Protein Engineering vol.5 no.3 pp.191–195, 1992

Quantitative analysis of protein far UV circular dichroism spectra by neural networks

Gerald Böhm[1], Rudolf Muhr and Rainer Jaenicke

Institut für Biophysik und Physikalische Biochemie, Universität Regensburg, Universitätsstrasse 31, D-8400 Regensburg, Germany

[1]To whom correspondence should be addressed

A new method based on neural network theory is presented to analyze and quantify the information content of far UV circular dichroism spectra. Using a backpropagation network model with a single hidden layer between input and output, it was possible to deduce five different secondary structure fractions (helix, parallel and antiparallel β-sheet, β-turn and random coil) with satisfactory correlations between calculated and measured secondary structure data. We demonstrate that for each wavelength interval a specific network is suitable. The remaining discrepancy between the secondary structure data from neural network prediction and crystallography may be attributed to errors in the determination of protein concentration and random noise in the CD signal, as indicated by simulations.

Key words: circular dichroism/neural network/protein secondary structure

Introduction

Circular dichroism (CD) spectroscopy is a valuable tool for the characterization of protein structures in solution (Schmid, 1989; Johnson, 1990). This is due to the inherent information content of the far UV CD spectra (between 180 and 250 nm) which depends predominantly on the difference in absorption of left-handed and right-handed circularly polarized light at the protein backbone. Thus the CD spectrum is sensitive for the secondary structure conformation of the protein under investigation. Aside from these chiral centers, disulfide bridges and aromatic side-chains (predominantly tryptophan) contribute to the CD spectrum. Several attempts to deconvolve the spectra with respect to the secondary structure information have been described in the past (Brahms and Brahms, 1981; Hennessey and Johnson, 1981; Compton and Johnson, 1986; Yang *et al.*, 1986; Manavalan and Johnson, 1987). Other methods are based on the assumption that the spectra are a linear combination of reference spectra for the five secondary structure types (helical conformation, parallel and antiparallel β-sheet, β-turn and random coil).

Recently, interest in neural network methods in structural biology has led to a number of applications focusing on secondary structure prediction (Qian and Sejnowski, 1988; Holley and Karplus, 1989), three-dimensional structure prediction (Bohr *et al.*, 1990) and prediction of ATP binding sites (Hirst and Sternberg, 1991). The present investigation was stimulated by attempts to apply homology modelling techniques to proteins from extreme halophiles. Comparative structure modelling of halophilic dihydrofolate reductase from *Halobacterium volcanii* may be effectively assisted by the assignment of precise secondary structure fractions. In this ongoing work on the fundamental principles of protein structure prediction we were therefore faced with the problem of assigning most precise secondary structure fractional indices to far UV CD spectra.

In order to develop a new method for the computation of secondary structure proportions, CD spectra are considered as a superposition of (i) secondary structure information, (ii) cystine absorption, (iii) aromatic side-chain absorption, (iv) random noise from CD measurement, and (v) non-random errors in protein concentration measurement. Based on this assumption, methods of assigning fractional indices for secondary structure types should be flexible with respect to the relationship between a specific CD spectrum and its corresponding structure, and should not necessarily imply linear combination of reference spectra. Also, pattern recognition algorithms require a large database which is not yet available. Neural networks (NNs) may be trained to tolerate noisy data, and they represent a most elegant method of non-algorithmic deconvolution of information.

Materials and methods

The data used in this work were taken from Compton and Johnson (1986). The authors compiled data for 15 proteins in the range 178–260 nm, at intervals of 2 nm. Similar data were reported by Yang *et al.* (1986). However, as has been pointed out and discussed by Provencher and Glöckner (1981) on thermolysin and subtilisin BPN′, there are inherent measurement errors in these data. We therefore used just 11 of these data sets and excluded data sets that obviously contradict our experience in CD spectroscopy.

For analysis of data, especially in the wavelength range between 200 and 250 nm, the data were interpolated by a cubic spline algorithm to obtain data at intervals of 1 nm and 0.5 nm. From the dataset cited above, we used 11 spectra (Table I): cytochrome *c*, hemoglobin, lactate dehydrogenase, lysozyme, myoglobin, ribonuclease A, flavodoxin, glyceraldehyde-3-phosphate dehydrogenase, subtilisin Novo and triosephosphate isomerase. In addition, poly-glutamate (a purely α-helical polypeptide) was used in the database. Also, hemerythrin and thermolysin data were included in the data set; these data are taken from a computer program distributed by Dr W.C.Johnson. Data are given in units of $\delta\epsilon$, e.g. the difference in absorption coefficient for right-handed and left-handed circularly polarized light. Before serving the data to the network, they were normalized to a range between −1.0 and +1.0.

