
PROTEIN FOLDS

A Distance-Based Approach

Edited by

Henrik Bohr and Søren Brunak



CRC Press

Boca Raton New York London Tokyo

Library of Congress Cataloging-in-Publication Data

Protein folds : a distance based approach / edited by Henrik Bohr and Søren Brunak.

p. cm.

Includes bibliographical references and index.

ISBN 0-8493-4009-8

1. Protein folding. 2. Protein binding. I. Bohr, Henrik. II. Brunak, Søren.

QP551.P695823 1995

574 19' 245--dc20

95-35601

CIP

This book contains information obtained from authentic and highly regarded sources. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

Neither this book nor any part may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, microfilming, and recording, or by any information storage or retrieval system, without prior permission in writing from the publisher.

All rights reserved. Authorization to photocopy items for internal or personal use, or the personal or internal use of specific clients, may be granted by CRC Press, Inc., provided that \$.50 per page photocopied is paid directly to Copyright Clearance Center, 27 Congress Street, Salem, MA 01970 USA. The fee code for users of the Transactional Reporting Service is ISBN 0-8493-4009-8/96/\$0.00+.50. The fee is subject to change without notice. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

CRC Press, Inc.'s consent does not extend to copying for general distribution, for promotion, for creating new works, or for resale. Specific permission must be obtained in writing from CRC Press for such copying.

Direct all inquiries to CRC Press, Inc., 2000 Corporate Blvd., N.W., Boca Raton, Florida 33431.

© 1996 by CRC Press, Inc.

No claim to original U.S. Government works

International Standard Book Number 0-8493-4009-8

Library of Congress Card Number 95-35601

Printed in the United States of America 1 2 3 4 5 6 7 8 9 0

Fitting 1-D Predictions into 3-D Structures

Burkhard Rost

EMBL, Meyerhofstrasse 1, D-69012 Heidelberg, Germany, rost@embl-heidelberg.de

Abstract

The experimental determination of protein structure cannot keep track with the rapid generation of new sequence information. Can theory contribute? The most successful prediction method — and the only one for prediction of 3-D structure — is homology modelling. It is applicable for about one quarter of the proteins. For the rest, the prediction task has to be simplified. An extreme simplification is to project 3-D structure onto 1-D strings of secondary structure or solvent accessibility. For these 1-D aspects of 3-D structure, prediction accuracy has been improved significantly by using evolutionary information as input to neural network systems. The gain in accuracy is based on the conservation of secondary structure and relative solvent accessibility within sequence families. Secondary structure and accessibility are conserved, as well, between remote homologues. This fact can be used by fitting 1-D predictions into 3-D structures to detect such remote homologues. In comparison to other threading approaches, 1-D threading is rather flexible. However, two factors decrease detection accuracy. Firstly, the loss of information by projecting 3-D structure onto 1-D strings (in particular the loss of distances between secondary structure segments). And secondly, the inaccuracy of predicting 1-D structure. A preliminary result is that every fifth remote homologue is detected correctly.

1 Introduction

Sequence determines structure, determines function. It is generally assumed that three-dimensional (3-D¹) protein structure is uniquely determined by sequence [1]. There are proteins assisting the folding *in vitro*, the chaperones. Are such assistants necessary for the majority of folding pathways? And do chaperones purely prevent unfolding, or are they actually necessary for folding? Answers remain open [2, 3, 4]. Anyway, the assumption that sequence determines structure constitutes a reasonable approximation. The function(s) of a protein are determined by its structure and environment (reagents). Thus, protein sequence

¹Abbreviations used: 3-D, three-dimensional; 2-D, two-dimensional; 1-D, one-dimensional; PDB, Protein Data Bank of experimentally determined 3-D structures of proteins; SWISSPROT, data base of protein sequences; DSSP, data base containing the secondary structure and solvent accessibility for proteins of known 3-D structure; HSSP, data base containing for each PDB protein of known 3-D structure the alignments of all SWISSPROT sequences homologue to the known structure; FSSP, data base of remote homologues of known 3-D structure; PHD, Profile based neural network prediction of secondary structure (PHDsec) and solvent accessibility (PHDacc).

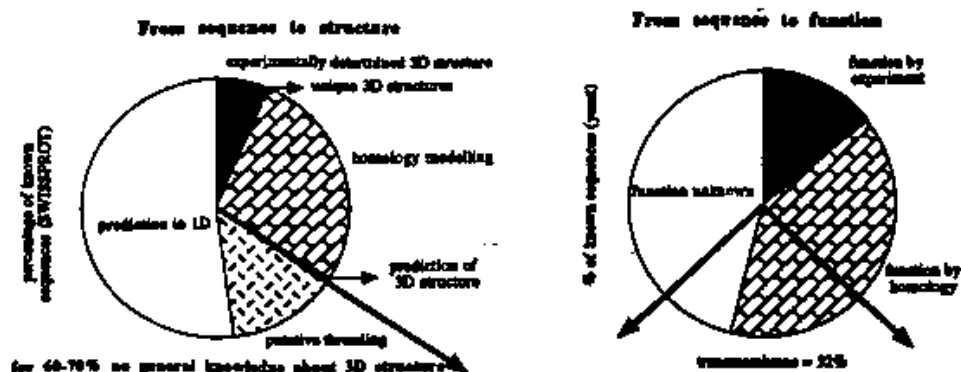


Figure 1: Prediction of protein structure and function — state of the art. *Prediction of structure:* For 70% of all known sequences, no knowledge about 3-D structure is available, in general. Threading methods may reduce this value to some 60% if they will become publicly available and sufficiently accurate for large scale sequence analysis. Today, for the majority of known protein sequences, the only successful sequence analyses are multiple alignments and predictions in 1-D. *Prediction of function:* For some 40-50% of the protein sequences obtained from sequencing whole genes (yeast III and VIII) some knowledge about function is either given by experiment or by homology modelling ([7]). Furthermore, for some 25-30% of the proteins transmembrane segments can be predicted (estimate derived from an analysis of 171 yeast VIII sequences, [68]).

determines function to a certain extent. Consequently, structure and function can, in principle, be induced from the sequence based on physico-chemical properties. In practice, this is prevented by the enormous complexity of the phase space. Can theory contribute to the advance of molecular biology, nevertheless?

The objective of theory is to reduce the sequence-structure and sequence-function gaps. Large scale gene-sequencing projects produce data of gene and consequently protein sequences at breathtaking pace [5, 6]. Knowledge about function and/or structure of the proteins plays a crucial role in designing experiments using such data. However, only for the minority of proteins with known sequence, the 3-D structure is also known. Experimental determination of protein function is easier than that of structure. Consequently, the sequence-function gap is less exposed than the sequence-structure gap [7]. In the near future, the gaps are unlikely to be reduced significantly by experiments. How does theory contribute today to reduce the gaps?

Homology modelling is the most powerful theoretical tool. The only reliable technique to predict both protein structure and protein function is homology modelling: for a search sequence its 3-D structure and function is predicted based on a sequence alignment to proteins of known 3-D structure and/or function [8, 9, 10, 11, 12]. This method increases the number of known 3-D structures by a factor of five [13], and the number of proteins with some knowledge about function by a factor of three [7, 14] (Fig. 1). What if there is no protein with significant sequence identity ($> 25\%$) in the data base of known structures?

Threading techniques may become a second comprehensive tool. Roughly, every fifth protein in PDB is remotely homologue to another PDB protein [15], i.e. has a homologous 3-D structures but no significant pairwise sequence similarity ($< 25\%$ [16]). Thus, theory

could reduce the sequence-structure gap by another 10-20% of the known sequences if remote homology were detected based on sequence information (as homology is by sequence alignment). One way to predict remote homology is by threading sequences into structures [17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27]. The principle objective is to define a criterion that enables to evaluate whether or not the sequence fits into the structure. If both homology modelling and threading fail, can theory predict 3-D structure from sequence?

