediction of secondary structure (PHDsec) and solvent accessibility (PHDacc); SOUS, protein sequence of unknown 3D structure (e.g. search sequence in alignment procedure).

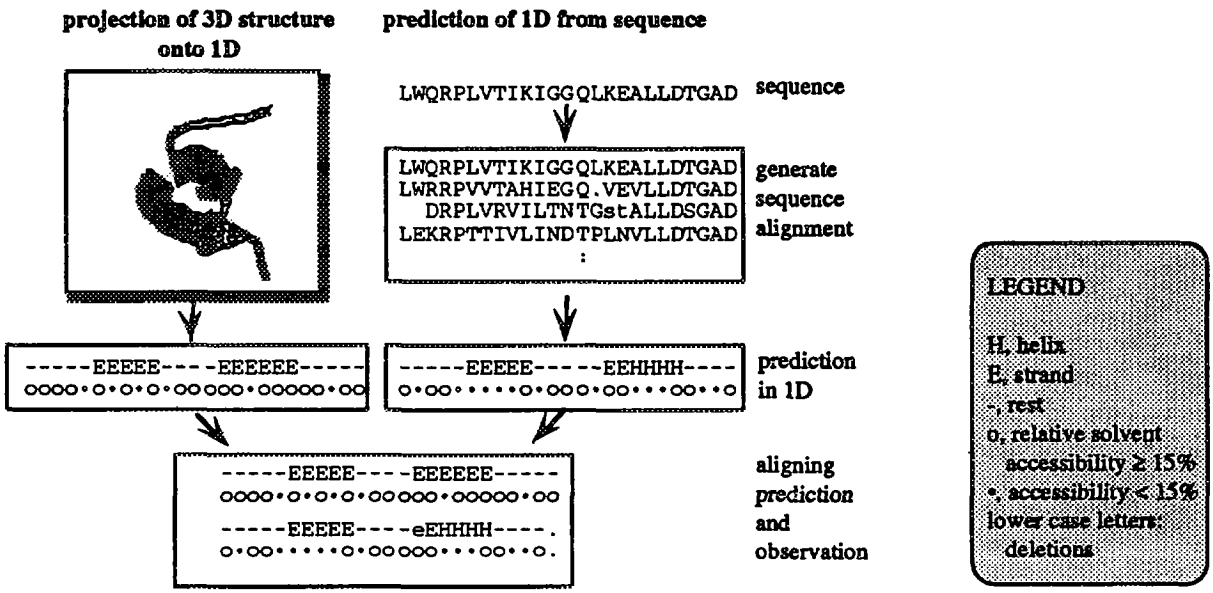


Figure 1: 1D threading: aligning 1D predictions with 3D structures.
First, for a list of proteins with known 3D structure (typically proteins unique in terms of pairwise sequence identity, (Hobohm and Sander 1994)) projections of 3D structure onto strings of secondary structure and relative solvent accessibility assignments were computed (DSSP, (Kabsch and Sander 1983)). Second, a search sequence of unknown structure (SOUS) was aligned against the sequence data base (SWISSPROT (Bairoch and Boeckmann 1994)). Third, the resulting multiple alignment was used as input to neural network systems (PHD (Rost and Sander 1994a, Rost and Sander 1994b)) that predicted both secondary structure and solvent accessibility for SOUS. Fourth, prediction and observation were extracted into strings composed of a six letter alphabet (Hb, He, Eb, Ee, Lb, Le). Fifth, the string predicted for the SOUS was aligned against the strings extracted from the data base of observed structures.

1D strings of secondary structure or solvent accessibility. The accuracy of predicting secondary structure or accessibility has been improved significantly using evolutionary information (Rost and Sander 1993b, Rost and Sander 1994a, Rost and Sander 1994b, Wako and Blundell 1994). 1D predictions are useful in many respects. In some cases, remote homology modelling can be based on 1D predictions. Is such a procedure restricted to selected cases, or can 1D predictions be used, in general, to detect remote homologues?

Fitting 1D predictions into 3D structures. Structural alignments show that secondary structure and accessibility is conserved between remote homologues (Rost 1995). In other words, there are conserved, global motifs of 1D structure. The idea of the 1D threading approach introduced here, is to detect remote homologues by aligning 1D predictions to known 3D structures. The approach is conceptually simple but there are many difficulties in details. Some strategies to optimise free parameters will be discussed and first results will be given.