There are two logical conditions that methods for secondary structure predictions have to fulfil. The first describes that any fractional index f for a secondary structure type k must be between zero and unity, i.e.

$$0 \leq f_k \leq 1 \quad (1)$$

Also, the sum of all fractions f of secondary structure types k must be unity if all possible secondary structures are taken into consideration:

$$\Sigma f_k = 1 \quad (2)$$

Backpropagation networks are feed-forward type networks that

Table I. Fractions of secondary structure from the proteins used in this work

Protein	Helix	Antiparallel β-sheet	Parallel β-sheet	β-turn	Random coil
Cytochrome *c*	0.38	0.00	0.00	0.17	0.45
Hemoglobin	0.75	0.00	0.00	0.14	0.11
Lactate dehydrogenase	0.41	0.06	0.11	0.11	0.31
Lysozyme	0.36	0.09	0.00	0.32	0.23
Myoglobin	0.78	0.00	0.00	0.12	0.10
Ribonuclease A	0.24	0.33	0.00	0.14	0.29
Flavodoxin	0.38	0.00	0.24	0.16	0.22
Glyceraldehyde-3-phosphate dehydrogenase	0.30	0.09	0.13	0.14	0.34
Subtilisin Novo	0.31	0.02	0.08	0.11	0.48
Triosephosphate isomerase	0.52	0.00	0.14	0.11	0.23
Poly-glutamate	1.00	0.00	0.00	0.00	0.00
Thermolysin	0.32	0.10	0.08	0.20	0.30
Hemerythrin	0.75	0.00	0.00	0.11	0.14

Data were taken from Compton and Johnson (1986). Some redundancies are in the dataset: the three-dimensional structures and thus the CD spectra of hemoglobin, myoglobin and hemerythrin are similar.

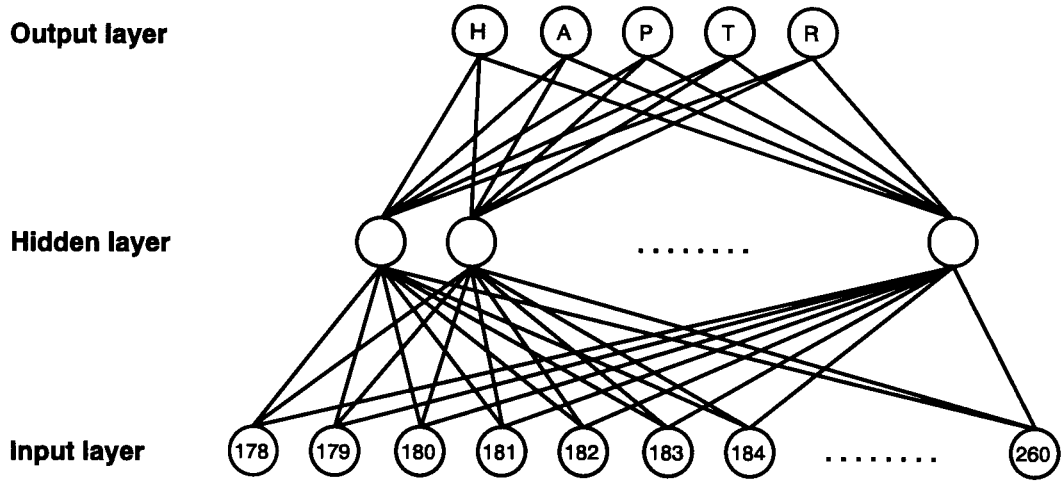

Fig. 1. Schematic topology of a NN. Data are presented to the input layer; for each data point (at each wavelength) there is a separate processing element ('neuron') which processes the data and sends the result via weighted connections to each neuron in the next layer (hidden layer). The adaptation of weights is performed in the learning phase. The output consists of five neurons that represent the five fractional states (helix, antiparallel and parallel β-sheet, β-turn and random coil). The neurons in the first two layers are only partly shown.

