J. Mol. Biol. (1993) 232, 584–599

Prediction of Protein Secondary Structure at Better than 70% Accuracy

Burkhard Rost and Chris Sander

European Molecular Biology Laboratory
Meyerhofstraße 1, D-6900 Heidelberg, Germany

(Received 1 February 1993; accepted 13 April 1993)

We have trained a two-layered feed-forward neural network on a non-redundant data base of 130 protein chains to predict the secondary structure of water-soluble proteins. A new key aspect is the use of evolutionary information in the form of multiple sequence alignments that are used as input in place of single sequences. The inclusion of protein family information in this form increases the prediction accuracy by six to eight percentage points. A combination of three levels of networks results in an overall three-state accuracy of 70·8% for globular proteins (sustained performance). If four membrane protein chains are included in the evaluation, the overall accuracy drops to 70·2%. The prediction is well balanced between α-helix, β-strand and loop: 65% of the observed strand residues are predicted correctly. The accuray in predicting the content of three secondary structure types is comparable to that of circular dichroism spectroscopy. The performance accuracy is verified by a sevenfold cross-validation test, and an additional test on 26 recently solved proteins. Of particular practical importance is the definition of a position-specific reliability index. For half of the residues predicted with a high level of reliability the overall accuracy increases to better than 82%. A further strength of the method is the more realistic prediction of segment length. The protein family prediction method is available for testing by academic researchers *via* an electronic mail server.

Keywords: protein secondary structure prediction; multiple sequence alignments; *secondary structure content; neural network*

1. Introduction

Large-scale sequencing projects produce an exploding number of known protein sequences. The current number is 26,000 (Bairoch & Boeckmann, 1992) sequences, but before the end of the century 100,000 will have to be dealt with. This is in contrast to the much slower increase in the number of known protein structures, currently at about 1000 (Bernstein *et al.*, 1977). Can theory help to narrow the widening gap? The most reliable prediction of the structure of new proteins is done by detection of significant similarities to proteins of known structure (Taylor & Orengo, 1989; Sander & Schneider, 1991; Vriend & Sander, 1991). But only about one-seventh of new sequences have sequence similarities to known structures (Bork *et al.*, 1992). What about the rest? Attempts to predict structure from sequence by physical simulation techniques, such as molecular dynamics (Momany *et al.*, 1975; Karplus & Petsko, 1990), have fallen far short of solving the task of finding the "hidden" relation between the primary and tertiary structure. Although the folding process may require catalysts such as chaperonins (Hubbard & Sander, 1991), the basic hypothesis that the three-dimensional (tertiary) structure of a protein is uniquely determined by its sequence of amino acids (primary structure) appears to remain valid (Anfinsen *et al.*, 1963; Ewbank & Creighton, 1992). A simple reduction of the prediction problem is the projection of the three-dimensional structure onto one dimension, i.e. onto a string of secondary structure assignments for each residue.

Secondary structure predictions have been performed by various methods (Szent-Györgyi & Cohen, 1957; Periti *et al.*, 1967; Ptitsyn, 1969; Pain & Robson, 1970; Robson & Pain, 1971), ever since Pauling suggested that proteins form certain local conformational patterns like helices and strands (Pauling & Corey, 1951; Pauling *et al.*, 1951). The different algorithms can be approximately grouped into those using (1) statistical information (Nagano, 1973; Chou & Fasman, 1974; Nagano & Hasegawa, 1975; Garnier *et al.*, 1978; Schulz & Schirmer, 1979; Levin *et al.*, 1986; Gibrat *et al.*, 1987; Biou *et al.*, 1988; Kanehisa, 1988; Levin & Garnier, 1988; Fasman, 1989; Garrett *et al.*, 1991; Muggleton *et al.*, 1992); (2) physico-chemical properties (Lim, 1974; Ptitsyn & Finkelstein, 1983); (3) sequence patterns

0022-2836/93/140584-16 $08.00/0

(Cohen *et al.*, 1983, 1986; Taylor & Thornton, 1983; Rooman *et al.*, 1989, 1991; Sternberg & King, 1990; Rooman & Wodak, 1991; Presnell *et al.*, 1992); (4) multi-layered (or neural) networks (Bohr *et al.*, 1988, 1990; Qian & Sejnowski, 1988; Holley & Karplus, 1989; Bossa & Pascarella, 1990; Kneller *et al.*, 1990; Hirst & Sternberg, 1992; Maclin & Shavlik, 1993; Stolorz *et al.*, 1992; Zhang *et al.*, 1992); and (5) evolutionary conservation (Maxfield & Scheraga, 1979; Zvelebil *et al.*, 1987; Frampton *et al.*, 1989; Benner & Gerloff, 1990; Barton *et al.*, 1991; Niermann & Kirschner, 1991; Ouzounis & Melvin, 1991; Musacchio *et al.*, 1992; Russell *et al.*, 1992; Gibson *et al.*, 1993). One of the problems of these prediction methods is that the formation of secondary structure elements is only to a certain degree due to sequentially local interaction of amino acids (Nagano & Hasegawa, 1975; Taylor, 1988; Zhong *et al.*, 1992). However, most methods known to date do rely on local information. For the last decade these methods have hovered around 60 to 64% in overall three-state accuracy. Some methods predicted, e.g. β-strands, only 12 percentage points better than the chance value of 33% (Biou *et al.*, 1988). Recently, the reported overall accuracy of 66·5% (Zhang *et al.*, 1992) and single examples of predictions of proteins of unknown structure have generated enthusiasm in the field (Barton *et al.*, 1991; Benner *et al.*, 1992; Rost & Sander, 1992; Russell *et al.*, 1992). Yet it was claimed that predictions cannot be better than 65($\pm$2)% (Garnier, 1992).

Here, we present the results of an in-depth analysis of the performance of multi-layered (neural) networks. By appropriately processing the information about structure contained in a multiple sequence alignment, it proves possible to increase the accuray of secondary structure prediction above 70%. Our system of networks outperforms previous methods in four respects. (1) The overall accuracy of 70·8% for globular water-soluble proteins is more than four percentage points higher than that of any other method published (and about 6 percentage points above a method tested with comparable rigour on the same data set). (2) The improvement in per residue accuracy is particularly significant for β-strands, with 65·4% of the observed β residues correctly predicted (compared to e.g. 46% for the well-established method GORIII). (3) The length distribution of predicted secondary structure elements is much more protein-like than that for other typical prediction methods. (4) The network allows the identification of residues predicted with higher than average reliability. More than one-fifth of all residues are predicted with an expected accuracy above 90%, more than half of all residues score above 82%.

2. Materials and Methods

(a) *Cross-validation technique and data set used*

It is impossible to accurately know in advance the accuracy of a prediction tool when applied to a new protein. How can the data bank of known structures be used to estimate the performance on new proteins? Two requirements are essential to derive a reasonable assessment of the method's generalization ability: (1) the pairwise identity of the protein chains used for developing the prediction tool and those for testing should be lower than the value sufficient for modelling tertiary structure by homology, and (2) a multiple cross-validation test (ideally jack-knife) has to be performed to exclude a potential dependency of the evaluated accuracy on the particular test set chosen.

For homologous proteins, alignment procedures predict secondary structure more accurately than any method using the sequence information only. Therefore, a tool not using the homology to a protein of known structure has to be tested on those cases for which it will be used, i.e. protein chains without significant pairwise homology to those used for developing the method. What is a significant homology? A length-dependent cut-off for relevant similarity is given by HSSP† (Sander & Schneider, 1991): e.g. for chains with more than 80 residues the mutual similarity should be $<25\%$. The restriction in accepted pairwise similarity reduced the number of protein chains that can be used from 700 (PDB: Protein Data Bank, 1992) to about 150 (Hobohm *et al.*, 1992). Table 1 gives the 130 chains we used. All chains are structures known at a resolution of at least 2·5 Å.

Jack-knife testing means use 129 chains for setting up the tool (training the network) and 1 for estimating the performance on new proteins (testing it). This has to be repeated 130 times until each protein has been used once for testing. The average over all 130 tests gives a reasonable estimate of the prediction accuracy. For networks, such a strategy is prohibited by the limitations of computational resources. The extreme contrast to jack-knife testing is to use only a single test set of, say, 20 proteins. Our experiences show that such an approach results in different accuracies for different test sets. The goal of testing the prediction tool is to assess the accuracy to be expected for any new protein sequence. Since different test sets yield different results, it is not sufficient to use only 1 set. As a compromise between jack-knife and a single test set, we worked with 7-fold cross-validation: 111 chains are taken for training, 19 for testing. This is repeated 7 times with different sets of 19 until all proteins have been used for testing exactly once (for one set the split was 114/16).

(b) *Measures of protein secondary structure prediction accuracy*

Once the data set is fixed, the problem arises of how to define a measure for the quality of a particular prediction. Most publications on secondary structure prediction compute ratios that reflect the number of properly predicted residues. All such coefficients can be derived from a 3×3 (for 3 secondary structure types) accuracy table A, with:

A_{ij} = number of residues predicted to be in structure type j and observed to be in type i.

† Abbreviations used: HSSP, database of Homology-derived Structures and Sequence alignments of Proteins; PDB, Protein Data Bank (of known 3-dimensional structures); DSSP, Dictionary of Secondary Structures of Proteins; PHD, Profile network from HeiDelberg (3 levels of networks); CD, circular dichroism spectroscopy.

Table 1

Representative set of 126 globular and 4 membrane protein chains with less than 25% pairwise similarity for lengths >80 used for training and testing the method (24,395 residues with 32% α, 21% β and 47% L, resolution ≤2·5 Å for crystal structures)

256b_A	2aat	8abp	6acn	1acx	8adh	3ait	1ak3_A	2alp	9api_A
9api_B	1azu	3b5c	1bbp_A	1bds	1bmv_1	1bmv_2	3blm	4bp2	2cab
7cat_A	1cbh	1cc5	2ccy_A	1cd4	1cdt_A	3cla	3cln	4cms	4cpa_I
6cpa	6cpp	4cpv	1crn	1cse_I	6cts	2cyp	5cyt_R	1eca	6dfr
3ebx	5er2_E	1etu	1fc2_C	1fdl_H	1fdx	1fkf	2fnr	2fxb	1fxi_A
4fxn	3gap_A	2gbp	2gcr	1gd1_O	2gls_A	2gn5	1gp1_A	4gr1	1hip
6hir	3hmg_A	3hmg_B	2hmz_A	5hvp_A	2il1b	3icb	7icd	1il8_A	9ins_B
1l58	1lap	5ldh	2lh4	2lhb	1lrd_3	2ltn_A	2ltn_B	5lyz	1mcp_L
2mev_4	2or1_L	1ovo_A	2pab_A	1paz	9pap	2pcy	4pfk	3pgm	2phh
1pyp	1r09_2	2mhu	1mrt	1ppt	1rbp	1rhd	4rhv_1	4rhv_3	4rhv_4
3rnt	7rsa	2rsp_A	4rxn	1s01	1sdh_A	4sgb_I	1sh1	2sns	2sod_B
2stv	2tgp_I	1tgs_I	3tim_A	6tmn_E	2tmv_P	1tnf_A	4ts1_A	2tsc_A	1ubq
2utg_A	9wga_A	2wrp_R	1wsy_A	1wsy_B	4xia_A	1prc_C	1prc_H	1prc_L	1prc_M

Protein Data Bank (PDB) identifier (first 4 characters) is followed by the chain identifier.

