CABIOS

Vol. 11 no. 3 1995
Pages 293–300

Analysis of transcription control signals using artificial neural networks

T.Murlidharan Nair, Sanjeev S.Tambe and B.D.Kulkarni[1]

Abstract

The role of the upstream region in controlling the transcription efficiency of a gene is well established. However, the question of predicting the extent of gene expressed given the upstream region has so far remained unresolved. Using an artificial neural network (ANN) to capture the internal representation associated with the transcription control signal, the present work predicts the rate of mRNA synthesis based on the pattern contained in the upstream region. Further, the model has been used to predict the transcription efficiency for all possible single base mutations associated with the β-globin promoter. The simulation results reveal that apart from the experimental observation that a -79G-A and -78G-A mutation increases the efficiency of transcription, mutation in these regions by C or T also causes an increase in transcription. Furthermore the simulation results verify that mutations in the conserved region, in general, decrease the transcriptional efficiency. However, the results also show that certain sequence elements, when mutated, either cause a marginal increase in the level of transcription or have no effect on transcription levels. The simulation results can be used as a guide in designing mutation experiments since an a priori estimate of the possible outcome of a mutation can be obtained.

Introduction

The basic property of all living cells is the ability to regulate the expression of their genes depending on some extracellular signals, and is mostly controlled at the level of RNA transcription. Transcription controlling elements are usually grouped into two classes: promoters and enhancers. These classes can overlap both physically and functionally. Single base mutations in prokaryotic promoters and regulatory sequences have helped in unravelling the complexities underlying the mechanism of transcription initiation and gene regulation (Meyer *et al.*, 1980; Youderian *et al.*, 1982). The transcription of bacterial genes wherein the regulatory mechanism is operative at two distinct levels, namely (i) the regulation at the initiation step of RNA synthesis, and (ii) premature termination of an elongating RNA chain (attenuation) are also likely to be used in the regulation of eukaryotic genes.

Genetic selection for promoters and regulatory mutations in higher eukaryotes were established by studying the *in vivo* or *in vitro* transcription of mutated regions of cloned DNA. This molecular genetic approach has been used to characterize the eukaryotic promoters of the herpes virus thymidine kinase (tk) (McKnight *et al.*, 1982, 1984; Graves *et al.*, 1986), the SV40 T-antigen (Gidoni *et al.*, 1985), and mammalian β-globin genes (Efstratiadis *et al.*, 1980; Lacy and Maniatis, 1980; Hardison, 1983). These studies have focused on DNA sequences close to the site where RNA synthesis starts, and on the sequences which lie too far removed from the initiation site to play a role in the regulation of structural gene expression. The visual analysis of active genes (McKnight and Miller, 1976, 1979) and biochemical measurements of the RNA synthesis rate of specific genes (Harpold *et al.*, 1979; Derman *et al.*, 1981) provide evidence that eukaryotic genes also contain signals for establishing transcription efficiency. The control of transcription accuracy through site-specific initiation is represented by an evolutionarily conserved sequence called the TATA-box which is found 20–30 nucleotides upstream from the transcription start site. The other control sequences which play a vital role in the expression of eukaryotic structural genes occur upstream from the TATA homology, however, such sequences exhibit no evolutionary conservation. If a greater part of the information (signals that regulate transcription) is encoded in these *cis*-acting elements, then the question arises whether it would be possible to extract that information from these elements by *in computo* experiments.

In order to answer the above question, we have chosen the globin gene as a model system for our study. The globin gene has been extensively studied and there are many genetic disorders associated with the mutations in this gene and its upstream regions. With a view to extract information from these *cis*-acting elements, our efforts were focused on the upstream region of the globin gene for developing strategies to analyse the transcription signals which resulted in the differential expression of the globin gene. These strategies could then be extended to other systems. Towards understanding the transcription code,

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

[1]*To whom correspondence should be addressed*

we have exploited the artificial neural network (ANN) modelling approach.

The ANNs can be conceptualized as mathematical approximations of the biological synapse, and can be visualized as a massively parallel computational device composed of a large number of simple computational units (neurons). The neurons communicate through a set of interconnections with variable strengths (weights), in which the learned information is stored. ANNs with error back-propagation (EBP) (Rumelhart *et al.*, 1986) currently represent the most popular learning paradigm, and have been successfully used to perform a variety of input–output mapping tasks for pattern recognition, generalization and classification. In fact the application of EBP networks in solving computational problems both in biology and other sciences exceeds its biological significance (Hirst and Sternberg, 1992; Nair *et al.*, 1994; Liebman, 1992; Rawlings and Fox, 1994; Sternberg *et al.*, 1994). Most of the previous work concerning DNA promoter has been directed towards recognizing promoter sequences from non-promoter sequences (Lukashin *et al.*, 1989; Demeler and Zhou, 1991; O'Neil, 1991, 1992). There have also been attempts to compare ANNs with statistical approaches in predicting promoters (Horton and Kanchisa, 1992). Advanced methods like knowledge-based artificial neural networks (KBANNs) have also been developed to recognize promoter sequences (Shavlik *et al.*, 1992). So far little effort has gone towards capturing the signals responsible for transcription efficiency from the promoter region. In this paper, we present the results on the development of a multilayered feedforward network using an EBP algorithm (Rumelhart *et al.*, 1986) for predicting the relative transcription levels (RTLs) of a eukaryotic gene and its use in the analysis of the transcription signals associated with the promoter region.

