ELSEVIER

Computers and Chemistry 24 (2000) 699–711

Computers & Chemistry

www.elsevier.com/locate/compchem

ANN modeling of DNA sequences: new strategies using DNA shape code

Rupali V. Parbhane, Sanjeev S. Tambe, Bhaskar D. Kulkarni *

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

Received 18 November 1999; accepted 8 February 2000

Abstract

Two new encoding strategies, namely, wedge and twist codes, which are based on the DNA helical parameters, are introduced to represent DNA sequences in artificial neural network (ANN)-based modeling of biological systems. The performance of the new coding strategies has been evaluated by conducting three case studies involving mapping (modeling) and classification applications of ANNs. The proposed coding schemes have been compared rigorously and shown to outperform the existing coding strategies especially in situations wherein limited data are available for building the ANN models.

Keywords: DNA sequence analysis; Twist angle; Wedge angle; Direction of the deflection angle; Error-back-propagation algorithm

1. Introduction

In the last decade, artificial neural networks (ANNs) have been extensively used in the analysis of nucleic acid sequences (see review by Nair, 1996); the main reason being their ability of recognizing and classifying patterns not only from the quantitative data but also from the qualitative data such as DNA sequences. These ANN abilities have been used in various classification applications in biological sciences e.g. analysis of *E. coli* promoter structures (Mahadevan and Ghosh, 1994), prokaryotic transcription terminator prediction (Nair et al., 1994), and identification of *E. coli* ribosome binding sites (Bisant and Maizel, 1995). ANNs have also been used in mapping (modeling) applications, for instance, in the analysis of transcription control signals (Nair et al., 1995) and DNA curvature (Parbhane et al., 1998), where the objective was to identify the functional relationship(s) between a DNA sequence and its property.

Of all the different ANN architectures, the one with a multilayered feed-forward structure and trained using the error-back-propagation (EBP) algorithm (Rumelhart et al., 1986) represents the most widely used network paradigm. The EBP network (EBPN) mostly comprises three layers (input, hidden and output) of interconnected neurons (also termed as 'processing elements' or 'nodes') and learns the relationship between its inputs and outputs via a procedure called 'network training'. The peculiarity of the EBP algorithm is that it trains non-linear multilayered networks wherein a non-linear activation function is used for the computation of the outputs of the hidden and/or output nodes. The non-linear neural networks are preferred over the linear ones for modeling high dimensional systems since the input–output relationships in such systems are often non-linear. In a typical ANN-based mapping (classification) application, the network input–output is an appropriately coded DNA sequence and its property (class), respectively. The EBPN training involves minimization of an error function using a steepest descent strategy (see e.g. Rumelhart and McClelland, 1986; Rumelhart et al., 1986; Freeman and Skapura, 1992)

* Corresponding author. Fax: +91-20-5893041.
E-mail address: bdk@che.ncl.res.in (B.D. Kulkarni).

0097-8485/00/$ - see front matter
PII: S0097-8485(00)00072-3

such as the *generalized delta rule* (GDR) wherein the network output is compared with its desired (target) value and the difference (error) is used to iteratively modify the strengths (weights) of the interneuron connections. Network training, following convergence, produces weights that can be considered to be the parameters of the converged ANN model. These weights can then be used to make predictions corresponding to the new DNA sequences, which were not part of the data employed during the development of the network model.

In ANN-based DNA sequence studies, an individual nucleotide of a sequence is represented using three main coding strategies viz., CODE-2, CODE-4 and the EIIP. The first two of these mononucleotide-based coding schemes use binary representation and, therefore, possess purely empirical character. The EIIP code (Nair et al., 1994) on the other hand uses a nucleotide-specific physical property, namely, the Electron Ion Interaction Potential (EIIP) for input coding and, therefore, has a sound theoretical basis. In CODE-2 and CODE-4 approaches, each nucleotide is represented by two (00 = A, 01 = T, 10 = G, and 11 = C) and four (0001 = C, 0010 = G, 0100 = A, and 1000 = T) binary digits, respectively, whereas in EIIP code nucleotides are characterized by their unique EIIP values (0.1260 = A, 0.1335 = T, 0.0806 = G, 0.1340 = C). Thus, CODE-2, CODE-4 and EIIP strategies require two, four and one neurons, respectively, to represent a nucleotide. Since the data requirement to train an ANN increases as the number of neurons in the network increases, the CODE-4 strategy needs maximum number of data points as compared to the CODE-2 and EIIP strategies with the EIIP code needing minimum number of data points. According to a thumb rule, the number of data points required for network training equals the number of network connection weights although reasonably satisfactory results have been obtained with lesser data points. This may be due to the intrinsic dimensionality of the system being much lower than its apparent dimensionality. More often than not, the available training data is insufficient and, hence, schemes requiring fewer neurons to code a nucleotide sequence are desirable. With this objective, we introduce here two coding strategies, namely, the wedge code and the twist code requiring just one value for the representation of a dinucleotide, in the ANN-based modeling of DNA sequences. The performance of the proposed strategies has been tested by conducting three case studies: (i) prediction of DNA curvature; (ii) prediction of the promoter strength of various promoters transcribed by *E. coli* RNA polymerase and (iii) prokaryotic transcription terminator prediction. While the first two case studies are the mapping applications of ANNs, the third one involves an ANN-based classification.