The sums over the columns of A give the number of residues predicted to be in structure i:

$$a_i = \sum_{j=1}^{3} A_{ji}, \quad \text{for } i = \alpha, \beta, \text{L}.$$

The sums over the rows give the number of residues observed to be in structure i:

$$b_i = \sum_{j=1}^{3} A_{ij}, \quad \text{for } i = \alpha, \beta, \text{L}.$$

The sum over all elements of A is the number of residues in the data bank used, abbreviated by b:

$$b = \sum_{j=1}^{3} b_j = \sum_{j=1}^{3} a_j.$$

The term observed refers to the experimental structure determined by X-ray or nuclear magnetic resonance as represented in the Dictionary of Secondary Structures of Proteins (DSSP: Kabsch & Sander, 1983), which distinguishes 8 secondary structure classes. These can be grouped into 3 classes according to the following conventions: H (α-helix), G (3_{10}-helix), I (π-helix)→helix (α), E (extended strand)→strand (β), and B (residue in isolated β-bridge), T (turn), S (bend), _ (rest, coil)→loop (L), with the corrections: B_→ββ, but B_B→LLL.) We shall frequently use the abbreviations:

$$Q_i = Q_i^{\%\text{obs}} = \frac{A_{ii}}{b_i} \times 100, \quad \text{for } i = \alpha, \beta, \text{L}, \tag{1}$$

which describe for class i the percentage of residues correctly predicted to be in class i relative to those observed to be in class i (for simplicity referred to as Q_i). The percentages of residues correctly predicted to be in class i from all residues predicted to be in i are given by:

$$Q_i^{\%\text{pred}} = \frac{A_{ii}}{a_i} \times 100, \quad \text{for } i = \alpha, \beta, \text{L}. \tag{2}$$

This percentage supplies an estimate of the conditional probability of correct prediction, given a predicted state. Most authors use the overall 3-state accuracy:

$$Q_{\text{total}} = \frac{\sum_{i=1}^{3} A_{ii}}{b} \times 100 \tag{3}$$

or, in other words, the percentage of all correctly predicted residues. These percentages describe the performance accuracy for a prediction tool averaged over all residues in the data bank. The expected accuracy for a single protein is best described by averaging over all chains. The mean accuracy per protein chain is given by:

$$\langle Q \rangle_{\text{chain}} = \frac{1}{N^{\text{chain}}} \sum_{c=1}^{N^{\text{chain}}} Q_{\text{total}}^{c}, \tag{4}$$

where N^{chain} is the number of all chains in the data bank, and Q_{total}^{c} the accuracy defined by eqn (3) for chain c. The standard deviation σ of this per chain accuracy can supply an estimate of the range of accuracy to be expected. $\langle Q \rangle_{\text{chain}}$ tends to be higher than Q_{total} if short chains are predicted more accurately than longer ones.

As the data bank used (Table 1) contains 32% α, 21% β and 47% L residues, Q_{total} tends to be dominated by the accuracy for loop prediction. However, since the user is primarily interested in the performance on the structure types α and β, the percentages given by eqns (1) and (2) reveal important additional information. The random prediction of 3 classes (if weighted by the percentage of occurrence) would be $Q_{\text{total}} = 36{\cdot}3\%$.

A more complicated measure of accuracy is given by the correlation coefficient introduced by Matthews (1975):

$$C_i = \frac{p_i n_i - u_i o_i}{\sqrt{(p_i + u_i)(p_i + o_i)(n_i + u_i)(n_i + o_i)}}, \quad \text{for } i = \alpha, \beta, \text{L}, \tag{5}$$

with p_i being the number of properly predicted residues in conformation i; n_i the number of those correctly not assigned to structure i; u_i the number of underestimated, and o_i that of overestimated conformations. In terms of the accuracy table A:

$$p_i = A_{ii}, \quad n_i = \sum_{j \neq i}^{3} \sum_{k \neq i}^{3} A_{jk}, \quad \text{for } i = \alpha, \beta, \text{L},$$

$$o_i = \sum_{j \neq i}^{3} A_{ji}, \quad u_i = \sum_{j \neq i}^{3} A_{ij}, \quad \text{for } i = \alpha, \beta, \text{L}.$$

Here, we should like to introduce an entropy-related information that merges the different percentages to a single number with all elements of the accuracy matrix contributing equally. The information can be defined as:

$$I = \ln \left\{ \frac{P_{\text{pred}}}{P_{\text{obs}}} \right\},$$

where P_{obs} is the probability for finding one particular string of b residues with b_i residues being in structure i out of all combinatorial possible ones, and P_{pred} is the probability for a particular realization of the prediction table A:

$$\frac{1}{P_{obs}} = \frac{b!}{\prod_{j=1}^{3} b_j!} \quad \text{and} \quad \frac{1}{P_{pred}} = \prod_{j=1}^{3} \left\{ \frac{b!}{\prod_{j=1}^{3} a_j!} \right\},$$

where $x!$ stands for x factorial. The information can be rewritten in the form:

$$I = \ln\left(\frac{b!}{\prod_{i=1}^{3} b_i!}\right) - \sum_{i=1}^{3} \ln\left(\frac{a_i!}{\prod_{j=1}^{3} A_{ij}!}\right). \tag{6}$$

This expression can be approximated by the Stirling formula (reasonably accurate for $A_{ij} > 10$):

$$\begin{aligned} I = &- \sum_{i=1}^{3} \Bigg((b_i + 0{\cdot}5) \times \ln b_i + (a_i + 0{\cdot}5) \times \ln a_i \\ &- \sum_{j=1}^{3} (A_{ij} + 0{\cdot}5) \times \ln A_{ij} \Bigg) + (b + 0{\cdot}5) \times \ln b - 2 \times \ln 2\pi \\ \cong &- \sum_{i=1}^{3} \Bigg(b_i \times \ln b_i + a_i \times \ln a_i \\ &- \sum_{j=1}^{3} A_{ij} \times \ln A_{ij} \Bigg) + b \times \ln b. \end{aligned}$$

For a perfect prediction the second term in eqn (6) vanishes, thus in order to normalize the entropy to 1 for perfect prediction we redefine:

$$I = IP_{obs},$$

and with the Stirling approximation:

$$I = 1 - \frac{\sum_{i=1}^{3} a_i \times \ln a_i - \sum_{i,j=1}^{3} A_{ij} \times \ln A_{ij}}{b \times \ln b - \sum_{i=1}^{3} b_i \times \ln b_i}. \tag{7}$$

This information is related to the probability of deviation of table A from a random distribution:

$I = 0$, if: $A_{ij} = 1/9$, for $i, j = 1, 2, 3$,
$I = 1$, if: $A_{ij} = 0$, for $i \neq j$ and
$A_{ii} = b_i$, $i, j = 1, 2, 3$.

The latter means a completely correct prediction. An artifact of the construction is that the same deviation of $I = 1$ results if all helices are predicted as loop, and all loops as helix (for $i = 1$, helix; $i = 3$, loop). This shortcoming is acceptable, since the probability for such a prediction is very small for any reasonable prediction method. The advantage of such an entropy is that, e.g., over- and underpredictions equally decrease the value of I. Therefore, a method performing well only for, e.g. loop, might yield a relatively high level of overall accuracy, but the entropy I will be low.

Measures for single-residue accuracy do not completely reflect the quality of a prediction (Thornton *et al.*, 1992). Suppose the following 2 predictions were to be compared:

Observed:	ααααααααααLLL
Prediction 1:	LααaLααLαLLL
Prediction 2:	LLαααααααααaL

Although prediction 1 results in a higher level of overall accuracy than prediction 2, the latter better predicts the occurrence of the long helix. It is important to estimate the number of helices, strands and loops, their lengths (number of residues in segment) and locations. Such an estimate is not covered by single-residue measures, but by quantities such as the number of predicted secondary elements, their average length, and the length distribution. The length distribution of a prediction is given by counting, e.g., all predicted helices of length n, with $n = 1, 2, \ldots, n_{max}$ (length of the longest helix predicted). This predicted distribution can be compared to the distribution obtained from the observed structure. A simpler measure of how good the length of segments is predicted is the average segment length $\langle L_i \rangle$, $i = \alpha, \beta, L$:

$$\langle L_i \rangle = \frac{\text{Sum of the lengths over all segments of structure } i}{\text{Number of all segments of structure } i}. \tag{8}$$

(c) *Classifications by a layered network*

Various pattern recognition-related problems have led to widespread research on neural networks in general, and on multi-layered feed-forward networks in particular (history, Cowan, 1990; theoretical background, Amit, 1989; Hertz *et al.*, 1991; applications, Rumelhart & McClelland, 1986; Müller & Reinhardt, 1990). The simplest network is a perceptron, as shown in Fig. 1 (Minsky & Papert, 1988). The signal from the 3 input nodes is fed forward to the output node, which performs a 2-step procedure: the first is a multiplication of the vector of junctions **J** (describing the connections between the nodes) and the input vector $\mathbf{s}^0$: $h = \mathbf{J}\mathbf{s}^0$; the second is a non-linear trigger, which can be a step function of the form:

$$s^1 = \begin{cases} 1, \text{ if } h > 0 \\ 0, \text{ if } h \leq 0. \end{cases}$$

(For our networks we used the sigmoid given in the legend to Fig. 1 and eqn (11).)

An extension of the perceptron is the introduction of more output nodes (by which **J** becomes a matrix, and h a vector) and an additional layer of units that are "hidden" in the sense that this layer is directly related neither to the input nor to the output. A further generalization of the concept is to introduce connections from, e.g., the output to the input layer (backwards).

A feed-forward network can be regarded as a statistical method that is able to classify patterns according to their intrinsic correlation, i.e. the characteristic information they contain. The network performs a simple task: map a vector $\mathbf{s}^0$ of dimension N^0 (number of input nodes) onto another one $\mathbf{s}$ of dimension N (number of output nodes). The collective effect of an entire network is a classification of patterns. Provided the number of units and layers of the network suffices, an arbitrary pattern classification can be performed.