Systems and methods

The simulation programmes were written in FORTRAN77 and compiled using the Microsoft Fortran 5.0 compiler for the IBM PC and compatibles under the DOS 5.0 operating system.

Data

The data for network modelling was taken from the mutation studies carried out by Myers *et al.* (1986), wherein saturation mutagenesis (Myers *et al.*, 1985) has been used to introduce random single base substitutions into the mouse β-globin promoter region. The effects of single base substitutions in the β-globin promoter have been determined by comparing the levels of correctly initiated RNA derived from the test and reference plasmids co-transfected into HeLa cells and expressed as the RTL of each mutant. RTL has been calculated using the following expression:

$$RTL = \frac{M/R_1}{WT/R_2} \quad (1)$$

where

M = signal of the mutant test gene

WT = signal from the wild-type test gene

R_1 = signal from the reference gene co-transfected with the mutant test gene

R_2 = signal from the reference gene co-transfected with the wild-type test gene

From a total of 129 mutants obtained with mutations between −101 and +20, 117 were used as the network training data set and the remaining 12 were used as the test data set.

Data representation

The input data, which consisted of the β-globin promoter sequence with appropriate mutations, have been coded in binary, similar to that used by Borries and Guangwen (Demeler and Zhou, 1991), called CODE-4 (0001 = C, 0010 = G, 0100 = A, 1000 = T). The target to each input sequence was the RTL (normalized) corresponding to each mutant.

Neural network simulation

The neural network simulations were performed on a 386 AT equipped with a maths coprocessor. A network training programme which featured a multilayered feed-forward type network has been developed in FORTRAN and used for simulation purposes (similar to the one previously used: Nair *et al.*, 1994). The neural network architecture is shown in Figure 1. The network consist of three layers: input, hidden and output. The number of neurons in the input and output layers are defined by the problem being studied. However, there is no easy way to assign a fixed number of neurons to the hidden layer, which is responsible for internal representation. Network training is done in two stages called the forward and reverse pass.

In the forward pass, the output of each neuron is computed as the weighted sum of its inputs passed through a non-linearity; however, the neurons in the input layer are simple distributive nodes, which do not alter the input value at all. In the reverse pass the network adjusts the connection strengths between the neurons using the generalized delta rule. The EBP algorithm with momentum term, which has been used for training the network, attempts to minimize an objective function,

