Volume 241, number 1,2, 223–228 FEB 06576 December 1988

# Protein secondary structure and homology by neural networks

## The α-helices in rhodopsin

Henrik Bohr, Jakob Bohr[+], Søren Brunak, Rodney M.J. Cotterill, Benny Lautrup[°], Leif Nørskov[†], Ole H. Olsen[†] and Steffen B. Petersen[†]

*The Technical University of Denmark, Building 307, DK-2800 Lyngby, [+]Risø National Laboratory, DK-4000 Roskilde, [°]Niels Bohr Institute, Blegdamsvej 17, DK-2100 København Ø* and *[†]NOVO Research Institute, Novo Allé, DK-2880 Bagsværd, Denmark*

Received 12 September 1988; revised version received 19 October 1988

Neural networks provide a basis for semiempirical studies of pattern matching between the primary and secondary structures of proteins. Networks of the perceptron class have been trained to classify the amino-acid residues into two categories for each of three types of secondary feature: α-helix or not, β-sheet or not, and random coil or not. The explicit prediction for the helices in rhodopsin is compared with both electron microscopy results and those of the Chou-Fasman method. A new measure of homology between proteins is provided by the network approach, which thereby leads to quantification of the differences between the primary structures of proteins.

Secondary structure; Neural network; Protein folding; Rhodopsin α-helix; Protein homology; Perceptron

## 1. INTRODUCTION

One of the current aims of molecular biology is the prediction of the folded (tertiary) structure of a protein from the knowledge of the sequence of its constituent amino acid residues. Such an ability would promote the possibility of tailoring artificially synthesized proteins to fit a particular enzymatic, signalling, structural or other function, and the recent successful hybridization of elements taken from different protein molecules indicates that such de novo protein synthesis may not be far off [1,2]. A step on the road to this goal would be the unequivocal prediction of the secondary structure of proteins. The traditional approach to this aspect of the folding problem has involved consideration of the various inter-residue forces and also the interactions of the residues with their environment, be that aqueous, lipid, or a combination of both [3–6]. Augmenting this ab initio method, one has the statistical approach [7–12] typified by the efforts of Chou and Fasman [13], which is capable of correctly predicting the secondary structure of up to about 50% of a protein's residues [14].

We present here some new results of a study which exploits a quite different approach [15] to secondary-structure prediction, that uses a neural network of the perceptron class [16–18]. Although this method lacks the physical transparency of the traditional line of attack, its frank pragmatism is surprisingly successful. Its ability to predict α-helix structure appears to be superior to any other method. In order to illustrate the potential of this method, we have predicted the second order structure of the α-helix-rich protein, rhodopsin. The method has even revealed new features of protein structure, and in addition it introduces a novel measure of homology.

## 2. MATERIALS AND METHODS

We have used a feed-forward neural network, see fig.1, with

*Correspondence address:* H. Bohr, The Technical University of Denmark, Building 307, DK-2800 Lyngby, Denmark

*Published by Elsevier Science Publishers B.V. (Biomedical Division)*
00145793/88/$3.50 

one layer of hidden units, and have adjusted its inter-layer connections according to the back-propagation algorithm originally prescribed by Werbos [19], by Parker [20], and more recently by Rummelhart et al. [18]. An artificial neural network is essentially a trainable device, which produces an output as a response to an input. Such devices have a superficial resemblance to the architecture of biological neural systems, and in the terminology of artificial neural networks it is therefore normal practice to borrow rather freely from the vocabulary of the science of neurobiology.

Apart from the units of the input layer, which perform no processing, each unit in the network – each neuron – computes a weighted linear sum of its real-valued inputs, $I_n$, $n = 1,2,\ldots,N$. Afterwards the neuron passes the weighted sum through a non-linear S-shaped (sigmoidal) function, $\sigma$, to produce the final real-valued output

$$O = \sigma\left(\sum_{n=1}^{N} w_n I_n - t\right), \quad (1)$$

where the synaptic weights belonging to the neuron have been denoted $w_n$ and the parameter $t$ plays the role of a threshold for the non-linearity. At the threshold the sigmoidal function makes a rapid transition from values near zero to values near one. The precise form of the sigmoidal function is not important. The feed-forward network is then constructed by connecting the output of one layer of neurons to the input of the next as seen in fig.1.