Materials and Methods

Aligning predicted and observed 1D strings

1D motifs indicative for fold. The principle idea of threading 1D predictions into 3D structures (TOPITS) is the following. Protein folds can often be described by secondary structure motifs (Orengo, et al. 1993, Richardson and Richardson 1989). Mostly, this requires some information about inter-segment distances. But, in some cases, comparisons of known secondary structure and accessibility are sufficient to at least formulate the hypothesis that the search sequence (SOUS) is remotely homologous to a certain structure (Bowie, et al. 1990, Bowie, et al. 1991, Sheridan, et al. 1985). Here, the idea was to compare 1D strings of predictions and observations to detect remote homologues. A precondition for this

concept to be successful is that 1D strings are conserved between remote homologues (Rost 1995).

Detecting remote homologues by 1D structure alignments. Dynamic programming (Needlman and Wunsch 1970, Smith and Waterman 1981) was used to align predicted and observed 1D strings (Figure 1). The string predicted for SOUS was aligned against all proteins in a set of unique PDB structures. The program used was a modified version of the multiple sequence alignment program *MaxHom* (Sander and Schneider 1991, Sander and Schneider 1994).

Merging sequence alignments and 1D threading. Below about 25% pairwise sequence identity, sequence alignments become blurred. However, sequence information is important even for very low identity. Unfortunately, the signal is obfuscated by noise. The 1D TOPITS threading was combined with sequence alignments by simply extending the six-letter alphabet to strings with 6×20 different symbols. The relative contribution of sequence information was tuned by adding a 1D threading comparison matrix and a usual sequence alignment profile matrix with varying relative weights.

1D predictions by combining evolutionary information and neural networks

Alignment set of known structures. 3D structure was projected onto 1D strings of secondary structure and solvent accessibility using DSSP (Kabsch and Sander 1983). For secondary structure three states were used: helix (H), strand (E), and loop (L); for relative solvent accessibility two states were distinguished: buried (b, relative accessibility <16%), and exposed (e, relative accessibility <16%). The 1D projections were extracted for a set of unique 3D structures into six-letter-strings (Hb, He, Eb, Ee, Lb, and Le) and stored as alignment set.

1aai_A	1ald	1atn_A	1bmv_1	1bov_A
1c2r_A	1cc5	1cms	1dhr	1dri
1ego	1etu	1gmf_A	1gp1_A	1hsc
1hsd_A	1ifb	1lh3	1mbn	1min_A
1npx	1nsb_B	1omf	1omp	1p04_A
1phh	1r09_3	1rbp	1sgt	1tie
1tlk	1wsy_A	2aza_A	2gbp	3adk
3ccp	3cro_L	3dfr	3er3_E	3grs
3icb	4i1b	5ldh	5p21	5tnc
6nn9				

Table 1: Data set for evaluation
As 'true' remote homologues all pairs were taken that are listed in the FSSP data base files (Holm and Sander 1994a): pdbid.fssp, and that are not contained in the corresponding data base of sequence alignments (HSSP; Sander and Schneider 1994). For all 46 proteins the PDB identifier (pdbid), and if one chain was used the chain identifier (_X) are given..

1D predictions by PHD. Both secondary structure and accessibility are conserved between 3D homologues with significant pairwise sequence identity (Rost and Sander 1994b, Rost, et al. 1994). This conservation can successfully be used for predictions. One way is to feed information derived from multiple sequence alignments as input into neural network systems (Rost and Sander 1993b, Rost and Sander 1994a, Rost and Sander 1994b). The neural network predictions (PHDsec, PHDacc) were used as input to the threading procedure. (Note: the predictions used for threading were obtained by cross-validation, i.e. SOUS had not been used to train the neural networks used for 1D predictions.)