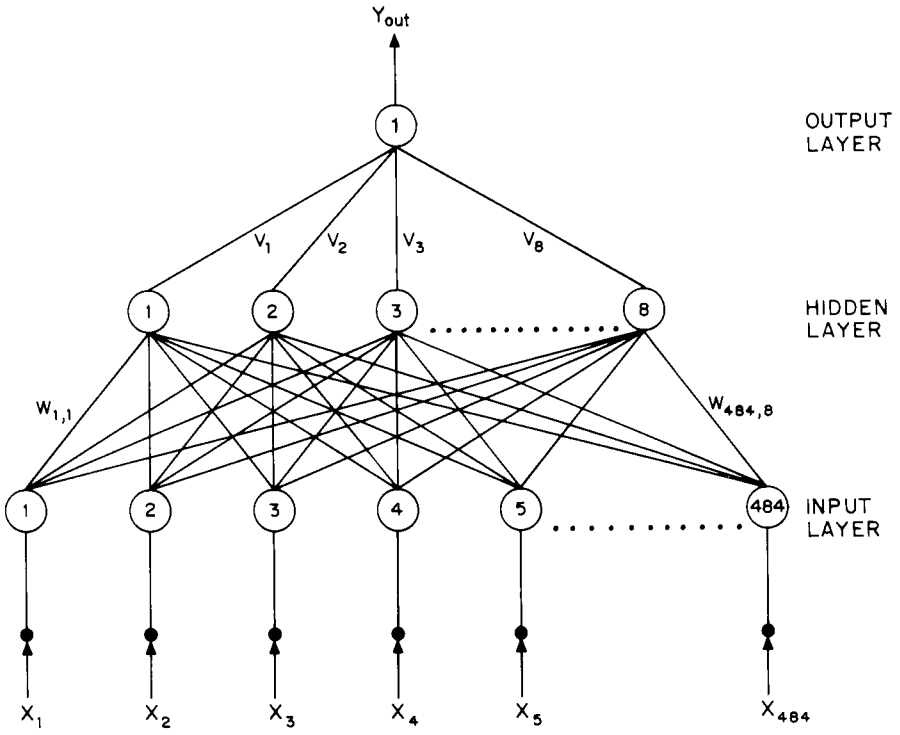

Fig. 1. Architecture of error back-propagation neural network used in the simulation: 484 neurons in the input layer, one hidden layer consisting of eight neurons, and one neuron in the output layer. The trained network approximates $y = f(x)$, where x and y represent the input and output.

namely the summed squared error, defined as:

$$E = \sum_p E_p = \sum_p \sum_i (t_{pi} - outnet_{pi})^2 \qquad (2)$$

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 coverged weights can be used for predicting the outcomes of the test data. The training process can be externally controlled by adjusting the two parameters, namely the learning rate and the momentum factor. The details of computations have been described by others in recent papers and monographs (Rumelhart *et al.*, 1986; Rumelhart and McClelland, 1986; Zupan and Gasterger, 1991, 1993).

Trajectory plots of the upstream region

The plots in Figure 3 were obtained using the Curvature programme (Shpigelman *et al.*, 1993), a generous gift from Drs E.N.Trifonov and E.Shpigelman. The programme uses the nearest-neighbour wedge model to calculate overall DNA path using local helix parameters: helix twist angle, wedge angle and direction (of deflection) angle. All the parameters were estimated from gel electrophoretic data (Bolshoy *et al.*, 1991). All the fragments have been projected on the same plane.

Results and discussion

We have analysed the transcriptional control signals of a eukaryotic protein-coding gene with a view to establish a relation between the site of mutation and its relative importance in the process of eukaryotic gene transcription using a modelling approach. Earlier studies on the effects of promoter substitutions on transcription of the mouse β-major globin gene were determined by transfecting the cloned mutant genes into HeLa cells on plasmids containing an SV40 transcription enhancer, and measuring the levels of correctly initiated β-globin transcripts after 2 days. These studies revealed that muation in three regions, namely the CACCC-box, located between -87 and -95, the CCAAT-box, located between -72 and -77, and the TATA-box, located between -26 and -30 relative to the transcription start site, decreased the level of transcription. While mutations in nucleotides immediately upstream from the CCAAT-box resulted in a 3- to 3.5-fold increase in transcription. These results corroborate the fact that the nucleotide sequences also carry other

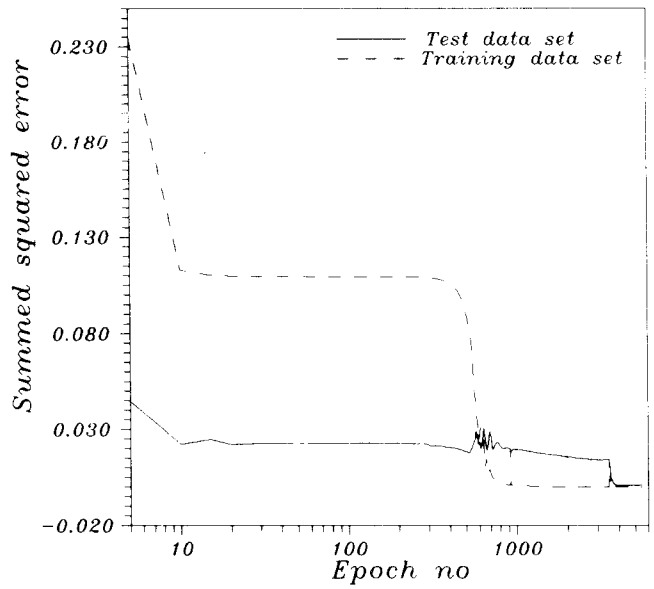

Fig. 2. Error profiles of training and test data set.

information in addition to the triplet code which are the instructions for protein synthesis (Nirenberg *et al.*, 1963; Sepyer *et al.*, 1963).