The purpose of the training of a neural network is to establish an association between definite inputs and desired outputs. The network is trained by exposing it to correlated pairs of inputs and outputs – the training examples – and iteratively updating it by adjusting its synaptic connections, until these associations have been correctly learned. The back-propagation algorithm is a steepest-descent method for readjusting the synaptic weights according to the output errors made by the network. An easily comprehensible technical exposition of the method can be found in [21].

We trained the network, by presenting it with the experimentally determined mapping between the sequence of amino acids and the secondary structure, for a number of proteins whose structures (in the crystalline form) have been determined by X-ray diffraction. We then monitored the performance of the network when it was presented with amino acid sequences of test proteins for which the corresponding secondary structures were also known from experiment. In all cases, the primary structure was presented to the network's input layer, and the secondary structure was either used as a target for the output layer (during the training period) or it simply emerged at the output layer (during the subsequent testing of performance). The network thus bore a superficial resemblance to Sejnowski and Rosenberg's remarkably successful NETtalk [22], with our

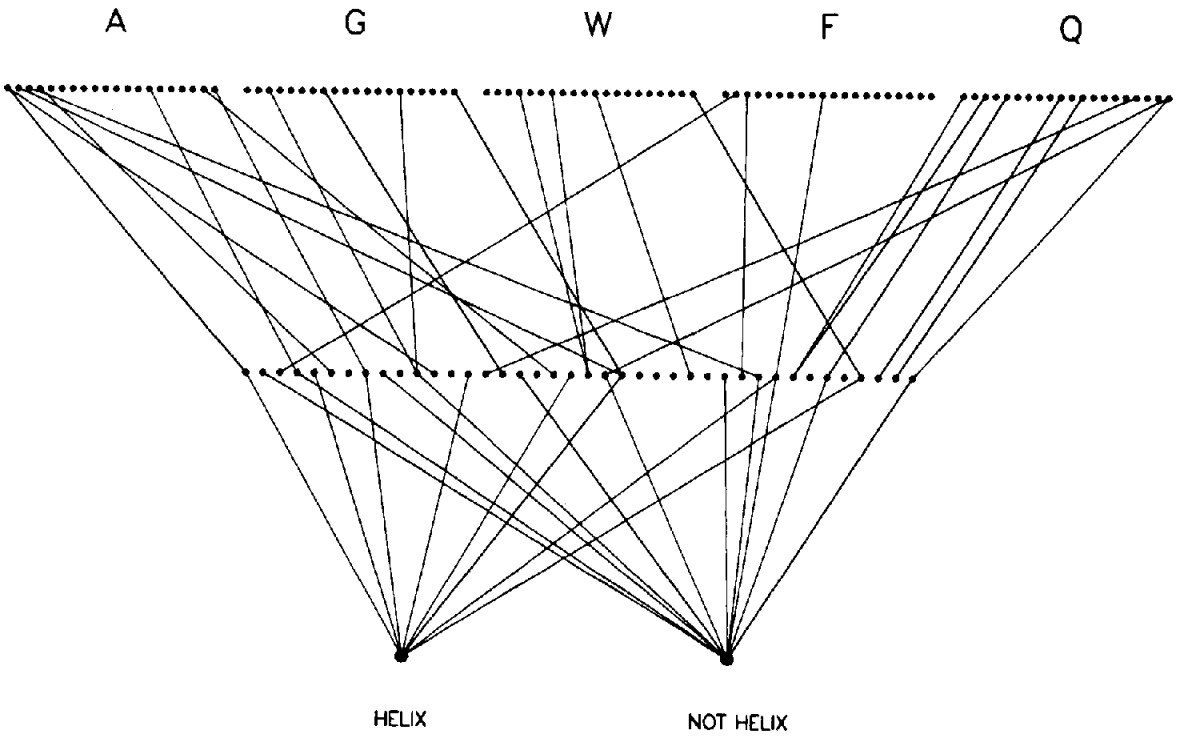


Fig.1. Architecture of a two-layer perceptron. Only a small fraction of the connections between neurons are shown. Information flows only forward (i.e. downward in the figure) from input towards output. In this drawing 5 amino acid residues are 'visible' to the network, which has the task of classifying the middle residue according to whether it participates in an $\alpha$ helix or not.

strings of amino acid residues replacing their strings of alphabetical letters, and with our secondary structure categories replacing their phonetic categories. Recently, Qian and Sejnowski [15] have reported a similar approach to the prediction of the secondary structure of proteins. They have shown that the neural-network method gives better values for the correlation coefficients [23] than those of previous methods. However, their overall ability to predict an assignment of a secondary structure barely exceeds 60% [15].

