

# Protein Structure by Distance Analysis

edited by

H. Bohr and S. Brunak

Center for Biological Sequence Analysis  
The Technical University of Denmark  
Lyngby, Denmark

1994

*IOS Press*

Amsterdam • Oxford • Washington DC



Ohmsha

Tokyo • Osaka • Kyoto

© The Authors mentioned in the Table of Contents.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted, in a form or by any means, without the prior written permission from the publisher.

ISBN 90 5199 161 4 (IOS Press)

ISBN 4-274-02263-3 (Ohmsha)

Library of Congress Catalog Card Number: 94-075946

*Publisher:*

IOS Press  
Van Diemenstraat 94  
1013 CN Amsterdam  
Netherlands

*Sole distributor in the UK and Ireland:*

IOS Press/Lavis Marketing  
73 Lime Walk  
Headington  
Oxford OX3 7AD  
England

*Distributor in the USA and Canada:*

IOS Press, Inc.  
P.O. Box 10558  
Burke, VA 2209-0558  
U.S.A.

*Distributor in Japan:*

Ohmsha, Ltd.  
3-1 Kanda Nishiki-Cho  
Chiyoda-Ku  
Tokyo 101, Japan

**LEGAL NOTICE**

The publisher is not responsible for the use which might be made of the following information.

**PRINTED IN THE NETHERLANDS**

# 1D Secondary Structure Prediction through Evolutionary Profiles

Burkhard Rost and Chris Sander

EMBL Heidelberg, Meyerhofstr. 1, D-69117 Heidelberg, Germany

## Abstract

For only about one third of the new proteins, 3D structure can be predicted. For the remaining two thirds, a compromise has to be made. An extreme simplification is the projection of 3D structure onto a string of 1D secondary structure assignments. Here, we report how neural networks can be configured such that strand is predicted significantly better, and that the prediction looks like native proteins in terms of the length of predicted segments. Using evolutionary information contained in multiple sequence alignments as input to neural networks, secondary structure can be predicted at significantly increased accuracy. Pre-processing the alignment information by using a position-specific conservation weight and the number of insertions and deletions in each alignment position is found to be advantageous. Addition of the global amino acid content yields a further improvement, mainly in predicting structural class. The final network system has a sustained overall accuracy of more than 72% evaluated on 250 sequence-unique chains. Of particular practical importance is the definition of a position-specific reliability index. For 40% of all residues the method has a sustained three-state accuracy of 88%, as high as the overall average for homology modelling.

## 1 Introduction

What is the goal of protein structure prediction? Two different objectives can be roughly distinguished. First, understanding the basic principles of protein folding. Second, making predictions that can be used in various fields of research in molecular biology. These objectives tend to partition the field of theoretical biology into two major schools that use different 'tool-boxes' to attack 'the protein folding problem', i.e. the question of how the three-dimensional (3D<sup>1</sup>) structure of a protein is determined in detail from the sequence of amino acids. That the sequence determines the 3D structure is well established [1, 2, 3], and appears to hold even in view of chaperones which play a rôle in assisting protein folding [4, 5]. The tools of the first school are physical and biochemical first principles. On this ground it is attempted to predict 3D structure from the sequence. The difficulty lies in the

<sup>1</sup> Abbreviations used: 3D three-dimensional; Swissprot: data bank of protein sequences; PDB: protein data bank of known structures; DSSP: dictionary of secondary structures of proteins; HSSP: data base of homology derived structure of proteins; SOS: protein sequence of unknown three-dimensional structure; PHD: profile network system from Heidelberg (three levels of networks for the prediction of secondary structure).

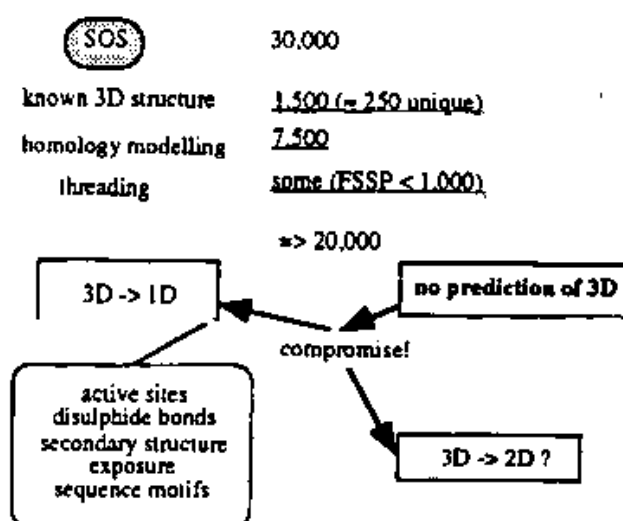


Figure 1: The scope of protein structure prediction: Numbers given are conservative estimates according to Swissprot (release 25) [51], PDB (December, 1993) [47, 52]. The number for homology modelling stems from the HSSP release of July, 1993 [53], that for threading from a data base of homologue structures, not exhibiting significant sequence homology [54].

complexity of the folding problem. Today, analyses based on first principles are at best restricted to very short protein sequences. The tool of the second school is application of statistics to a data bank of experimentally known protein structures. The difficulties mainly lie in that such attempts are restricted to features already common to the known examples, and that, so far, only simplified features of 3D structure can be dealt with. A third school might be made out as those who attempt combining the pros and cons of the two pillars, e.g. [6].

What is the scope of protein structure prediction in practice? Given a sequence of unknown 3D structure (SOS), how can theory help to predict features of its 3D structure? There is approximately a chance of 1:3 that the 3D structure of SOS can be predicted with sufficient accuracy (Fig. 1). However, for some two thirds of all known sequences prediction in 3D is not possible. These proceedings contributes various approaches to predict a 2D distance map from the sequence, but even such approaches do not yet operate in general. Thus, in most cases, we still have to accept an extreme compromise by projecting 3D structure onto 1D, i.e. by restricting the prediction to very simple features, like positions of active sites, disulphide bonds, secondary structure, surface exposure, sequence motifs or function of the protein.

Here, we shall present some aspects of a data base based prediction of secondary structure (assigned by DSSP [7], further projected onto 3 states: helix, strand, and rest — dubbed loop). The method uses a three-level system of neural networks, described in detail elsewhere [8, 9, 10, 11]. The main idea is that instead of single amino acid sequences, profiles of evolutionary conserved sequences are input to the network. The following questions will be addressed. How can neural networks be adapted to the special case of classifying sequence patterns into secondary structure? How can the information used as

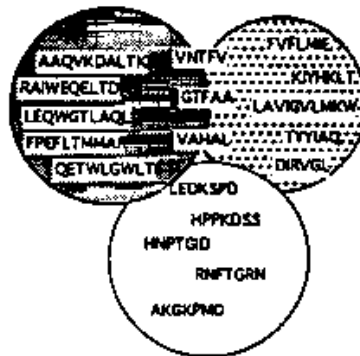


Figure 2: Secondary structure prediction as a pattern recognition problem: Certain oligopeptides have high preference to be in a particular secondary structure. Circles: upper left (dark shading): helix, upper right (light shading): strand, centre (no shading): loop. The 3 pentapeptides between the helix and strand circles are observed in both structures [55].

input be increased? Does pre-processing of the evolutionary profiles pay off in terms of prediction accuracy? Is the method accurate in comparison to the ultimate prediction, and in comparison to other prediction methods? What are the perspectives to predict the secondary structure of SOS?

