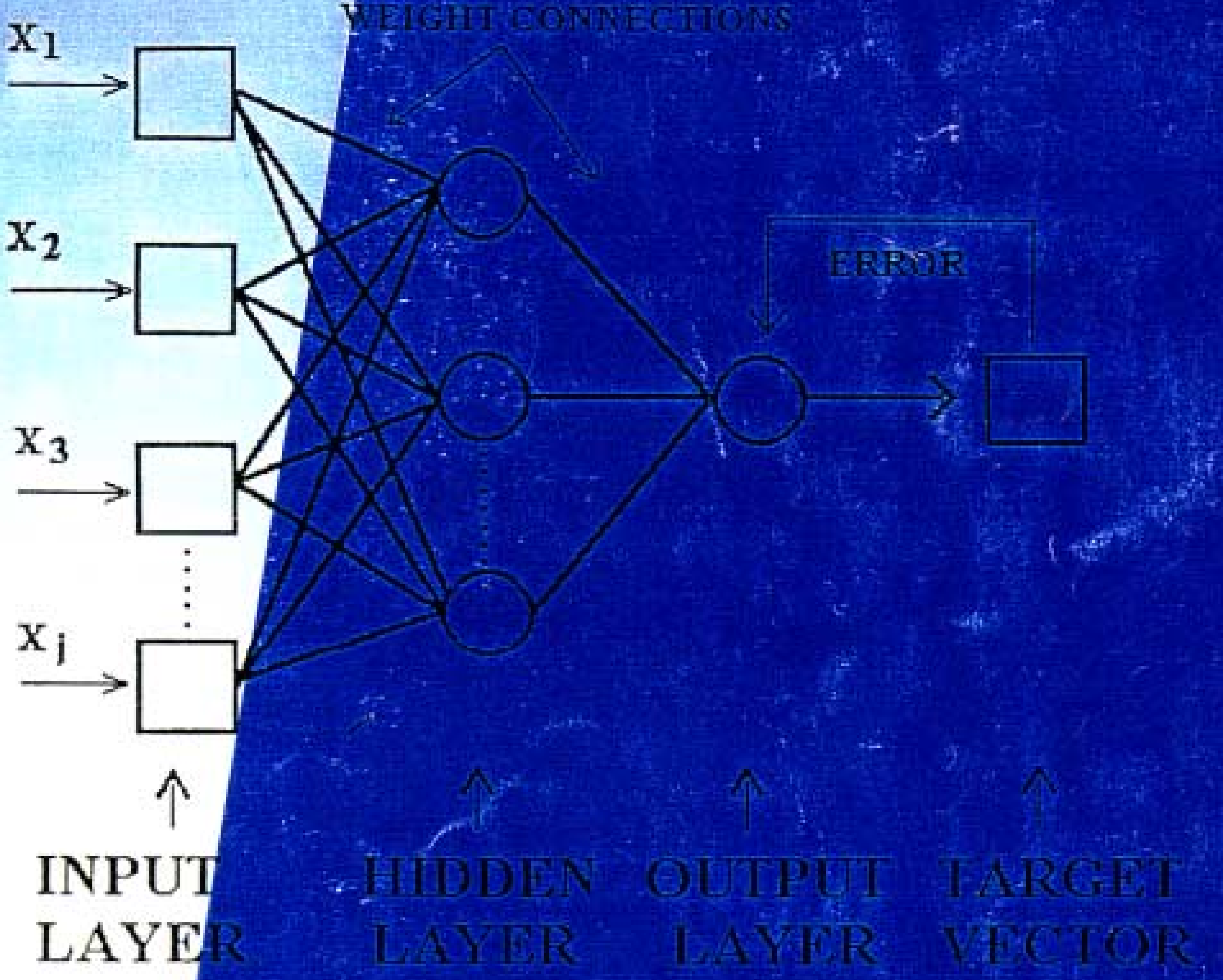
FEBS LETTERS

AN INTERNATIONAL JOURNAL FOR THE RAPID PUBLICATION OF SHORT REPORTS IN BIOCHEMISTRY, BIOPHYSICS AND MOLECULAR CELL BIOLOGY



Published by Elsevier on behalf of the Federation of European Biochemical Societies

Volume 346, Number 2,3, 13 June 1994

LAST NUMBER OF THIS VOLUME

FEBS Letters 346 (1994) 273–277

FEBS 14103

Application of artificial neural networks for prokaryotic transcription terminator prediction