In a folded protein, a given residue does not by itself correspond to one specific class of secondary structure (e.g. $\alpha$-helix, $\beta$-sheet, $\beta$-turn, or random coil). The residue-to-secondary-unit mapping is, in a non-trivial manner, determined by a residue's context, just as the pronunciation of a vowel, say, depends upon the types of other letters which surround it. Our input layer consequently consisted of a 'window' comprising a number of successive residues. The size of the window was chosen to be 51 units, thus providing 25 residues as context on each side of the central residue. Various window sizes between 7 and 91 units were initially tried. The smaller windows did not permit the network to learn all the examples, whereas the larger windows did not lead to any appreciable improvement in the predictive power of the network. A window size of 51 units may also be justified by noting that $\beta$-sheets include hydrogen-bond

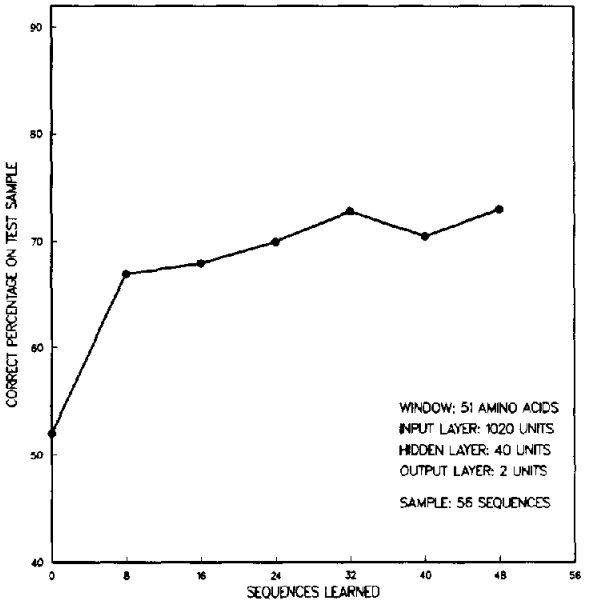


Fig.2. Learning and test curve. Percentage of correctly predicted secondary structure elements versus size of training sample. The network was trained on successively larger portions of the sample, consisting of 56 proteins, selected from the Brookhaven Protein Data Bank. The testing was done on the remaining part of the 56 proteins. The selection criterion was that the resolution of the X-ray coordinates should be better than 2 Å. The window size was 51 amino acids and the input layer of the network thus contained 1020 neurons. The hidden layer and the output layer contained 40 and 2 neurons, respectively.

formation between amino acid residues separated in the sequence by an amount which is typically of this order of magnitude. In special cases, such as structures containing the $\beta$-barrel motif, longer range interactions are of course involved. The neural network method for secondary-structure prediction is only based on local information and is thus unable to detect correlations spanning over larger amino acid sequences than can be seen by the network in the window.

Our choice of representation maximized the orthogonality between both the various input items and the output [22]. Because there are twenty different types of amino acid residues, each of these was therefore represented by nineteen zeros and a single one. Glycine for example was coded by 00000010000000000000, alanine by 00000000000000001000, and so on. A window of seven residues therefore required 140 input cells, and because there were 40 hidden units, the total number of adjustable synapses and thresholds in the network was close to 6000, while one with a 51-residue window contained about 40000. The advantage of such a sparse representation is that biochemically unjustified algebraic relationships between the coded forms of the amino acid residues are avoided. Our output layer, on the other hand, consisted of only two cells. (The state of a formal neuron is a non-negative real number between 0 and 1, which is called its activity, and may be interpreted as a measure of the confidence level for the presence of a feature, detected by this neuron.) The sum of the non-negative activities in the two output cells is always unity, and a dominant activity in the one cell denotes recognition of the feature in question, while a dominant activity in the other indicates that the feature is not present. We have used separate networks for each of three secondary-structure features (i.e. $\alpha$-helix, $\beta$-sheet, and random coil, in that we have not attempted to differentiate between coil and turn in the present study). In this respect, our approach differs from that used by Qian and Sejnowski [15].