For applications, a further intrinsic feature of networks is important: the ability to generalize. Suppose there is a certain rule according to which a number of examples S is grouped. This set might be split into two distinct sets S_{Train} and S_{Test}. Suppose S_{Train} is learned by the network, i.e. the net extracts internal rules for grouping S_{Train}. Then, generalization refers to the ability to correctly classify S_{Test} with the rules (junction) derived from S_{Train}. The better the classification, the better the hidden rule is deduced by the network. If the number of training

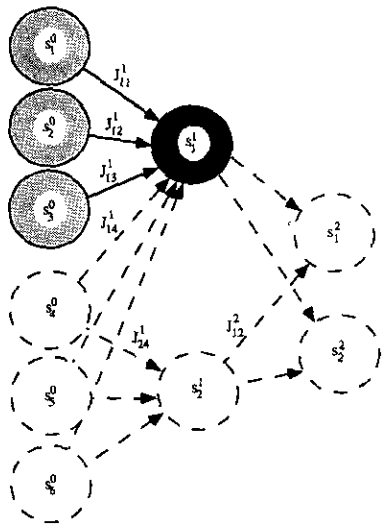

Figure 1. Function of a perceptron, the simplest neural network. A simple perceptron has only 1 output unit (black). Each of the left nodes receives a certain input signal (e.g. binary, i.e. =0 or 1). All units are connected to the output node by the junctions J^1, with e.g. J^1_{1j} connecting input unit j with output unit 1. The contribution of each left node (e.g. the jth) to the signal arriving at the right one is a product of the strength of the junction connecting the 2 units, and the input: e.g. $J^1_{1j}s^0_j$. All products (here 3) are summed by the right node (here s^1_1). This sum is then evaluated by a non-linear trigger function. The resulting map of the sum onto an interval between 0 and 1 is the actual output of the network. The broken-line nodes show a potential extension of the perceptron to a 2-layered feed-forward network. Stippled circles, input units, signal = 1 or 0. Black circle, output unit. Step 1, the input to this unit is summed according to:

$$h^1_i = \sum_{j=1}^{N^0+1} J_{ij}s^0_j \quad \text{(here, } i=1\text{)}.$$

Step 2, the output from this unit is computed by a sigmoid trigger function:

$$s^1_i = \frac{1}{1+\exp(-h^1_i)}.$$

Broken-line circles, the potential extension to a 2-layered feed-forward network.

examples is sufficient (and can be learned), generalization is perfect (Solla, 1991).

(d) *First level: sequence-to-structure net*

The system of networks we used consists of 3 levels. The first is a net classifying strings of adjacent residues (=*sequence pattern*) into the 3 secondary structure classes helix (α), strand (β) and loop (L) of the central residue. For representation of the input, a multiple sequence alignment is used (Fig. 2). The alignments are taken from the HSSP data bank (Sander & Schneider, 1991). One pattern is given by a window of $w=13$ consecutive residues (in Fig. 2, $w=7$). For each residue the frequency of occurrence of each of the 20 amino acids at one position in the alignment is computed. These 20 numbers represent a basic cell of the input layer in Fig. 2. Consequently, the whole input for a particular pattern extends over $20w$ input units. The target output is the secondary structure class of the central residue. (For the example of position 4 in Fig. 2, the target output is 1,0,0, which means that the central asparagine residue (N) is observed to be in a helix.) The window is shifted residue by residue through the protein chain, thus yielding N patterns for a chain with N residues. In order to allow a window to extend over the N terminus and the C terminus, a further unit has to be added for each residue. Suppose the secondary structure of the N terminus of a protein is to be predicted. Then this residue must be at position 4 (for $w=7$). The input for the following window positions is given by the residues of the protein. But there is no residue before the first. Therefore, an additional unit has to be added to each of the first 3 basic cells (Fig. 2). The value of these units is set to 1, those for the other 20 units for the first 3 basic cells are set to 0. Thus, finally $(20+1)w$ units are required for the input.

Two coding concepts were investigated: (1) the frequencies were directly used as the values of real input units, and (2) they were transposed into 4 binary units:

$$\begin{array}{ll} 0000 \text{ for frequency} & f<0{\cdot}02 \\ 0001 \text{ for} & 0{\cdot}02\le f<0{\cdot}33 \\ 0011 \text{ for} & 0{\cdot}33\le f<0{\cdot}66 \\ 0111 \text{ for} & 0{\cdot}66\le f<0{\cdot}98 \\ 1111 \text{ for} & f\ge 0{\cdot}98. \end{array} \qquad (9)$$

The value of output unit i of the network for sample v is computed according to:

$$s^{2,\nu}_i = f\left\{\sum_{j=1}^{N^1+1} J^2_{ij} f\left\{\sum_{k=1}^{N^0+1} J^1_{jk}s^{0,\nu}_k\right\}\right\}, \qquad (10)$$

with J^λ_{ij} as the junction between unit j in layer $\lambda-1$ and unit i in layer λ (for 2 layers of junctions, the layers of units are counted as: input layer, $\lambda=0$; hidden layer, $\lambda=1$; output layer, $\lambda=2$), the number of hidden N^1 and input N^0 units, the input of pattern ν to the kth input unit, $s^{0,\nu}_k$, and the sigmoid trigger function chosen as:

$$f(x) = \frac{1}{1+e^{-\beta x}}. \qquad (11)$$

β determines the slope of the sigmoid function.

The error E for pattern ν (the total error is a sum over all patterns) can be defined as:

$$E(\{J^1\},\{J^2\}) = \sum_{i=1}^{3} (s^{2,\nu}_i - d^\nu_i)^2, \qquad (12)$$

with $s^{2,\nu}_i$ being the output of the network for output unit i and sample ν, d^ν_i the observed secondary structure for sample ν and unit i (note, this is a binary quantity). The brackets { } emphasize that the error does not depend on 2 variables, but on 2 sets of variables given by the junction matrices of the first (J^1) and the second (J^2) layer.

A typical network we used contained 5000 to 15,000 junctions, i.e. free variables to be optimized. The number of examples used for the optimization of these variables was roughly 25,000.

For the training procedure, the straightforward gradient descent (with momentum term) was used (Press *et al.*, 1986; Rumelhart *et al.*, 1986). At each optimization time step t, the junctions J are changed such that the error decreases:

$$\Delta J(t+1) = J(t) - \varepsilon\frac{\partial E}{\partial J}(t) + \eta\Delta J(t-1), \qquad (13)$$

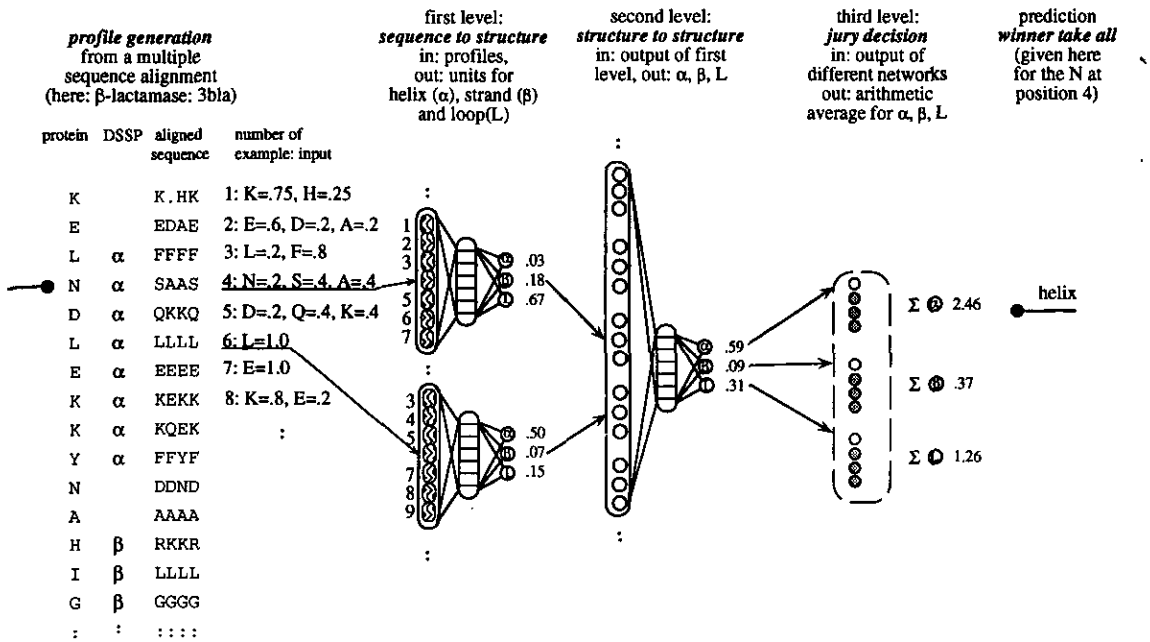

Figure 2. Our network system for secondary structure prediction. Our network system for predicting secondary structure consists of 3 layers: 2 network layers and 1 layer averaging over independently trained networks. ⊛, Basic cell containing $20+1$ units to code residues at position 1 to w of the input window; here, $w=7$. ⊜, Hidden units. Circled α, β and L, output units for helix, strand and loop. Stippled circles, output from architectures not shown here. —●, Example: residue N at position 4 predicted to be in helix ●—.

with the following choices: learning strength (step width of gradient descent) $\varepsilon = 0{\cdot}05$, momentum term $\eta = 0{\cdot}2$, initial junctions $J(0)$ chosen at random $\in [-0{\cdot}1, 0{\cdot}1]$. Training a network means trying to minimize the error E for all training examples. The partial derivative can easily be derived analytically. Suppressing the sample index ν, the iteration formulae for an arbitrary number of layers become, for the last layer L:

$$\frac{\partial E(t)}{\partial J_{ij}^{L}} = (s_i^L - d_i) s_i^L (1 - s_i^L) s_j^{L-1}, \tag{14}$$

for all previous layers $\lambda = 1, L-1$:

$$\frac{\partial E(t)}{\partial J_{ij}^{\lambda}} = (1 - s_i^{\lambda}) s_j^{\lambda - 1} \sum_{k=1}^{N^{\lambda+1}} \frac{\partial E(t)}{\partial J_{ki}^{\lambda+1}} J_{ki}^{\lambda+1}. \tag{15}$$

Usually, the patterns are picked at random from the stack of all training examples. Since the 3 secondary structure types are not equally distributed, a balanced training was also tested. For each time step, a pack of 3 patterns was picked at random with 1 example for each structural class (α, β, L). This implies a substitution of ∂E in eqns (13) to (15) by a sum over the error for each of the 3 samples. Thus, for 100 time steps of unbalanced training, 47% samples for loop, 32% for helix and 21% for strand, and for balanced training, 33 samples for each class are presented. The optimization procedure (eqn (13)) was terminated once the accuracy was higher than 70% for all training samples (75% for the second level network). Terminating in a minimum, i.e. running the minimization to completion, risks the deterioration of generalization by overtraining (unpublished data).

(e) *Second level: structure-to-structure net*

The first level network is trained to classify mutually independent segments of residues in terms of the state of a single central residue. There is no explicit representation of the fact that consecutive patterns are correlated, like for a helix consisting of at least 3 consecutive patterns. The correlation can be taken into account, at least in part, by using a second level, a structure-to-structure network. For this network the input is given by a window of $w = 17$ basic cells (Fig. 2). Each of these basic cells codes the output from the sequence-to-structure network for 1 example (i.e. the 3 output values for the prediction of secondary structure for the central residue). The target output is again the secondary structure class of the central cell (Fig. 2: output for the segment of position 4).