Table 1
Wedge and twist code values for different dinucleotides

Dinucleotide	Wedge code	Twist code
AA	−7.2	−35.62
AC	1.1	34.40
AG	8.4	27.70
AT	2.6	31.50
CA	−3.5	−34.50
CC	−2.1	−33.67
CG	6.7	29.80
CT	−8.4	−27.70
GA	5.3	36.90
GC	5.0	40.00
GG	2.1	33.67
GT	−1.1	−34.40
TA	0.9	36.00
TC	−5.3	−36.90
TG	3.5	34.50
TT	7.2	35.62

1.1. Philosophy of wedge and twist codes

There exist several helical parameters describing the DNA structure (Dickerson et al., 1989a,b) that are based on translation and rotation. In this study, we shall consider the parameters based on the wedge model, which are estimated from the experimental gel retardation data of Bolshoy et al. (1991). The DNA helical parameters characterizing the wedge (deflection) angle (σ), twist angle (Ω) and the direction of deflection angle (δ) are known as DNA shape code. These Eulerian angles are functions of the dinucleotides i.e. adjacent base pairs in a DNA molecule. The dinucleotides AA (5′-AA-3′ on one strand) and TT (on the opposing strand) together form two stacked A·T base pairs so that the wedge and twist angle values are equal for the AA and TT dinucleotides. Similarly, dinucleotide pairs AC and GT, AG and CT, CA and TG, CC and GG, and GA and TC have equal magnitudes for the wedge and twist angles. For a detailed discussion of the specific features of these angles, the reader is directed to Kabsch et al. (1982), Bolshoy et al. (1991), and Shpigelman et al. (1993). To have a unique dinucleotide-specific value for the wedge and twist codes, the sign of the direction of deflection angle can be ascribed to the values of the wedge and twist angles, since the direction angle δ changes its sign for the complementary dinucleotides. The wedge and twist code values obtained thereby are listed in Table 1. Since these codes incorporate the structural and physical properties of dinucleotides, they have a sound theoretical basis and, therefore, can be employed to replace the arbitrary coding strategies such as the CODE-2 and CODE-4. As compared to the EIIP code, which among the existing strategies requires the least number (i.e., one) of input

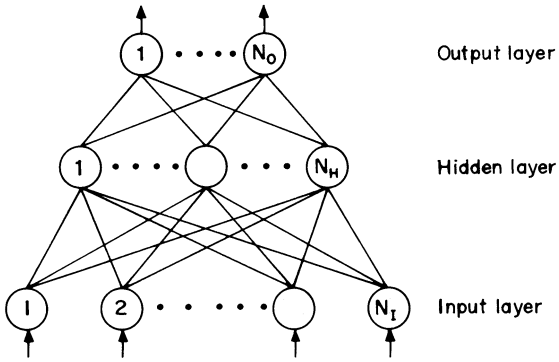

Fig. 1. General architecture of EBPN consisting of N_I, N_H and N_O neurons in the input, hidden and output layers, respectively. Each neuron in the input and hidden layers is connected to all the neurons in the next layer by means of 'weighted' links. In the present study, the input to an EBPN is an appropriately coded DNA sequence and the network output is either a functional property or the class (type) of the input sequence.