## 2 Configuring a system of neural networks to predict secondary structure

### 2.1 Neural networks as tools for pattern classification

How can secondary structure be predicted from the database? Some stretches of sequence exhibit preferences for specific secondary structure states (Fig. 2). Consequently, the objective of a secondary structure prediction method is to classify sequences into e.g. three classes: helix, strand, and loop. A rather intuitive approach to pattern classification is drawing lines in between groups. A simple mathematical tool to perform such a task is the projection of vector  $v_1$  onto another  $v_2$ , i.e. the computation in the product  $v_1 * v_2$ . Suppose, a stretch of  $w$  adjacent residues of a protein sequence is presented by a vector with  $w * 20$  (20 different amino acids) components. Such a vector determines one point in a space of  $w * 20$  dimensions. In analogy to the circles in Fig. 2, the task is to find a matrix that projects all points into three classes.

A simple tool that performs this task is a neural network [12, 13] (more precisely a multi-layered feed-forward network). The input is the sequence vector, i.e. each component of the vector determines the value of one input unit of the network. The output the secondary structure state (of the central residue in the stretch). In the simple case of only one layer, the output  $out$  generated by the input  $in$  is given by:  $out = f(J * in)$ , with  $J$  as the matrix for the junctions between each unit in the input and each unit in the output layer. The function  $f$  is a sigmoid trigger function (e.g. hyperbolic tangents), that assigns 0 to large negative values, 1 to large positive ones, and some value in between 0 and 1 else. From the data

base we know which residue is found in which conformation. Given the junctions  $J$ , the output of the network is uniquely determined. Thus, the error of the network  $E$  can simply be computed as the square of the difference between network and observed output (summed over the three states). A simple way to reduce the error is by changing the junctions for each example presented such that the error decreases for each example presented, i.e. by gradient descent (as well known as back-propagation [14]):  $\Delta J(t+1) \propto -\epsilon \partial E(t) / \partial J$ , in other words by computing the partial derivative of the error with respect to the junctions ( $t$  is the iterative time, i.e. the presentation of one example). The factor  $\epsilon$  determines width of the stepwise convergence to a small error. The concept can be extended by using a layer of units between output and input, called hidden units. All results presented here were derived from networks with 15 hidden units. Training was done with conjugate gradient descent with the learning strength  $\epsilon = 0.05$ , and the momentum term  $(\Delta J(t+1) \propto -\epsilon \partial E(t) / \partial J + \alpha \Delta J(t-1)) \alpha = 0.2$ . The training was terminated once 75% of the training examples were classified correctly.

## 2.2 Better prediction of strand residues by balanced training

Evaluated on a data set of 130 unique protein chains (126 globular + 4 membrane) a network achieves a typical overall accuracy in 3 states of some 61–62%. In detail, the prediction is best for loop residues and far worst for strand (Fig. 3). The prediction accuracy for each of the 3 classes approximately mirrors the observed occurrence of these classes in the data set (Fig. 3). Consequently, the idea is to improve the prediction for strand residues by simply increasing the frequency in training examples for strand. (Note: such a procedure has been used previously for non-network predictions [15].) That is, instead of presenting in 1000 iteration time steps 220 examples for strand, 310 for helix and 470 for loop (dubbed 'unbalanced training'), now at each time step one example for each class is used for training (dubbed 'balanced training'). The first result is that all three classes are predicted almost equally well (Fig. 3). The second result is that the overall accuracy is reduced. (The latter stems from the evidence that the unbalanced prediction is better predicts loop residues, which tend to dominate the overall accuracy.)

## 2.3 Better prediction of segment length by 2nd level of structure-to-structure network

The network described so far, learns the classification of mutually independent patterns. The result is that the average length of a helix predicted is about 4 residues, compared to an average of 10 observed. The reason is that the network (as introduced here) cannot learn to correlate the secondary structure of adjacent residues, i.e. that e.g. helices span over at least three adjacent residues. A simple way to correct this shortage is the introduction of a 2nd level network (Fig. 4). A 1st level sequence-to-structure network predicts secondary structure from sequence. The output of the 1st level net is input to a 2nd level structure-to-structure network that predicts secondary structure from stretches of predictions. The 2nd level network has almost no effect in terms of overall accuracy. (This was probably the reason why the concept was forgotten after having been used already in one of the first applications of neural networks to secondary structure prediction [16].) However, the structure-to-structure network fulfils the purpose for which we introduced it: the average predicted helix extends over some 7 residues, i.e. the prediction looks considerably more protein like than that of the 1st level network (Fig. 3).

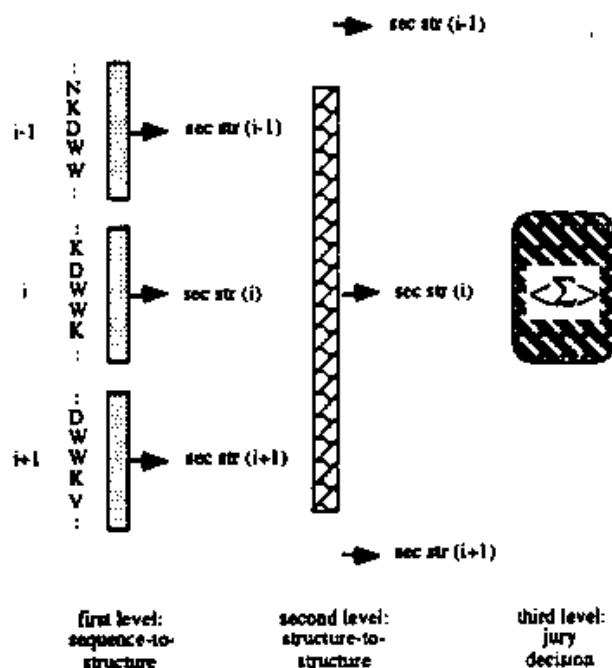
method	overall accuracy	■ helix	▨ strand	□ loop
unbalanced	2%		comparison: data bank distribution	
balanced	60%		comparison: 33:33:33	
1st level	HHH EEE	comparison:	HHHHH SSSS	
2nd level	HHHHHH EEEE	observed:	EEEE	
single network vs. jury decision				
PHD 3	72%		HHHHHHH EEEE	

Figure 3: Improvements on the network side: Neural networks can easily be adapted to specific tasks. The first example is balanced training by which the performance for strand is improved (shown in pie charts, for comparison the relative distribution of helix, strand, and loop in the data set). The second example is the improvement in predicting the average length of secondary structure segments (for comparison the observed averages). Networks divide patterns by introducing lines. Differently trained networks make different errors (here the task is to distinguish between filled and open circles). An arithmetic average over different network outputs (jury decision) is comparable to computing the centre of mass for all lines. The error of the jury decision is smaller than that for any particular network. For comparison the results for the final network system (PHD3) as presented in chapter 3 are given.

## 2.4 Better overall accuracy by compiling a jury decision over different networks

A network classifies patterns separating them by lines. One particular training results in a particular classification, with a particular (partially random) error.