As in coding the sequence profiles, the output units of the first level sequence-to-structure net can be encoded by either input units with real values, or by several binary input units for each. Here, we report only data for binary coding of the second level input, using 8 bits/real number. (The coding was done as in eqn (9).)

(f) *Third level: jury decision*

Training the network is a walk through a relatively complex space. The gradient descent is sensitive to minor changes in parameters: it is, for instance, important how the initial junctions are chosen, and how the parameters for the training dynamics (eqns (12) to (13)) are adjusted (such as step width ε, inertia parameter η, definition of the error E, slope and form of sigmoid decision function f). A particular realization of the classification task (which for simplicity will be referred to as a particular architecture) is associated with a particular error (corresponding to a particular local minimum). This error is caused, partly, by a random noise in the strength of the junctions J. Combining χ different architectures results in a reduction of the noise provided the networks are not completely correlated. The simplest way to combine

independent networks is to compute an arithmetic average (jury decision):

$$\langle s_i \rangle_{\text{jury}} = \frac{1}{\chi} \sum_{a=1}^{\chi} s_i^a, \quad \text{for } i = \alpha, \beta, \text{L}. \tag{16}$$

s_i^a is the value of output unit i for architecture a. For simple problems, the benefit of a jury decision has been shown (Hansen & Salamon, 1990; Lincoln & Skrzypek, 1990). A concept similar to the jury is to combine different secondary structure predictions. This combination has been claimed to be successful for non-network methods (Biou *et al.*, 1988; Nishikawa & Noguchi, 1991). A more elaborate approach was used by training a network to combine different secondary structure prediction methods (Zhang *et al.*, 1992).

We generated the different architectures in the following way: the training was done in a balanced and in an unbalanced fashion on the first and second level of the network system. In this way, 2×2 different architectures were trained independently. One such quartet was trained with a real coding of sequence profiles, another by adding conservation weights (next section). This led to the first 8 or 9 networks used for the jury decision. The other network was trained with coding the profiles by 4 bits (in balanced fashion).

(g) *Adding conservation weights to the sequence profiles*

The sequence profiles differ in the number of sequences in the family and in the similarity of the aligned sequences to the input sequence. The additional information about the alignment can be exploited by placing a higher weight on positions that are particularly well conserved. Such a weight is contained in the HSSP data base (Sander & Schneider, 1991). For position i in the sequence it is defined by Schneider & Sander (unpublished results):

$$CW_i = \frac{\sum_{r,s=1}^{N^{\text{ali}}} w_{rs} sim_{rs}^i}{\sum_{r,s=1}^{N^{\text{ali}}} w_{rs}}$$

with

$$w_{rs} = (1 - \tfrac{1}{100} \times \%\text{identity}_{rs}), \tag{17}$$

where N^{ali} is the number of alignments, $\%\text{identity}_{rs}$ the percentage of sequence identity (over the entire length of the sequence) of sequence r and s in the alignment, sim_{rs}^i a value from the similarity matrix between these 2 sequences at position i (e.g. the Dayhoff matrix: Dayhoff, 1978). The conservation weight is scaled such that the mean value averaged over all residues in any particular protein becomes unity: $\langle CW_i \rangle_{\text{sequence}} = 1$.

We used this quantity as an additional input unit for each residue. (The value defined by eqn (17) was divided by 2, the maximal size of CW_i we found in the data base, to make it smaller than 1.) Thus, for the networks using conservation weights, the input layer is extended to $(20+1+1) \times 13$ units for the first and to $(3+1+1) \times 17$ for the second level network.

(h) *Reliability index for the prediction*

The vector algebra given in eqn (10) describes a separation of the input vectors (a classification of the patterns). Two patterns can be separated more easily if they have a large mutual distance. The final step of the prediction is a winner-take-all decision (Fig. 2), i.e. the highest output value is chosen as the prediction. If the difference between the output values proves to be correlated to the probability of a correct prediction, such a quantity could be rather useful in practice. We define a reliability index as:

$$RI = \text{INTEGER}(10 \times (out_{\max} - out_{\text{next}})), \tag{18}$$

where $out_{\max}$ is the output of the unit with highest value, and out_{next} that of the unit with the next highest value. The factor 10 normalizes RI to integer values from 0 to 9, as the sigmoid trigger function (eqn (1)) permits maximal output values of 1. $RI = 9$ should correspond to a rather reliable prediction.

(i) *Filtering the prediction*

Helices have a minimum length of 3 (Kabsch & Sander, 1983). Although the second level structure-to-structure net corrects the tendency of the first level sequence-to-structure net to predict too short helix fragments, the final jury prediction still contains single Hs. The simplest way to exclude such an unrealistic prediction is to convert all helices of length 1 or 2 into loop. A slightly more elaborate alternative makes use of the reliability index (eqn (18)), according to the following prescription:

- if length of helix <3, and RI for all residues in that helix <4, then helix→loop;
- if length of helix <3, and at least for 1 residue in the helix $RI \geq 4$, then extend helix up to a final length of 3 (in the direction of the residue flanking the helix with lowest RI).

In practice, the precise details of the filter do not matter much.

(j) *Secondary structure content*

The knowledge of secondary structure content can contribute to the assessment of the folding type of a new protein. One experimental way to estimate secondary structure content is circular dichroism spectroscopy (Johnson, 1990). The accuracy of the secondary structure content predicted by the network system for protein chain c can simply be calculated as the difference between observed and predicted content averaged over all N^{chain} chains:

$$\Omega_i = \frac{1}{N^{\text{chain}}} \sum_{c=1}^{N^{\text{chain}}} |\omega_i^{\text{obs},c} - \omega_i^{\text{pred},c}|,$$

with:

$$\omega_i^{\text{obs},c} = \frac{A_{i4}^c}{A_{44}^c}, \quad \text{and} \quad \omega_i^{\text{pred},c} = \frac{A_{4i}^c}{A_{44}^c}, \quad \text{for } i = \alpha, \beta, \text{L}, \tag{19}$$

where $\omega_i^{\text{obs},c}$ is the content of secondary structure i as observed for chain c. $\omega_i^{\text{pred},c}$ is the one predicted (the accuracy matrices A^c are determined for each protein chain). Ω_i is a measure for the success of predicting structure content in general, and in particular for the correctness of grouping proteins into structural classes. An alternative measure, the Pearson correlation coefficient, is sometimes used when assessing the success of circular dichroism estimates (Johnson, 1990; Perczel *et al.*, 1991, 1992; Böhm *et al.*, 1992).

$$\text{Corr}\Omega_i = \frac{\langle \omega_i^{\text{obs},c} \times \omega_i^{\text{pred},c} \rangle - \langle \omega_i^{\text{obs},c} \rangle \times \langle \omega_i^{\text{pred},c} \rangle}{\sqrt{\langle (\omega_i^{\text{obs},c})^2 \rangle - \langle \omega_i^{\text{obs},c} \rangle^2} \times \sqrt{\langle (\omega_i^{\text{pred},c})^2 \rangle - \langle \omega_i^{\text{pred},c} \rangle^2}} \tag{20}$$

Table 2

Set of 24 all-helical protein chains containing pairwise similarities above 25% (3468 residues with 64% α, 3% β and 33% L)

155c	1cc5	1ccr	1cdp	1eca	1hmz	1l21	1lh5	1lrd_A	1lrd_B
1pmb	1prc_A	1prc_B	1prc_C	1wsy_C	256b	2ccy_A	2ccy_B	2wrp_A	3hhb_A
3hhb_B	3icb	3mba	5tnc						

A chain is labelled all-α if more than 85% of repetitive secondary structure is helix and the length of the chain is >74 residues (Kneller *et al.*, 1990).

with the abbreviation (x stands for any variable in eqn (20) enclosed by $\langle \ldots \rangle$):

$$\langle x^c \rangle = \frac{1}{N^{\text{chain}}} \sum_{c=1}^{N^{\text{chain}}} x^c.$$

For proteins of unknown structure, the structural class is, of course, also not known. Three questions arose. Is the network prediction accurate enough to sort the proteins into structural classes? Can the performance of a network be increased by training only on one structural class? If the prediction accuracy can be increased by the classification into structural classes, does the increase vanish if the proteins are classified, not according to the experimentally known content of secondary structure, but according to the predicted one? To help answering these questions, we have looked at a list of helical proteins. We used the prediction of the network system to classify the proteins as helical. Not only proteins with exclusively helix (Levitt & Chothia, 1976) were classified as all-helical, but according to the rules used by, e.g., Kneller *et al.* (1990), all chains were classified as all-helical if the sequence length is >74 residues and, if at least 85% of the repetitive secondary structure (helix, strand) is helix (Kneller *et al.*, 1990). A jack-knife test on 24 all-helical chains (Table 2) was performed, i.e. in this case a 24-fold cross-validation check.

3. Results

(a) *More than six percentage points gained by use of sequence profiles*

Four results can be summarized: multiple cross-validation shows that some of the previous work on the performance of neural networks in secondary structure prediction overestimated the expected accuracy; use of sequence profiles in binary coding increases the overall accuracy by about three

Table 3

Prediction accuracies, average segment lengths and information for various networks

Type of network	Q_{total} (4)	Q_α (1)	$Q_\alpha^{\%\text{pred}}$ (2)	C_α (5)	$\langle L_\alpha \rangle$ (8)	Q_β (1)	$Q_\beta^{\%\text{pred}}$ (2)	C_β (5)	$\langle L_\beta \rangle$ (8)	Information (7)
Reference 1st	61·7	56	59	0·39	4·2	41	52	0·34	2·9	0·12
Reference 2nd	62·6	57	62	0·42	6·2	42	53	0·35	3·8	0·13
Balanced 2nd	60·6	58	62	0·43	6·9	57	45	0·36	4·6	0·13
Prof binary	65·3	69	67	0·53	7·3	63	51	0·44	4·8	0·19
Prof real unbal	68·2	64	72	0·55	9·2	55	59	0·46	4·8	0·20
Prof real bal	67·4	70	68	0·54	8·1	63	55	0·47	4·9	0·21
Prof cons	68·0	71	71	0·57	8·3	68	54	0·48	5·1	0·22
Jury	70·2	71	72	0·59	9·2	66	58	0·51	5·1	0·24
Jury†	70·8	72	73	0·60	9·3	66	60	0·52	5·0	0·25
Observed					9·0				5·1	

Given are averages over 7 test sets chosen such that each chain of Table 1 is used exactly once for testing, and that the ratio of helix/strand/loop is not the same as for all 130 chains. (References to the equation numbers in the text are given in parentheses.) The networks can be described by the following Table.

† Results excluding the membrane protein 1prc.