require supervised learning. They are characterized by overlap properties and resemble a common topology for neural network applications. Although proposed in 1974 by P.Werbos (Schöneburg *et al.*, 1990), the backpropagation algorithm has been in use only since its further development by Rumelhart and coworkers in 1986. As described schematically in Figure 1, a simple neural network consists of processing elements in several layers. In the topology used here, all processing elements ('neurons') of a layer are connected to each neuron of the next layer. Information and signals are transferred through these connections and processed in the neurons. The connections are numerically weighted; the weights are gradually changed and adapted periodically in the 'learning phase' or 'training phase', until each pattern presented to the input layer is correctly projected to the corresponding pattern of the output layer. Error propagation during learning is performed via the generalized delta-rule

$$\Delta w_{ij}(t) = \sigma \cdot \delta_i \cdot o_j + \mu \cdot \Delta w_{ij}(t-1) \qquad (3)$$

where w is the weight between two connected elements i and j at the learning cycle t, δ_i is the error at the processing element i, o_i is the output of element j, σ is the learning constant (usually of the order of 0.5), and μ is the momentum (0.5–1.0). At the end of each cycle, the new weights for the next evaluations are calculated according to

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \qquad (4)$$

The 'recall phase' is then used to serve input data to the NN which were not used during the training phase. The network calculates the corresponding output according to the adapted weights. For a more comprehensive treatise, cf. Rumelhart *et al.* (1986a,b). Meanwhile, some modifications of the above, simple model are used; cf Schöneburg *et al.* (1990). For the solution of the neural network model, a computer program distributed by Schöneburg *et al.* (1990) has been used. Validation of the results was performed by Neural Network programs from Neural Ware Inc. (USA).

All calculations described in this work were performed with an industry standard personal computer (based on Intel 80386SX-processor, 13 MHz clock) equipped with the operating system DOS. A floating point unit and 1 MByte of memory is highly recommended. The average calculation time of a training phase is then between 6 and 18 h, depending on convergence, size of

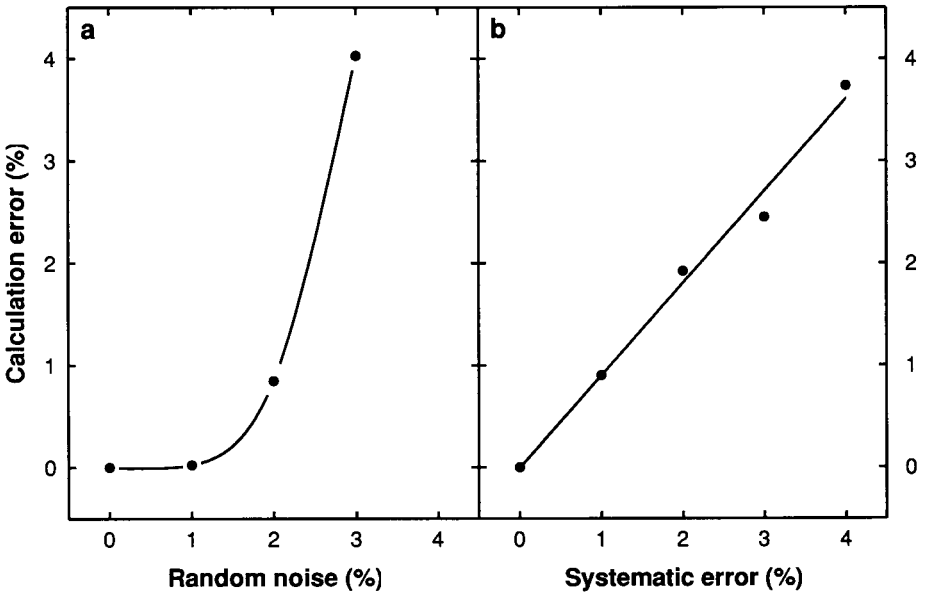

Fig. 2. (**a**) Dependence of the relative error in the output on the random noise which is superimposed onto the input data. This is the result of a simulation with linear combinations of reference spectra as input data; learning was performed with 50 spectra, recall was done with 250 spectra. The error describes the average relative difference between expected and calculated output. (**b**) Effect of systematic errors on the result of the NN calculation. Again, 50 simulated spectra were in the training set and 250 spectra were used for recall.

the input dataset and topology of the network; recall usually needs less than a second. Learning on a PC with an Intel 80486DX-processor (33 MHz clock) still takes between 1 and 3 h.