If we assume that majority of transcription control signals are encoded in the upstream sequences, then the analysis of these sequences should reveal valuable biological information. In order to decipher these signals a neural network has been used to capture the internal representation of these sequences. A net with 484 neurons (sequence length × 4) in the input layer, eight neurons in the hidden layer, and one neuron in the output layer has been used for training purposes. The network in its training phase was presented with 117 patterns. After every epoch, which corresponds to the presentation of all the 117 training patterns once to the net, the weights were extracted and used for predicting the outcomes of test data set containing 12 patterns. The number of neurons in the hidden layer which gave best approximations has been found to be eight. However, using nine neurons or more did not increase the prediction capability of the net, while fewer than eight neurons in the hidden layer hampered its prediction capability. The momentum term was optimized to 0.9 and the learning rate to 0.6. The error profiles of the training and the test data sets corresponding to the optimized network parameters are shown in Figure 2. The weights after 4194 epochs corresponded to the minimum error with respect to the test data set, and hence were taken as optimal. Further training, however, did not improve the prediction capability. The results of the predictions are shown in Table I. It is noteworthy that the network-predicted outputs (RTLs) provide a good fit to the experimental

Table I. Network-predicted values of the relative transcription level

Mutation	Experimental RTL	Network-predicted RTL
−101 C-T	0.93	0.99008
−87 C-A	0.41	0.48265
−81 A-G	0.98	0.96633
−65 C-T	1.10	1.13651
−63 C-T	0.99	0.90017
−58 T-C	0.96	0.98198
−54 G-T	0.95	1.03684
−42 C-A	1.01	1.00160
−38 G-A	1.12	1.04128
−35 G-A	1.11	1.04421
−32 C-G	1.09	0.92708
−27 A-G	0.40	0.49693

RTL values (correlation coefficient > 0.95). The difference in the network-predicted output and the experimental value of RTLs can be accounted for by assuming that additional signals which may also play a role in regulation of transcription may be encoded in other *cis*- and *trans*-acting factors, are not presented to the net. The data set chosen for testing the generalization capability of the net contained mutations spanning almost the entire upstream region, namely −101 to −27, and was not part of the training data set. We would like to emphasize here that the test data set so chosen also contained signals which resulted in both low (−87C-A, RTL = 0.410; −27A-G, RTL = 0.40) and high (−101C-T, RTL 0.930, −38G-A, RTL = 1.1201) levels of transcription. There was no significant change in the prediction capability of the network with different sets of test and training data. The ability of the net to predict both high and low levels of transcription with a very good degree of accuracy essentially points to the fact that the net in its training phase has learned to recognize the signals associated with the upstream region and decipher it in a manner analogous to that by RNA polymerase during the process of transcription.

The trained and validated neural net was then used as a heuristic device to predict the RTL in the case of hitherto unknown mutations associated with the β-globin promoter. The results of the simulations are shown in Figure 4. It is known from earlier studies (Myers *et al.*, 1986) that two different mutations in nucleotides upstream from the CCAAT-box (−79G-A, RTL = 3.5; −78G-A, RTL = 2.9) resulted in a 3.0- to 3.5-fold increase in transcription. Our simulation result verifies this fact. In addition, the results of our simulation obtained by mutating these positions by C and T also indicate an increase in the level of transcription (−79G-C, RTL = 2.043; −79G-T, RTL = 2.090; −78G-C, RTL = 1.276; −78G-T, RTL = 1.713), a fact that has not yet been reported in the experimental study. These positions

Table II. Simulated RTL values in the case of mutation in the conserved regions

Mutation in the CACCC box (−95 to −87)								
*	−95G-T	0.19	*	−95G-A	0.14		−95G-C	0.31
*	−94C-T	0.23		−94C-A	0.39		−94C-G	0.39
*	−93C-T	0.27	*	−93C-A	0.25		−93C-G	0.69
	−92A-T	**0.94**		−92A-G	0.76		−92A-C	0.66
	−91C-T	0.44	*	−91C-A	0.11		−91C-G	0.30
	−90A-T	0.24	*	−90A-G	0.26		−90A-C	0.25
	−89C-T	**1.14**		**−89C-A**	**0.86**		**−89C-G**	**1.03**
	−88C-T	0.30	*	−88C-A	0.26	*	−88C-G	0.87
	−87C-T	0.36	*	−87C-A	0.48	*	**−87C-G**	**0.92**
Mutation in the CCAAT box (−77 to −72)								
	−77C-A	0.36		−77C-T	0.08	*	−77C-G	0.10
	−76C-A	0.30		−76C-T	0.79	*	−76C-G	0.19
	−75A-C	0.26		−75A-T	0.34		−75A-G	0.12
	−74A-C	**0.81**		−74A-T	0.34	*	−74A-G	0.33
	−73T-C	0.43	*	−73T-A	0.22		−73T-G	0.38
	−72C-A	0.71	*	−72C-T	0.17		−72C-G	0.55
Mutation in the TATA box (−30 to −26)								
*	−30T-A	0.37		−30T-G	0.77	*	−30T-C	0.37
	−29A-T	**0.93**	*	−29A-G	0.51		−29A-C	0.79
	−28T-A	**0.95**		−28T-G	0.65		−28T-C	0.66
	−27A-T	0.43	*	−27A-G	0.50		−27A-C	0.52
	26A-T	0.74	*	−26A-G	0.55		−26A-C	0.64