Adjusting free parameters for alignment

Basing the comparison matrix on data base counts. Two strings can be aligned by simply matching identical character pairs. However, to effectively align protein sequences, matches have to be weighted (Altschul 1991, Dayhoff 1978, Henikoff and Henikoff 1992, McLachlan 1971, Pearson and Lipman 1988). Which is the best weighting matrix? Some of the proposed comparison matrices appear favourable, but there is no clearly best method (Henikoff and Henikoff 1993). The same applies for the comparison matrix used to align predicted and observed secondary structure and accessibility. Various strategies were explored to find the optimal matrix (Rost 1995). The most successful (data not shown) was to use data base counts:

$$M_{ij} = \ln F_{ij} - \langle \ln F \rangle_{\{i,j\}}, \text{ with } F_{ij} = \frac{f_{ij}}{f_i \times f_j} \qquad (1)$$

for all i, j = Hb, He, Eb, Ee, Lb, Le, i.e., all states for the first and second string in the alignment. $\langle x \rangle$ gives the average of x over all states i and j, and f are the counts from the data base. Which data set should be chosen for counting? Two factors influence TOPITS. First, the information that is lost by projecting 3D structure onto 1D. Second, the inaccuracy of 1D predictions. Thus, it is a priori not clear whether counts should be based on (i) a set of structurally aligned homologous pairs (resulting matrix dubbed M3 in Table 2), or (ii) on a set of predicted and observed strings for the same protein (i.e. the full matrix defining the per-residue prediction accuracy of a prediction method (Rost and Sander 1993a), called M1 in Table 2).

Choosing gap penalties according to preference in coverage or correctness. One complication with dynamic programming is the optimal adjustment of gap penalties (Vingron and Waterman 1994). The optimal choice depends on the context, i.e., the particular protein family. In general, there is a trade-off between coverage and correctness of detection for the choice of the free gap parameters *gap_open* (penalty for opening a gap) and *gap_elongation* (penalty for continuing an open gap). The relative values were not so crucial; all results will be given for: $gap\_elongation = 0.1 \times gap\_open$.

Sorting the threading list and choosing cut-off thresholds is crucial. The simple alignment score is not sufficient to sort the final results, as it depends for example on the length of the search sequence and the secondary structure content. To render a value independent of the specificity of one alignment, the alignment score was normalised (Sippl and Jaritz 1994):

$$zE_k = \frac{E_k - \langle E \rangle_{\{\kappa\}}}{\sigma_{\{\kappa\}}} \quad (2)$$

where E_i is the alignment score E for the hit at position i of the threading list.

Measuring expected detection accuracy

Definition of simple measures for accuracy in detecting remote homologues. Given a list of 'true' remote homologues and another of predicted homologues, various measures for accuracy can be defined. The most important is the simplest: how many of the first hits are correct?

$$Q_{1st} = 100 \times \frac{\text{number of correct first hits}}{\text{number of all proteins}} \quad (3)$$

Less strict is to cut the alignment list for a given score e.g. zE at a given threshold θ and to count the percentage of correct hits in the remaining list:

$$Q_\theta = 100 \times \frac{\text{N correct first hits with } zE > \theta}{\text{N hits with } zE > \theta} \quad (4)$$

Data set for evaluation

The expected prediction accuracy depends on the data set chosen for evaluation. So far, most publications on threading used only several 'favourite test cases'. Such a procedure is rather arbitrary. A more reasonable approach would be to select a list of unique proteins (Hobohm and Sander 1994, Holm and Sander 1994b), to compile for each of these proteins all remote homologues by structural alignment, and then to compare the threading results with the list of all remote homologues found. Here, a preliminary realisation of this concept was pursued. The results given were based on structural alignments of 46 protein chains (with 384 aligned pairs in total, Table 1; the list is still too small for a standard set; a more comprehensive list is being collected).

			Q_{1st} eq. (3) for rank		Q_θ (eq. (4) for θ =		Q_{1st} eq. (3) for filter with: length alignment / length seq 2 >				
	matrix	go	1	2	3	2	0.8	0.7	0.6	0.5	0.4
PDB	**M3**	**3**	**26**	**26**	**59**	**41**	**26**	**24**	**26**	**28**	**31**
PDB	M3	5	18	27	52	36	22	16	13	13	13
PHD	M1	1	10	21	25	21	23	19	15	10	10
PHD	M1	3	15	30	47	30	28	30	23	15	15
PHD	**M1**	**5**	**21**	**28**	**41**	**34**	**26**	**28**	**26**	**19**	**17**
PHD	M1	8	23	17	44	34	21	28	26	21	23
PHD	M1sym	3	19	19	39	25	23	28	28	23	23
PHD	M1inv	3	17	17	36	27	19	23	23	23	21
PHD	M3sep	3	19	21	59	22	17	15	15	15	21
PHD	M3	3	17	23	25	20	21	23	28	21	19
PHD	M3	5	15	19	41	29	19	19	17	13	17
PHD	M3	8	15	17	41	35	15	13	10	8	13