In general, prediction is limited to 1-D. Despite advances in detail, the ability to predict 3-D structure from sequence by theoretical means has not much improved. In general, 3-D structure cannot be predicted ab initio. The prediction problem has to be simplified. One extreme simplification is projecting 3-D structure onto 1-D strings of secondary structure assignments (e.g. helix, strand, rest). Improving the accuracy of secondary structure prediction has been protruded for the last three decades. Another 1-D feature of 3-D structure that had been subject to prediction methods is the position of a residue with respect to the surface of a protein, i.e. its solvent accessibility. Predictions of such 1-D features are of limited accuracy, but are applicable in general. Can predictions of simplified 1-D aspects of 3-D structure be used successfully as a starting point to predict structure or function?

In some cases, 1-D predictions can be used to infer aspects of function and 3-D structure. Supposed, the pattern of predicted secondary structure elements, e.g. "helix-strand-helix-strand", were similar to either the secondary structure pattern of a protein with known 3-D structure, or to the pattern of a protein for which secondary structure is predicted and not experimentally determined. Then, the hypothesis that the two proteins are remote homologues may be used to predict 3-D structure and/or function. For selected cases, remote homology modelling has been used to predict 3-D structure (e.g. [28]). Is such a procedure necessarily restricted to selected cases, or can it be done automatically?

Can 1-D predictions seed an automatic prediction of remote homologues? Here, some preliminary results will be given to answer this question. The goal is to recognise remote homologues based on alignments of a projection of 3-D structure onto 1-D strings of secondary structure and solvent accessibility assignments. First, the results of the underlying methods for predicting secondary structure and solvent accessibility will be sketched briefly. Second, the principles of the threading method will be described, and some strategies for the optimisation of free parameters will be given.

2 Combining evolutionary information and neural networks

2.1 Prediction of secondary structure

Prediction of secondary structure based on single sequences is limited to < 65% accuracy. The basic idea of most prediction methods is that stretches of adjacent residues from protein sequences are unique. E.g. given a pentapeptide of five consecutive residues, is the central residue of all equal pentapeptides always observed in the same secondary structure? For pentapeptides, this is indeed not the case [29]; peptides of e.g. 13 residues are unique. However, for prediction purposes it is not sufficient to just search for an identical peptide of 13 residues, as it may not be contained in the data base. Instead, the goal is to classify similar patterns in either of the three classes: helix, strand, or rest. Secondary structure prediction based on pattern classification has been pursued even before the first X-ray structure was determined [30, 31, 32]. A decade ago, prediction accuracy reached some 50-55% three-state accuracy [33] (percentage of residues predicted correctly in either of the three states). More advanced algorithms and increased data bases pushed the accuracy to 60-65%, a mark that was long taken as insurmountable [34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. The main

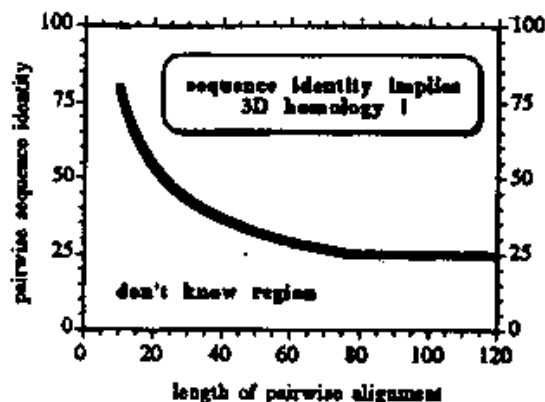


Figure 2: Information about 3-D structure contained in evolutionary records. All protein pairs in the current data base (PDB [69]) that have 30 out of 100 aligned residues in common have homologous 3-D structure [16]. Thus, any sequence pair above the line for which one sequence has a known 3-D structure, can be used for prediction of 3-D structure by homology modelling. All 3-D homologues that fall below the line are referred to as remote homologues; some 4,000 remote homologues are known today (FSSP, [15]). Remote homologues could be detected based on sequence information by successful threading techniques. To illustrate the problem of threading by numbers: more than one billion pairs fall under the line, of these only some thousands of pairs are remote homologues.

difficulty was that the input information contained in stretches of 13-21 consecutive residues is not sufficient. Do protein sequences contain any additionally information about 3-D structure that could be used for prediction?

Sequence families contain much more information than single sequences. Protein sequence determines protein structure. But, how unique is this relation? How many residues can be exchanged in a protein without changing the 3-D structure? Evolutionary pressure to maintain protein function has explored paths to exchange about 75% of all residues without changing the 3-D structure (Fig. 2, [16]). Can any three out of four amino acid be exchanged against any other? Not at all, instead a random exchange of some residues will often be sufficient to destabilize a structure. Thus, the detailed pattern of amino acid exchanges found in sequence families is highly informative about the structure of that family [45]. Are 1-D projections of 3-D structure, such as secondary structure and solvent accessibility, conserved between 3-D homologues?

Secondary structure was conserved to some 90% between 3-D homologues. Sequence alignments can be used to predict secondary structure for proteins for which there is a known structure with significant sequence identity (all above the line in Fig. 2). Within sequence families, the pairwise identity of secondary structure segments was some 90% (Fig. 3, [46, 47]). In other words, two proteins can adopt the same 3-D structure and yet differ in the secondary structure assignment by about 10%. Conservation was higher in the core than at the surface [46]. Can evolutionary information be used to improve secondary structure predictions?

Evolutionary information improved prediction accuracy to > 72%. The way to use evolutionary information for prediction was the following. First, the data base of known

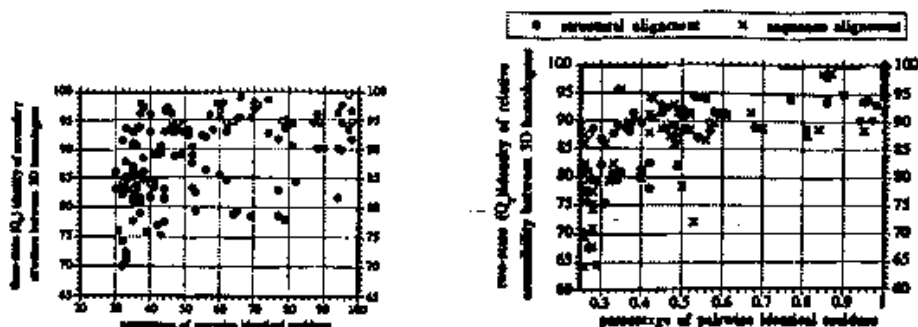


Figure 3: Conservation of 1-D structure between 3-D homologues. (a) Conservation of secondary structure (averaged over protein pairs) vs. the percentage of identical residues between the sequences of the pair. Homologues were aligned on sequence-based [16] comparisons. The identity is given as the overlap of secondary structure segments between the pairs [46]. (b) Conservation of relative solvent accessibility (averaged over protein pairs) vs. the percentage of identical residues between the sequences of the pair. Homologues were aligned both based on structure-based [15] and on sequence-based [16] comparisons. The identity is given as the percentage of residues in either of the two states for relative solvent accessibility (buried, exposed, [53]).

sequences was scanned by sequence alignment methods [16] for similar sequences. Second, the list of sequences found was filtered by the length-dependent threshold for significant sequence identity (Fig. 2). Third, for all probable 3-D homologues, a profile of amino acid exchanges was compiled. Fourth, this profile was used for prediction. The first method that has been proven in a cross-validation experiment based on 250 unique protein chains to predict secondary structure at a sustained level of > 72% three-state accuracy was a neural network [42, 48]. For this method the profiles, and additional information derived from the multiple sequence alignments, were fed as input into the neural network system. Various further details of the architectures of a network system composed of two layers of networks and a final compilation of an arithmetic average over independently trained networks, were important to yield a prediction with a high overall accuracy, a relative accurate prediction of strands, and succeeded to correctly predict secondary structure segments rather than single residues (Table 1, [42, 43]). For the 40% of all residues predicted with higher reliability, the method (dubbed PHDsec) reached a value of 90%, i.e. was as accurate as homology modelling would be if applicable. Almost ten percentage points of the improvement in overall accuracy stemmed from using evolutionary information. Does evolutionary information also improve the prediction of other aspects of 3-D structure?