A different training (e.g. unbalanced vs. balanced training) results in a different realisation of the classification task. Which network to choose in practice? A simple way out of the dilemma is to compile an arithmetic average over different network realisations (dubbed 'jury decision'). The procedure can be expected to yield a better result than each particular network if the errors of each of the networks are to a certain degree uncorrelated (Fig.

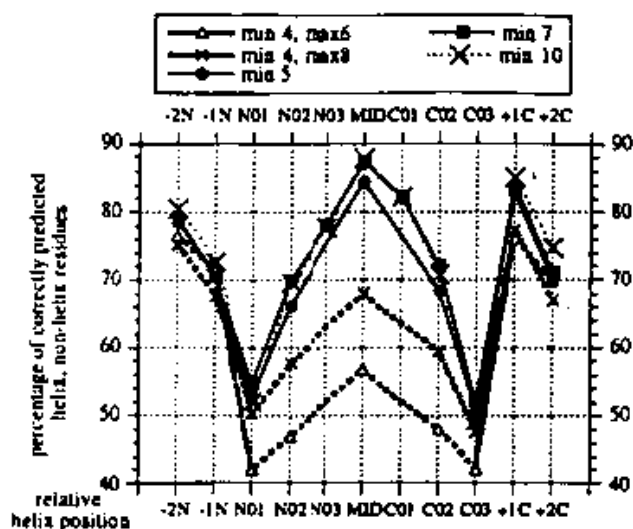


**Figure 4: Three level system for prediction of secondary structure:** First level, sequence-to-structure network: a window of 13 adjacent residues is shifted through all proteins. For each window the task of the network is to predict the secondary structure state of the central residue (here: D, W, W). Second level, structure-to-structure network: a window of 17 adjacent residues is shifted through all proteins. Again the task is to predict the secondary structure for the central residue. But the input is the output values, i.e. the predictions, of the 1st level net (as shown the 2nd level predicts the secondary structure for W at position i). Third level, jury decision: the output from differently trained networks (not sketched) for the same sequence position are summed. The secondary structure prediction for residue W at sequence position i is assigned to the unit with the maximal sum.

3). More precisely, the gain to be expected by the jury decision is inversely proportional to the correlation of the errors of particular networks. We use a jury decision on some 10 differently trained networks as the 3rd level of the network system (Fig. 4). The result is that the overall accuracy of the 3rd level is some 1–2 percentage points superior to any network on the 2nd level. Additionally, the 3rd level enables the combination of differently focused networks (e.g. unbalanced training: focus lies on predicting loop, balanced training: focus on predicting strand).

## 2.5 A tip to further improvements: core of helices better predicted than caps

Initiated by a discussion during the conference in Denmark, we investigated the following question: How does the prediction accuracy depend on the position of a residue in a helix? It



**Figure 5: Two-state accuracy in dependence of position relative to helices:** The two-state accuracy (helix, non-helix) is plotted for various positions relative to a helix: -2N, two residues before the begin of the helix; -1N, one residue before; N01, first residue in helix; N02, second residue; N03, third; MID, core of the helix = all other residues in the helix; C01, three residues before end of helix; C02, two before helix end; C03, last residue in helix; +1C, one residue after helix end; +2C, two residues after helix end. The curves differ according to the minimal and maximal lengths of helices analysed. Apparently, the network did not learn the N- and C-cap preferences inside a helix. This is less clear for shorter helices. One possible explanation is that for these the preferences must be more exposed, as only few residue types can bend a helix so sharply that a short helix is stable.

is known that there is a very strong preference for N-terminal caps of helices (first residues in a helix), and a still strong preference for C-terminal caps (last residues) [17]. The observation is slightly blurred by the difficulty to automatically determine begin and end of helices [18, 19, 20, 21, 22, 23]. But, the tendency is strong enough to allow the conclusion that cap residues exhibit a unique pattern. It should be expected that the network system learns such preferences.

However, the result indicates the opposite. The prediction accuracy is clearly worse for N- and C-caps than for the core of helices (Fig. 5). The tendency is that this is less so for shorter helices than for longer ones (Fig. 5). This observation indicates that the prediction can probably be increased by combining in a jury decision either a network trained explicitly on caps with the one described here, or a prediction method explicitly focusing on cap preferences [17] (see as well contribution of Brunak & Engelbrecht in this issue) with the network system.

secondary structure	DSSP	E	E	E	E	E	E	E	E	S	E	E	E	E	E	H	H	H						
SH3	N	S	T	N	K	D	W	W	K	V	E	V	N	D	R	Q	C	F	V	P	A	A	Y	
alignment	s1	K	K	S	N	P	D	W	W	E	G	E	L	N	G	Q	R	C	V	F	P	A	S	Y
	s2	E	E	K	.	G	E	W	W	K	A	K	.	.	K	R	E	G	F	I	P	S	N	Y
	s3	R	S	T	.	G	D	W	W	L	A	r	v	T	G	R	E	G	Y	V	P	P	S	N
	s4	F	E	.	.	.	F	F	G	V	.	v	D	D	L	Q	V	F	V	P	P	A	Y	Y
profile	V	0	0	0	0	0	0	0	0	40	0	40	0	0	0	0	20	20	40	0	0	0	0	0
	L	0	0	0	0	0	0	0	0	20	0	0	20	0	0	20	0	0	0	0	0	0	0	0
	I	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0
	N	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	F	20	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
	W	0	0	0	0	0	0	80	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Y	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	0	0	0	0	80
	G	0	0	0	0	50	0	0	0	20	0	0	0	40	0	0	40	0	0	0	0	0	0	0
	A	0	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	0	40	40	0
	P	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	100	20	0	0
	S	0	40	25	0	0	0	0	0	0	0	20	20	0	0	0	0	0	0	0	0	40	20	0
	T	0	0	50	0	0	0	0	0	0	0	0	20	0	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	H	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	R	20	0	0	0	0	0	0	0	0	20	0	0	0	40	20	0	0	0	0	0	0	0	0
	K	0	20	0	0	25	0	0	40	0	20	0	0	20	0	0	0	0	0	0	0	0	0	0
	Q	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20	40	0	0	0	0	0	0	0
	Z	20	20	0	0	25	0	0	20	0	40	0	0	0	40	0	0	0	0	0	0	0	0	0
additional information	N	40	0	0	100	0	0	0	0	0	0	0	40	0	0	0	0	0	0	0	0	0	0	40
	D	0	0	0	0	0	25	0	0	0	0	0	0	20	40	0	0	0	0	0	0	0	0	0
	$N_{i,m}$	0	0	0	0	0	0	0	0	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0
	$N_{i,ss}$	0	0	1	3	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	$CW_i$	1.0	0.8	0.7	0.8	0.6	1.3	1.5	1.5	0.8	0.9	1.0	0.7	0.7	0.9	0.9	0.7	1.5	1.0	1.2	1.5	0.9	0.7	1.5