T. Murlidharan Nair, Sanjeev S. Tambe, B.D. Kulkarni*

Chemical Engineering Division, National Chemical Laboratory, Pune 411 008, India

Received 21 March 1994; revised version received 2 May 1994

Abstract

Artificial neural networks (ANN) to predict terminator sequences, based on a feed-forward architecture and trained using the error back propagation technique, have been developed. The network uses two different methods for coding nucleotide sequences. In one the nucleotide bases are coded in binary while the other uses the electron–ion interaction potential values (EIIP) of the nucleotide bases. The latter strategy is new, property based and substantially reduces the network size. The prediction capacity of the artificial neural network using both coding strategies is more than 95%.

Key words: Artificial neural network; Electron–ion interaction potential; Terminator

1. Introduction

Terminators are sequences which primarily regulate gene expression by providing stop signals at the end of transcription units, and thus allow adjacent genes and/or operons to be transcribed and regulated independently [1]. To understand the control mechanism, it is imperative to identify template codes involved in the site-specific recombination process. Not much information is known about the DNA sequences of terminators. Earlier studies [2,3] have shown that factor-independent terminators shared features like G/C-rich dyad symmetry followed by a stretch of 4–8 adjacent thymine residues immediately upstream of the last nucleotide incorporated into the RNA chain. It should also be noted that there are many independent terminators that do not comply with the consensus pattern of dyad symmetry and T-stretch [3]. As a result, the conditions for termination are not well defined. Therefore, it has become important to develop methods to identify terminators due to the inconsistent consensus patterns that they contain. Essentially, the task involves recognizing the hidden pattern in the terminator region. Towards that goal, a modelling approach known as the 'artificial neural network' (ANN), which possesses the ability to learn and generalize nonlinear functional relationship(s), can be exploited.

ANNs have been used to solve a variety of problems in biology and have been extensively reviewed [4]. The paradigm was originally developed to simulate the brain's learning process by modelling its fundamental unit, i.e. the nerve cells and their interconnections. In ANN, the nerve cells are replaced by computational units called neurons, and the axons by symbolic connections. The synaptic strengths are represented by weights and threshold (biases) which are applied to the connections and neurons, respectively. Network modelling involves a training stage, during which a training set of inputs and its corresponding targets are presented to the neural net. The network, by a process of iterative learning, attempts to minimize an objective function (error function), usually the difference between the network computed output and the desired output. Learning or training is said to be complete when the network satisfies some convergence criteria. A converged or a trained net has the ability to recognize and generalize the patterns intrinsic to the training set, and this ability has been exploited to recognize hidden patterns in the DNA and protein sequence [4,5]. In this paper we present our results on the development of a multilayered feedforward network using backpropagation algorithm [6] for terminator prediction. We have also employed a novel coding strategy.

2. Experimental

2.1. Data

The terminator sequences were taken from the compilation by Volker Brendel et al. [7]. From a total of 128 terminators of length 51, 88 were chosen for training the network. The remaining 40 terminators were used as the test data set. A pseudo-random number generator was used for constructing random sequences with equal composition of A, T, G and C. These random sequences were combined with the terminator sequences in the ratio 1:3 (one terminator followed by three random sequences).

2.2. Data representation

The input data was coded using two different strategies. In one case

*Corresponding author. Fax: (91) (212) 330 233.
E-mail: bdk@ncl.ernet.in

This is NCL communication no. 5980.

SSDI 0014-5793(94)00489-I

the network was presented with data coded in binary, similar to that used by Borries and Guangwen [8], called CODE-4 (0001 = C, 0010 = G, 0100 = A, 1000 = T). The target to each input sequence was coded 1 for a terminator sequence and 0 for the random sequence. The second type of coding is based on the informational spectra method (ISM) [9–12], which is a mathematical and physical method for the analysis of informational content of DNA and protein sequences [13,14]. In this form of coding the electron–ion interaction potential (EIIP) associated with each nucleotide is calculated using the following equation:

$$W = 0.252Z^{*}\mathrm{Sin}(1.04\pi Z^{*})/2\pi \qquad (1)$$

where Z^{*} is the quasi valence number and is determined as:

$$Z^{*} = \sum^{m} n_i Z_i / N \qquad (2)$$

where

Z_i = valence number of the i^{th} atomic component, n_i = number of atoms of the i^{th} component, m = number of atomic components in the molecule, and N = total number of atoms.