Asterisks denote the mutations that were part of the training data set. Mutations within the conserved region which showed little or no effect on the amount of transcript formed are shown in bold type.

correspond to the region just upstream of the CCAAT-box. Earlier studies on base substitution in these region in promoters of the herpes virus tk gene and β-globin gene (mouse/rabbit) have also shown an increase in the level of transcription (see Table II, Myers *et al.*, 1986). This further adds credence to the simulation results obtained by the network. Furthermore, simulation studies on the three highly conserved regions show that not all mutations in these regions decrease the level of transcription as is generally believed (based on the available experimental results). Table II summarizes the simulation results of all the possible mutations in the three conserved regions (−87 to −95, CACCC-box; −72 to −77, CCAAT-box; −26 to −30 TATA-box). The mutations in these regions which did not seem to affect the efficiency of transcription significantly are in bold type. It can be seen that in certain instances even a marginal increase is observed. Thus the simulation results in general give us an intuitive understanding of the relative importance of certain sequence elements in the process of transcription, within the conserved region and upstream region.

The relationship between DNA curvature and transcriptional activity *in vivo* has been suggested in a number of cases (Perez-Martin *et al.*, 1994). Although most of them are in prokaryotic systems, the same could hold true for eukarotic systems as well (Kerppola and Curran, 1991a,b). A mutation in the upstream region in structural terms means a change of a dinucleotide 'wedge'. Since DNA curvature is facilitated when a dinucleotide residue occurs in phase with the B-DNA helical repeat (10.5 bp/turn), a mutation could alter the superstructure either by severing curvature, or by contributing to it. The superstructures associated with the upstream region were investigated by calculating the DNA path using the Curvature programme (Shpigelman *et al.*, 1993). The overlap of the DNA axis (a curve passing through the centre of the base pairs) in the case of mutations which resulted in very high and very low levels of transcriptional activity are shown in Figure 3. The analysis reveals that a change in the superstructure results in the alteration of transcriptional activity. However, it has also been observed that two similar superstructures may give rise to different levels of transcriptional activity. The results point to the fact that the network is able to capture this relationship between superstructure and transcriptional activity.

At this point it is important to realize the difference between human language and DNA language. The DNA language permits the sequences to encode one or more additional messages. Only few words in the human language can imitate the overlapping of message (as exemplified by Trifonov, 1989). The string '*togetherno-where*' can be read in four different ways: '*together*

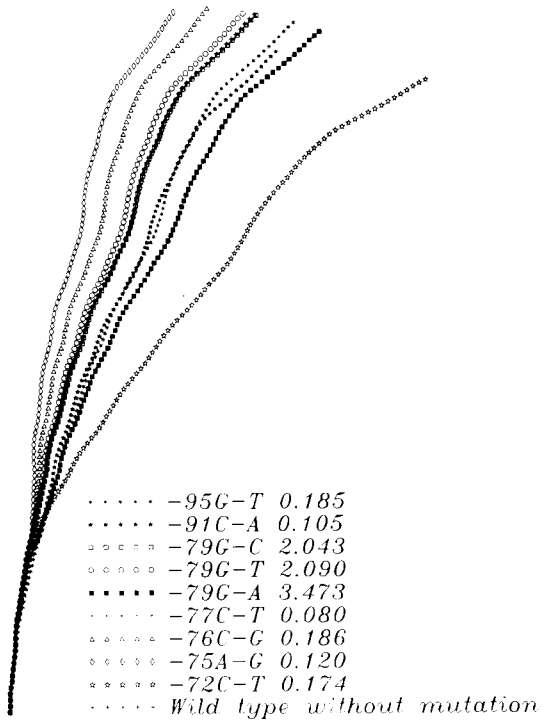

Fig. 3. Overlap of the DNA axis of the upstream region calculated using the Curvature programme. It is the curve passing through the centre of the base pair (see Shpigelman *et al.*, 1994).