The power of most methods published to date is measured by the Pearson correlation coefficient r, which describes the success of prediction of each state k determined by

$$r_k = \frac{(\Sigma x_i \cdot y_i - N^{-1} \cdot \Sigma x_i \cdot \Sigma y_i)}{(\Sigma x_i^2 - N^{-1} \cdot (\Sigma x_i)^2)^{1/2} \cdot (\Sigma y_i^2 - N^{-1} \cdot (\Sigma y_i)^2)^{1/2}} \quad (5)$$

where all summations range from $i = 1$ to N, with N being the number of measurements. x_i is the fraction of the secondary structure element k from dataset i, whereas y_i is the calculated value for the respective secondary structure element.

Results

Choice of network topology

We used a series of standard topologies described in the literature: perception, adaline and madaline, counter-propagation and back-propagation networks (Schöneburg *et al.*, 1990). It turned out that only the backpropagation network was able to generalize in the recall phase. All other topologies were able to perform pattern recognition, e.g. spectra used in the dataset were recognized with perfect confidence and no error in the recall; however, spectra not included in the training phase had incorrectly assigned secondary structure parameters in the recall. This allows the conclusion that these topologies are incompetent to generalize the learned rules in the recall for the problem under investigation. We thus concentrated on back-propagation networks. A simple network with no hidden layer also led to bad results, but with one hidden layer the predictions were successful. A second hidden layer did not improve the functionality, but may be important for random noise filtering in future investigations.

The standard topology (Figure 1) finally consisted of a net with 83 processing elements in the input layer (one for each data point in the wavelength range of 178–260 nm), a hidden layer with 45 neurons, and an output layer with five neurons representing the five secondary structure types under investigation. The topology varies when different wavelength intervals or ranges are used. It should be noted that the reliability drops significantly when the net contains less than ~ 100 neurons. All elements in each layer were fully connected to the neighboring layer.

Choice of transfer function

Usually, back-propagation networks use a special sigmoidal function for the transfer of data between layers. We found that this sigmoidal function gave slightly poorer results than a simple linear function. Linear transfer functions, however, are usually not appropriate for multilayer networks since two layers connected by linear transfer may be equally represented by a single layer with appropriate weights. In this case, however, another learning rule must be applied. In contrast to this, results were worse with no hidden layer used than with one hidden layer and linear transfer. Therefore, linear transfer functions with one hidden layer were used for the results described below.

Choice of learning rate

Once the appropriate topology for the calculations was found, the influence of the learning rate σ and the momentum μ on the results and the convergence behavior of the network was investigated. Some dependence of these parameters on the success of the method was observed; best results were obtained by using $\sigma = 0.1$ for the first 30 000 training steps, then decreasing σ to 0.05 and finally to 0.01. The momentum (μ) may vary between 0.1 and 0.6. The prediction was always successful when the learning rate was below 0.3 but rarely converged when σ was 0.5 or higher.

The preceding protocol allows a simulated search on the energy hyperspace (covered by the network weight parameters) using a decreasing resolution of the search parameter, to find the global minimum on the energy hyperspace. Note that the network calculations did not converge when the learning rate was chosen above 0.3. Convergence was defined as a total of 0.00001 when reproducing the training set, or at least 300 000 cycles of learning when no further reduction in the total error was noticed.

Simulation of linear combination of reference spectra

To assess the quality of the method and to get information about possible problems, datasets of simulated spectra were created. These sets consisted of random linear combination of reference spectra for the five different secondary structure types; data for the reference spectra were taken from Compton and Johnson (1986). Six to eight spectra in the training set were sufficient to deconvolve the simulation spectra into the respective reference spectra with perfect agreement in the recall phase; the average error in the determination of each of the five fractional indices was <0.001. This demonstrates that the designed network is able to deconvolve appropriately linear combinations of spectra.

Simulation of random noise

The simulated spectra were then perturbed by adding random noise values to the (ideal) linear combination of the reference spectra. Random terms were chosen as 0.01, 0.02 and 0.03; the normalized amplitude of the helical reference spectrum was between -1.0 and 1.0. Thus, the random terms represent a reasonable signal to noise ratio commonly observed for CD measurements. As expected, the ability of the network to extract the correct fractional indices from the noisy spectra decreased with increase of the random noise (Figure 2a). More complicated topologies of networks will be used in future simulations to see if these topologies tolerate noisy data.