The EIIP values for the nucleotides obtained using Eqns. 1 and 2 are: A, 0.1260; T, 0.1335; G, 0.0806; and C, 0.1340. Thus, each nucleotide, irrespective of its position, is represented by a definite number, and the numerical series so obtained are finite length deterministic discrete signals. The normalized signals are presented to the neural net as the input data. The targets are represented in an analogous manner to CODE-4. This representation will henceforth be referred to as EIIP code.

2.3. Back propagation network simulation

All the computations were performed on a 386 AT model equipped with a math co-processor. A network training programme in FORTRAN was developed and featured a multilayered feed-forward type network (Fig. 1). These networks contain one or more layers of neurons, called hidden layers, between the output and the network's input. The output of each neuron is the weighted sum of its inputs passed through a non-linear activation function. The networks learn by modifying the strength of the interconnections (the weights) between neurons, according to some specified rule called the 'learning algorithm'. We have used the most popular network training algorithm, 'back propagation' (BP) [6], which attempts to minimize the error function, namely, the summed squared error, and is defined by:

$$E = \sum_{p} E_p = \sum_{p} \sum_{i} (t_{pi} - \mathrm{outnet}_{pi})^2 \qquad (3)$$

where the index p ranges over the set of input patterns, i ranges over the set of output units, E_p represents the error on pattern p, t_{pi} is the target, and outnet_{pi} is the actual output of the ith output unit when pattern p has been presented. When the summed squared error falls below the prescribed threshold, the network is said to be trained and the converged weights can be used for predicting the outcomes of the test data. A detailed description of the BP algorithm can be found in several references (e.g. [6,15–17]).

Further we have tried to pinpoint that region in the terminator which is crucial for recognition by the network. This was done by randomizing a calliper of 10 bases (one turn of the helix) and moving the calliper from one end of the sequence to the other and presenting these sequences to the net for prediction.

3. Results and discussion

We have trained two separate neural nets using the above-mentioned coding strategies. In one case we coded the nucleotide bases as per CODE-4. Since this method did not reflect any intrinsic property of the bases, we have also used the EIIP coding strategy which reflects a physical property of the system under study. The informational spectra method [11], which uses the EIIP values, is a tool for the analysis of the informational content of proteins and nucleotide sequences [12,13] and has also been used to obtain consensus spectra for different sequences [14]. It aims to establish a relationship between a sequence and its biological activity. Biological processes that take place in nature are highly specific and result from the selective interactions between macromolecules. These interactions are based on an efficient recognition which takes place over a relatively larger separation. The basis of molecular recognition has been attributed to the electric forces determined by the electrostatic potential around a molecule [18]. The electrostatic potential depends upon the distribution and the energy state of the valence electrons. EIIP values are the physical parameters that influence the delocalized electrons. Earlier studies have established a correlation between EIIP values of organic molecules and their biological activity [19–21]. This suggests that some of the information responsible for the biological activity of DNA and protein sequences may be encoded in their primary structure in terms of the distribution of EIIP values of constitutive elements (in our case nucleotides). Hence, the network was presented with a numerical series of EIIP values which are finite length deterministic discrete signals corresponding to the terminator sequence.

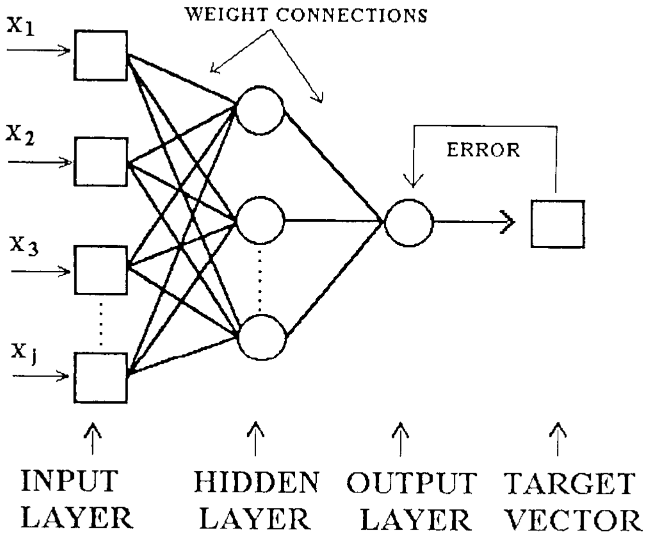


Fig. 1. General architecture of error back propagation neural network.