Type of network	Number of levels	Profiles?	Coding of input	Conservation weight?	Balanced training?
Reference 1st	1	No	Binary	No	No
Reference 2nd	2	No	Binary	No	No
Balanced 2nd	2	No	Binary	No	Yes
Prof binary	2	Yes	Binary	No	Yes
Prof real unbal	2	Yes	Real/binary	No	No
Prof real bal	2	Yes	Real/binary	No	Yes
Prof cons	2	Yes	Real/binary	Yes	Yes
Jury	2	Yes	Mixed	Mixed	Mixed

The column headed Coding of input refers to the alternatives of coding real numbers either by binary or by real input units. Real/binary means that the first level network is coded with real numbers, the second level net with binary ones.

percentage points; the real coding of profiles results in an additional increase of about three percentage points; and using conservation weights adds a further half percentage point to the overall accuracy.

When applying a sevenfold cross-validation test on a data set without significant pairwise similarity (Table 1), we found that a network comparable to those used in earlier studies reached an overall accuracy of 61·7% (Table 3) instead of 63·6 to 65% (Qian & Sejnowski, 1988; Holley & Karplus, 1989; Kneller *et al.*, 1990; Stolorz *et al.*, 1992). For the final network system with three levels, the accuracies for the seven different test sets are shown in Figure 3. The differences between the best and the worst of the seven sets (each comparable in size to those used in previous publications not performing multiple cross-validation) is about seven percentage points. Therefore, a fortuitous choice of a single test set is a probable cause of overestimate. The conclusion is that a simple network is not as good in predicting secondary structure as empirical-statistical methods such as COMBINE (Biou *et al.*, 1988). However, the same network when trained on multiple sequence alignments outperforms all previously published methods (Fig. 4). Evolutionary information is extremely useful in predicting secondary structure. Sequence profiles are one way of using evolutionary information (and the simplest).

The question of whether to code the real valued profiles by four bits or real numbers is answered clearly by the comparison of the overall accuracy of the first level networks: 64·9% (binary) to 65·9% (real). For the coding of the input to the second level, the binary coding proves to be superior: 66·7% (real) *versus* 67·4% (binary). Thus, the strategy used in the end is to code the profiles on the first level (sequence-to-structure) by real numbers, the input to the second level (structure-to-structure) by eight bits per unit.

A further half percentage point can be gained by using the conservation weight CW (eqn (17)) as additional input unit (Table 3). Overall, the first two levels of the sequence profile networks reach more than 68% in three-state accuracy.

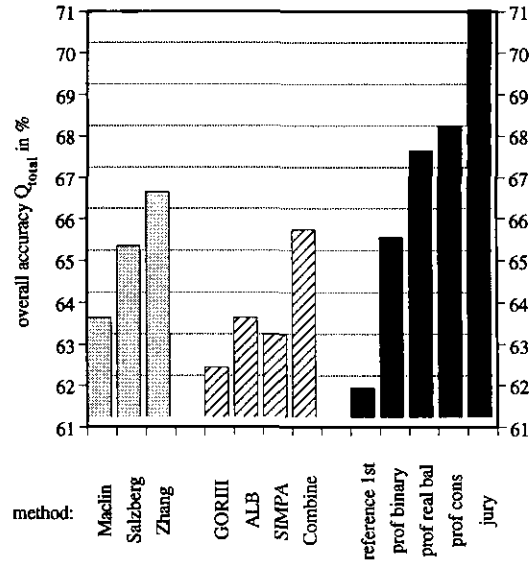

Figure 4. Overall accuracy of various methods. The methods shown used multiple cross-validation. The first 3 methods used data sets with similarities >30% between test and training sets (seq. hom.): Maclin (Maclin & Shavlik, 1993), and Salzberg (Salzberg & Cost, 1992) multi-layered network approach, Zhang (Zhang *et al.*, 1992) an approach combining networks in combination with other methods. Of the middle 4 methods, Combine (Biou *et al.*, 1988), SIMPA (Levin & Garnier, 1988) and GORIII (Gibrat *et al.*, 1987) are reported to have had no significant similarities between test and training sets (no seq. hom.); ALB (Ptitsyn & Finkelstein, 1983) was evaluated on the data set of Table 1. The last 5 bars show our networks; for the abbreviations, see Table 3. Stippled bars, network methods, seq. hom.; hatched bars, non-network methods, no seq. hom.; filled bars, our network methods, no seq. hom.

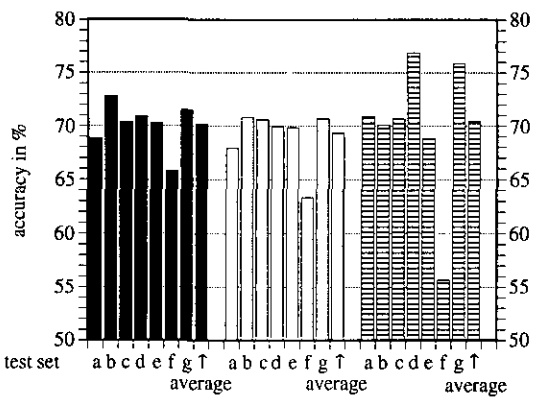

Figure 3. Variation of prediction accuracy with choice of test set. The accuracies are given for each of the 7 test sets used for the 7-fold cross-validation analysis. The last column in each set gives the arithmetic averages over all residues in all 7 test sets. The sets were chosen to be of about the same size (18 chains with some 3500 residues). The content of secondary structure differed between the sets (α, 25 to 39%; β, 16 to 27%; and L, 42 to 52%) to reflect the fact that this ratio is known for the current data bank but not for structurally unknown proteins. Q_{helix} and Q_{strand} are the percentages of observed structures (eqn (1)). Filled bars, Q_{total}; stippled bars, Q_{helix}; hatched bars, Q_{strand}.

(b) *Further two percentage points by the jury decision*

How does the expected overall accuracy depend on: the choice of the test sets; the details of the training procedure; and on the particular protein to be predicted?

Seven-fold cross-validation yields an estimate for the overall accuracy to be expected that is relatively independent on the choice of the test sets (Fig. 3).

The average over all seven test sets depends on the particular type of neural network used: networks trained in balanced or alternatively in unbalanced fashion differ in the overall accuracy by

about one to three percentage points (Table 3). The use of conservation weights does not produce a substantial increase in overall accuracy (0·5 percentage point), but it helps to extract the information given by the training samples in a different way from a network without conservation weights (the increase in information for the conservation weight net is more clear than that in overall accuracy, Table 3). An arithmetic combination of different architectures (II f) improves the performance if the architectures are not fully correlated in the error of their predictions. The jury decision over nine networks we used as the third level of the network system improves the overall accuracy by two percentage points. The final network system is the first tool ever to exceed the "magical" 70% in overall three-state accuracy: 70·8% (for soluble chains). (The final 3-level network system will be referred to as PHD for Profile network from HeiDelberg.)

What is to be expected as accuracy for a single new protein? The accuracy averaged over chains (rather than single residues, eqn (4) reaches 71·0(±9·3)% (± standard deviation). This means that the expected three-state accuracy most likely lies between 62 and 80% (Fig. 5). The fact that the per chain average is slightly higher than the average over the whole data bank indicates that shorter chains are predicted slightly more accurately (the average length of the chains in Table 1 is about 190 residues).

(c) *Performance worse for membrane proteins and single sequences*

Membrane proteins have a different physical environment from water-soluble globular proteins and, hence, different rules have to be learned to predict the structure. We included four chains of the membrane protein photosynthetic reaction centre (1prc_C, 1prc_H, 1prc_L and 1prc_M) in our data set to see how accurately these chains are predicted. The prediction accuracy was, as expected, below

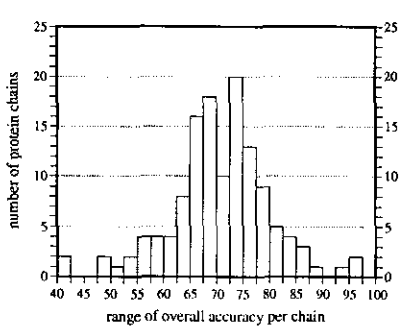

Figure 5. Expected variation of prediction accuracy with protein chain. The distribution of the per chain 3-state accuracy (eqn (4)) can be interpreted as the expected variation of prediction accuracy for protein sequences of unknown structure. The standard deviation is 9·3%. The chains predicted worst are 1crn, 1fc2_C and 2mev_4; those predicted best are 9api_B, 1ppt and 2utg_A.

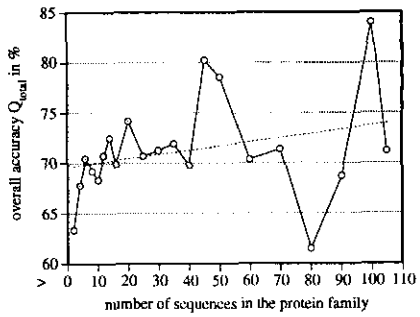

Figure 6. Improvement of accuracy with increasing family size. The points give the 3-state accuracy averaged over protein chains for which the number of sequences in the sequence profile falls into a certain interval. The intervals have been chosen such that each point represents an average over some 6 protein chains. The broken line gives a simple fit as a visual guide. The overall trend toward higher accuracy with increasing size of the protein family is partly masked by the strong variation of accuracy for individual protein chains.

average. The conclusion is that the results presented in this study apply strictly only to water-soluble globular proteins. However, an interesting side result of the inclusion of 1prc in training is that the four membrane chains are predicted with an overall accuracy of 58%. The network tends to overpredict strands and to underpredict helices (data not explicitly shown here). When 1prc is included in the computation of the overall three-state accuracy, the result drops from 70·8% to 70·2%.

The network using sequence profiles scores six to eight percentage points higher than a network using single sequences. What is to be expected if a network trained on sequence profiles is tested without providing the information of the multiple alignment? It turns out that the gain is almost completely lost. This leads to the question of how the increase in overall accuracy gained by using sequence profiles depends on the number of sequences in the multiple alignment. The increase of accuracy *versus* the number of sequences in the multiple alignments (Fig. 6) is partially obscured by the ±10% standard deviation of the per chain averages. The more sequences in the alignment the better, but how does the similarity of the aligned sequence to the input sequence influence the prediction? There is as well an effect of the distribution of the similarities of the alignments. Our experience is that it is best for the prediction if the alignment has a large number of sequences ranging from 30 to 100% similarity to the target sequence.