Table 2: Accuracy in detecting remote homologues
Both for an alignment of PDB with PDB (optimal 1D threading) and a TOPITS alignment of PHD with PDB (prediction with observation), the detection accuracy is given for various comparison matrices, gap open penalties and different thresholds for cutting off the list. All values are percentage averages over 46 threading experiments (Table 1).

Abbreviations for comparison matrices: *M3* , counts based on structural alignments; *M1* , counts based on comparing predictions of 1D with observations; $M1inv_{ij} = M1_{ji}$, i.e., transposed of *M1*; $Msym1_{ij} = (M1_{ji}+M1_{ij})/2$, i.e. symmetrical average between *M1* and *M1inv*; *M3sep* , matrix optimising the separation between random alignments and remote homologues (Rost 1995).

Abbreviations for scores: *go* ; gap open penalty; *rank 1* and *rank 2* , percentage of proteins for which the correct hit was at rank 1 or 2 of the threading list; *norm(z) > n* , all hits *i* , with z(i) > n + 0.006 * length of alignment (where 0.006 is an empirical constant derived to linearly fit the distribution of the z-score vs. the alignment length), numbers given are: percentage of all correct hits for that cut-off; filter for length of aligned sequence: percentages of correct first ranking hits for that filter.

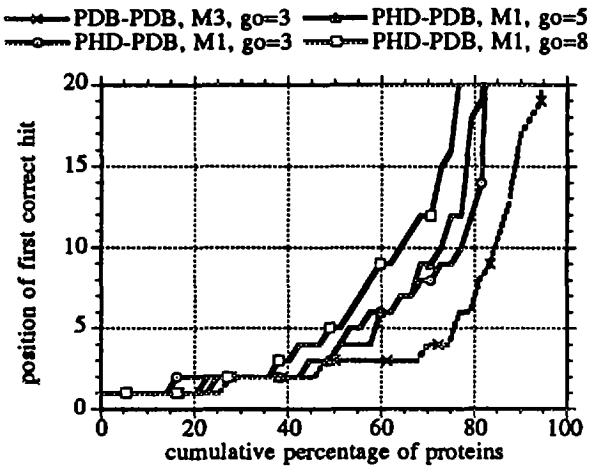


Figure 2: Rank of first correct hit
The rank of the first correctly detected remote homologue vs. the cumulative percentage of proteins for which that rank had been first. Results for an alignment of PDB with PDB (matrix M3 (Table 2), gap-open =3); and of PHD with PDB (matrix M1 (Table 2), gap-open = 3, 5, 8). For example, for half of the proteins, the correct hit ranked among the first three hits detected by TOPITS (go=5).