Figure 6: Extracting evolutionary information: HSSP alignment of four sequences (s1–s4) to a region of SH3 [29]. Lower case letters in the alignment indicate insertions. The secondary structure is assigned with DSSP [7], the co-ordinates were kindly provided by Andrea Musacchio [56]. The HSSP profile is purely a computation of residue frequencies. The number of insertions and deletions are simple counts. The conservation weight is computed according to:  $CW_i = \sum_{r,s}^N w_{r,s} sim_{r,s}^i / \sum_{r,s}^N w_{r,s}$ , with  $w_{r,s} = (1 - 0.01 * \%identity_{r,s})$ , where  $N$  is the number of sequences in the alignment,  $identity_{r,s}$  the percentage of sequence identity (over the entire length) of sequence  $r$  and  $s$  in the alignment, and  $sim_{r,s}^i$  a value from the similarity matrix between sequence  $r$  and  $s$  at position  $i$  (e.g. Dayhoff [57])

### 3 Increasing input information by using profiles of evolutionary conservation

#### 3.1 Input information insufficient to benefit from neural network

In principle, neural networks can achieve to classify highly correlated patterns. However, without hidden units the networks decompose only first order correlation. The network system as described so far, yields about the same results, no matter of whether or not hidden units are used [16, 24, 25]. In other words, the input information is not sufficient to use the capacity of the networks. But how can the input information be increased? One way that has been investigated without success was to increase the number of adjacent residue used for one example (here we chose:  $w = 13$  for the 1st level, and  $w = 17$  for the 2nd level). The problem with this appears to be the too small size of the data base, i.e. the longer the

stretch, the less examples will be found in the data base for one particular stretch.

### **3.2 Specific information contained in pattern of evolutionary conservation**

How can information be compiled that is specific about a certain structure? Studying multiple alignments reveals that structure is more conserved than sequence [26, 27, 28]. Different protein sequences can adopt the same 3D structure. By the pressure of selection (survival of appropriate function) evolution has explored the sequence variation that is possible without changing the 3D structure of a protein (and by that the function): a pair of native proteins has similar 3D structure if 30% residues are pairwise identical [27, 29]. Not any two residues can be exchanged. Instead, the substitution pattern is highly specific for a certain structure. The idea to use such information for prediction is not new [30, 26, 31, 32, 33]. Recently, the same idea has been used to predict secondary structure for single cases (for summary [10, 34]).

### **3.3 Compiling profiles of evolutionary conservation from multiple alignments**

How can the information contained in the multiple alignment be extracted? The program used for generating the multiple alignment is MaxHom/HSSP [29]. The algorithm builds up the alignment in essentially two steps. In sweep 1, the sequences are aligned consecutively to the guide sequence by standard dynamic programming [35]. After each sequence has been added to the alignment an alignment profile is compiled and used to align the next sequence. In sweep 2, after all sequences have been added, the profile is recompiled, and the dynamic programming algorithm is repeated, this time using the constant profile, as derived after completion of sweep 1. The profile gives a simple count of the frequency of occurrence for each amino acid (Fig. 6). Additional information are the numbers of insertions and deletions in the alignment, and the compilation of a specific conservation weight (defined in caption of Fig. 6).

## **4 The more specific the input information, the better the prediction accuracy**

### **4.1 Coarse-grained vs. fine-grained profile: from 68% to 70%**

Do the details in the alignment matter? First, we projected the profile frequencies onto a grid of 4 with the intervals: 0–2, 3–33, 34–66, 67–100. A 2nd level network reaches an overall three-state accuracy of some 65%, compared to some 60% for a network similarly trained but using single sequences instead of multiple alignments as input (Fig. 7). The 3 level system reaches some 68%. Second, we used all details contained in the multiple alignment. The result is an increase of accuracy to almost 70%. This finding is slightly surprising. It bears the question: how does the improvement depend on the particular alignment, i.e. on the number of sequences aligned and on the variety? The answer is not clear cut. There is a tendency that the more various sequences in the alignment, the larger the improvement [11].

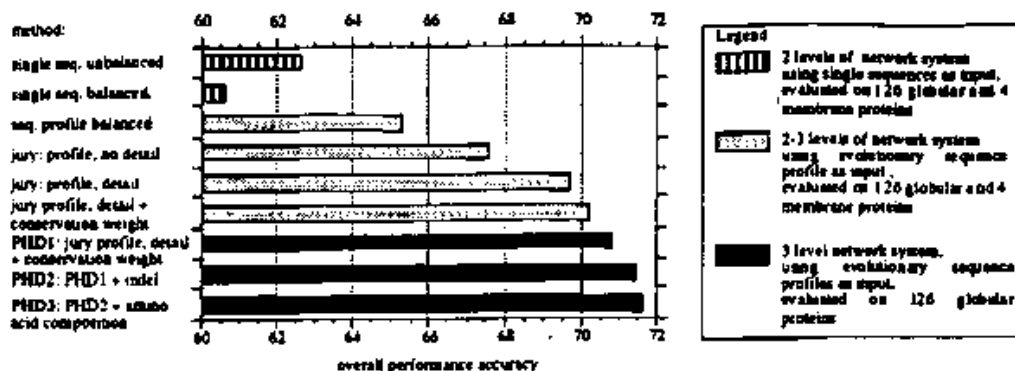


Figure 7: Stepwise improvement of overall accuracy: The stepwise improvement of secondary structure prediction based on multiple alignments in terms of overall accuracy.

However, this effect is partially blurred by a considerable variation of prediction accuracy with the protein chain (one standard deviation in order of 10%).

#### 4.2 Adding conservation weight and indels: >71%

The position-specific conservation weight (Fig. 6) constitutes a knowledge-based focus on important sites. In principle, the network should be able to learn such a weight. The question at hand is: does it pay-off to assist the network by explicitly adding information to the input? It does. Using the conservation weight as additional input increases performance accuracy to above 70% (Fig. 7, the network system using detailed profiles and conservation weight will be dubbed 'PHD1').

Insertions and deletions are likely to occur in loop regions. Consequently, adding the number of insertions and deletions (dubbed 'indels') at each alignment position to the input should improve the performance for loop. Indeed, it does. The overall three-state accuracy becomes 71.4% compared to 70.8% without using indels (note: based on a set of 126 globular proteins). Furthermore, the tendency for over-prediction of helix and strand is reduced and the accuracy in predicting loop regions is increased.

#### 4.3 Adding amino acid composition: better prediction of structural class

All information incorporated so far has been local in sequence, i.e. the input is compiled from a window of 13–17 consecutive residues. Of course, global information is introduced implicitly by the profiles, as a particular residue substitution pattern can depend on interactions between residues far apart in sequence. How can global information be explicitly used? A straightforward, technically simple way is to add units that code for the frequencies of amino acid occurrence in the protein outside of the stretch of 13–17 residues under investigation. The improvement in terms of overall accuracy is, at most, marginal. The network predictions do not differ much in terms of overall accuracy, average length predicted (and further measures, not given here) between a three-level network system using profile details, conservation weight, indels (dubbed 'PHD2') and a comparable system additionally using amino acid composition (dubbed 'PHD3'). However, the predictions differ in detail.