(d) *Reliability index helps to evaluate the prediction*

All results reported use the final winner-take-all projection of the three real output units onto one secondary structure (Fig. 2). A part of a protein cannot be in a helix, say, 65% of the time. But the prediction that the part is in a helix can have a probability of being accurate of 65%. The results

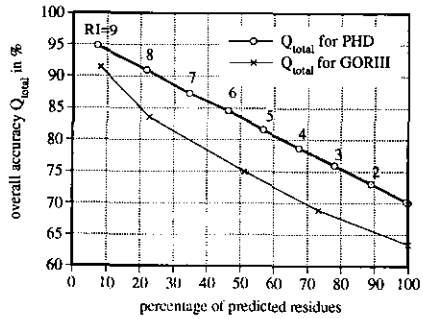

Figure 7. Expected prediction accuracy for residues with a reliability index above a given cut-off. Plotted are averages of the 3-state accuracy over all those residues with reliability index $RI > n$, $n = 0, \ldots, 9$ (eqn (18)). For comparison, the reliability is given for the well-known GORIII method, which uses 5 rather than 10 reliability intervals. For example, about 28% of all residues have $RI > 7$ and of these 92% are correctly predicted in PHD; for GORIII, 23% have $RI_{GORIII} > 4$, 83% of these are correctly predicted.

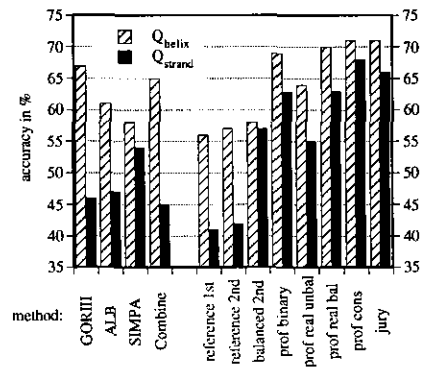

Figure 9. Comparison of helix/strand accuracies for various methods. Separate accuracies for helix and strand (eqn (1)) are not available for all methods. Here, 4 of these are compared to the performance of various networks tested by us. For the abbreviations, see Fig. 4 and Table 3.

shown in Figure 7 prove that the differences between the real output values (eqn (18)) supply an effective measure for the reliability of a prediction. The higher the difference between highest and next highest output unit, the more reliable the prediction, e.g. 57% of the residues have an accuracy of 82% ($RI \geq 5$), and 22% score at 91% ($RI \geq 8$). The averages of Figure 7 are cumulative, i.e. averages over all residues with $RI \geq n$.

Suppose a residue is predicted to be in a strand with $RI = 5$. How many of all residues predicted to be in the strand with $RI = 5$ are predicted correctly? To answser this question, an alternative non-cumulative average has to be computed: only the residues are averaged that have $RI = n$, with $n = 1, \ldots, 9$. Figure 8 shows that, e.g., the expected accuracy for a strand residue with $RI = 5$ is 76%.

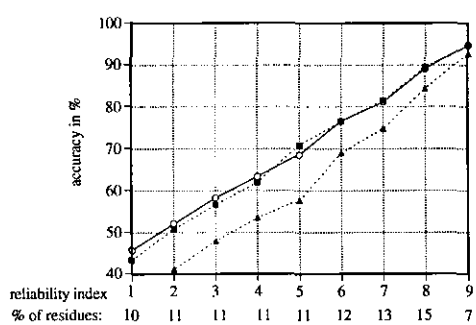

Figure 8. Expected prediction accuracy for residues with a reliability index equal to a given value. Similar to Fig. 7, except that here the non-cumulative accuracies are given, i.e. the accuracy of all residues with reliability index $RI = n$, $n = 1, \ldots, 9$. The fraction of residues that are predicted with $RI = n$ are also given. For example, 7% of all residues have $RI = 9$ and 95% of these are correctly predicted, 10% of all residues have $RI = 1$ and only 46% of these are correctly predicted. The linearity of this function and of that in Fig. 7 is surprising. (⊙ Q_{total}; (■) $Q^{\%pred}_{helix}$; (▲) $Q^{\%pred}_{strand}$.

(e) *Balanced prediction by balanced training*

The three-state accuracy Q_{total} poorly reflects the quality of predicting strand, since there are only 21% strand residues in the data banks. For most methods and for the reference network the percentage of the observed strands that were predicted correctly Q_β (eqn (1)) is below 45% (Fig. 9, note that the probability of correctly predicting strand residues at random in a 3-state prediction is 33%). The balanced training procedure provides an elegant way to correct the poor performance in predicting strand (Table 3). The final system PHD correctly predicts two-thirds of all observed strand residues, which is about ten percentage points superior to previous methods. $Q^{\%obs}_{strand}$ (eqn (1)) is about ten percentage points better than that of, e.g., ALB, and $Q^{\%pred}_{strand}$ (eqn (2)) is about six percentage points better than that of ALB. However, this gain is obtained at the expense of overprediction for strand.

The result of a more balanced prediction suggests that the low level of accuracy of previous methods in predicting strand residues, possibly, was not only caused by the experimental fact that strand formation is more dominated by long-distance interactions than, e.g., helix formation. The network technique reveals that the particular representation of the data might have been another reason for a poor performance on strand.

(f) *Substantial improvement in predicting segment lengths*

Computing per residue scores is only one way to evaluate the quality of a prediction. A detailed analysis of a few particular examples, like the cAMP-dependent protein kinase 1cpk (Benner & Gerloff, 1990; Thornton *et al.*, 1992; Rost *et al.*, 1993) or the Src-homology 3/2 domains SH3/SH2 (Barton *et al.*, 1991; Benner *et al.*, 1992; Musacchio *et al.*, 1992; Rost & Sander, 1992; Russell *et al.*, 1992) indicates that single-residue comparisons do not

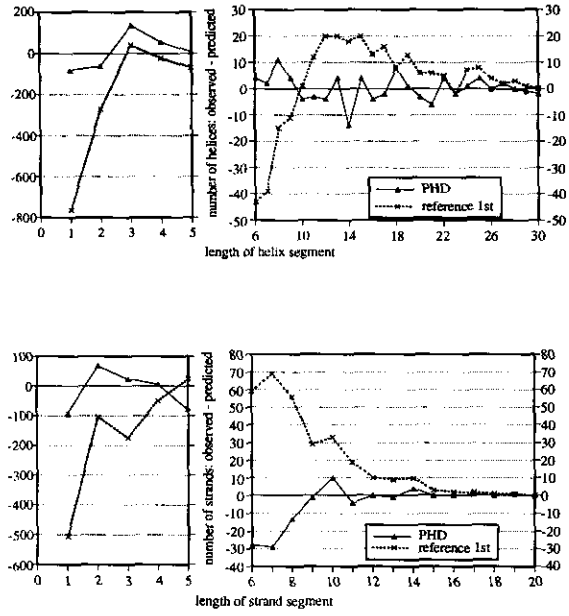

Figure 10. Comparison of length distribution of observed and predicted segments. Given are the differences between the number of helix and strand segments observed and predicted to have a certain length. The first level reference network (reference 1st) has much stronger deviations than the final network systen PHD (PHD). Helices of length 1 or 2 are artifacts (no filter was applied).

capture the possibility of using a secondary structure prediction for predicting some main features of the protein's tertiary structure. An alternative to single-residue scores is the comparison between the average length of the predicted and that of the observed segments. The reference net, e.g., predicts by far too short segments (Table 3, Fig. 10). The predictions of this network appear fragmented compared to typical globular proteins. The second level of the structure-to-structure network proves to be successful in learning the correlation between consecutive residues. The PHD prediction impressively reproduces the length distribution of secondary structure elements (Fig. 10), the average lengths (Table 3).

Here, we use two components for evaluation of the prediction quality: the per residue scores (summarized in a single number by the entropy, eqn (7)) and the length distribution of the predicted segments. However, this is still not sufficient. The ends of secondary structure elements cannot be defined uniquely. Different assignments according to the three-dimensional co-ordinates differ in the length of the assigned segments (Woodcock *et al.*, 1992). In addition, variation in segments between different crystal forms is non-negligible (Brändén & Jones, 1990). For example, a method predicting all strands correctly except that all are one residue too long, is worse in terms of average length than PHD, although the prediction is almost perfect. Unfortunately, a convincing measure for assessing how well the main secondary structure elements are predicted is missing still (Rost *et al.*, 1993). Such a measure has to reflect the potential to predict main features of the protein's tertiary structure, given the secondary structure prediction (B. Rost, C. Sander & R. Schneider, unpublished results).

(g) *Secondary structure content predicted successfully*

How accurately does PHD predict the content of secondary structure? The answer is: with less than 10% error. The error in predicting secondary structure content (eqn (19)) is: $\Omega_\alpha = 8{\cdot}5\%$ ($\sigma = 8{\cdot}3\%$) and $\Omega_\beta = 8{\cdot}1\%$ ($\sigma = 7{\cdot}8\%$). This is not as good as what was reported by using an alternative theoretical prediction by a "tandem-network" specialized on predicting secondary structure content (Muskal & Kim, 1992). The result of the elaborate "tandem" is: $\Omega_\alpha = 5{\cdot}0\%$ ($\sigma = 3{\cdot}4\%$) and $\Omega_\beta = 5{\cdot}6\%$ ($\sigma = 4{\cdot}9\%$). Unfortunately, however, that analysis did not perform multiple cross-validation. Moreover, the data set contained pairwise similarities. Thus, it is difficult to fully evaluate the result. For the best of the seven test sets used here, the performance was: $\Omega_\alpha = 5{\cdot}5\%$ ($\sigma = 4{\cdot}8\%$) and $\Omega_\beta = 7{\cdot}0\%$ ($\sigma = 6{\cdot}5\%$), comparable to the result of Muskal & Kim (1992).

How does the prediction of secondary structure content compare to experimental methods like circular dichroism (CD)? The comparison is complicated because there is no analysis of CD on a comparable data set at hand, and since the CD results we used are evaluated on the basis of distinguishing five structure types: helix, antiparallel sheet, parallel sheet, turn and loop. On a set of 15 or 16 proteins, CD reaches values of Corr$\Omega \approx 0{\cdot}95$ to 1·0 (helix), 0·4 to 0·9 (strand), 0·61 to 0·96 (loop). The ranges are due to different frequency ranges in CD in different publications (Böhm *et al.*, 1992). On all 130 chains, PHD resulted in Corr$\Omega = 0{\cdot}84$, 0·73, 0·73. On a smaller set of 26 new protein chains with recently solved structure (see next paragraph, and Fig. 11), the values were: 0·92, 0·86, 0·90. To compare equal data sets, we checked the performance on a set of 22 proteins used by Perczel *et al.* (1992). We used either the same protein, or those in our data set (Table 1) similar to that used by Perczel *et al.* (1992). CD spectroscopy yields Corr$\Omega = 0{\cdot}84$ (helix), 0·41 (antiparallel sheet), 0·37 (parallel sheet), 0·56 (loop). PHD does better with: Corr$\Omega = 0{\cdot}86$ (helix), 0·88 (sheet), 0·68 (loop). The conclusion is that the PHD prediction of secondary structure content is comparable, at least, to an intermediate level of CD analysis (not measuring the entire frequency range). Therefore, the theoretical prediction of secondary structure content is competitive with CD in some cases, in particular for strand.