To train the network, we have chosen 56 proteins (which comprised a total of about ten thousand residues) from the Brookhaven Protein Data Bank [24] by the criterion that the structural coordinates be known to better than 2 Å. The training was performed on a subset consisting of $n$ of these proteins (the training set) which was extended sequentially to ultimately consist of all 56 proteins. During this process the remaining of the $56 - n$ proteins (the test set) were used to test the performance of the network to predict the specified category of secondary structure. Fig.2 shows the network's ability (expressed as a percentage) to correctly predict the assignments of the $\alpha$-helix structure to amino acids in the test set; after completion of the training, the score approaches 73%. The correlation coefficient [23] $C_\alpha$ was 0.38. Window sizes smaller than 51 amino acids gave smaller correlation coefficients. For larger window sizes, the value of $C_\alpha$ did not exceed its saturation level of 0.38.

## 3. RESULTS AND DISCUSSION

Several interesting things have come to light during these studies. We have examined the membrane-bound rhodopsin protein in the light-harvesting membrane of *Halobacterium halobium*, which was the subject of the classic electron-microscopical structure determination carried out

by Henderson and Unwin [25]. Their results indicated the presence of a trimer configuration, each monomer apparently consisting of seven $\alpha$-helices (as well as the associated turn and coil fragments). Examination of the electron density profiles of their derived structure, however, shows that one of these helices is not quite so well defined; its electron density is below that of the other six. The primary structure of this protein was presented to our trained network, and it responded by predicting seven $\alpha$-helices, although one of these is indeed less well defined than the other six; see fig.3. The figure shows the $\alpha$-helix activity for each of the 249 amino acids. An activity approaching unity means that an $\alpha$-helix structure is predicted with maximal confidence, whereas an activity close to zero indicates the absence of this structure. On the basis of this activity diagram, we have tentatively assigned $\alpha$-helix structures to the amino acid stretches (0–21), (32–62), (73–87), (91–122), (127–160), (166–178) and (182–217), these assignments being shown as horizontal bars. Also shown in fig.3 is the prediction by Argos et al. [26] using the Chou and Fasman method. Clearly, the ability of our method to predict the $\alpha$-helix is superior in its discriminatory power. Our results endorse the original interpretation of the electron microscopical work [25] and they do not corroborate the more recent interpretation of investigations of this protein, using ultraviolet circular dichroism spectroscopy [27], which found evidence of only five helices.

This study has also revealed an interesting property of a set of homologous proteins. If one trains the network on one member of a pair of such proteins and then tests its predictive performance