2.2 Prediction of solvent accessibility

Solvent accessibility depends on residue type and structural environment. The solvent accessibility of a residue embedded in a protein can be measured by the number of water (solvent) molecules that can be placed around it [49]. Values typically vary between 0 and 200 Å². To compare residues of different sizes, a relative solvent accessibility has to be compiled

Table 1: Accuracy of secondary structure prediction.

Abbreviations of methods: HM, homology modelling; RAN, random sequence alignments; PHDsec, neural network prediction using multiple alignment information as input [42, 48]; SIMPA, similarity based statistical prediction method using single sequence information only [70] (note: SIMPA has not the highest accuracy reported for classical methods, but it scored better than others on the particular data set used here [42]); LPAG, statistical prediction method using multiple alignment information [71].

Abbreviations of measures: Nprot, number of proteins used for testing (all prediction results hold for test proteins with less than 25% sequence identity to the proteins used to derive the prediction method); set, different test sets are numbered to indicate identical sets (1:[46], 2, 3, 4:[42], 5 and 6: proteins of recently determined 3-D structure); Q_3 , three-state overall per-residue accuracy, i.e. number of residues predicted correctly in either of the three states helix, strand, rest; I, information, entropy measure for prediction accuracy [46, 48]; Sov_3 , three-state overall per-segment accuracy, i.e. the overlap of predicted and observed secondary structure segments (rather than single residues) [46]; date, whenever new protein structures are added to the data base (PDB) the neural network prediction was re-evaluated.

method	set	Nprot	Q_3	I	Sov_3	date
HM	set 1	80	88.4	0.62	89.7	
RAN	set 1	80	35.2	0.01	30.6	
PHDsec	set 2	126	71.6	0.27	72.8	Jun, 93
PHDsec	set 3	124	72.5	0.28	75.6	Aug, 93
SIMPA	set 3	124	60.7	0.12	61.7	
PHDsec	set 4	60	74.8	0.34	76.8	
LPAG	set 4	60	68.5			
PHDsec	set 5	27	72.0	0.28	72.4	May, 94
PHDsec	set 6	59	73.0	0.30	75.7	Nov, 94

[51, 52, 53]. Conservation of solvent accessibility between 3-D homologous pairs is best analysed by computing the correlation between the relative solvent accessibility of the two [53]. Here, a simpler measure is used, a two-state description of solvent accessibility (buried < 16% solvent accessible; exposed \geq 16% solvent accessible, [54, 55, 56]). If protein cores were conserved completely between 3-D homologues, the two-state identity should be 100%. Is solvent accessibility conserved that well within protein families?

Solvent accessibility was conserved to some 85% between 3-D homologues. Conservation of solvent accessibility in two states is found to be clearly less than 100% (Fig. 3, [53]). A possible reason may be that sequences were not correctly aligned at low levels of sequence identity (sharp decrease around 40%, Fig. 3). Indeed, the decrease in conservation was less significant for structural alignments of the same protein pairs (Fig. 3). However, even for structural alignments, solvent accessibility was conserved to only 85% in two states. Furthermore, conservation dropped for low levels of sequence identity (Fig. 3). This observation sheds a light on why the accuracy of automatic homology modelling decreases with decreasing sequence identity. When three states were distinguished for relative solvent accessibility (buried,

Table 2: Accuracy of solvent accessibility prediction.

Abbreviations of methods: HM: SeqAli, homology modelling, alignments based on sequence comparisons [53]; HM: StrAli, homology modelling, alignments based on structural comparisons [53]; PHDacc, neural network prediction based on multiple alignments [53]; Wako, statistical prediction method based on multiple sequence alignments [72].

Abbreviations of measures: The protein sets are as in Table 1 (set 3a is a subset of set 3); Q_3 , three-state overall per-residue accuracy, i.e. percentage of residues correctly predicted in either of the three-states buried, intermediate, exposed [53]; Q_2 , two-state overall accuracy (buried, exposed); Corr, correlation between predicted and observed relative solvent accessibility [53].

method	set	Nprot	Q_3	Q_2	Corr	date
HM: SeqAli	set 1	80	71.6	83.8	0.68	
HM: StrAli	set 1	80	73.6	84.8	0.77	
RAN	set 1	80	33.9	52.0	0.01	
PHDacc	set 2	126	57.9	75.0	0.54	
PHDacc	set 3a	112	57.9	74.7	0.54	Mar. 94
PHDacc	set 7	13	60.8	79.2	0.61	
Wako	set 7		76.5			
PHDacc	set 5	27	57.6	73.4	0.55	May, 94
PHDacc	set 6	59	57.0	74.0	0.54	Nov, 94

intermediate, exposed), the identity for 3-D homologues was reduced to an average of < 75% [53]. As the distribution of these three states is comparable to that of the three secondary structure states, the conclusion is that solvent accessibility is less well conserved in evolution than is secondary structure. Does this imply that the prediction of solvent accessibility is less accurate than the prediction of secondary structure?

Solvent accessibility was predicted at 75% two-state accuracy. The profile based neural network system used (named PHDacc) was similar to PHDsec [53]). A cross-validation experiment on 238 unique protein chains yielded an expected two-state accuracy of 75% (residues correctly predicted as either buried or exposed), and an expected three-state accuracy (buried, intermediate, exposed) of < 60% (Table 2). Thus, solvent accessibility was predicted less accurately than secondary structure. Does this imply that the network has intrinsically more difficulties with the accessibility prediction? To find out, the prediction accuracy was normalised such that on the normalised scale a prediction by homology yielded 100% and a random prediction 0%. On this normalised scale solvent accessibility prediction turned out to be more accurate than secondary structure prediction [53]. In other words, PHDacc came closer to the optimal prediction performance given by the conservation of solvent accessibility than PHDsec to the mark given by the conservation of secondary structure (Table 1, Table 2). Thus, the low level of accuracy for predicting solvent accessibility mainly stems from the fact that this aspect of 3-D structure is less conserved between 3-D homologues.

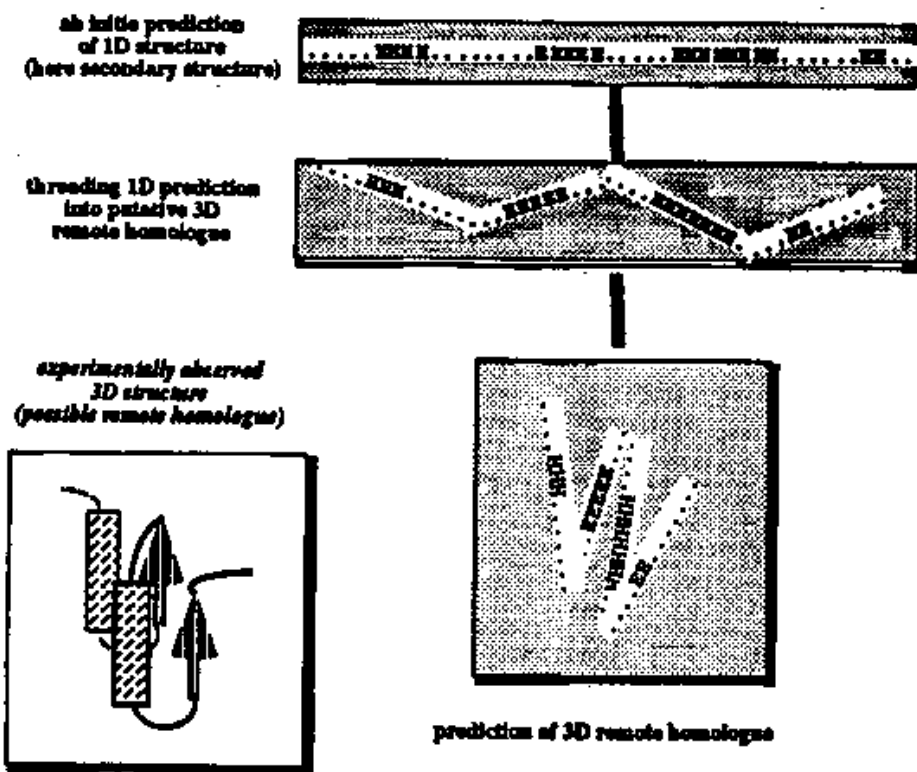


Figure 4: Threading 1-D strings into 3-D structures. The principle idea of threading 1-D predictions into 3-D structures is the following. First, for a search sequence secondary structure and solvent accessibility are predicted. Second, the known 3-D structures are projected onto 1-D strings of secondary structure and solvent accessibility (DSSP, [50]). Third, the predicted 1-D strings are aligned to the 1-D strings of known structures. The best match of the dynamic programming algorithm constitutes the predicted remote homologue for the search sequence.