What can be profited by predicting secondary structure content? Is the network accurate enough to classify the proteins correctly into structural classes like, e.g., all-helical chains? PHD correctly classified 14 of the 24 chains of Table 2 as helical. Two were falsely classified as all-helical. Does it pay off to train a network only on all-helical proteins?

The network trained on 24 all-helical protein chains performed for these about three percentage points (Q_{total}) better than a comparable network trained on all chains of Table 1. The conclusion is that if the structural class is known, specialized training on all-helical proteins increases the overall accuracy. (*Note: this result should not be confused with that* of reports published on the advantage of learning secondary structure prediction for, e.g., exclusively all-helical proteins that use a 2-state overall accuracy (Kneller *et al.*, 1990; Muggleton *et al.*, 1992; Rost & Sander, 1993). A 2-state accuracy scores generally higher than a 3-state accuracy (Maxfield & Scheraga, 1976). To investigate the benefit of restricting a prediction method to one structural subclass, two methods have to be compared on the same testing set and the same number of secondary structure states: the one extracting rules from all proteins and the one extracting the rules from, e.g., exclusively all-helical proteins.) Is this conclusion valid if the class is not given *a priori* but has to be determined by prediction with some error? PHD identified two chains falsely as helical. For these the helix net performed substantially worse than the general one. Due to these misclassifications, the resulting overall average over all helical chains (including the misclassified ones, excluding those not having been identified as helical) was about the same for the helix net as for the general one. Using a network specialized on helical chains might slightly improve the performance, but the risk is that the accuracy is substantially reduced for proteins misclassified as all-helical.

What about β-chains, or alternative structural classes? We did not further investigate this problem because PHD misclassified eight chains as all-β (according to the definition used by Kneller *et al.*, 1990), and only one-half of the observed β-chains were identified by PHD as being all-β. Circular dichroism spectroscopy is not superior to PHD in estimating the content of β-strand. Thus, it appears not to be of practical interest to investigate the performance of a profile network on β-chains before more accurate prediction methods or experimental techniques for the assessment of β-strand content are available.

(h) *No decrease in overall accuracy by filtering the prediction*

Filtering the prediction by substituting one or two-residue helices by loops does not effect the overall accuracy. It yields $Q_{total} = 70{\cdot}3\%$ (compared to 70·2% without filter). The increase in information to $I = 0{\cdot}24$ (0·238) stems from the fact that the tendency of overpredicting helices is slightly reduced. Alternative filtering procedures were tested without any significant difference. As a consequence, we apply the filtering procedure by default when the secondary structure of a new protein is predicted. Since the effect is small and has nothing to do with the principal technique of PHD, all results presented, except in this paragraph, are related to the performance without filter.

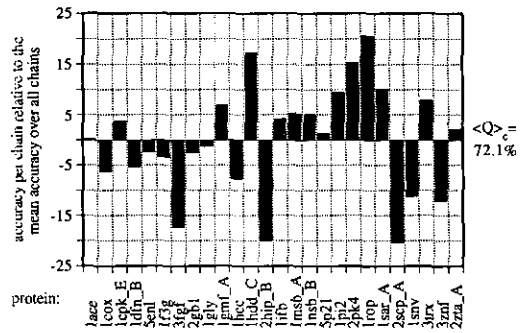

Figure 11. Prediction accuracy for 26 new protein chains. The 26 protein chains chosen from a Protein Data Bank prerelease have less than 25% (for length >80) similarity to any of the chains used for training PHD (Table 1). Given is the deviation of per chain accuracy from the mean value (eqn (4)) over all 26 chains (72·1%). The proteins are: acetylcholinesterase (1ace); cholesterol oxidase (1cox); cAMP-dependent protein kinase (1cpk); defensin (1dfn); enolase (5enl); phosphocarrier (1f3g); basic fibroblast growth factor (3fgf); protein gB1 domain (2gb1); glucoamylase (1gly); granulocyte-macrophage colony-stimulating factor (1gmf); 16th complement control protein of factor h (1hcc); engrailed homeodomain complex (1hdd); high potential iron sulphur protein (2hip); intestinal fatty acid binding protein (1ifb); mannose binding protein A (lectin domain) (1msb); neuraminidase sialidase (1nsb); c-h-ras p21 protein (5p21); Bowman Birk proteinase inhibitor pi-ii (1pi2); human plasminogen kringle (2pk4); rop: ColE1 repressor of primer (1rop); hydrolase-ribonuclease sa (1sar); sarcoplasmic calcium binding protein (2scp); Sindbis virus capsid protein (1snv); thioredoxin (3trx); zinc finger (3znf); gcn4 leucine zipper (2zta).

(i) *Marginal influence of free parameters and potential improvements*

The large number of tests we performed leads to a certain experience in how to choose the free parameters of the network (window size w, number of bits used to code real values, number of layers and hidden units, criterion for stopping the optimization procedure, dynamical constants ε, η, interval and distribution of the junctions used as the starting point, slope of the sigmoid trigger function f, and definition of the error E). How does the choice of these parameters influence the performance of the network? We were surprised by the simple answer: not much. Therefore, we recommend spending less time on optimizing such parameters than on attempts to reformulate the problem. An optimal choice of the free parameters might increase the result by of the order of one percentage point. In comparison, the usage of multiple sequence alignments instead of single sequences is a straightforward change by which more than six percentage points are gained.

4. Discussion

The overall three-state accuracy above 70% by using multiple sequence alignments as input to multi-layered networks is rather convincing. It is reflected in an average overall accuracy of at least four percentage points, and a strand accuracy ten percentage points better than that of any previous method. Moreover, the network reproduces well the length of the observed secondary structure elements, although these variables were never explicitly used for training the system. The difference between the observed and the predicted length distribution is a kind of consistency check for how well the network learned to extract the relevant features of the problem. A further consistency check was the use of the reliability index, which promises to be rather useful in practice: residues predicted more reliably than others can be identified. The fifth of all residues with highest reliability is predicted with an accuracy >90%. Four chains of the membrane protein photo-reaction centre (1prc) were included in testing and training. The prediction for these four is about 12 percentage points below the average performance. But since the physical environment for such proteins is completely different from that of water-soluble ones, it is surprising that the prediction yields a reasonable result at all. Whether this result will persist for future membrane proteins is not clear.

For the data set of 130 protein chains, the magic barrier of 70% overall three-state accuracy has been broken. But will the method score as high for the next 130 proteins? After having performed all analyses, we investigated the performance on 26 new proteins with a recently solved structure (Fig. 11). They were selected so that none had significant sequence similarity to any protein in the training set used. For these 26 chains the results were: $Q_{\text{total}} = 71{\cdot}5\%$, with $Q_\alpha = 71\%$, $Q_\beta = 64\%$ (and $Q_\alpha^{\%\text{pred}} = 74\%$, $Q_\beta^{\%\text{pred}} = 59\%$). The per chain average was $\langle Q \rangle_c = 72{\cdot}1\%$ (Fig. 11). This success indicates that the quality of the network system described here is probably not overestimated. What about the next 100 chains? Further tests will show. Publically particularly effective are blind tests, i.e. predictions of proteins with yet unknown structure (Benner & Gerloff, 1990; Benner *et al.*, 1992; Rost & Sander, 1992; Russell *et al.*, 1992). A better check will be the long-term use of PHD as an everyday prediction tool. The PHD method is available for fully automatic use. Send the word *help* by electronic mail to the internet address *PredictProtein @Embl-Heidelberg.de* for detailed instructions on how to automatically obtain a predicted secondary structure for your sequence.

We emphasize our gratitude to three colleagues from our group who contributed substantial ideas and help: Reinhard Schneider, Michael Scharf and Gerrit Vriend. Further thanks to Christos Ouzonis (EMBL, Heidelberg), Sara Solla (AT&T, Holmdel), Françoise Fogelman-Soulie (Mimetics, Chatenay-Malabry), Pierre Nadal (ENS, Paris), Søren Brunak (Techn. Univ., Copenhagen) and Andreas Herz (Caltech, Pasadena).

References

Amit, D. J. (1989). *Modeling Brain Function: The World of Attractor Networks*. Cambridge University Press, Cambridge.

Anfinsen, C. B., Epstein, C. J. & Goldberger, R. F. (1963). The genetic control of tertiary protein structure: studies with model systems. *Cold Spring Harbor Symp. Quant. Biol.* **28**, 439–449.

Bairoch, A. & Boeckmann, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **20**, 2019–2022.

Barton, G. J., Newman, R. H., Freemont, P. S. & Crumpton, M. J. (1991). Amino acid sequence analysis of the annexin super-gene family of proteins. *Eur. J. Biochem.* **198**, 749–760.

Benner, S. A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enz. Reg.* **31**, 121–181.

Benner, S. A., Cohen, M. A. & Gerloff, D. (1992). Correct structure prediction? *Nature (London)*, **359**, 781.

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

Biou, V., Gibrat, J. F., Levin, J. M., Robson, B. & Garnier, J. (1988). Secondary structure prediction: combination of three different methods. *Protein Eng.* **2**, 185–191.

Böhm, G., Muhr, R. & Jaenicke, R. (1992). Quantitative analysis of protein far UV circular dichroism spectra by neural networks. *Protein Eng.* **5**, 191–195.

Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M. J., Lautrup, B., Nørskov, L., Olsen, O. H. & Petersen, S. B. (1988). Protein secondary structure and homology by neural networks. *FEBS Letters*, **241**, 223–228.

Bohr, H., Bohr, J., Brunak, S., Fredholm, H., Lautrup, B. & Petersen, S. B. (1990). A novel approach to prediction of the 3-dimensional structures of protein backbones by neural networks. *FEBS Letters*, **261**, 43–46.

Bork, P., Ouzonis, C., Sander, C., Scharf, M., Schneider, R. & Sonnhammer, E. (1992). What's in a genome? *Nature (London)*, **358**, 287.

Bossa, F. & Pascarella, S. (1990). PRONET: a microcomputer program for predicting the secondary structure of proteins with a neural network. *CABIOS*, **5**, 319–320.

Brändén, C.-I. & Jones, T. A. (1990). Between objectivity and subjectivity. *Nature (London)*, **343**, 687–689.

Chou, P. Y. & Fasman, U. D. (1974). Prediction of protein conformation. *Biochemistry*, **13**, 211–215.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1983). Secondary structure assignment for α/β proteins by a combinatorial approach. *Biochemistry*, **22**, 4894–4904.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D. & Fletterick, R. J. (1986). Turn prediction in proteins using a pattern-matching approach. *Biochemistry*, **25**, 266–275.

Cowan, J. D. (1990). Neural networks: the early days. In *Neural Information Processing Systems 2* (Touretzky, D. S., ed.), pp. 828–842, Morgan Kaufmann, San Mateo, CA.

Dayhoff, M. O. (1978). Editor of *Atlas of Protein Sequence*

and Structure. National Biomedical Research Foundation, Washington, DC.

Ewbank, J. J. & Creighton, T. E. (1992). Protein folding by stages. *Curr. Opin. Struct. Biol.* **2**, 347–349.