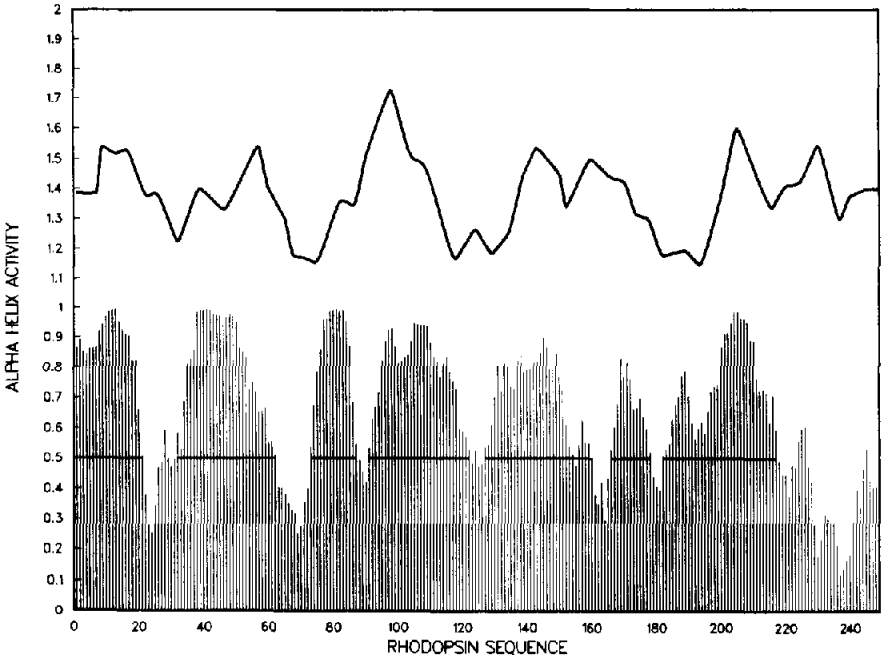


Fig.3. Each of the 249 vertical lines corresponds to the activity of the $\alpha$-helix signaling neuron for a particular amino acid residue in the primary structure of rhodopsin. The activities shown were averaged over nearest neighbours and next-nearest neighbours. In extended regions where the activity exceeds 0.5 a horizontal bar is drawn, giving the positions of the helices, which are (0–21), (32–62), (73–87), (91–122), (127–160), (166–178) and (182–217). Above the vertical lines are shown the confidence levels for the Chou-Fasman prediction adapted from reference [25]. (The $\alpha$-helix signaling confidence level for this curve is 0 at both the beginning and the end.) With the one-letter coding for amino acids, the sequence of rhodopsin, including the 13 amino acid rhodopsin precursor, is MLELLPTAVEGVSQAQITGRPEWIWLALGTALMGLGTLYFLVKGMGVSDPDAKKFYAITTLVPAIAFTMYLSMLLGYGLT - MVPFGGEQNPIYWARYADWLFTTPLLLLDLALLVDADQGTILALVGADGIMIGTGLVGALTKVYSYRFVWWAISTAAMLY- ILYVLFFGFTSKAESMRPEVASTFKVLRNVTVVLWSAYPVVWLIGSEGAGIVPLNIETLLFMVLDVSAKVGFGLILLRSRAIFG- EAEAPEPSAGDGAAATSD. The sequence proper starts with QAQITGR.

on the other member of the pair, one gets a measure of the degree to which the primary-to-secondary mapping in the first protein resembles the corresponding mapping in the second protein. Furthermore the resulting overlap, $q_{12}$, again measured as a percentage of correctly-forecast secondary structures, is close to (or even identical to) the result obtained if the training molecule and the testing molecule are interchanged. In order to have a general measure of homology rather than one biased towards a specific category of secondary structure, the homology studies have been carried out with a network with three output cells, one for $\alpha$-helices, one for $\beta$-sheets and one for random coils.

Fig.4 presents the results of this type of analysis for a total of seven different proteins; the symmetry is apparently a general one, irrespective of which homologous pair is used. Since the triangular inequality, $(100 - q_{12}) + (100 - q_{23}) \geq (100 - q_{13})$, is also obeyed, this measure seems to possess the basic properties of a metric; the quantity $(100 - q_{ij})$ can be looked upon as measuring a distance reflecting the departure from perfect homology between proteins $i$ and $j$. This way of looking at the issue could prove to be fruitful as a general approach to the question of homology,

| | Window: 15 Amino Acids | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2KAI | 1TGN | 3PTP | 1SGC | 3SGB | 1EST | 2GCH |
| 2KAI | 100 | 61 | 63 | 35 | 55 | 53 | 49 |
| 1TGN | 59 | 100 | 99 | 43 | 52 | 53 | 52 |
| 3PTP | 58 | 100 | 100 | 44 | 52 | 53 | 54 |
| 1SGC | 43 | 36 | 37 | 100 | 70 | 50 | 41 |
| 3SGB | 51 | 49 | 48 | 65 | 100 | 49 | 54 |
| 1EST | 50 | 53 | 55 | 46 | 44 | 100 | 48 |
| 2GCH | 50 | 51 | 52 | 45 | 52 | 55 | 100 |

| | Window: 51 Amino Acids | | | | | | |
|---|---|---|---|---|---|---|---|
| 2KAI | 100 | 69 | 71 | 44 | 45 | 57 | 59 |
| 1TGN | 75 | 100 | 100 | 43 | 50 | 59 | 64 |
| 3PTP | 74 | 100 | 100 | 42 | 49 | 60 | 64 |
| 1SGC | 33 | 38 | 38 | 100 | 70 | 55 | 48 |
| 3SGB | 45 | 49 | 48 | 72 | 100 | 51 | 56 |
| 1EST | 55 | 49 | 49 | 48 | 54 | 100 | 62 |
| 2GCH | 50 | 58 | 59 | 46 | 58 | 66 | 100 |