Simulation of error in protein concentration determination

Apart from random noise, CD spectra contain systematic errors that arise from errors in the protein concentration measurements. These determinations are most often performed by colorimetry and therefore usually contain an error in the order of $1-5\%$. It is expected that these errors cause even more trouble to networks than random noise. The results of simulations on the effect of these errors on the deconvolution of the linear combined simulation spectra is shown in Figure 2b. This problem may be circumvented by using more precise protein concentration measurements for the spectra used in the training and recall phase, or by significantly extending the training dataset.

Table II. Comparison of correlation coefficients for the determination of secondary structure fractions from CD measurements, as calculated for several methods published to date

Method	Wavelength region (nm)	Helix	Antiparallel β-sheet	Parallel β-sheet	β-turn	Random coil
[a]Hennessey and Johnson (1981)	178–260	0.98	0.55	0.63	0.30	0.61
[b]Manavalan and Johnson (1987)	178–260	0.97	0.78	0.67	0.49	0.86
[a]Provencher and Glöckner (1981)	178–260	0.96	0.23	0.39	0.51	0.64
[b]Provencher and Glöckner (1981)	187–260	0.98	0.63	0.56	0.65	0.83
Backpropagation NN (this work)	178–260	1.00	0.91	0.63	0.64	0.96
[a]Hennessey and Johnson (1981)	190–260	0.98	0.40	0.00	0.18	0.24
[a]Manavalan and Johnson (1987)	190–260	0.95	0.57	0.47	0.54	0.69
Backpropagation NN (this work)	200–250	1.00	−0.36	0.84	0.59	0.99

Data other than this work was taken from [a]Manavalan and Johnson (1987) or [b]Johnson (1990). Coefficients [b] were calculated with an identical data set of 16 proteins. For a discussion of the differences between the correlation coefficients of the two authors cited above, see Johnson (1990).

Table III. Example result of a NN calculation

	Helix	Antiparallel β-sheet	Parallel β-sheet	β-turn	Random coil	
Lactate dehydrogenase						
measured	0.41	0.06	0.11	0.11	0.31	sum: 1.00
calculated	0.396	0.008	0.123	0.170	0.305	sum: 1.002
error	−0.014	−0.052	0.013	0.060	−0.005	
Myoglobin						
measured	0.78	0.00	0.00	0.12	0.10	sum: 1.00
calculated	0.721	−0.036	0.008	0.139	0.172	sum: 1.004
error	−0.059	−0.036	0.008	0.019	0.072	
Glyceraldehyde-3-phosphate dehydrogenase						
measured	0.30	0.09	0.13	0.14	0.34	sum: 1.00
calculated	0.330	0.035	0.076	0.182	0.379	sum: 1.002
error	0.030	−0.055	−0.054	0.042	0.039	
Triosephosphate isomerase						
measured	0.52	0.00	0.14	0.11	0.23	sum: 1.00
calculated	0.493	0.003	0.116	0.121	0.268	sum: 1.002
error	−0.027	0.003	−0.024	0.011	0.038	

The learning dataset consisted of nine proteins, recall was performed with the remaining four proteins; measured data are crystallographically derived fractions from X-ray structure. Removal of hemoglobin and hemerythrin from the learning dataset did not affect the result for the homologous protein myoglobin. The average error per prediction is 3.3% in this case.

Application to spectra deconvolution

Table II shows the result of the NN method compared with previously published algorithms, as measured by Equation 5. For the determination of the correlation coefficients, the training set consisted of nine randomly selected spectra, and the recall set contained the four remaining spectra. This methodology avoids misinterpretation due to pattern recognition and represents a more reliable approach for realistic results. Independently of the datasets used for learning and recall, the values given in Table II (which are averaged over four distinct calculations) are better than those previously published. In Table III, results for a single calculation are presented as an example; there, four proteins were chosen with different structural topology to assess a wide range of input data for deconvolution. The outcome is quite similar when hemoglobin and hemerythrin (which is homologous to myoglobin that is used for recall) is excluded from the training set; structural homology between proteins in the learning and recall datasets is therefore not necessary for the success of the method.

Discussion

Table II demonstrates the excellent performance of the method described in this work, as compared with previous methods. It is shown that the method works well in the wavelength range between 178 and 260 nm. However, β-sheets are not determined with sufficient precision if wavelengths range only from 200 to 250 nm. The remaining discrepancy between calculations and measurements may have several reasons. One possible explanation could be that the fractions of secondary structure derived from crystallography may be inadequate; a more precise approach could be a reassignment of secondary structure types according to the Kabsch and Sander method. This has been discussed in detail by Perczel *et al.* (1991). As can be seen from Table III, for all four results Equation (2) is surprisingly well satisfied. However, Equation (1) does not hold in the case of β-sheets of myoglobin where a negative fraction is obtained. This is even worse when a limited data set between 200 and 250 nm is used (data not shown); in this case, several fractions of secondary structure are predicted to be negative. Thus, a restriction should be implemented into the currently used NN to fulfil both Equations (1) and (2).