Fasman, G. F. (1989). Protein conformation prediction. In *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., ed.), pp. 193–316, Plenum, New York and London.

Frampton, J., Leutz, A., Gibson, T. J. & Graf, T. (1989). DNA-binding domain ancestry. *Nature (London)*, **342**, 134.

Garnier, J. (1993). Prediction of protein structure. In *Biological Sequences: Finding Structure and Function by Neural Networks* (Brunak, S., ed.), Institute for Scientific Interchange Foundation, Torino, Italy, in the press.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97–120.

Garrett, R. C., Thornton, J. M. & Taylor, W. R. (1991). An extension of secondary structure prediction towards the prediction of the tertiary structure. *FEBS Letters*, **280**, 141–146.

Gibrat, J.-F., Garnier, J. & Robson, B. (1987). Further developments of protein secondary structure prediction using information theory. New Parameters and consideration of residue pairs. *J. Mol. Biol.* **198**, 425–443.

Gibson, T. J., Thompson, J. D. & Abagyan, R. A. (1993). Proposed structure for the DNA-binding domain of the helix-loop-helix family of eukaryotic gene regulatory proteins. *Protein Eng.* **6**, 41–50.

Hansen, L. K. & Salamon, P. (1990). Neural network ensembles. *IEEE Trans. Pattern Anal. Machine Intel.* **12**, 993–1001.

Hertz, J. A., Krogh, A. & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, CA.

Hirst, J. D. & Sternberg, M. J. E. (1992). Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, **31**, 615–623.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Holley, H. L. & Karplus, M. (1989). Protein secondary structure prediction with a neural network. *Proc. Nat. Acad. Sci., U.S.A.* **86**, 152–156.

Hubbard, T. J. P. & Sander, C. (1991). The role of heat-shock and chaperone proteins in protein folding: possible molecular mechanisms. *Protein Eng.* **4**, 711–717.

Johnson, C. W. J. (1990). Protein secondary structure and circular dichroism: a practical guide. *Proteins: Struct. Funct. Genet.* **7**, 205–214.

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

Kanehisa, M. (1988). A multivariate analysis method for discriminating protein secondary structural segments. *Protein Eng.* **2**, 87–92.

Karplus, M. & Petsko, G. A. (1990). Molecular dynamics simulations in biology. *Nature (London)*, **347**, 631–639.

Kneller, D. G., Cohen, F. E. & Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**, 171–182.

Levin, J. M. & Garnier, J. (1988). 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. Acta*, **955**, 283–295.

Levin, J. M., Robson, B. & Garnier, J. (1986). An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Letters*, **205**, 303–308.

Levitt, M. & Chothia, C. (1976). Structural patterns in globular proteins. *Nature (London)*, **261**, 552–558.

Lim, V. I. (1974). Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. *J. Mol. Biol.* **88**, 857–872.

Lincoln, W. P. & Skrzypek, J. (1990). In *Neural Information Processing Systems 2* (Touretzky, D. S., ed.), pp. 650–657, Morgan Kaufmann, San Mateo, CA.

Maclin, R. & Shavlik, J. W. (1993). Using knowledge-based neural networks to improve algorithms: refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, in the press.

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Maxfield, F. R. & Scheraga, H. A. (1976). Status of empirical methods for the prediction of protein backbone topography. *Biochemistry*, **15**, 5138–5153.

Maxfield, F. R. & Scheraga, H. A. (1979). Improvements in the prediction of protein topography by reduction of statistical errors. *Biochemistry*, **18**, 697–704.

Minsky, M. & Papert, S. (1988). *Perceptrons*. MIT Press, Cambridge, MA.

Momany, F. A., McGuire, R. F., Burgess, A. W. & Scheraga, H. A. (1975). Energy parameters in polypeptides. *J. Phys. Chem.* **79**, 2361–2381.

Muggleton, S., King, R. D. & Sternberg, M. J. E. (1992). Protein secondary structure prediction using logic-based machine learning. *Protein Eng.* **5**, 647–657.

Müller, B. & Reinhardt, J. (1990). *Neural Networks*. Springer, Berlin.

Musacchio, A., Gibson, T., Lehto, V.-P. & Saraste, M. (1992). SH3 – an abundant protein domain in search of a function. *FEBS Letters*, **307**, 55–61.

Muskal, S. M. & Kim, S.-H. (1992). Predicting protein secondary structure content. A tandem neural network approach. *J. Mol. Biol.* **225**, 713–727.

Nagano, K. (1973). Logical analysis of the mechanism of protein folding. *J. Mol. Biol.* **75**, 401–420.

Nagano, K. & Hasegawa, K. (1975). Logical analysis of the mechanism of protein folding. *J. Mol. Biol.* **94**, 257–281.

Niermann, T. & Kirschner, K. (1991). Improving the prediction of secondary structure of "TIM-barrel" enzymes (Corrigendum). *Protein Eng.* **4**, 359–370.

Nishikawa, K. & Noguchi, T. (1991). Predicting protein secondary structure based on amino acid sequence. *Methods Enzymol.* **202**, 31–44.

Ouzounis, C. A. & Melvin, W. T. (1991). Primary and secondary structural patterns in eukaryotic cytochrome P-450 families correspond to structures of the helix-rich domain of *Pseudomonas putida* cytochrome P-450cam. *Eur. J. Biochem.* **198**, 307–315.

Pain, R. H. & Robson, B. (1970). Analysis of the code relating sequence to secondary structure in proteins. *Nature (London)*, **227**, 62–63.

Pauling, L. & Corey, R. B. (1951). Configurations of polypeptide chains with favored orientations around

single bonds: two new pleated sheets. *Proc. Nat. Acad. Sci., U.S.A.* **37**, 729–740.

Pauling, L., Corey, R. B. & Branson, H. R. (1951). The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Nat. Acad. Sci., U.S.A.* **37**, 205.

Perczel, A., Hollósi, M., Tusnády, G. & Fasman, G. D. (1991). Convex constraint analysis: a natural deconvolution of circular dichroism curves of proteins. *Protein Eng.* **4**, 669–679.

Perczel, A., Park, K. & Fasman, G. D. (1992). Deconvolution of the circular dichroism spectra of proteins: the circular dichroism spectra of the antiparallel β-sheet in proteins. *Proteins: Struct. Funct. Genet.* **13**, 57–69.

Periti, P. F., Quagliarotti, G. & Liquori, A. M. (1967). Recognition of α-helical segments in proteins of known primary structure. *J. Mol. Biol.* **24**, 313–322.

Presnell, S. R., Cohen, B. I. & Cohen, F. E. (1992). A segment-based approach to protein secondary structure prediction. *Biochemistry*, **31**, 983–993.

Press, W. H., Flamery, B. P., Tenholsky, S. A. & Vetterling, W. T. (1986). *Numerical Recipes: The Art of Scientific Computing*, University Press, Cambridge, MA.

Ptitsyn, O. B. (1969). Statistical analysis of the distribution of amino acid residues among helical and non-helical regions in globular proteins. *J. Mol. Biol.* **42**, 501–510.

Ptitsyn, O. B. & Finkelstein, A. V. (1983). Theory of protein secondary structure and algorithm of its prediction. *Biopolymers*, **22**, 15–25.

Qian, N. & Sejnowski, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**, 865–884.

Robson, B. & Pain, R. H. (1971). Analysis of the code relating sequence to conformation in proteins: possible implications for the mechanism of formation of helical regions. *J. Mol. Biol.* **58**, 237–259.

Rooman, M. J. & Wodak, S. (1991). Weak correlation between predictive power of individual sequence patterns and overall prediction accuracy in proteins. *Proteins: Struct. Funct. Genet.* **9**, 69–78.

Rooman, M. J., Wodak, S. & Thornton, J. M. (1989). Amino acid sequence templates derived from recurrent turn motifs in proteins: critical evaluation of their predictive power. *Protein Eng.* **3**, 23–27.

Rooman, M. J., Kocher, J. P. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments: influence of local interactions. *J. Mol. Biol.* **221**, 961–979.

Rost, B. & Sander, C. (1992). Jury returns on structure prediction. *Nature (London)*, **360**, 540.

Rost, B. & Sander, C. (1993). Secondary structure prediction of all-helical proteins in two states. *Protein Eng.* in the press.

Rost, B., Sander, C. & Schneider, R. (1993). Progress in protein structure prediction? *Trends Biochem. Sci.* **18**, 120–123.

Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition.* MIT Press, Cambridge, MA.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating error. *Nature (London)*, **323**, 533–536.

Russell, R. B., Breed, J. & Barton, G. J. (1992). Conservation analysis and structure prediction of the SH2 family of phosphotyrosine binding domains. *FEBS Letters*, **304**, 15–20.

Salzberg, S. & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.* **227**, 371–374.

Sander, C. & Schneider, R. (1991). Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Schulz, G. E. & Schirmer, R. H. (1979). *Principles of Protein Structure*, Springer, New York.

Solla, S. A. (1991). A theory of supervised learning. In *Neural Networks from Biology to High Energy Physics* (Benhar, O. *et al.*, eds), pp. 11–28, Elba, Italy.

Sternberg, M. J. E. & King, R. D. (1990). Machine learning approach for the prediction of protein secondary structure. *J. Mol. Biol.* **216**, 441–457.

Stolorz, P., Lapedes, A. & Xia, Y. (1992). Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* **225**, 363–377.

Szent-Györgyi, A. G. & Cohen, C. (1957). Role of proline in polypeptide chain configuration of proteins. *Science*, **126**, 697.

Taylor, W. R. (1988). Pattern matching methods in protein sequence comparison and structure prediction. *Protein Eng.* **2**, 77–86.

Taylor, W. R. & Orengo, C. A. (1989). A holistic approach to protein structure alignment. *Protein Eng.* **2**, 505–519.

Taylor, W. R. & Thornton, J. M. (1983). Prediction of super-secondary structure in proteins. *Nature (London)*, **301**, 540–542.

Thornton, J. M., Flores, T. P., Jones, D. T. & Swindells, M. B. (1992). Prediction of progress at last. *Nature (London)*, **354**, 105–106.

Vriend, G. & Sander, C. (1991). Detection of common three-dimensional substructures in proteins. *Proteins: Struct. Funct. Genet.* **11**, 52–58.

Woodcock, S., Mornon, J.-P. & Henrissat, B. (1992). Detection of secondary structure elements in proteins by hydrophobic cluster analysis. *Protein Eng.* **5**, 629–635.

Zhang, X., Mesirov, J. P. & Waltz, D. L. (1992). Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225**, 1049–1063.

Zhong, L., Johnson, W. & Curtis, J. (1992). Environment affects amino acid preference for secondary structure. *Proc. Nat. Acad. Sci., U.S.A.* **89**, 4462–4465.

Zvelebil, M. J., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using alignment of homologous sequences. *J. Mol. Biol.* **195**, 957–961.

Edited by A. R. Fersht