Results

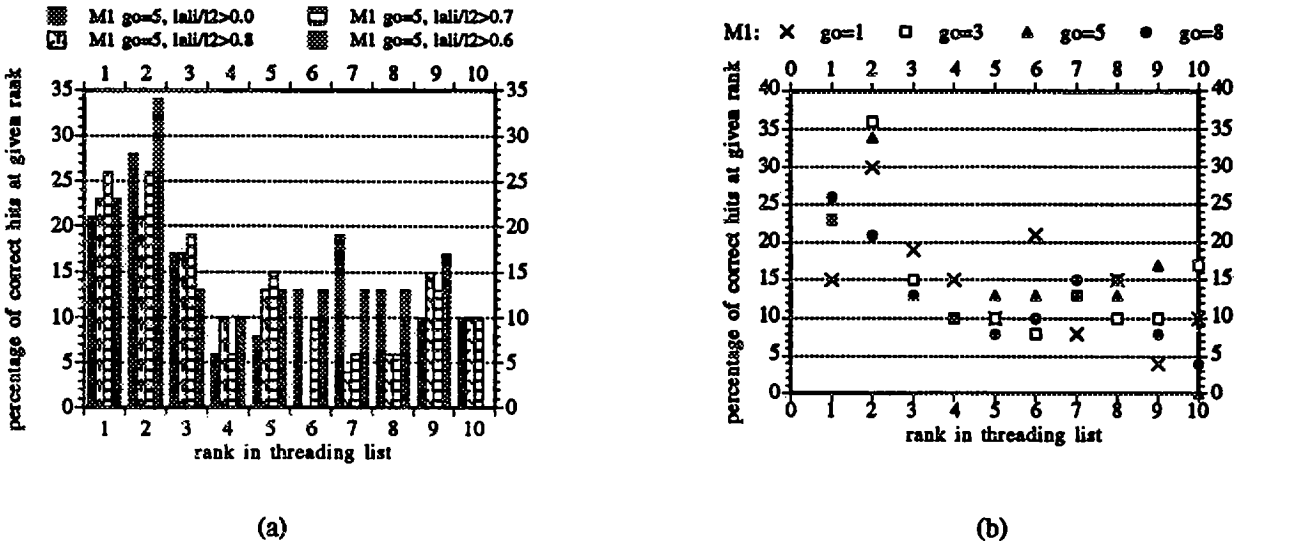
Loss of distance information by 1D projection is crucial. When the threading was performed using known strings of secondary structure and relative solvent accessibility as search strings (optimal prediction scenario: PDB against PDB), the percentage of correct hits above a given z-score cut-off (eq. 4) was at best some 60% (Table 2). Without applying a filter, the percentage of correct first hits (eq. 3) was only some 25%. Thus, the fold motif was captured by 1D information, but for the majority of cases the projection onto 1D reduced the information too drastically. Prediction of secondary structure and solvent accessibility rates at some 60-80% accuracy. This error margin was reflected, as well, in 1D threading (PHD against PDB): the percentage of correct hits above a given z-score cut-off (eq. 4) was about 40%, and the percentage of correct first hits about 20%, i.e. the accuracy falls into the range of 60-80% of the optimal performance. For more than 40% of the proteins, the first or second hit was correct (Table 2), for about 60% the correct hit was among the first six predicted homologues (Figure 2).

Most successful comparison matrix reflects inaccuracy of 1D prediction. For different scoring matrices, detection accuracy varied by four percentage points for the first correct hit; the percentage of correctly detected homologues for a given z-score cut-off varied by about 11 percentage points (Table 2). Three results stick out from analysing the performance for various comparison matrices. First, the most successful matrix was the one for which the counts (eq. 1) had been based on the performance of the 1D neural network predictions (M1 in Table 2). Second, a dramatically incorrect strategy such as simply using the transposed of the best matrix (M1) had a relatively small effect on the detection accuracy. This was surprising as the best matrix was not symmetric (indeed the symmetric average between the best matrix M1 and its transposed resulted in the lowest accuracy, Table 2). Third, the highest percentage of correct hits above a given z-score cut-off (eq. 4), yielded a matrix that had been optimised to distinguish between random alignments and 3D homologues (Rost 1995).

Best results for gap open penalty of five. Changing the gap open penalty from one to eight increased the percentage of correct first hits from 10 to 23% (Table 2). These numbers over-emphasise the influence of the gap penalty. When considering a histogram of the correct ranking among the first 10 alignment hits, the alignment proved relatively stable (Figure 3b). Overall, a gap open penalty of five was found to be best (Figure 3b, Table 2). One encouraging result was that too low gap penalties were obvious without knowledge about the 'true' homologues, e.g., a gap open penalty of one resulted in alignments with too many insertions, on average, every third residue was a gap.

z-scores of correct hits do depend on alignment length. The rational behind introducing the alignment z-score (eq. 1) had been to render scores independent of for example the alignment length. However, the z-scores of the first correct hits did depend on alignment length. The z-scores could be fitted by a linear function of alignment length; the ratio z-score/alignment length was 0.006 (a similar result appears to hold for other threading approaches, as well, Michael Braxenthaler, Manfred Sippl, David Eisenberg, private communications). This linear fit was used to determine the thresholds for computing the percentage of correct predictions above a given cut-off (eq. 4). The percentage of correct hits above a z-score of 3 (+0.006 * alignment length) was about 40% for the best comparison matrix (M1) with a gap penalty of five (the highest value was 47%, Table 2).