The network architecture for CODE-4 representation consisted of 204 (sequence length × 4) neurons in the input layer, a single hidden layer with 7 neurons, and an output layer with 1 neuron. The network architecture for the sequences coded with their EIIP values consisted of an input layer with 51 neurons, a single hidden layer with 7 neurons, and an output layer with 1 neuron. The number of neurons in the hidden layer was fixed to 7 by trying different combinations of neurons. The optimal prediction capability was obtained with 7 neurons in the hidden layer. However, using 8 neurons or more did not increase

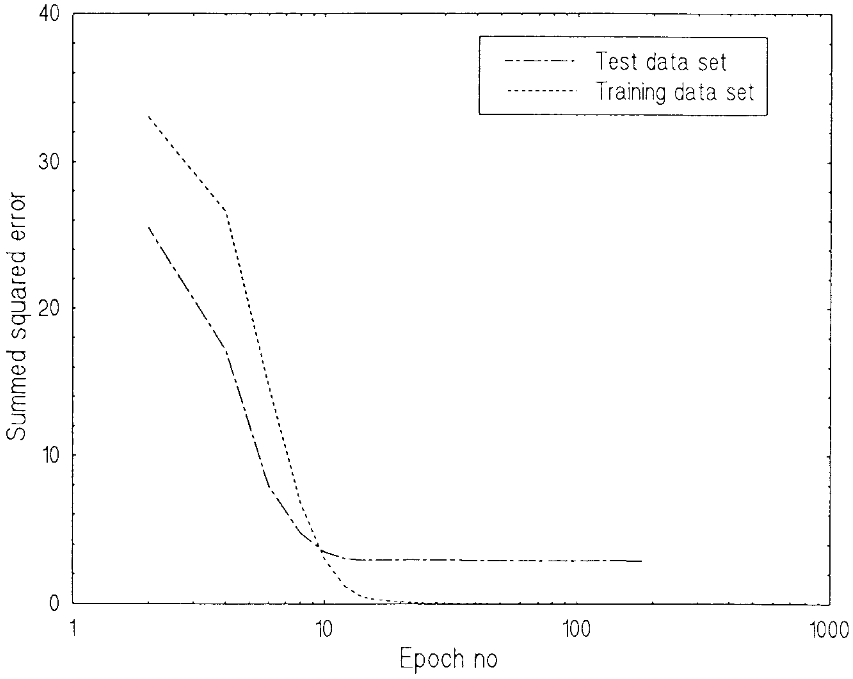


Fig. 2. Error profiles of the training and test data set using CODE-4

the prediction capability of the net, whilst less than 7 neurons in the hidden layer hampered the prediction capability of the net. The momentum term α was optimized to 0.6 and the learning rate to 0.2. The net was presented with 352 patterns consisting of 88 terminators and 264 random sequences for training. After every epoch, which corresponds to the presentation of all the 352 training patterns once to the net, the weights were extracted and used for predicting the test data set of 160 patterns (40 terminators and 120 random sequences). A network output lying between 0.5 and 1 indicates that the input pattern is a terminator, and an output of less than 0.5 corresponds to a random sequence. Extracting the weights after every epoch and using them for predicting