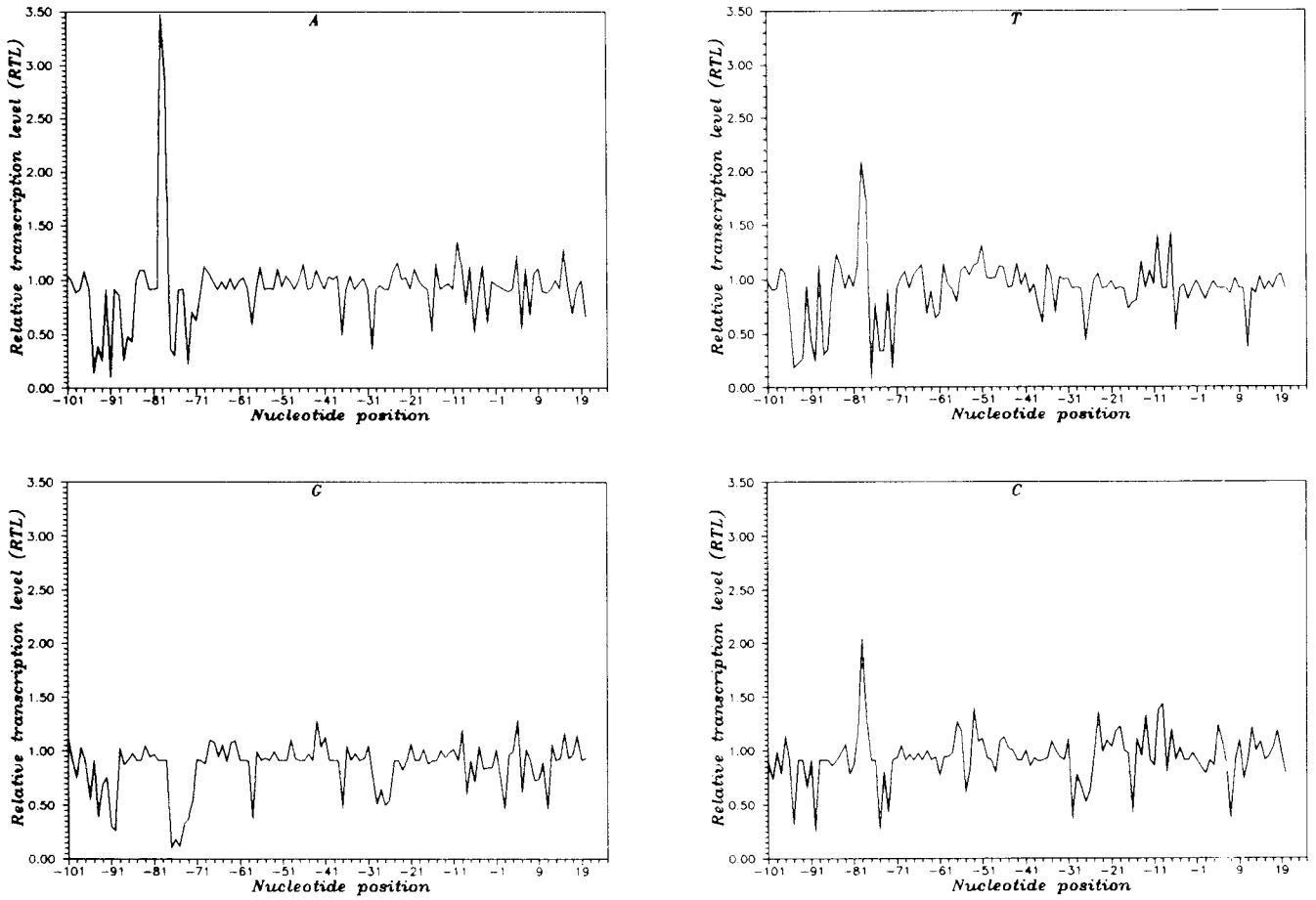

Fig. 4. Simulated values of RTLs for all possible single base substitution in the β-globin promoter obtained by using the converged weights.

nowhere', '*together now here*', '*To get her now here*' and '*To get her nowhere*'. Neural networks are capable of capturing all the possible messages hidden in such an input. They can approximate linear as well as non-linear dependencies and also possess a strong capability of class separation (Gallinari *et al.*, 1988; Asoh *et al.*, 1990; Webb and Lowe, 1990). Thus the net in the present case has been able to capture all the overlapping messages contained in the upstream region which act as signals in the process of transcription. It is worth mentioning here that the eukaryotic transcription is a complex phenomenon which involves an interplay of many factors (proteins) which are not represented in the input. The ability of the network to capture overlapping messages and the results of our *in computo* experiments, evinces the presence of different overlapping signals in the upstream region of the gene that dictate the process of transcription by way of communicating with the many different factors involved. The property of the sequence to encode multiple signals may be thought of as the activity associated with that sequence. Thus it is seen that neural networks are also capable of establishing the sequence–activity relationship.

The ANN model was built using mutation data obtained by saturation mutagenesis of cloned DNA fragments (Myers *et al.*, 1985). The resulting duplex DNA fragments obtained using this method contain random, single base substitutions. In general it is not possible to direct mutation to a particular site using this procedure. In order to obtain a desired mutant, one would have to resort to site-directed mutagenesis (Zoller and Smith, 1982). However, the ANN model that we have built can be used to predict the expression of hitherto unknown mutations. Thus the model also serves as a tool to design mutation experiments.