3 Aligning 1-D strings of secondary structure and solvent accessibility

3.1 Principle idea of 1-D threading

How can 1-D predictions be used to predict more aspects about 3-D structure? The simple idea is the following. Protein folds often are described by motifs of secondary structure (e.g. H-E-H-E) in combination with a table of inter segment distances (e.g. "H1 near to H2, E1 near to E2, strands on top of helices"). But even without the distance table, a prediction of "H-E-H-E" combined with a certain pattern of solvent accessibility states should suffice to, at least, formulate the hypothesis that the structure sketched in figure 4 is homologue to the search sequence.

Detection of motifs by alignment of 1-D strings. The mostly studied technique to detect similar motifs in 1-D strings is dynamic programming [57, 58, 59, 60]. The application to

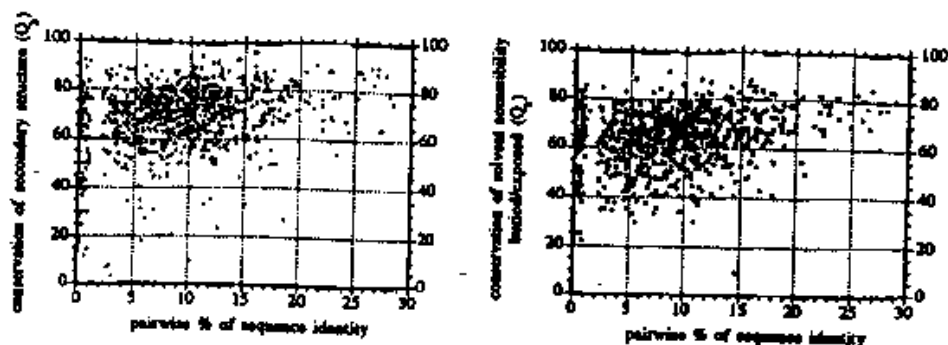


Figure 5: Conservation of secondary structure and solvent accessibility between remote homologues. The conservation of secondary structure is given as the percentage of identical residues in either of the three states helix, strand and rest; the conservation of solvent accessibility as the percentage of residues identical in either of the two states buried (relative accessibility $< 16\%$) and exposed (relative accessibility $\geq 16\%$). Each point reflects one of a total of 2211 protein pairs structurally aligned in the FSSP data base [15].

1-D threading worked as follows. First, for a list of proteins of known 3-D structure (typically unique set, [61]) projections of 3-D structure onto strings of secondary structure and relative solvent accessibility assignments were computed (DSSP, [50]). Second, a search sequence of unknown structure was aligned against the sequence data base (SWISSPROT [62]), and used as input to neural network systems that predicted secondary structure and solvent accessibility. Third, prediction and observation were extracted into strings composed of a six letter alphabet (3×2), for the three secondary structure states, and the two relative accessibility states. Fourth, the string predicted for the search sequence was aligned with the strings extracted from the data base of observed structures. (The program used for the alignment procedure was a modified version of the sequence alignment program *MaxHom* [13, 16]).

Secondary structure and solvent accessibility was conserved between remote homologues. 1-D threading is possible only if secondary structure and solvent accessibility are conserved, not only within sequence families (Fig. 3), but also between remote homologues. The average conservation of secondary structure was 69% (Fig. 5, overall three-state identity, one standard deviation 13%); the average conservation of relative solvent accessibility 65% (Fig. 5, overall two-state identity, one standard deviation 10%). Thus, both secondary structure and relative solvent accessibility were largely conserved between remote homologues. An interesting detail was that, although both 1-D aspects were less conserved for remote homologues than for homologues with significant pairwise sequence identity (Fig. 3), the conservation did not decrease significantly with sequence identity.

3.2 Adjusting free parameters for alignment

How can free alignment parameters be optimised? The success of alignment algorithms depends on two important groups of free parameters. First, the values for the comparison matrix: matches or mismatches between the six states of the strings aligned have not equal significance with respect to the structural motif searched for. Second, the penalty given for

introducing a gap ("gap open penalty") and for continuing a gap ("gap elongation penalty"). Less important (in terms of less sensitive to the stability of the resulting alignment) are the minimal and maximal value for the scoring matrix (chosen as: maximal match score = 1; minimal (mis-)match score = -1). Optimisation of the free parameters is a highly non-trivial task; the main obstacle being the lack of a clear-cut optimisation criterion. In the following some optimisation strategies and some particular choices for the free parameters will be discussed.

Distinguishing random alignments and true remote homologues. Some relations between the elements of the scoring matrix are evident from expert knowledge. For example, a mismatch between buried helix and exposed strand is much worse than a mismatch between buried helix and buried loop. But how should the values be chosen exactly? What is the optimisation criterion for the choice of the scoring matrix? The alignment goal is to detect true remote homologues. This implies not only that remote homologues are detected, but also that the alignment procedure distinguishes between remote homologues (termed "correct hits") and non-homologous proteins (termed "false positives" or "incorrect hits"). Thus, one idea for the optimisation of the scoring matrix is to choose the values such that scores for correct and incorrect hits become maximally separated.

Different concepts of maximal separation of distributions. For some thousand pairs, the scores were computed for random alignments (incorrect hits) and structurally aligned remote homologues (correct hits). The distributions do overlap (Fig. 6). Various matrices were generated by expert driven trial and error. No choice resulted in an overlap of zero. Thus, the next question is what 'maximal separation' means. Various concepts are feasible: (i) maximize the difference between the averages of the distributions, (ii) maximize the quotient between the differences of averages normalised with the product of the standard deviations, or (iii) minimize the area of overlap. The alternatives (i)-(iii) did not result in the same choice for the free variables of the scoring matrix. The separation shown in Fig. 6 was optimal with respect to criterion (ii) (and almost best for (i) and (iii)).

Secondary structure and solvent accessibility superior to either of the two. Would the separation between correct and incorrect hits be less distinct, if either only secondary structure or only accessibility were used, instead of both combined? Indeed, the separation was less for either of the two 1-D features alone than for a combination of the two. However, the degree of separation that was gained by adding relative solvent accessibility states to the description in terms of secondary structure assignments, was rather small (Fig. 6).

Optimisation of gap penalties with respect to remote homologues detected. There are three reasons to introduce gaps. First, loop regions can be inserted between regular secondary structure segments without influencing the basic fold [63]. Second, slightly elongated or shortened helices or strands can form similar motifs [46]. Third, predictions are generally more accurate for the core of predicted segments than at the ends, in other words it is less likely that a whole segment is predicted falsely than that some residues at the ends of segments are predicted wrongly [43]. The optimisation criterion for gap penalties is different than that for the scoring matrix in that gap penalties have to be optimised with respect to a given alignment list (3.3.). If the goal is to minimize the percentage of false positives (3.3.), then the optimal values for a search with PDB against PDB (3.4.) were: gap open penalty = 1.0; gap elongation penalty = 0.1. However, if the goal is to have one correct prediction at first rank, higher values (3/0.3, 5/0.5) are better.

3.3 Measuring prediction accuracy for threading

Sorting the alignment list by normalised alignment scores. For one sequence (search sequence), the predicted strings are aligned against a list of typically some 300 proteins of known structure.