Proteins can be classified into four structural classes based on the relative content in secondary structure [36]: all- $\alpha$ , all- $\beta$ ,  $\alpha/\beta$ , and rest. Such a classification is not clear cut [10]. Consequently, different schemes have been used in literature. The results given here use the scheme of [37]: all- $\alpha$  =  $\alpha$  content  $\geq 45\%$ ,  $\beta$  content  $< 5\%$ ; all- $\beta$  =  $\alpha$  content  $< 5\%$ ,  $\beta$  content  $\geq 45\%$ ; and  $\alpha/\beta$  =  $\alpha$  content  $\geq 30\%$ ,  $\beta$  content  $\geq 20\%$ .

The task is to predict the structural class from the amino acid sequence. (Similar work has been done before, unfortunately the results were either not at all cross-validated or did depend on data sets which did not exclude homologue proteins [38, 39, 40, 41, 42, 43, 37, 44].) PHD1 and PHD2, i.e. the networks not using global information successfully classify 70% of the 126 proteins into one of the four classes. PHD3, i.e. the network additionally using the global amino acid composition achieves 75%. The conclusion is that including global information does not pay-off in terms of local measures (per-residue accuracy *asf.*), but it does pay-off in terms of a global measure like the relative content of secondary structure.

## 5 Network system stands competition with alternatives

### 5.1 Prediction between 35% (random) and 88% (homology modelling)

What is the goal of predicting secondary structure? It appears to be simple: predict secondary structure such, that it is essentially compatible with the true 3D structure. But what is essentially compatible? One practical line to tackle the concept of compatibility is given by homology modelling. Homology modelling allows for predicting accurately the complete 3D structure of a protein, if applicable. This implies that it offers a perfect prediction of secondary structure [45, 46]. In an analysis of more than 100 homologue sequences, we found that the expected three-state accuracy for homology modelling is some 88%. What is the worst prediction one can get? One answer is given by random alignments: 35%. All prediction methods operate in between these two values: 35% — 88%. Consequently, it is meaningful to evaluate prediction accuracy additionally with respect to these two lines. In the following we shall use the term 'normalised overall three-state accuracy'. What we mean is a scaling of the prediction accuracy such, that a random prediction scores at 0, and homology prediction scores at 100%.

For a few selected methods the normalised accuracy is given along with the number of protein chains used for the evaluation (Fig. 8). Can these numbers be compared? Most of them were evaluated on different data sets. For some the data sets were too small. Only some of the methods did exclude homologue sequences in evaluation. Furthermore, as there is a considerable variation in prediction accuracy with protein chain, some methods might, by chance, have chosen chains that are easier to be predicted. Thus it is crucial to estimate prediction methods on the same data set.

### 5.2 Comparison of performance for different data sets

The levels of accuracy given so far based on a cross-validation test with 126 globular, and either with or without 4 additional membrane chains. How do the results depend on the data set? The set of 130 proteins was the maximal set of unique proteins in PDB [47] in 1992 [48]. Meanwhile, the number of unique proteins has risen to more than 250. This provides an excellent opportunity to investigate how PHD performs on a set of 124 new proteins

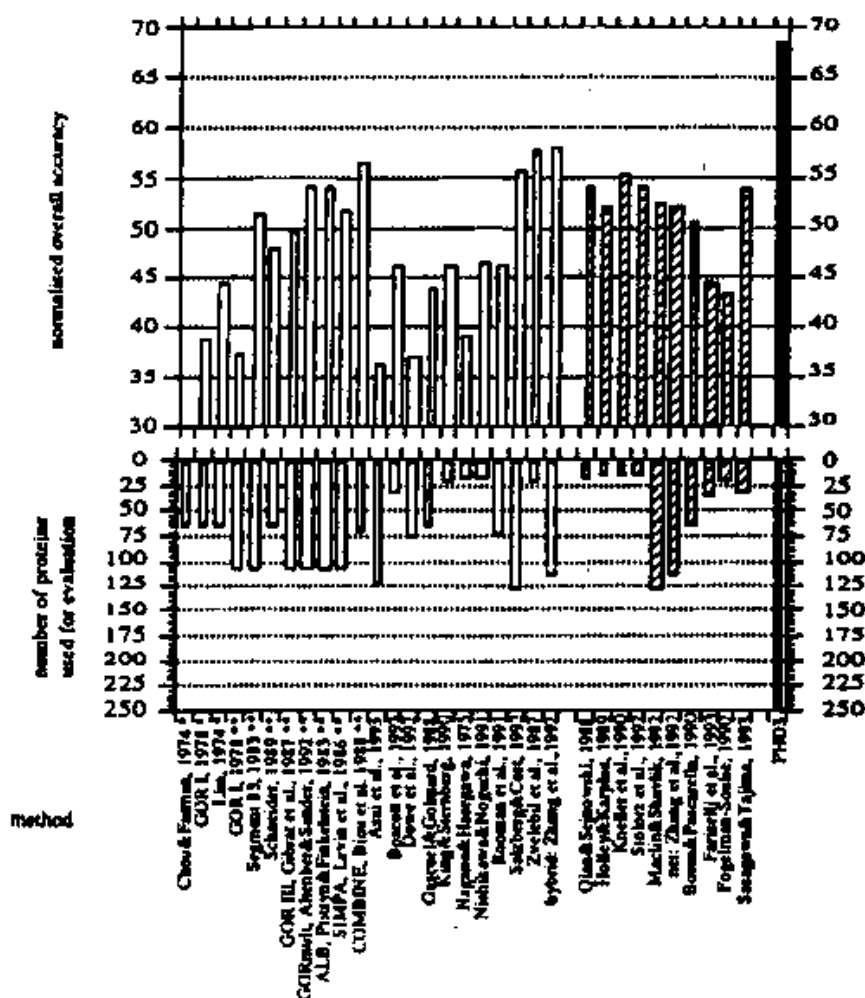


Figure 8: Normalised overall accuracy for various prediction methods: The results labelled by '\*' were taken from [58], those labelled by '\*\*' from [59] (note: these are based on the same set, but it is not excluded that some proteins had been used for setting up the methods, only the result for ALB a method based on physical principles is reliable). The other results are taken from the publications, referenced as in the literature list [49, 60, 61, 62, 63, 64, 65, 66, 33, 67, 15, 24, 17, 68, 69, 70, 71, 72, 73, 59, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83]. The values are normalised such that 0% = random prediction, and 100% = homology modelling.

(dubbed 'Set 2') not homologue to the previously used set of 130. Furthermore, Set 2 can be used to compare the network with methods like GORIII and COMBINE, as these base on proteins not homologue to Set 2 (Fig. 9). To render a comparison with other methods, (like the Chou-Fasman method [49, 50] that still is one of the most often used program in practice) we performed a cross-validation experiment on the data sets used for evaluating these method (Fig. 9). The conclusion is that PHD is some 10 percentage points superior in terms of normalised overall accuracy to any other method, including those that use multiple alignments on the ground of statistics. Furthermore, the additional tests indicate that the network performance is likely to be >60% for more than 90% of the proteins.