It is further important to realize that the model is not intended to replace laboratory work, but should be used only as a guide in designing experiments. The results of the predictions are purely theoretical and should be subjected to experimental validation. Furthermore, since the training data set contained a majority of single base substitutions, the ability of the network to predict the outcomes of double mutations or more could not be ascertained. Moreover data of this kind are scarce.

In conclusion, we would like to state that, although it is

a well-established fact that the nucleotide sequences, apart from carrying the triplet code, also carry other information in the form of the sequence pattern, an approach like ANNs can be used in the identification and quantification of such patterns. The simulation results point to the fact that the model can be used in the identification of important sequence elements within conserved regions. The results presented here also establish a sequence–activity relationship. It also underlines the use of a modelling approach in determining the extent of functionality associated with a particular regulatory region. Finally, we would like to suggest that there are many more messages that are lying buried in the sequence and are waiting to be deciphered. Neural networks can be used as tools in decoding these DNA Morse codes. ANNs may not provide a direct solution to the problem but would facilitate in eliminating unnecessary sidetracks, and thus would assist in providing valuable insights into the fundamental mechanisms.

Acknowledgements

T.M.N. thanks Professor Niranjan V.Joshi, CES,IISc, for his critical comments and useful suggestions. T.M.N. also thanks Professors E.N.Trifonov and E.S.Shpigelman for the Curvature programme. The authors thank the referees for their valuable comments on the previous version of the manuscript. This project was supported by the Council of Scientific and Industrial Research (CSIR), India.

References

Asoh,H. and Otou,N. (1990) An approximation of nonlinear discriminant analysis by multilayer neural networks. *Proc. Int. Joint Conference on Neural Networks III*, pp. 211–216.

Bolshoy,A., McNamara,P., Harrington,R. and Trifonov,E.N. (1991) Curved DNA without AA: experimental estimation of all 16 DNA wedge angles. *Proc. Natl. Acad. Sci. USA*, **88**, 2312–2316.

Demeler,B. and Zhou,G. (1991) Neural network optimization for *E. coli* promoter prediction. *Nucleic Acids Res.*, **19**, 1593–1599.

Derman,E., Krauter,K., Walling,L., Weinberger,C., Ray,M. and Darnell,J.E. (1981) Transcription control in the production of liver specific mRNAs. *Cell*, **23**, 731–739.

Efstratiadis,A., Poskony,J.W., Maniatis,T., Lawn,R.M., O'Connell,C., Spritz,R.A., De Riel,J.K., Forget,B.G., Weissman,S.M., Slighton,J.L., Blechi,A.E., Smithies,O., Baralle,F.E., Shoulders,C.C. and Proudfoot,J.N. (1980). The structure and evolution of the human β-globin gene family. *Cell*, **21**, 653–668.

Gallinari,P., Thiria,S. and Soulie,F.F., (1988) Multilayer perceptrons and data analysis. *Proc. Int. Conference on Neural Networks I*, pp. 391–399.

Gidoni,D., Kadonaga,J.T., Barrera-Saldana,H., Takahashi,K., Chambom,P. and Tjian,R. (1985) Bidirectional SV40 transcription mediated by tandem Sp1 binding interactions. *Science*, **230**, 511–517.

Graves,B.J., Johnson,P.F. and McKnight,S.L. (1986) Homologous recognition of a promoter domain common to the MSV LTR and the HSV tk gene. *Cell*, **44**, 565–576.

Hardison,R.C. (1983) The nucleotide sequence of the Rabbit embryonic globin gene β4. *J. Biol. Chem.*, **258**, 8739–8744.

Harpold,M.H., Evans,R.M., Salditt-Georgieff,M. and Darnell,J.E. (1979) Production on mRNA in chinese hamster cells: Relationship of the rate of synthesis to cytoplasmic concentration of nine specific mRNA sequences. *Cell*, **17**, 1025–1035.

Hirst,J.D. and Sternberg,J.E.M. (1992) Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry*, **31**, 7211–7218.

Horton,P.B. and Kanchisa,M., (1992) An assessment of neural network and statistical approaches for prediction of *E. coli* promoter sites. *Nucleic Acids Res.*, **20**, 4331–4338.

Kerppola,T.K. and Curran,T., (1991a) Fos–Jun heterodimers and Jun homodimers bend DNA in opposite orientations: implications for transcription factor cooperativity. *Cell*, **66**, 317–326.

Kerppola,T.K. and Curran,T. (1991b) DNA bending by Fos and Jun: the flexible hinge model. *Science*, **254**, 1210–1214.

Lacy,E. and Maniatis,T. (1980) The nucleotide sequence of a rabbit β-globin pseudogene. *Cell*, **21**, 545–553.