Improvement of correct first hits to 25-30% by filtering the list. Visual inspection of the threading alignments revealed that false positives could often be spotted simply by ignoring all hits for which the aligned protein was much longer than the alignment. This was successfully used to automatically filter the threading list. The effect of such a simple procedure was an improvement to some 25-30% correct first hits. This result raised the hope that by appropriately filtering the threading list, detection accuracy could be improved furthermore. However, preliminary results were negative, so far.



(a) (b)

Figure 3: Correct predictions for first ten hits
The percentage of correct hits (averaged over all 46 proteins Table 1) at each of the first ten positions of the final TOPITS alignment is given for (a) different filtering values (length of alignment / length of aligned sequence, values given > 0, i.e. no filter; > 0.8, > 0.7, and >0.6); and (b) for gap-open penalties between one and eight. For all results the comparison matrix was M1 (Table 2).

Further improvement by combining TOPITS and sequence alignment. The anticipated scenario for using a combination of 1D structure and sequence alignments was to start with a sequence alignment, and then to tune on the 1D structure contribution stepwise. The hope was that correct hits would rank on, say, the first 10 positions in all alignment lists, i.e., that the stability of a hit - not necessarily the first - could be used to gain detection reliability. Only two examples are picked out to illustrate the possible benefit of a combination: the TU-elongation factor (1etu) and the tryptophan synthase (1wsy). For both, TOPITS found a correct hit at position four of the list (5p21 for 1etu, and 2liv for 1wsy). Using a 50:50 mixture of 1D structure:sequence, ras (5p21) appeared at the top of the TU-elongation factor list, and the periplasmic binding protein ranked at position three of the tryptophan synthase list. Furthermore, in both cases, the combination of the two different alignment lists (structure only, structure + sequence), would have given not only the correct first hit, but as well a correct second and third hit. (Note: all hits were in the range of <15% pairwise sequence identity, i.e., undetectable for sequence alignment, alone.) For such a 50:50 mixture the percentage of correct hits above a z-score of three was more than 65%.

Conclusions

Yet another threading approach? The quick reader may be confused by the similarity of TOPITS and the first threading approaches (Bowie, et al. 1990, Bowie, et al. 1991, Scharf 1989) in terms of states (six) and underlying descriptions of 3D structure (secondary structure, accessibility). However, the difference is a principle one. Conventional threading methods describe 3D structures in terms of profiles or potentials to fit locally (over a given range of residues) into a given 3D structure. The alignment of predicted and observed secondary structure and accessibility (TOPITS) starts from a profile-independent, full prediction of aspects of 3D structure. This has two consequences for the strength of TOPITS. First, the cooperativity of protein structure is taken into account, as the alignment is not done by gliding windows. Second, TOPITS enables the prediction of remote homology for a pair of sequences both unknown in 3D structure, thus permitting to possibly predict function of two proteins that cannot be aligned based on sequence information, only.

Are 1D structure motifs sufficient to detect remote homologues? Aligning strings of observed secondary structure and solvent accessibility mutually (optimal

prediction) yielded more than 50% correctly predicted remote 3D homologues, when the alignment list were cut off at a pre-defined (length dependent) threshold for the alignment z-score. However, only every fourth hit that ranked first in the alignment list was correct. Thus, even the best-case scenario of 1D threading is rather unreliable. Aligning 1D predictions and observed 1D strings, decreased the detection accuracy to about 20%. The reduction from 25% to 20% compares well with the expected accuracy of 1D predictions. Encouraging was that the correct remote homologue ranked in more than 70% of the cases among the first ten hits. As for all threading approaches, the obstacle for TOPITS was to reduce the number of false positives, or in other words to appropriately sort the resulting alignment list.

Can the alignment list be filtered or resorted? The most simple and effective way to filter the alignment list was to consider as correct hits only those for which the aligned sequence contained less than 0.6-0.8 times more residues than the alignment of that sequence with the search sequence. The resulting accuracy for a correctly detected remote homologue ranking at the first position of the alignment reached 30%. A further improvement by similar filters and a resorting of the result list based on a differently compiled z-score (Sippl and Jaritz 1994) appears to be feasible.