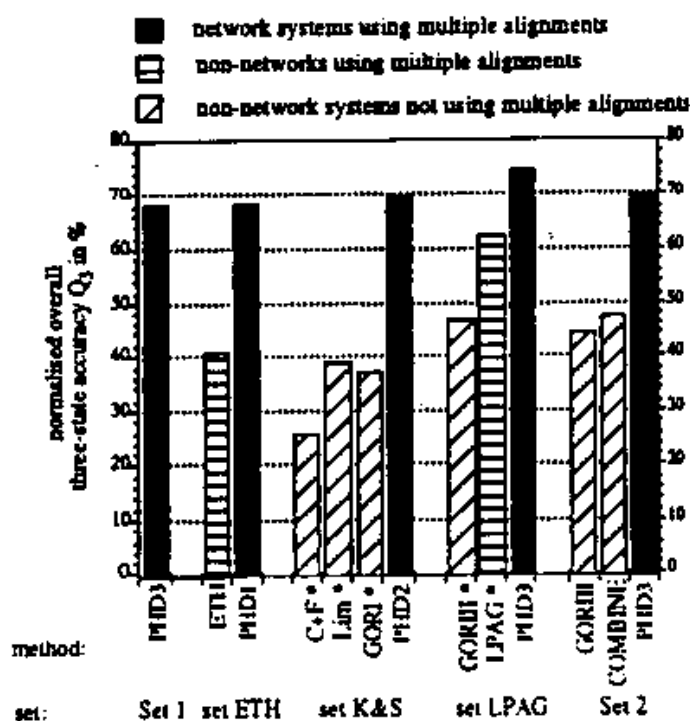


Figure 9: Various prediction methods evaluated on the same data sets: Normalised overall accuracy (0% = random prediction, 100% = homology modelling). The different sets are: 'Set 1', 126 unique globular protein chains (explicitly given in [11]); 'set ETH', 5 proteins for which expert predictions from Benner, Gerloff et al. were published [84, 85, 86, 87, 88]; 'set K&S', 62 proteins used for an analysis of secondary structure prediction methods a decade ago [58]; 'set LPAG', 82 protein fragments (from less than 20 structure families) used for evaluation of the GOR method with multiple alignments [81]; 'Set 2', 124 unique globular protein chains with no homology to Set 1 (explicitly given in [10]). The results labelled by '\*' were taken from the literature.

## 6 Predictions useful for research in molecular biology

### 6.1 Variation of prediction accuracy with protein chain

How good is the prediction of PHD on SOS? After all the tests described, this question is still not an easy one. The expected overall accuracy is above 72% evaluated on all 250 unique globular protein chains. But, the variation with the protein is considerable, i.e. one standard deviation is some 9%. (And an analysis of the prediction accuracy of homology modelling suggests, that this value is not a disadvantage of the prediction method. On the contrary, the variation reflects the variation between different proteins. Thus, if a prediction method is reported to have a small variation this might indicate that the test set was too small.) The expected accuracy is therefore  $72 \pm 9\%$ . (But can be worse for exceptional cases.)

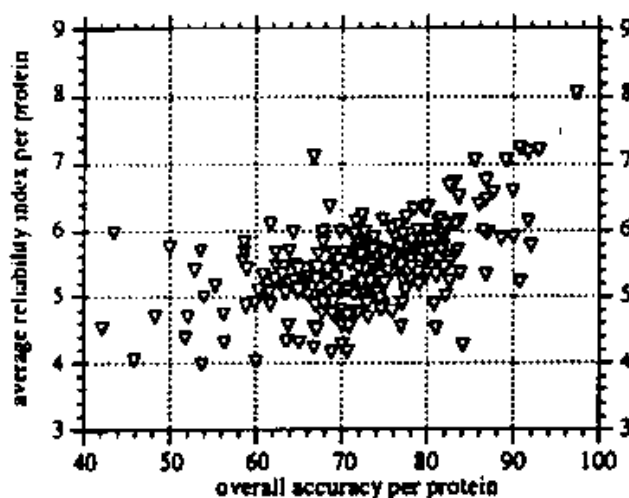


Figure 10: Average reliability index: The reliability index is averaged for each protein chain and plotted versus the overall accuracy per chain. A more complicated  $z$ -score-like measure did not yield a better distinction. The basic message is that if the prediction for SOS has a reliability index  $> 5$ , then the likelihood that the prediction is less accurate than 60% is approximately halved.

## 6.2 Definition of prediction reliability

Is it predictable, how accurate the prediction is for SOS? At first sight this question might look absurd. But, the answer is yes. Why? The PHD output consists of three values for helix, strand, and loop. Usually, the prediction is assigned to the output unit with maximal value. However, the difference between the output unit with highest value, and that of the unit with the next highest value, provides an estimate for the reliability of the prediction. Our experience is that this value correlates well with the prediction accuracy. About 40% of all residues are predicted at an accuracy of 88%, i.e. comparable to what would be possible if homology modelling could be done [10]. A similar index can be defined by statistical methods such as COMBINE. Compared on the same set of 124 proteins, COMBINE (without alignments) predicts 10% of the residues at a level of 80% accuracy, PHD3 more than 70% of the residues [10]. The average reliability per protein provides an additional help to estimate the quality of the prediction for SOS: the majority of proteins predicted at  $< 60\%$  accuracy has an average reliability index  $< 5$  (Fig. 9).

## 6.3 Availability of PHD prediction

The PHD predictions are available for fully automatic use. Send the word *help* (either as subject, or as only word in text) by electronic mail to the internet address *PredictProtein@EMBL-Heidelberg.de* for detailed instructions. Should the answer take more than 2 days, please repeat the procedure, or feel free to send a message to the internet address *Predict-Help@EMBL-Heidelberg.de*. (Note: *PredictProtein* is an automatic server with no ability to read letters, thus, send any personal message to *Predict-Help*.) The prediction will either be based on an alignment done by the server, or on an alignment provided by you in appropriate format. Is there an interest in obtaining secondary structure predictions? By early

December, 1993, more than 8,000 requests have been processed by the PHD server. The number of requests per month is increasing. Will the predictions be useful for 'scientific progress'? We shall see, meanwhile much can be done for a fruitful dialogue between theory and experiment.

## Acknowledgements

First of all, we thank Søren Brunak and Henrik Bohr for the marvellous organisation of a vivid conference in a relaxed, inspiring and personal atmosphere. Thanks to two colleagues at the EMBL: Reinhard Schneider and Gert Vriend, for valuable ideas, discussion and assistance. Furthermore, we wish to thank Jean Garnier (INRA, Paris) for having supplied the programs GORIII and COMBINE. Last, not least, we wish to express our gratitude to all those who make co-ordinates of experimentally determined protein 3D structure available.

## References

- [1] C. B. Anfinsen, E. Haber, M. Sela and F. H. White Jr., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, *Proc. Natl. Acad. Sc. U.S.A.*, **47**, 1309-1314 (1961).
- [2] 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).
- [3] C. B. Anfinsen, Principles that govern the folding of protein chains, *Science*, **181**, 223-230 (1973).
- [4] T. E. Creighton, Up the kinetic pathway, *Nature*, **356**, 194-195 (1992).
- [5] J. J. Ewbank and T. E. Creighton, Protein folding by stages, *Curr. Biol.*, **2**, 347-349 (1992).
- [6] M. J. Sippl, Calculation of Conformational Ensembles from Potentials of Mean Force. An Approach to the Knowledge-based Prediction of Local Structures of Globular Proteins, *J. Mol. Biol.*, **213**, 859-883 (1990).
- [7] W. Kabsch and C. Sander, Dictionary of Protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features, *Biopolymers*, **22**, 2577-2637 (1983).
- [8] B. Rost and C. Sander, Exercising Multi-layered Networks on Protein Secondary Structure. In: O. Benhar, S. Brunak, P. DelGiudice and M. Grandolfo (eds.), *Neural Networks: From Biology to High Energy Physics*. Elba, Italy: International Journal of Neural Systems, 1992, 209-220.
- [9] 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).
- [10] B. Rost and C. Sander, Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins, Journal of Chemical Biology*, submitted, September 24, 1993.