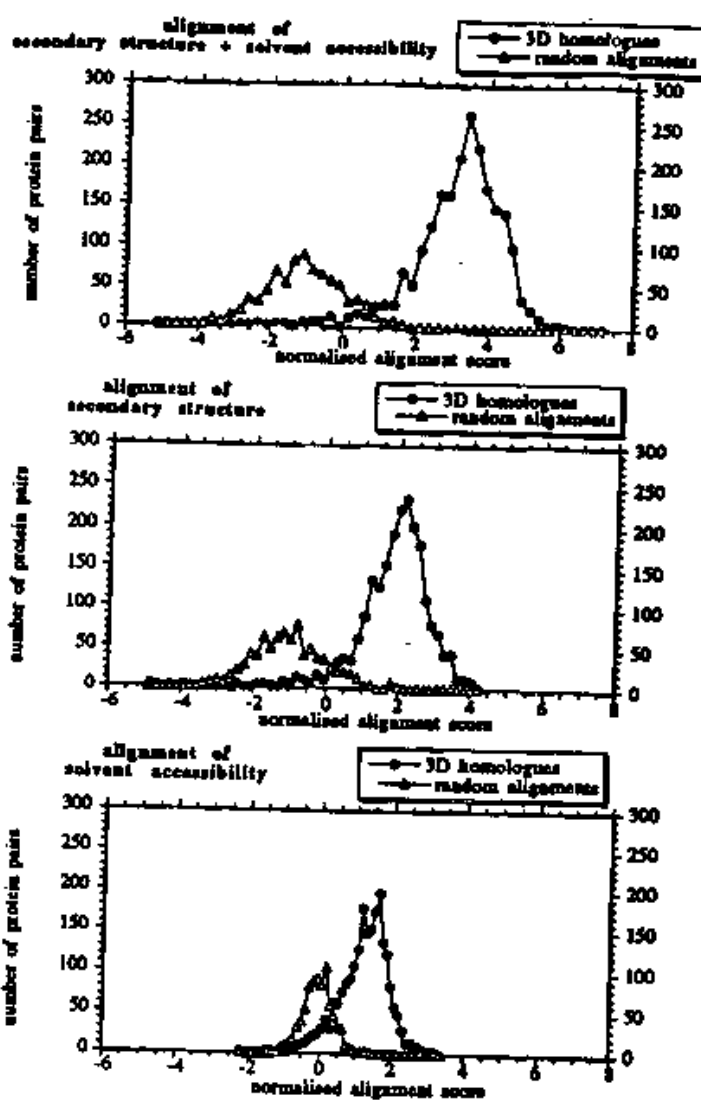


Figure 6: Separation of 3-D homologues and random alignments. For three different scenarios, the separation of random alignments (generated by sequence alignments of non-homologous PDB pairs with no significant pairwise sequence identity) and structural alignments of 3-D homologues (taken from FSSP, [15]) are compared. First, strings of secondary structure and relative solvent accessibility were aligned (six-letter alphabet: 3×2 , for three secondary structure types helix, strand, rest and two states for relative solvent accessibility, buried and exposed). Second, strings of secondary structure were aligned (3 states). And third, strings of relative solvent accessibility were aligned (2 states).

Since alignments can involve different sub-regions of a protein, the result of one threading experiment is a list of several hundreds of alignments. Previously published threading methods uncovered that it is crucial for the success, to adequately sort the alignment hit list. Sorting by scores, comparable to the alignment score compiled here, is not sufficient [64, 65]. Instead, only if the score is normalised with respect to the background given by the data base, the distinction between correct and false positives becomes feasible [64]. The normalisation used here does not alter the simple sorting of the hit list by the alignment score, but enables a comparison of scores between different search sequences:

$$zE_i = \frac{E_i - \langle E \rangle}{\sigma} \quad (1)$$

where E_i is the alignment score for the i 'th hit in the list of all alignments obtained for a search protein, $\langle E \rangle$ is the average over all hits for that sequence and σ the standard deviation of the distribution of all hits. In other words, zE_i describes the significance of a given hit i , for a search sequence.

Lack of widely accepted standards to measure accuracy. Due to historical reasons, threading results in the past have put too much emphasis on "find-self" measures, i.e. the ability of a threading potential to correctly find the search sequence among a large number of decoys [22, 23, 64, 65, 66]. The measures defined here are restricted to analysing the detection of remote homologues. Another important issue is the correctness of the alignment (Holm, Ouzounis, Sander, Sippl, manuscript in preparation).

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 (Table 3). The most important is the simplest: how many of the first hits are correct (eq. (T3a))? Less strict is to cut the alignment list at a given threshold and to count the percentage of correct hits in the remaining list (eq. (T3c)). Such a number is of interest if an expert user could manage to reduce the list further, or if it were more important for the user to detect a remote homologue than to rely on the correctness of a given hit.

More than some "favourable cases" are necessary for evaluating threading. So far, most publications on the threading issue used a couple of "favourite remote homologues" to test the method. Such a procedure is rather arbitrary. A more reasonable approach would be to start with a list of unique proteins [61, 67], and to compile for each of these proteins all remote homologues. A first version of such an approach has been explored here. All results given are based on structural alignments of 46 protein chains. (A more comprehensive list is currently being collected.)

3.4 Preliminary results

Expert knowledge helps to separate correct hits from false positives. For some 40% of all proteins, one correct hit was found in the alignment list among the first four hits. As an explicit example for a 1-D threading the first 15 hits for the search sequence TU-elongation factor (1etu) are shown (Fig. 7). Due to the high number of inserted residues (*IDEL* in Fig. 7) an expert would have probably ignored all hits before the fourth (dihydropteridine reductase, 1dhr), or even before the sixth (ras-p21, 5p21). Both are correctly detected remote homologues. For the TU-elongation factor many of the first hits were correct and the only hits without exceptionally long insertions were correct. The optimistic conclusion is that 1-D threading is successful in detecting remote homologues. Does this hold in general? Is it always possible to distinguish between correct and incorrect hits based on some expert criteria?

Three scenarios: PDB vs. PDB; PHD vs. PDB; and PHD vs. PHD. Three different scenarios were compared to analyse threading performance. Firstly, the case of an error-

Table 3: Measures for accuracy in detecting remote homologues. For each prediction of a remote homologue (each search sequences) there are two lists of alignments, one summarising the true remote homologues (as given in e.g. FSSP [15]), and the other the predicted homologues (sorted by any score). Q_{1st} estimates how often a correct answer will come out if only the first hit is used as prediction; $Q_{1stcorrect}$ gives the average position of a correct hit; $Q_{zE(n)}$ gives the percentage of correct hits if the alignment list is cut according to a given criterion (e.g. how many of the hits with a z-score (Eq. (1)) $zE > n$ are correct?); $Cov_{zE(n)}$ coverage of true positives (i.e. how many of the true remote homologues are predicted at a given cut-off threshold). Note: all counts refer to lists, which are purged off from trivial hits (such as hits with significant pairwise sequence identity). In practice, this is accomplished by simply excluding all hits from the list, which are contained in HSSP files [13], i.e. could have been detected by simple sequence alignment.

$$Q_{1st} = 100 \times \frac{\text{number of correct first hits}}{\text{number of all proteins}} \quad (T3a)$$

$$Q_{1stcorrect} = 100 \times \frac{1}{\text{position of first correct hit}} \quad (T3b)$$

$$Q_{zE(n)} = 100 \times \frac{\text{number of correct first hits with } zE > n}{\text{number of hits with } zE > n} \quad (T3c)$$

$$Cov_{zE(n)} = 100 \times \frac{\text{number of correct first hits with } zE > n}{\text{number of true remote homologues } (N_r)} \quad (T3d)$$

free prediction of secondary structure and solvent accessibility represented by alignments of observed with observed strings (PDB-PDB). Secondly, predicted strings are aligned with observed ones (PHD-PDB). Thirdly, predicted strings for the search sequence are aligned with predicted strings for the putative homologues. The first case of an optimal prediction was tested to analyse the degree to which the reduction of the information by predicting 3-D structure onto 1-D strings (and by losing distance information) results in ambiguities. The last case of aligning predictions with predictions was motivated by the hypothesis that prediction errors may to a certain degree be non-random, i.e. the prediction may e.g. always fail to predict a certain twisted strand in a fold. In that case the prediction for the search sequence and for the remote homologue may align better than the prediction for the search sequence and the observed structure for the remote homologue.