Fig.4. Matrices showing the primary-to-secondary overlap in per cent between seven proteinases. The random score is 33% and signals no homology. The symmetry increases with increased window size up to 51 amino acid residues. The data bank abbreviations for the seven proteinases stand for: 2KAI = kallikrein A, 1TGN = trypsinogen, 3PTP = $\beta$-trypsin, 1SGC = proteinase A, 3SGB = proteinase B, 1EST = elastase, 2GCH = $\gamma$-chymotrypsin A.

| | Window: 15 Amino Acids | | | | | | |
|---|---|---|---|---|---|---|---|
| | 2KAI | 1TGN | 3PTP | 1SGC | 3SGB | 1EST | 2GCH |
| 2KAI | 98 | 10 | 10 | 10 | 8 | 9 | 9 |
| 1TGN | 11 | 95 | 95 | 12 | 9 | 11 | 16 |
| 3PTP | 13 | 94 | 96 | 12 | 12 | 10 | 11 |
| 1SGC | 6 | 9 | 9 | 93 | 33 | 8 | 11 |
| 3SGB | 9 | 8 | 8 | 37 | 94 | 10 | 14 |
| 1EST | 9 | 10 | 11 | 9 | 8 | 96 | 14 |
| 2GCH | 7 | 13 | 10 | 10 | 12 | 13 | 96 |

| | Window: 51 Amino Acids | | | | | | |
|---|---|---|---|---|---|---|---|
| 2KAI | 100 | 21 | 21 | 8 | 10 | 9 | 13 |
| 1TGN | 20 | 100 | 100 | 8 | 8 | 15 | 22 |
| 3PTP | 20 | 97 | 100 | 10 | 8 | 15 | 21 |
| 1SGC | 8 | 10 | 10 | 100 | 48 | 10 | 11 |
| 3SGB | 9 | 7 | 7 | 49 | 100 | 9 | 9 |
| 1EST | 11 | 17 | 17 | 13 | 9 | 100 | 18 |
| 2GCH | 13 | 19 | 20 | 12 | 6 | 16 | 100 |

Fig.5. As in fig.4, but with primary-to-primary overlaps. The random score is in this case 5%. There is no a priori connection between the primary-to-secondary and primary-to-primary overlaps, but the data do indicate the existence of such a connection.

embracing also the traditional measures of that quantity. It is a tantalizing question whether or not a dimension of the embedding space for these proteins can be defined, and if so what it is.

In an extension of this study, we have taught the network to determine a particular amino acid residue in a sequence on the basis of the surrounding residues in the chain. This produces a network with a 'spell-checking' skill with respect to a given protein. Again by measuring the predictive ability of this approach on other proteins, homology solely based on the primary structure of proteins can also be quantified; see fig.5.

Both of these homology measuring methods are linear in their computational cost (once the network has learned the reference sequence). This

| | 2KAI | 1TGN | 3PTP | 1SGC | 3SGB | 1EST | 2GCH |
|---|---|---|---|---|---|---|---|
| 2KAI | 100 | 43 | 43 | 23 | 24 | 34 | 38 |
| 1TGN | 43 | 100 | 100 | 28 | 25 | 36 | 43 |
| 3PTP | 43 | 100 | 100 | 27 | 26 | 41 | 44 |
| 1SGC | 23 | 28 | 27 | 100 | 65 | 27 | 28 |
| 3SGB | 24 | 25 | 26 | 65 | 100 | 25 | 28 |
| 1EST | 34 | 36 | 41 | 27 | 25 | 100 | 41 |
| 2GCH | 38 | 43 | 44 | 28 | 28 | 41 | 100 |

Fig.6. The primary-to-primary homology as determined by conventional (PIR) techniques. Although these percentages differ in magnitude from the ones found by neural networks, the overall structure of the homology in this set of proteases remains the same.

contrasts with the conventional methods which appear to be quadratic, due to the time-consuming process of cross detection of insertions and deletions [28]. In fig.6 the results of a conventional analysis using the Protein Information Resource (PIR) method [29] is shown in the same format as the previous figure. One should note the striking qualitative agreement between the various approaches in the prediction of homology.