How stable are the alignments with respect to free parameters? The results enabled to favour one comparison matrix and one value for the gap open penalty over others. As final gap-open penalty five was chosen, and as final comparison matrix the one that was based on counts reflecting the 1D (PHD) prediction accuracy. The structure-derived matrix was inferior to that matrix. This was surprising, as the dominant shortcoming for TOPITS was not the limited accuracy of 1D prediction, but the loss of information by projecting 3D structure onto 1D. Although, optimal values could be distinguished, performance accuracy was not too sensitive with respect to a small change of the free parameters.

Can TOPITS be combined with sequence alignments? Preliminary results indicate that a combination of 1D structure and sequence alignment could improve the accuracy of detecting remote homologues. This promising trend will have to be confirmed. Different relative contributions of sequence and 1D structure alignments could be combined by a set of logical rules to increase the reliability of the method further.

Can 1D predictions be used to predict other aspects of 3D structure? Yes, remote 3D homologues can be detected, but only at best every third automatically detected hit is correct. Mean-force potential based threading approaches, at least, the more elaborated ones (Sippl and Jaritz 1994, Sippl and Weitckus 1992), may be more accurate (no analysis of any threading method has so far been based on a larger data set, thus, there is no data for comparisons available, yet). However, this does not at all imply that TOPITS is useless. Three scenarios could make TOPITS become an important contribution to improved threading methods. First, conventional threading consumes approximately ten times more CPU time. As the correct hit ranks in some 80% of the cases among the first 20 hits (Figure 2), TOPITS could be used as a fast pre-screening. Second, even when conventional approaches implicitly predict secondary structure and accessibility, TOPITS uses completely different information than potential based threading. Various independent threading methods could be combined to yield much more reliable threading results, but this has yet to be proven. Third, in contrast to other threading methods, TOPITS could predict remote homology for a pair of proteins both unknown in 3D structure. Thus, it could be of use to extent the range of sequence alignments for the prediction of, e.g., protein function.

Acknowledgements. First of all, thanks to Chris Sander (EMBL) for his intellectual, emotional, and financial support. Second, thanks to Reinhard Schneider (EMBL) for valuable ideas, important discussions, and for having tailored his alignment program MaxHom to the purpose of threading 1D predictions into 3D structures. Furthermore, thanks to Michael Braxenthaler (Washington) and Manfred Sippl (Salzburg) for fruitful discussions about threading details. Last not least, thanks to Séan O'Donoghue (EMBL) for his helpful comments on the manuscript.

References.

Abagyan, R.; Frishman, D.; and Argos, P. 1994. Recognition of distantly related proteins through energy calculations. *Proteins* 19:132-140.

Altschul, S. F. 1991. Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.* 219:555-565.

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

Bernstein, F. C. et al. 1977. The Protein Data Bank: a computer based archival file for macromolecular structures. *J. Mol. Biol.* 112:535-542.

Bowie, J. U. et al. 1990. Identification of protein folds: matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins* 7:257-264.

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

Bryant, S. H.; and Lawrence, C. E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins* 16:92-112.

Dayhoff, M. O. 1978. *Atlas of Protein Sequence; and Structure*. Washington, D. C., U. S. A.: National Biomedical Research Foundation.

Eisenberg, D.; Lüthy, R.; and McLachlan, A. D. 1991. Secondary structure-based profiles: use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. *Proteins* 10:229-239.

Greer, J. 1991. Comparative modeling of homologous proteins. *Meth. Enzymol.* 202:239-252.

Henikoff, S.; and Henikoff, J. G. 1992. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sc. U.S.A.* 89:10915-10919.

Henikoff, S.; and Henikoff, J. G. 1993. Performance evaluation of amino acid substitution matrices. *Proteins* 17:49-61.

Hobohm, U.; and Sander, C. 1994. Enlarged representative set of protein structures. *Prot. Sci.* 3:522-524.

Holm, L.; and Sander, C. 1994a. The FSSP database of structurally aligned protein fold families. *Nucl. Acids Res.* 22:3600-3609.

Holm, L.; and Sander, C. 1994b. Parser for protein folding units. *Proteins* 19:256-268.