Liebman,M.N., (1992) Application of neural networks to the analysis of structure and function in biologically active macromolecules. In Lim,H.A., Fickett,J.W., Cantor,C.R. and Robins,R.J. (eds), *Proc. 2nd Int Conference on Bioinformatics, Supercomputing and Complex Genome Analysis*. World Scientific, Singapore, pp. 331–347.

Lukashin *et al.* (1989) Neural network models for promoter recognition. *J. Biomol. Struct. Dyn.*, **6**, 1123–1133.

McKnight,S.L., Kingsbury,R., Spence,A. and Smith,M. (1984) The distal transcription signals of the herpesvirus tk gene share a common hexanucleotide control sequence. *Cell*, **37**, 253–262.

McKnight,S.L. and Miller,O.L. (1976) Ultrastructural pattern of RNA synthesis during early embryogenesis of *Drosophila melanogaster*. *Cell*, **8**, 305–319.

McKnight,S.L. and Miller,O.L., (1979) Post-replicative nonribosomal transcription units in *D. melanogaster* embryos. *Cell*, **17**, 551–563.

McKnight,S.L. and Kingsbury,R. (1982) Transcription control signals of a eukaryotic protein coding gene. *Science*, **217**, 316–324.

Meyer,B.J., Maurer,R. and Ptashne,M. (1980) Gene regulation at the right operator (O_R) of bacteriophage λ II. O_R1, O_R2, and O_R3: their roles in mediating the effects of repressor and cro. *J. Mol. Biol.*, **139**, 163–194.

Myers,R.M., Lerman,S.L. and Maniatis,T. (1985) A general method for saturation mutagenesis of cloned DNA fragments. *Science*, **229**, 242–247.

Myers,R.M., Tilly,K. and Maniatis,T. (1986) Fine structure genetic analysis of a β-globin promoter. *Science*, **232**, 613–618.

Nair,T.M., Tambe,S.S. and Kulkarni,B.D. (1994) Application of artificial neural networks for prokaryotic transcription terminator prediction. *FEBS Lett.* **346**, 273–277.

Nirenberg,M.W., Jones,O.W., Leder,P., Clark,B.F.C., Sly,W.S. and Pestka,S., (1963) On the coding of genetic information. *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 549–557.

O'Neil,M.C. (1991) Training back-propagation neural network to define and detect DNA-binding sites. *Nucleic Acids Res.*, **19**, 313–318.

O'Neill,M.C. (1992) *E. coli* promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res.*, **20**, 3471–3477.

Perez-Martin,J., Rojo,F. and De Lorenzo,V. (1994) Promoters responsive to DNA bending: a common theme in prokaryotic gene expression. *Microbiol. Rev.*, **58**, 268–290.

Rawlings,C.J. and Fox,J.P. (1994) Artificial intelligence in molecular biology: a review and assessment. *Phil. Trans. R. Soc. Lond. B*, **344**, 353–363.

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

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

Shavlik,J.W., Towell,G.G. and Noordewier,M.O. (1990) Using neural networks to refine existing biological knowledge. *Int. J. Genome Res.*, **1**, 81–107.

Shpigelman,E.S., Trifonov,E.N. and Bolshoy,A. (1993) Curvature: software for the analysis of curved DNA. *Comput. Applic. Biosci.* **9**, 435–440.

Speyer,J.F., Lengyel,P., Basilio,C., Wahba,A.J., Gardner,R.S. and Ochoa,S. (1963) Synthetic polynucleotides and amino acid Code. *Cold Spring Harbor Symp. Quant. Biol.*, **28**, 559–567.

Sternberg,J.E.M., King,R.D., Lewis,A.R. and Muggleton,S. (1994) Application of machine learning to structural molecular biology. *Phil. Trans. R. Soc. Lond. B*, **344**, 365–371.

Trifonov,E.N. (1989) The multiple codes of nucleotide sequences. *Bull. Math. Biol.*, **51**, 417–432.

Webb,A.R. and Lowe,D. (1990) The optimised internal representation of multilayered classifier networks performs nonlinear discriminant analysis. *Neural Networks*, **3**, 367–375.

Youderian,P., Bouvier,S. and Susskind,M.M. (1982) Sequence determinants of promoter activity. *Cell*, **30**, 843–853.

Zoller,M.J. and Smith,M. (1982) Oligonucleotide directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any fragment of DNA. *Nucleic Acids Res.*, **10**, 6487–6500.

Zupan,J. and Gasteiger,J. (1991) Neural networks: a new method for solving chemical problems or just a passing phase? *Anal. Chim. Acta*, **248**, 1–30.

Zupan,J. and Gasteiger,J. (1993) Neural networks in chemistry. *Angew. Chem. Int. Ed. Engl.* **32**, 503–527.

Received on November 22, 1994; revised on February 15, 1995; accepted on February 20, 1995