- [11] B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *J. Mol. Biol.*, **232**, 584-599 (1993).
- [12] B. Rost, Neural networks and evolution — advanced prediction of protein secondary structure. Dep. of Physics and Astronomy, University of Heidelberg, F.R.G., 1993.
- [13] B. Rost and G. Vriend, Neural Networks in Chemistry, *CDA News*, **8**, 24-27 (1993).
- [14] D. E. Rumelhart, G. E. Hinton and R. J. Williams, Learning representations by back-propagating error, *Nature*, **323**, 533-536 (1986).
- [15] 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).
- [16] N. Qian and T. J. Sejnowski, Predicting the Secondary Structure of Globular Proteins Using Neural Network Models, *J. Mol. Biol.*, **202**, 865-884 (1988).
- [17] R. Schneider, Sekundärstrukturvorhersage von Proteinen unter Berücksichtigung von Tertiärstrukturaspekten. Department of Biology, Univ. Heidelberg, FRG, 1989.
- [18] W. R. Taylor and C. A. Orengo, A holistic approach to protein structure alignment, *Prot. Engin.*, **2**, 505-19 (1989).
- [19] C. M. Wilmot and J. M. Thornton,  $\beta$ -Turns and their distortions: a proposed new nomenclature, *Prot. Engin.*, **3**, 479-493 (1990).
- [20] S. Brunak, Non-linearities in training sets identified by inspecting the order in which neural networks learn. In: O. Benhar, C. Bosio, P. Del Giudice and E. Tabet (eds.), *Neural Networks From Biology to High Energy Physics*. Elba, Italy: 1991, 277-88.
- [21] A. Perczel, K. Park and G. D. Fasman, Deconvolution of the Circular Dichroism Spectra of Proteins: The Circular Dichroism Spectra of the Antiparallel  $\beta$ -Sheet in Proteins, *Proteins*, **13**, 57-69 (1992).
- [22] S. Woodcock, J.-P. Moron and B. Henrissat, Detection of secondary structure elements in proteins by hydrophobic cluster analysis, *Prot. Engin.*, **5**, 629-635 (1992).
- [23] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat and J.-P. Moron, Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment, *Prot. Engin.*, **6**, 377-382 (1993).
- [24] H. L. Holley and M. Karplus, Protein secondary structure prediction with a neural network, *Proc. Natl. Acad. Sc. U.S.A.*, **86**, 152-156 (1989).
- [25] S. B. Petersen, H. Bohr, J. Bohr, S. Brunak, R. M. J. Cotterill, H. Fredholm and B. Lautrup, Training neural networks to analyse biological sequences, *TIBTECH*, **8**, 304-308 (1990).
- [26] R. E. Dickerson, R. Timkovich and R. J. Almassy, The Cytochrome Fold and the Evolution of Bacterial Energy Metabolism, *J. Mol. Biol.*, **100**, 473-491 (1976).
- [27] C. Chothia and A. M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO J.*, **5**, 823-826 (1986).

- [28] A. Pastore and A. M. Lesk, Comparison of the Structures of Globins and Phycocyanins: Evidence for Evolutionary Relationship, *Proteins*, **8**, 133-55 (1990).
- [29] R. Schneider and C. Sander, Database of Homology-Derived Structures and the Structurally Meaning of Sequence Alignment, *Proteins*, **9**, 56-68 (1991).
- [30] E. Zuckerkandl and L. Pauling, Evolutionary Divergence and Convergence in Proteins. In: V. Bryson and H. J. Vogel (eds.), *Evolving Genes And Proteins*. New York and London: Academic Press, 1965, 97-166.
- [31] F. R. Maxfield and H. A. Scheraga, Improvements in the Prediction of Protein Topography by Reduction of Statistical Errors. *Biochem.*, **18**, 697-704 (1979).
- [32] K. Nishikawa, Assessment of secondary structure prediction of proteins: Comparison of computerized Chou-Fasman method with others, *Biochim. Biophys. Ac.*, **748**, 285-299 (1983).
- [33] M. J. Zvelebil, G. J. Barton, W. R. Taylor and M. J. E. Sternberg, Prediction of protein secondary structure and active sites using alignment of homologous sequences, *J. Mol. Biol.*, **195**, 957-961 (1987).
- [34] B. Rost, C. Sander and R. Schneider, Progress in protein structure prediction?, *TIBS*, **18**, 120-123 (1993).
- [35] T. F. Smith and M. S. Waterman, Comparison of biosequences, *Adv. Appl. Math.*, **2j**, 482-489 (1981).
- [36] M. Levitt and C. Chothia, Structural patterns in globular proteins, *Nature*, **261**, 552-558 (1976).
- [37] C.-T. Zhang and K.-C. Chou, An optimization approach to predicting protein structural class from amino acid composition, *Prot. Sci.*, **1**, 401-408 (1992).
- [38] R. P. Sheridan, J. S. Dixon, R. Venkataghavan, I. D. Kuntz and K. P. Scott, Amino acid composition and hydrophobicity patterns of protein domains correlate with their structures, *Biopolymers*, **24**, 1995-2023 (1985).
- [39] P. Klein, Prediction of protein structural class by discriminant analysis, *Biochim. Biophys. Acta*, **874**, 205-215 (1986).
- [40] P. Klein and C. DeLisi, Prediction of protein structural class from the amino acid sequence, *Biopolymers*, **25**, 1659-1672 (1986).
- [41] P. Klein, J. A. Jacquez and C. DeLisi, Prediction of protein function by discriminant analysis, *Mathematical Biosciences*, **81**, 177-189 (1986).
- [42] G. Deleage and B. Roux, An algorithm for protein secondary structure prediction based on class prediction, *Prot. Engin.*, **1**, 289-294 (1987).
- [43] G. Deleage and B. Roux, Use of class prediction to improve protein secondary structure prediction. In: F. G. D. (eds.), *Prediction of protein structure and the principles of protein conformation*. New York: Plenum Press, 1989, 587-597.