Johnston, M. et al. 1994. Complete nucleotide sequence of saccaromyces cerevisiae chromosome VIII. *Science* 265:2077-2082.

Jones, D. T.; Taylor, W. R.; and Thornton, J. M. 1992. A new approach to protein fold recognition. *Nature* 358:86-89.

Kabsch, W.; and 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* 18:103-106.

May, A. C. W.; and Blundell, T. L. 1994. Automated comparative modelling of protein structures. *Curr. Opin. Biotech.* 5:355-360.

McLachlan, A. D. 1971. Tests for comparing related amino acid sequences. *J. Mol. Biol.* 61:409-424.

Needlman, S. B.; and Wunsch, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* 48:443-53.

Nishikawa, K.; and Matsuo, Y. 1993. Development of pseudo-energy potentials for assessing protein 3-D-1D compatibility; and detecting weak homologies. *Prot. Engin.* 6:811-820.

Oliver, S., et al. 1992. The complete DNA sequence of yeast chromosome III. *Nature* 357:38-46.

Orengo, C. A., et al. 1993. Recurring structural motifs in proteins with different functions. *Curr. Biol.* 3:131-139.

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

Pearson, W. R.; and Lipman, D. J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sc. U.S.A.* 85:2444-2448.

Richardson, J. S.; and Richardson, D. C. 1989. Principles; and patterns of protein conformation. In Fasman, G. D. ed. *Prediction of protein structure; and the principles of protein conformation.*, 1-98. New York, London: Plenum Press.

Rost, B. 1995. Fitting 1D predictions into 3D structures. In Bohr, H., and Brunak, S. eds. *Protein structure by distance analysis*. CRC Press. Forthcoming.

Rost, B.; and Sander, C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles; and neural networks. *Proc. Natl. Acad. Sc. U.S.A.* 90:7558-7562.

Rost, B.; and Sander, C. 1993b. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232:584-599.

Rost, B.; and Sander, C. 1994a. Combining evolutionary information; and neural networks to predict protein secondary structure. *Proteins* 19:55-72.

Rost, B.; and Sander, C. 1994b. Conservation; and prediction of solvent accessibility in protein families. *Proteins* 20:216-226.

Rost, B.; Sander, C.; and Schneider, R. 1994. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235:13-26.

Sander, C.; and Schneider, R. 1991. Database of homology-derived structures; and the structurally meaning of sequence alignment. *Proteins* 9:56-68.

Sander, C.; and Schneider, R. 1994. The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* 22:3597-3599.

Scharf, M. 1989. Analyse von Paarwechselwirkungen in Proteinen. Ph.D. diss., Department of Physics, University of Heidelberg.

Sheridan, R. P.; Dixon, J. S.; and Venkataraghavan, R. 1985. Generating plausible protein folds by secondary structure similarity. *Int. J. Peptide Protein Res.* 25:132-143.

Sippl, M. J. 1993a. Boltzmann's principle, knowledge based mean fields; and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided Mol. Design* 7:473-501.

Sippl, M. J. 1993b. Recognition of errors in three-dimensional structures of proteins. *Proteins* 17:355-362.

Sippl, M. J.; and Jaritz, M. 1994. Predictive Power of Mean Force Pair Potentials. In Bohr, H., and Brunak, S. eds. *Protein structure by distance analysis*, 113-134. Amsterdam, Oxford, Washington DC: IOS Press.

Sippl, M. J.; and Weitckus, S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* 13:258-271.

Smith, T. F.; and Waterman, M. S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.

Vingron, M.; and Waterman, M. S. 1994. Sequence alignment; and penalty choice. *J. Mol. Biol.* 235:1-12.

Wako, H.; and Blundell, T. L. 1994. Use of amino acid environment-dependent substitution tables; and conformational propensities in structure prediction from aligned sequences of homologous proteins I. Solvent accessibility classes. *J. Mol. Biol.* 238:682-692.

Wilmanns, M.; and Eisenberg, D. 1993. Three-dimensional profiles from residue-pair preferences: Identification of sequences with β/α-barrel fold. *Proc. Natl. Acad. Sc. U.S.A.* 90:1379-1383.

Wodak, S. J.; and Rooman, M. J. 1993. Generating; and testing protein folds. *Curr. Opin. Str. Biol.* 3:247-259.