High accuracy in predicting the 1-D strings is crucial. The following four results can be summarised. First, a high accuracy in predicting secondary structure and solvent accessibility is very important for the 1-D threading to work out (Table 4): an alignment of PDB with PDB (i.e. observation with observation) is almost twice as accurate as an alignment of PHD with PDB (i.e. prediction with observation). Second, the projection of 3-D structure onto 1-D results in a detection accuracy of about 50%. Third, an alignment of PHD with PDB is slightly superior to an alignment of PHD with PHD, i.e. the errors in secondary structure and solvent accessibility prediction are to a certain extent not random. Fourth, the accuracy in detecting remote homologues was about 20%.

Refinement by filtering the alignment list. The single prediction example discussed (Fig. 7) suggested that there may be other criteria than the z-score (eq. (1)) to distinguish between correct and false positives. Preliminary results confirm what has been uncovered by others before (e.g. Braxenthaler and Eisenberg, these Proceedings): an adequate combination of

POS	E	LEN	IDEL	ZE	%IDE	STRH	OK	ID2	NAME2
1	80.5	141	135	2.23	0.11	0.96		2ctc	CARBOXYPEPTIDASE A
2	79.7	139	103	2.20	0.07	0.91	*	1dri	D-RIBOSE-BINDING PROTEIN
3	78.6	141	73	2.15	0.08	0.90	*	1dhr	DIHYDROPTERIDINE REDUCTASE
4	78.0	141	152	2.12	0.09	0.92		1gsc	HEAT-SHOCK PROTEIN
5	79.0	139	123	2.11	0.10	0.94		1bil	LEUCINE AMINOPEPTIDASE
6	77.9	122	49	2.11	0.09	0.91	*	5p21	RAS P21 PROTEIN
7	76.4	141	120	2.04	0.06	0.91		1gal	GLUCOSE OXIDASE
8	76.4	141	138	2.04	0.11	0.91	*	1gd1	GLYCERALDEHYDE-PHOSPHATE DEHYDROGENASE
9	76.0	141	65	2.02	0.07	0.82	*	1far	FERRDOXIN
10	75.6	141	111	2.00	0.08	0.87	*	1ipd	3-ISOPROPYLGLUTATE DEHYDROGENASE
11	75.3	141	118	1.98	0.10	0.90		1fba	FRUCTOSE-BISPHOSPHATE ALDOLASE
12	75.1	135	90	1.98	0.08	0.87	*	1pft	PHOSPHOFRUCTOKINASE
13	74.8	141	107	1.96	0.09	0.88		1pda	2-ORPHEBILINOGEN DEAMINASE
14	74.7	140	134	1.96	0.08	0.93		1omp	D-HALTODEXTRIN-BINDING PROTEIN
15	74.7	141	134	1.96	0.06	0.90	*	8abp	ARABINOSE-BINDING PROTEIN

Figure 7: Example for a threading list. Search sequence was 1etu, the TU-elongation factor. Abbreviations: POS, position in list; E, alignment score; LEN, length of alignment; IDEL, number of residues inserted in either of the two aligned strings; ZE, z-score for alignment score (Eq. (1)); %IDE, ratio of pairwise sequence identity; STRH, structural homology (identical symbols in six-letter alphabet); OK = *, if the pair is a true 3-D homologue, i.e. is contained in the FSSP data base [15]; ID2, PDB identifier of aligned 3-D structure; NAME2, SWISSPROT name of aligned 3-D structure.

information from the alignment list (number of residues inserted, length of alignment, length of search sequence and of remote homologue) improves the accuracy in detecting remote homologues. In general, there is a trade-off between a high level of reliability of the prediction (eq. (T3c)) and the likelihood that an existing remote homologue is detected (eq. (T3d)). Different threshold criteria can shift the balance between higher coverage and higher accuracy (Fig. 8).

4 Conclusions

Evolutionary information is the key to accurate predictions in 1-D. Predictions of secondary structure and solvent accessibility were improved significantly by tailoring the neural network systems to the problem. However, the most significant improvement resulted from including evolutionary information (Fig. 4, Table 1, Table 2). The final predictions of secondary structure were very accurate: three thirds of all segments were predicted correctly, and for the 40% of the residues predicted at higher reliability the prediction accuracy was comparable to homology modelling [42]. Solvent accessibility was more difficult to predict than secondary structure. This was largely due to a lower degree in conservation of solvent accessibility (Fig. 3). This suggests that prediction methods should focus on features of protein structure that are

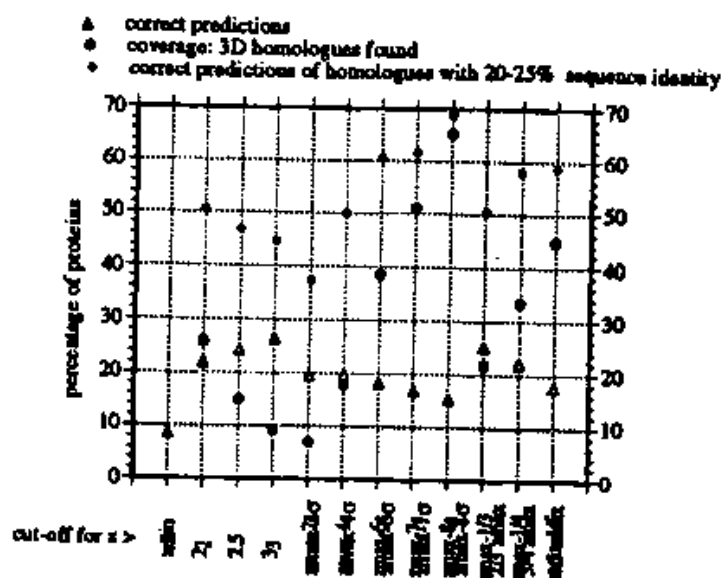


Figure 8: Preliminary results for detection of remote homologues. The alignment list is cut-off at various values for the alignment z-score (eq. (1)), i.e. only hits with a value larger than the threshold are taken as predicted remote homologues. The cut-off values given are: $zE > z_{min}$, i.e. all hits are taken; $zE > 2, 2.5, 3$, i.e. hits for which the z-score exceeds a value independent of the search sequence; $zE > (z_{max} - n \times \sigma)$, with $n = 2, 4, 6, 7, 8$, i.e. hits for which the z-score exceeds a value defined by the standard deviation and the maximal z-score for all alignments with a search sequence; and $zE > r \times z_{max}$, with $r = 0.66, 0.75, 0.8$, i.e. hits for which the z-score has a maximal difference to the maximal z-score for the search sequence. For all cut-off thresholds n the percentage of correct hits ($Q_{z,E}$, eq. (T3c), open triangles) and the coverage, i.e. the percentage of true homologues detected at that threshold ($Cov_{z,E}(n)$, eq. (T3d), filled circles) are plotted. Using additionally sequence information for the alignment improves the detection accuracy for pairs with a pairwise sequence identity above the average for remote homologues (20-25%, circles with point), but below the average of a significant detection of homologue based on sequence information alone (> 25%, for more elaborate alignment procedures).

conserved in evolution. Are 1-D predictions useful in practice?

Conservative evaluations of accuracy are more productive than best-case expectations. A common fault in evaluating prediction methods is that free parameters are fitted on the data sets used for evaluation. However, the opposite is more useful: the performance on a set of new proteins should tend to be higher rather than lower than the published expected accuracy (Table 1, Table 2). The neural network based predictions of secondary structure and solvent accessibility are being used frequently (more than 80 requests per day to an automatic prediction service, for information, 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/predictprotein.html>). Users find the 1-D pre-

Table 4: Accuracy of 1-D threading.