Another reason for the still limited success of the method could be a superposition of random noise (from CD signal measurement) and/or systematic errors (from protein concentration determination) onto the real spectra; it has been shown by simulations using artificial spectra that the results obtained are indistinguishable from the results with the spectra from Table I. However, the results described in this work may be satisfactory for a wide range of applications. The average error for the determination of a single fraction is $\sim 3.5\%$, thus our current model provides a very good starting point for further investigations. These will include research on the topology of networks, transfer functions and parameters. As shown by the present data, the NN approach may be successfully used to analyze far UV CD spectra in a quantitative fashion. However, the structure prediction obtained by the NN method should be taken with care, as long as neural networks still lack fundamental characterization.

Note

A program, suitable for Personal Computers in the MS-DOS environment equipped with the Microsoft Windows™ System (3.0 or higher), will be made available at the end of 1992, and will be distributed freely. To receive a copy of the program, readers are requested to use file transfer access to Internet (ftp) and log-in as 'anonymous' to the machine with the Internet address 132.199.1.42 (rbisg1.biologie.uni-regensburg.de). Enter 'binary' and 'cd cd__spectroscopy', then 'get nncalc.exe' and 'get nndocu.exe' to transfer the respective files. The program and documentation binaries are self-extracting, compressed files.

Acknowledgements

The authors wish to thank Dr Fritz Wünsch (Physical Department) and Martin Stetter (Biophysical Department, University of Regensburg) for discussion and their helpful comments. The generous gift of CD spectra by Dr W.C.Johnson is gratefully acknowledged. This work was supported by the Deutsche Forschungsgemeinschaft, grant Ja 78/29-1.

References

Bohr,H., Bohr,J., Brunak,S., Cotterill,R.M.J., Fredholm,H., Lautrup,B. and Petersen,S.B. (1990) *FEBS Lett.*, **261**, 43–46.

Brahms,S. and Brahms,J. (1980) *J. Mol. Biol.*, **138**, 149–178.

Compton,L.A. and Johnson,W.C.,Jr (1986) Anal. Biochem., **155**, 155–167.

Hennessey,J.P. and Johnson,W.C.,Jr (1981) *Biochemistry*, **20**, 1085–1094.

Hirst,J.D. and Sternberg,M.J.E. (1991) *Protein Engng*, **4**, 615–623.

Holley,L.H. and Karplus,M. (1989) *Proc. Natl Acad. Sci. USA*, **86**, 152–156.

Johnson,W.C.,Jr (1990) *Proteins: Struct. Funct. Genet.*, **7**, 205–214.

Manavalan,P. and Johnson,W.C.,Jr (1987) *Anal. Biochem.*, **167**, 76–85.

Perczel,A., Hollosi,M., Tusdady,G. and Fasman,G.D. (1991) *Protein Engng*, **4**, 669–680.

Provencher,S.W. and Glöckner,J. (1981) *Biochemistry*, **20**, 33–37.

Qian,N. and Sejnowski,T.J. (1988) *J. Mol. Biol.*, **202**, 865–884.

Rumelhart,D.E., Hinton,G.E. and Williams,R.J. (1986a) *Nature*, **323**, 533–536.

Rumelhart,D.E., McClelland,J.L. and the PDP Research Group (1986b) *Parallel Distributed Processing, Explorations in the Microstructure of Cognition. Volume 1: Foundations; Volume 2: Psychological and Biological Models*. MIT Press, Cambridge, USA.

Schöneburg,E., Hansen,N. and Gawelczyk,A. (1990) *Neuronale Netzwerke*. Markt und Technik Verlag AG, Haar.

Schmid,F.X. (1989) In Creighton,T.E. (ed.), *Protein Structure: A Practical Approach*. IRL Press, Oxford, pp. 251–285.

Yang,J.T., Wu,C.-S.C. and Martinez,H.M. (1986) *Methods Enzymol.*, **130**, 208–269.

Received on October 10, 1991; revised and accepted on January 24, 1992