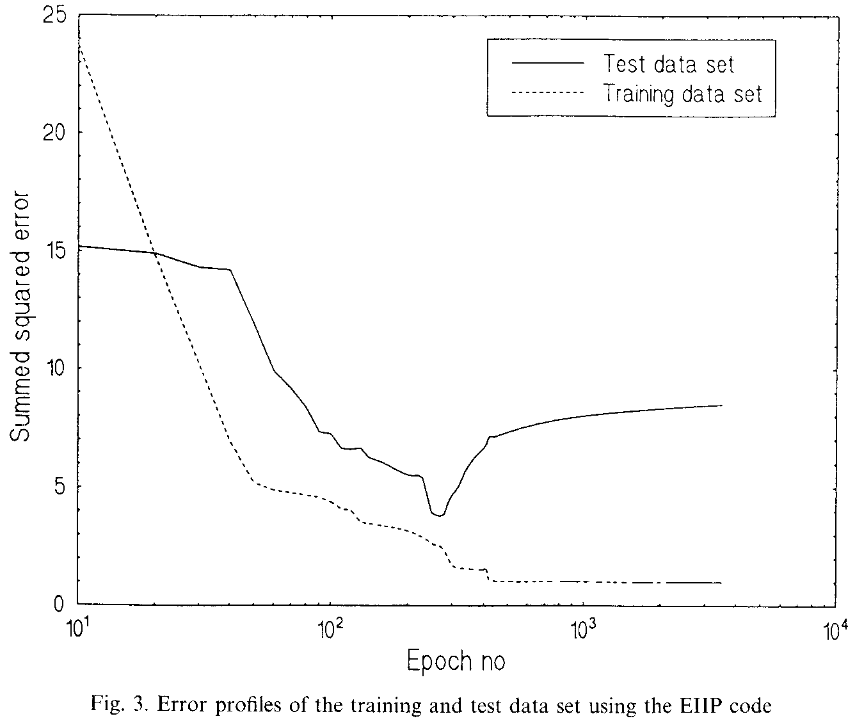


Fig. 3. Error profiles of the training and test data set using the EIIP code

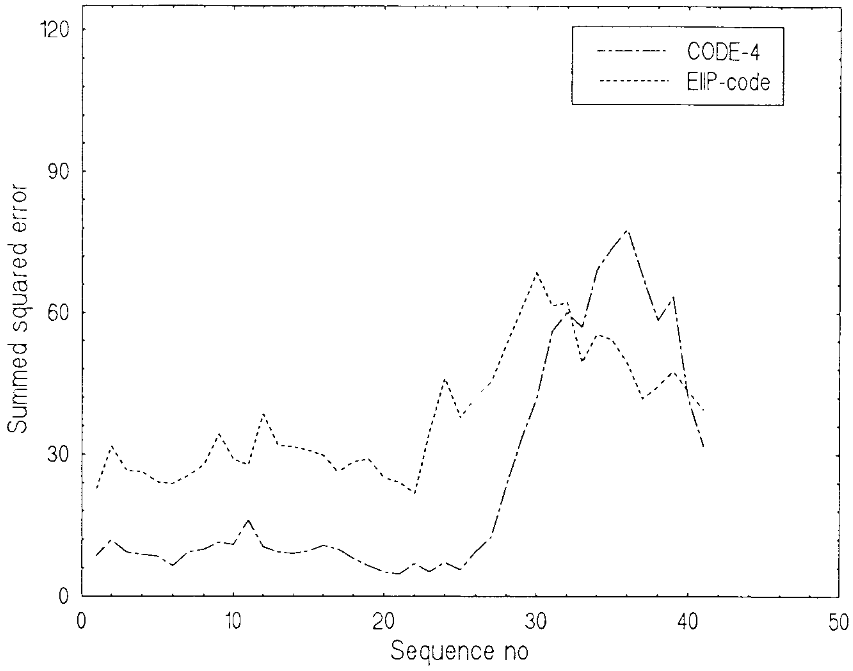


Fig. 4. Error profile of the sequences when randomised at different regions.

the outcomes of the test data is known as cross validation, and tests the generalization capability of the network. The weights corresponding to the minimum error for the test data set were taken as the optimal weights.

The results of the prediction for the test set with optimal weights show that out of the 160 test patterns, 157 (98.125%) patterns were correctly predicted with CODE-4 and 153 (95.625%) patterns were predicted by EIIP code. It should also be noted that neither of the coding strategies predicted false positives, i.e. none of the random sequences were predicted as terminators. Furthermore, a network was trained by coding the four nucleotides by arbitrary, properly spaced numbers (A, 0.25; T, 0.50; G, 0.75; and C,1.0). The network architecture for this net was the same as used earlier (viz. 51 input neurons, 7 hidden neurons and 1 output neuron). The network prediction using this arbitrary coding strategy was much more inferior as compared to EIIP code and CODE-4 (only 147 patterns out of 160 were predicted correctly). The net also took a longer time to converge (>1,500 epochs).

Figs. 2 and 3 show the error profiles of the training and test data when CODE-4 and the EIIP code, respectively, were used. It can be observed that the network attains the best prediction capability in 273 epochs with the EIIP code. Prolonged training with the EIIP code did not increase the prediction capability any further; instead the network lost its generalization capability and started memorizing the training patterns. With CODE-4 the error profiles show a continuous decrease. However, prolonged training did not increase the prediction capability of the network, and, therefore, the weights at the end of 200 iterations are considered to be optimal. The marginally lower prediction capability of the EIIP code could be attributed to the smaller network size, which means less parameter space as compared to CODE-4. However, coding sequences by their EIIP values has the advantages that it reduces the network size to one-fourth as compared to CODE-4 and also involves less training time. The optimal weights and the programme to simulate the network may be obtained from the authors on request by E-mail.