Values are given as percentages. Abbreviations of measures as in Table 3. Two cut-off thresholds are compared: $zE > 3$; and $zE > 3/4$ of the maximal value for zE for a given alignment list. Abbreviations of methods: PDB-PDB, the aligned strings of the search sequence and the putative homologue both originate from experimentally determined structures, i.e. these values describe the detection accuracy for ideal predictions of secondary structure and solvent accessibility; PHD-PDB, alignment of predicted (search sequence) and observed (to be detected homologues) strings (the predictions were done by cross-validation, i.e. no protein used to train the neural network systems did have more than 25% pairwise sequence identity to the search sequence of the threading); PHD-PHD, the aligned strings of the search sequence and the putative homologue both originate from predictions. E.g. if only those alignments between a 1-D prediction for a search sequence and 1-D strings extracted from a data base of experimentally determined 3-D structures with $zE > 3$ were taken, then 27% of these hits had been correct hits, and 8% of all true positives had been found above this threshold (21% of all first hits had been correct).

method	Q_{1st}	$Q_{zE(n)}$	$Cov_{zE(n)}$	$Q_{zE(n)}$	$Cov_{zE(n)}$
comment:		$zE > 3$	$zE > 3$	$zE > \frac{3}{4} \times z_{max}$	$zE > \frac{3}{4} \times z_{max}$
PDB-PDB	28	46	10	42	28
PHD-PDB	21	27	8	22	28
PHD-PHD	19	21	13	19	28

dictions useful. But, can 1-D predictions be used to predict more aspects of 3-D structure?

Standards for evaluating threading techniques are becoming required. How can threading experiments be evaluated? Appropriate evaluation consists of two parts, first, the choice of a representative data set, and second, the definition of adequate measures for prediction accuracy. The data set used here, was a first attempt to validate threading experiments on more than just a handful of proteins. The measures defined here, were restricted to analysing the detection of remote homologues. The most stringent measure is the percentage of proteins for which the first predicted 3-D homologue is correct (Table 3). Given these definitions, does 1-D threading work?

1-D threading possible, but so far very inaccurate. The positive message is "sometimes 1-D threading does work" (Fig. 7). How often? Three important results are. First, erasing the inter-segment distances reduces the 3-D information too drastically (PDB vs. PDB, Fig. 6, Table 4). Second, the inaccuracy of the neural network predictions reduces the accuracy in detecting 3-D homologues further, but is relatively less harmful than the loss of distance information (PHD vs. PDB, Table 4). Third, the accuracy in detecting remote homologues is about 20%. Does this imply that 1-D threading is a flop?

1-D threading is intrinsically more flexible than other threading techniques. The threading approach presented here differs from threading tools based on potentials of mean-force in that it is more coarse-grained. Although, a large scale comparison is not yet available, it appears to be very likely that mean-force based approaches, at least, the better ones (Braxenthaler,

these Proceedings and (64)), are more accurate than 1-D threading. However, this still does not imply that 1-D threading is of academic interest, only. Fitting 1-D predictions into 3-D structures uses completely different information than mean-force based threading techniques. This raises the hope, that the two could be combined to more reliably filter out false positives from an alignment list.

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 1-D predictions into 3-D structures. Furthermore, thanks to Michael Braxenthaler(Washington) and Manfred Sippl(Salzburg) for fruitful discussions about threading details. Thanks also to Christos Ouzounis(EMBL); Henrik Bohr, Søren Brunak, Jacob Engelbrecht, and Jan Hansen(all four Copenhagen), and Tim Hubbard(Cambridge) for helpful discussions. Last not least, I should like to express my gratitude to all those who contributed by human or financial resources to organise a rather inspiring workshop embedded into an outstanding program for the discussions outside the halls of the Royal Danish Academy of sciences. To mention only three: thanks to Johanne Keiding, Henrik Bohr and Søren Brunak.

References

- [1] C. J. Epstein, R. F. Goldberger and C. B. Anfinsen, The genetic control of tertiary protein structure: studies with model systems, *Cold Spring Harbour Symp. Quant. Biol.*, 28, 439-449, 1963.
- [2] J. Martin and F. U. Hartl, Protein folding in the cell: molecular chaperones pave the way, *Structure*, 1, 161-164, 1993.
- [3] T. J. P. Hubbard and C. Sander, The role of heat-shock and chaperone proteins in protein folding: possible molecular mechanisms, *Prot. Engin.*, 4, 711-717, 1991.
- [4] F.-U. Hartl, R. Hlodan and T. Langer, Molecular chaperones in protein folding: the art of avoiding sticky situations, *TIBS*, 19, 20-25, 1994.
- [5] S. Oliver, *et al.*, The complete DNA sequence of yeast chromosome III, *Nature*, 357, 38-46, 1992.
- [6] M. Johnston, *et al.*, Complete Nucleotide Sequence of *Saccharomyces cerevisiae* Chromosome VIII. *Science*, 265, 2077-2082, 1994.
- [7] P. Bork, C. Ouzounis and C. Sander, From genome sequences to protein function, *Curr. Opin. Str. Biol.*, 4, 393-403, 1994.
- [8] J. Greer, Comparative Modeling of Homologous Proteins, *Meth. Enzymol.*, 202, 239-252, 1991.
- [9] M. S. Johnson, J. P. Overington and T. L. Blundell, Alignment and Searching for Common Protein Folds Using a Data Bank of Structural Templates, *J. Mol. Biol.*, 231, 735-752, 1993.
- [10] A. C. W. May and T. L. Blundell, Automated comparative modelling of protein structures, *Curr. Opin. Biotech.*, 5, 355-360, 1994.
- [11] A. Sali and T. Blundell, Comparative Protein Modelling by Satisfaction of Spatial Restraints, in: H. Bohr and S. Brunak eds., *Protein Structure by Distance Analysis*, Amsterdam, Oxford, Washington: IOS Press, 64-87, 1994.

- [12] G. Vriend and C. Sander, Detection of Common Three-Dimensional Substructures in Proteins, *Proteins*, 11, 52-58, 1991.
- [13] C. Sander and R. Schneider, The HSSP database of protein structure-sequence alignments, *Nucl. Acids Res.*, 22, 3597-3599, 1994.
- [14] A. Bairoch, The PROSITE Dictionary of Sites and Patterns in Proteins, its Current Status, *Nucl. Acids Res.*, 21, 3097-3103, 1993.
- [15] L. Holm and C. Sander, The FSSP database of structurally aligned protein fold families, *Nucl. Acids Res.*, 22, 3600-3609, 1994.
- [16] C. Sander and R. Schneider, Database of Homology-Derived Structures and the Structurally Meaning of Sequence Alignment, *Proteins*, 9, 56-68, 1991.
- [17] R. Abagyan, D. Frishman and P. Argos, Recognition of distantly related proteins through energy calculations, *Proteins*, 19, 132-140, 1994.
- [18] T. L. Blundell and M. S. Johnson, Catching a common fold, *Prot. Sci.*, 2, 877-883, 1993.
- [19] J. U. Bowie, N. D. Clarke, C. O. Pabo and R. T. Sauer, Identification of Protein Folds: Matching Hydrophobicity Patterns of Sequence Sets With Solvent Accessibility Patterns of Known Structures, *Proteins*, 7, 257-264, 1990.
- [20] G. M. Crippen and V. N. Maiorov, A Potential Function that Identifies Correct Protein Folds, in: H. Bohr and S. Brunak eds., *Protein Structure by Distance Analysis*, Amsterdam, Oxford, Washington: IOS Press, 158-174, 1994.
- [21] K. Nishikawa and Y. Matsuo, Development of pseudoenergy potentials for assessing protein 3-D-1-D compatibility and detecting weak homologies, *Prot. Engin.*, 6, 811-820, 1993.
- [22] C. Ouzounis, C. Sander, M. Scharf and R. Schneider, Prediction of protein structure by evaluation of sequence-structure fitness: Aligning sequences to contact profiles derived from 3-D structures, *J. Mol. Biol.*, 232, 805-825, 1993.
- [23] M. J. Sippl and S. Weitckus, 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, 1992.
- [24] M. Wilmanns and D. Eisenberg, 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, 1993.
- [25] D. Eisenberg, R. Lüthy and A. D. McLachland, Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities, *Proteins*, 10, 229-239, 1991.
- [26] J. U. Bowie, R. Lüthy and D. Eisenberg, A Method to Identify Protein Sequences That Fold into a Known Three-Dimensional Structure, *Science*, 253, 164-169, 1991.
- [27] D. T. Jones, W. R. Taylor and J. M. Thornton, A New Approach to Protein Fold Recognition, *Nature*, 358, 86-89, 1992.
- [28] T. Meitinger, A. Meindl, P. Bork, B. Rost, C. Sander, M. Haasemann and J. Murken, Molecular modelling of the Norrie disease protein predicts a cysteine knot growth factor tertiary structure, *Nature Gen.*, 5, 376-380, 1993.
- [29] W. Kabsch and C. Sander, On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations, *Proc. Natl. Acad. Sc. U.S.A.*, 81, 1075-1078, 1984.
- [30] A. G. Szent-Györgyi and C. Cohen, Role of Proline in Polypeptide Chain Configuration of Proteins, *Science*, 126, 697, 1957.
- [31] J. C. Kendrew, R. E. Dickerson, B. E. Strandberg, R. J. Hart, D. R. Davies and D. C. Phillips, Structure of myoglobin: a three-dimensional Fourier synthesis at 2 Å resolution, *Nature*, 185, 422-427, 1960.