- [44] B. A. Metfessel, P. N. Saurugger, D. P. Connelly and S. S. Rich, Cross-validation of protein structural class prediction using statistical clustering and neural networks, *Prot. Sci.*, **2**, 1171-1182 (1993).
- [45] R. B. Russell and G. J. Barton, The limits of protein secondary structure prediction accuracy from multiple sequence alignment, *J. Mol. Biol.* submitted May 1993.
- [46] B. Rost, R. Schneider and C. Sander, Redefining the goals of protein secondary structure prediction, *J. Mol. Biol.*, **235** in the press (1994).
- [47] 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).
- [48] U. Hobohm, M. Scharf, R. Schneider and C. Sander, Selection of representative protein data sets, *Prot. Sci.*, **1**, 409-17 (1992).
- [49] P. Y. Chou and U. D. Fasman, Prediction of protein conformation, *Biochem.*, **13**, 211-215 (1974).
- [50] P. Y. Chou and G. D. Fasman, Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.*, **47**, 45-148 (1978).
- [51] A. Bairoch and B. Boeckmann, The SWISS-PROT protein sequence data bank, *Nucl. Acids Res.*, **20**, 2019-2022 (1992).
- [52] E. E. Abola, F. C. Bernstein and T. F. Koetzle, The Protein Data Bank. In: E. Lesk A. M. (eds.). Computational molecular biology. Sources and methods for sequence analysis. Oxford: Oxford University Press, 1988, 69-81.
- [53] R. Schneider and C. Sander, The HSSP data base of protein structure-sequence alignment, *Nucl. Acids Res.*, **21**, 3105-3109 (1993).
- [54] L. Holm, C. Ouzounis, C. Sander, G. Tuparev and G. Vriend, A database of protein structure families with common folding motifs, *Prot. Sci.*, **1**, 1691-1698 (1993).
- [55] 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).
- [56] A. Musacchio, M. Noble, R. Paupit, R. Wierenga and M. Saraste, Crystal structure of a Src-homology 3 (SH3) domain, *Nature*, **359**, 851-855 (1992).
- [57] M. O. Dayhoff, Atlas of Protein Sequence and Structure. Washington, D.C., U.S.A.: National Biomedical Research Foundation, 1978.
- [58] W. Kabsch and C. Sander, How good are predictions of protein secondary structure?, *FEBS Lett.*, **155**, 179-182 (1983).
- [59] B. Altenberg and C. Sander, Current Quality of Secondary Structure Prediction, EMBL, 1992.

- [60] V. I. Lim, Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure, *J. Mol. Biol.*, **88**, 857-872 (1974).
- [61] K. Nagano and K. Hasegawa, Logical Analysis of the Mechanism of Protein Folding, *J. Mol. Biol.*, **94**, 257-281 (1975).
- [62] J. Garnier, D. J. Osguthorpe and B. Robson, Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins, *J. Mol. Biol.*, **120**, 97-120 (1978).
- [63] W. Kabsch and C. Sander, Segment83. unpublished 1983.
- [64] O. B. Ptitsyn and A. V. Finkelstein, Theory of protein secondary structure and algorithm of its prediction, *Biopolymers*, **22**, 15-25 (1983).
- [65] 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).
- [66] J.-F. Gibrai, J. Garnier and B. Robson, Further Developments of Protein Secondary Structure Prediction Using Information Theory. New Parameters and Consideration of Residue Pairs, *J. Mol. Biol.*, **198**, 425-443 (1987).
- [67] V. Biou, J. F. Gibrai, J. M. Levin, B. Robson and J. Garnier, Secondary structure prediction: combination of three different methods, *Prot. Engin.*, **2**, 185-91 (1988).
- [68] F. Bossa and S. Pascarella, PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network, *CABIOS*, **5**, 319-320 (1990).
- [69] F. Fogelman-Soulié and C. Mejía, Incorporating knowledge in multi-layer networks: the example of proteins secondary structure prediction. In: F. Fogelman-Soulié and J. Héroult (eds.), *Neuro Computing, Algorithms, Architectures and Applications*. Berlin, Heidelberg: Springer, 1990, 185-194.
- [70] R. D. King and M. J. Sternberg, Machine Learning Approach for the Prediction of Protein Secondary Structure, *J. Mol. Biol.*, **216**, 441-457 (1990).
- [71] D. G. Kneller, F. E. Cohen and R. Langridge, Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network, *J. Mol. Biol.*, **214**, 171-182 (1990).
- [72] K. Nishikawa and T. Noguchi, Predicting Protein Secondary Structure Based on Amino Acid Sequence, *Meth. Enz.*, **202**, 31-44 (1991).
- [73] M. J. Rooman, J. P. Kocher and S. J. Wodak, Prediction of Protein Backbone Conformation Based on Seven Structure Assignments: Influence of Local Interactions, *J. Mol. Biol.*, **221**, 961-979 (1991).
- [74] D. L. Dowe, J. Oliver, T. I. Dix, L. Allison and C. S. Wallace, A Decision Graph Explanation of Protein Secondary Structure Prediction. Dep. Computer Science, Monash Univ., Clayton 3168, Australia, 1992.
- [75] S. Salzberg and S. Cost, Predicting Protein Secondary Structure with a Nearest-neighbor Algorithm, *J. Mol. Biol.*, **227**, 371-374 (1992).

- [76] P. Stolorz, A. Lapedes and Y. Xia, Predicting Protein Secondary Structure Using Neural Net and Statistical Methods, *J. Mol. Biol.*, **225**, 363-377 (1992).
- [77] X. Zhang, J. P. Mesirov and D. L. Waltz, Hybrid System for Protein Secondary Structure Prediction, *J. Mol. Biol.*, **225**, 1049-63 (1992).
- [78] K. Asai, S. Hayamizu and K. Handa, Prediction of protein secondary structure by the hidden Markov model, *CABIOS*, **9**, 141-146 (1993).
- [79] P. E. Boscott, G. J. Barton and W. G. Richards, Secondary structure prediction for modelling by homology, *Prot. Engin.*, **6**, 261-266 (1993).
- [80] P. Fariselli, M. Compiani and R. Casadio, Predicting secondary structures of membrane proteins with neural networks, *Eur. Biophys. J.*, **22**, 41-51 (1993).
- [81] J. M. Levin, S. Pascarella, P. Argos and J. Garnier, Quantification of Secondary Structure Prediction Improvement Using Multiple Alignments, *Prot. Engin.*, in press (1993).
- [82] R. Maclin and J. W. Shavlik, Using Knowledge-Based Neural Networks to Improve Algorithms: Refining the Chou-Fasman Algorithm for Protein Folding, *Machine Learning*, **11**, 195-215 (1993).
- [83] F. Sasagawa and K. Tajima, Prediction of protein secondary structures by a neural network, *CABIOS*, **9**, 147-152 (1993).
- [84] S. A. Benner and D. Gerloff, Patterns of Divergence in Homologous Proteins as Indicators of Secondary and Tertiary Structure of the Catalytic Domain of Protein Kinases, *Adv. Enz. Reg.*, **31**, 121-181 (1990).
- [85] S. A. Benner, Predicting de novo the folded structure of proteins, *Curr. Opin. Str. Biol.*, **2**, 402-412 (1992).
- [86] S. A. Benner, M. A. Cohen and D. Gerloff, Correct structure prediction?, *Nature*, **359**, 781 (1992).
- [87] S. A. Benner, M. A. Cohen and D. Gerloff, Predicted Secondary Structure for the Src Homology 3 Domain, *J. Mol. Biol.*, **229**, 295-305 (1993).
- [88] D. L. Gerloff, T. F. Jenny, L. J. Knecht, G. H. Gonnert and S. A. Benner, The nitrogenase MoFe protein, *FEBS Lett.*, **318**, 118-124 (1993).