The analysis of the terminators, wherein a region of sequences of fixed length (10 bps) was randomized, revealed that the sequences between 30 and 51 were the most important in the recognition process. Fig. 4 shows the error profile of the network when different regions in the sequence were randomized. It is noteworthy that the region corresponding to the maximum error also corresponds to the region in the sequence which contains the dyad symmetry and T-run, which are the well-known features of the terminators.

Our results also corroborate the fact that functionally related sequences have some common information hidden in them. This hidden information may be extracted by analyzing the finite length deterministic discrete signals by presenting them to a network.

To conclude, the results presented here suggest that ANN can be successfully employed for terminator prediction. It is also shown that in addition to the commonly used CODE-4, an alternative approach, in the form of the EIIP code for representing nucleotide bases, can be exploited. This approach has the advantage that it reduces the network size and training time significantly. We would like to suggest here that this coding strategy could be employed directly or indirectly as one of the sensor algorithms in the coding recognition module (CRM) [22–24] used for analyzing coding and non-coding regions in DNA sequences. Furthermore, the calliper randomization strategy used here can be exploited as an alternative method to localize the consensus of a sequence responsible for a particular biological activity.

Acknowledgements: We would like to thank the Council of Scientific and Industrial Research (CSIR, New Delhi) for supporting this project. We also thank the referees for their comments on the previous version of the manuscript.

References

1 von Hippel, P.H., Bear, D.G., Morgan, W.D. and McSwiggen, J.A. (1984) Annu. Rev. Bioch. 53, 389–446.

2 Rosenberg, M. and Court, D. (1979) Annu. Rev. Genet. 13, 319–353.

3 Brendel, V. and Trifonov, E.N. (1984) Nucleic Acids Res. 12, 4411–4427.

4 Hirst, J.D. and Sternberg, J.E.M. (1992) Biochemistry 31, 7211–7218.

5 Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Norskov, L., Olsen, O.H. and Peterson, S.B. (1990) FEBS Lett. 261, 43–46.

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

7 Brendel, V., Hamm, H.G. and Trifonov, E.N. (1986) J. Biomol. Struct. Dyn. 3, 705–723.

8 Demeler, B. and Zhou, G. (1991) Nucleic Acids Res. 18, 1593–1599.

9 Veljković, V. and Slavić, I. (1972) Phys. Rev. Lett. 29, 105–106.

10 Veljković, V., Cosić, I, Dimitrijević, B. and Lalović, D. (1985) IEEE Trans. Biomed. Eng. 32, 337–341.

11 Veljković, V. and Cosić, I. (1987) Cancer Biochem. Biophy. 9, 139–148.

12 Veljković, V. and Metlas, R. (1987) in: Proceedings in Protein Engineering, Oxford, p. 102.

13 Veljković, V. and Metlas, R. (1988) Cancer Biochem. Biophys. 10, 191–206.

14 Cosić, I., Nesić, D., Pavlović, M. and Williams, R. (1986) Biochem. Biophy. Res. Commun. 141, 831–834.

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

16 Zupan, J. and Gasterger, J. (1991) An. Chem. Acta 248, 1–30.

17 Zupan, J. and Gasterger, J. (1993) Angew. Chem. Int. Ed. Engl. 32, 503–527.

18 Harrison, A.W. (1970) Solid State Theory, McGraw Hill, New York.

19 Cosić, I. and Nesić, D. (1987) Eur. J. Biochem. 170, 247–252.

20 Veljković, V., Lalović, D. (1976) Cancer Biochem. Biophys. 1, 295–298.

21 Veljković, V., (1980) A Theoretical Approach to Preselection of Carcinogens and Chemical Carcinogenesis, Gorden Breach, New York.

22 Mural, R.J., Einstein, R.J., Guan, X., Mann, R.C. and Uberbacher, E.C. (1992) TibTech 10, 66–69.

23 Uberbacher, E.C. and Mural, R.J. (1991) Proc. Natl. Acad. Sci. USA 88, 11261–11265.

24 Roberts, L. (1991) Science 254, 805.