- [32] M. F. Perutz, M. G. Rossmann, A. F. Cullis, G. Muirhead, G. Will and A. T. North. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis, *Nature*, 185, 416-422, 1960.
- [33] W. Kabsch and C. Sander. How good are predictions of protein secondary structure?. *FEBS Lett.*, 155, 179-182, 1983.
- [34] J. Garnier and J. M. Levin. The protein structure code: what is its present status?. *CABIOS*, 7, 133-142, 1991.
- [35] J. M. Levin and J. Garnier. Improvements in a secondary structure prediction method based on a search for local sequence homologies and its use as a model building tool. *Biochim. Biophys. Ac.*, 955, 283-295, 1988.
- [36] V. Biou, J. F. Gibrat, J. M. Levin, B. Robson and J. Garnier. Secondary structure prediction: combination of three different methods, *Prot. Engin.*, 2, 185-91, 1988.
- [37] O. Gascuel and J. L. Golmard. A simple method for predicting the secondary structure of globular proteins: implications and accuracy, *CABIOS*, 4, 357-365, 1988.
- [38] N. Qian and T. J. Sejnowski. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.*, 202, 865-884, 1988.
- [39] H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, B. Lautrup, L. Nørskov, O. H. Olsen and S. B. Petersen. Protein secondary structure and homology by neural networks, *FEBS Lett.*, 241, 223-228, 1988.
- [40] X. Zhang, J. P. Mesirov and D. L. Waltz. Hybrid System for Protein Secondary Structure Prediction, *J. Mol. Biol.*, 225, 1049-63, 1992.
- [41] B. Rost, C. Sander and R. Schneider. Progress in protein structure prediction?, *TIBS*, 18, 120-123, 1993.
- [42] B. Rost and C. Sander. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins*, 19, 55-72, 1994.
- [43] B. Rost and C. Sander. 1-D secondary structure prediction through evolutionary profiles. in: H. Bohr and S. Brunak eds., *Protein Structure by Distance Analysis*, Amsterdam, Oxford, Washington: IOS Press, 257-276, 1994.
- [44] R. M. Sweet. Evolutionary similarity among peptide segments is a basis for prediction of protein folding, *Biopolymers*, 25, 1565-1577, 1986.
- [45] B. Rost and C. Sander. Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Proc. Natl. Acad. Sc. U.S.A.*, 90, 7558-7562, 1993.
- [46] B. Rost, C. Sander and R. Schneider. Redefining the goals of protein secondary structure prediction, *J. Mol. Biol.*, 235, 13-26, 1994.
- [47] R. B. Russell and G. J. Barton. The limits of protein secondary structure prediction accuracy from multiple sequence alignment, *J. Mol. Biol.*, 234, 951-957, 1993.
- [48] B. Rost and C. Sander. Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.*, 232, 584-599, 1993.
- [49] B. K. Lee and F. M. Richards. The interpretation of protein structures: Estimation of static accessibility, *J. Mol. Biol.*, 55, 379-400, 1971.
- [50] W. Kabsch and C. Sander. Dictionary of Protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features, *Biopolymers*, 22, 2577-2637, 1983.
- [51] C. Sander, M. Scharf and R. Schneider. Design of protein structures, in: A. R. Rees, M. J. E. Sternberg and R. Wetzel eds., *Protein Engineering*, Oxford: IRL Press, 89-115, 1992.
- [52] G. D. Rose, A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus. Hydrophobicity of Amino Acid Residues in Globular Proteins, *Science*, 229, 834-838, 1985.
- [53] B. Rost and C. Sander. Conservation and prediction of solvent accessibility in protein

- families, *Proteins*, 20, 216-226 1994.
- [54] T. J. P. Hubbard and T. L. Blundell, Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling, *Prot. Engin.*, 1, 159-171, 1987.
- [55] J. Janin, Surface and inside volumes in globular proteins, *Nature*, 277, 491-492, 1979.
- [56] S. Miller, J. Janin, A. M. Lesk and C. Chothia, Interior and Surface of Monomeric Proteins, *J. Mol. Biol.*, 196, 641-656, 1987.
- [57] T. F. Smith and M. S. Waterman, Identification of common molecular subsequences, *J. Mol. Biol.*, 147, 195-197, 1981.
- [58] S. F. Altschul, Amino Acid Substitution Matrices from an Information Theoretic Perspective, *J. Mol. Biol.*, 219, 555-565, 1991.
- [59] A. D. McLachlan, Tests for comparing related amino acid sequences, *J. Mol. Biol.*, 61, 409-424, 1971.
- [60] S. B. Needleman and C. D. Wunsch, A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins, *J. Mol. Biol.*, 48, 443-53, 1970.
- [61] U. Hobohm and C. Sander, Enlarged representative set of protein structures, *Prot. Sci.*, 3, 522-524, 1994.
- [62] A. Bairoch and B. Boeckmann, The SWISS-PROT protein sequence data bank: current status, *Nucl. Acids Res.*, 22, 3578-3580, 1994.
- [63] A. M. Lesk, *Protein Architecture — A Practical Approach*, Oxford, New York, Tokyo: Oxford University Press, 1991.
- [64] M. J. Sippl and M. Jaritz, Predictive Power of Mean Force Pair Potentials, in: H. Bohr and S. Brunak eds., *Protein Structure by Distance Analysis*, Amsterdam, Oxford, Washington DC: IOS Press, 113-134, 1994.
- [65] M. J. Sippl, 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, 1993.
- [66] M. J. Sippl, Recognition of Errors in Three-Dimensional Structures of Proteins, *Proteins*, 17, 355-362, 1993.
- [67] L. Holm and C. Sander, Parser for Protein Folding Units, *Proteins*, 19, 256-268, 1994.
- [68] B. Rost, R. Casadio, P. Fariselli and C. Sander, Prediction of helical transmembrane segments at 95% accuracy, *Prot. Sci.*, in press 1995.
- [69] F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: a computer based archival file for macromolecular structures, *J. Mol. Biol.*, 112, 535-542, 1977.
- [70] J. M. Levin, B. Robson and J. Garnier, An algorithm for secondary structure determination in proteins based on sequence similarity, *FEBS Lett.*, 205, 303-308, 1986.
- [71] J. M. Levin, S. Pascarella, P. Argos and J. Garnier, Quantification of Secondary Structure Prediction Improvement Using Multiple Alignments, *Prot. Engin.*, 6, 849-854, 1993.
- [72] H. Wako and T. L. Blundell, 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, 1994.