*Acknowledgements:* This work was supported by the Danish Natural Science Council. Two of us (J.B. and R.M.J.C.) also wish to thank the Thomas B. Thrige Foundation.

## REFERENCES

[1] Blow, D.M., Fersht, A.R. and Winter, G. (1986) Phil. Trans. R. Soc. Lond. A 317, 291.
[2] Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E. and Thornton, J.M. (1987) Nature 326, 347.
[3] Brünger, A.T., Clore, G.M., Gronenborn, A.M. and Karplus, M. (1986) Proc. Natl. Acad. Sci. USA 83, 3801.
[4] Van Gunsteren, W.F. and Berendsen, H.J.C. (1982) in: The 600th Meeting at the University of Oxford on Computer Prediction of Molecular Conformation (Thompson, J.S. and Robson, B. eds) Biochem. Soc. Trends 10, 301.
[5] Paine, G.H. and Scheraga, H.A. (1987) Biopolymers 26, 1125; (1986) Biopolymers 25, 1347; (1985) Biopolymers 24, 1391.
[6] Nadler, W., Brünger, A.T., Schulten, K. and Karplus, K. (1987) Proc. Natl. Acad. Sci. USA 84, 7933.
[7] Lim, V.I. (1974) J. Mol. Biol. 88, 873.
[8] Kotelshuck, D. and Scheraga, H.A. (1969) Proc. Natl. Acad. Sci. USA 62, 14.
[9] Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97.
[10] Robson, B. and Pain, R.H. (1974) Biochem. J. 141, 883.
[11] Robson, B. and Garnier, J. (1986) Introduction to Proteins and Protein Engineering, Elsevier, Amsterdam, New York.
[12] Levitt, M., Sander, C. and Stern, P.S. (1985) J. Mol. Biol. 181, 423.
[13] Chou, P.Y. and Fasman, G.D. (1978) Annu. Rev. Biochem. 7, 251.
[14] Kabsch, W. and Sander, C. (1983) FEBS Lett. 155, 179.
[15] Qian, N. and Sejnowski, T.J. (1988) J. Mol. Biol. 202, in press.
[16] Rosenblatt, F. (1962) Principles of Neurodynamics, Spartan, NY.
[17] Minsky, M.L. and Papert, S. (1969) Perceptrons, MIT Press.
[18] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986) in: Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Rumelhart, D.E., McClelland, J.L. and the PDP Research Group eds) vol.1: Foundations. MIT Press, Cambridge, MA.
[19] Werbos, P.J. (1974) Thesis in Applied Mathematics, Harvard University.
[20] Parker, D.B. (1982) Invention Report, S81-64, Stanford University.
[21] Lippmann, R.P. (1987) IEEE ASSP Magazine, April 1987, p.4.
[22] Sejnowski, T.J. and Rosenberg, C.R. (1987) Complex Systems 1, 145.
[23] Mathews, B.W. (1975) Biochim. Biophys. Acta 405, 442.
[24] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F. jr, Brice, M.D., Rogers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535.
[25] Henderson, R. and Unwin, P.N.T. (1975) Nature 257, 28–32.
[26] Argos, P., Rao, J.K. and Hargrave, P.A. (1982) Eur. J. Biochem. 565.
[27] Jap, B.K., Maestre, M.F., Hayward, S.B. and Glaeser, R.M. (1983) Biophys. J. 43, 81–89.
[28] Sankoff, D. and Kruskal, J.B. (1983) Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, Addison-Wesley.
[29] George, D.G., Barker, W.C. and Hunt, L.T. (1986) Nucleic Acids Res. 14.