neurons to represent a nucleotide, the use of wedge and twist codes reduce the input space of an ANN by half thereby leading to a smaller network and, consequently, requiring a smaller data set for training the network. This paper is organized as follows. Firstly, procedural details of the ANN-based modeling along with the strategies for optimizing the network architecture and weights are outlined. Secondly, the results of three ANN-based case studies wherein the proposed codes have been utilized for the dinucleotide representation are presented. Specifically, the results obtained by using the wedge and twist codes are compared with those obtained using the CODE-4 and EIIP coding strategies. The CODE-2 scheme has not been considered for comparison since the CODE-4 strategy has been found to outperform the CODE-2 strategy (Demeler and Zhou, 1991). The performance of the two new codes is also compared with a dinucleotide-based random strategy wherein the 16 possible dinucleotide combinations are coded by equally spaced real numbers in [0, 1] range as given by: 0.0625 = AA, 0.125 = AC, 0.1875 = AG, 0.25 = AT, 0.3125 = CA, 0.375 = CC, 0.4375 = CG, 0.50 = CT, 0.5625 = GA, 0.625 = GC, 0.6875 = GG, 0.75 = GT, 0.8125 = TA, 0.875 = TC, 0.9375 = TG and 1.00 = TT. In all the case studies, the network training and simulation procedures for the random dinucleotide coding approach are same as that for the wedge and twist codes.

2. Materials and methods

The neural networks considered are three-layered feed-forward type trained using the EBP algorithm. The logistic sigmoid transfer function has been employed at the hidden and also at the output nodes of all the networks. In a situation where sufficient training data are available for network training, all the coding schemes are likely to perform equally well. The efficiency of the proposed codes, therefore, has been tested using limited training data (case studies I and II).

A generalized EBPN architecture for the mapping and classification applications of DNA sequences is shown in Fig. 1. The computer code for training such an EBPN was written in FORTRAN-77 and compiled using the Microsoft FORTRAN compiler for the IBM PC and compatibles. The sequence data used for training the networks in the case studies, the optimal weights obtained thereby and the program for making new predictions, are available on request from the corresponding author.

2.1. Neural network simulation

The neural network simulations were performed on a 486 (66MHz) PC. The error function used during the network training was root-mean-squared-error (RMSE) defined as:

$$\mathrm{RMSE} = \sqrt{\sum_{p=1}^{P} E_p^2 (P \times N_O)} = \sqrt{\sum_{p=1}^{P} \sum_{i=1}^{N_O} (t_{pi} - o_{pi})^2 (P \times N_O)} \quad (1)$$

where index p ranges over the number of input patterns (P); i ranges over the number of output units (N_O); E_p represents the error on pattern p and t_{pi} and o_{pi} are the target and actual output values of the ith output unit when the pth pattern is presented to the network.

Although the objective of network training is to minimize the RMSE with respect to the training set, it does not guarantee that the trained network possesses satisfactory generalization ability. Such an ability ensures that the network is capable of predicting accurately the outputs when new inputs, which do not belong to the training set, are presented to the network. Since the weights resulting in the minimum RMSE for a representative test set ensure satisfactory generalization performance, these are considered to be the optimal weights in practice.

In general, network training (more specifically the RMSEs with respect to the training and test sets) shows sensitivity towards the number of network hidden nodes (N_H) and the GDR parameters, namely, the momentum coefficient (α), and learning rate (η). To obtain the overall optimal weights resulting in the least

RMSE for the test set, several independent training runs were performed by systematically varying the number of hidden nodes and the magnitudes of the GDR parameters (α and η). For each combination of the stated parameters, additionally, the effect of the random number generator seed was examined. This is necessary for studying the effect of the randomly initialized weights whose sequence depends on the seed value of the random number generator. By changing the seed value, a different sequence of random numbers is generated and, consequently, the starting point in the weight space of an ANN gets shifted. This helps in rigorous exploration of the non-linear error surface possessed by the EBP networks.

3. Case study I: prediction of DNA curvature

According to the junction model, the principal sequence feature responsible for the intrinsic DNA curvature is generally assumed to be the runs of adenines. On the other hand, the wedge model of DNA curvature considers that each dinucleotide step is associated with a characteristic deflection of the local helix axis (Bolshoy et al., 1991). It may however be noted that the generality of such first principle models for predicting the curvature is still being debated (Dlakic and Harrington, 1996). Thus, a practical and simpler approach is to develop an empirical model correlating a nucleotide sequence of DNA and its effective curvature. The use of ANNs for developing such models has an advantage in that they can approximate non-linear relationships even between qualitative and quantitative data. Accordingly, this case study aims at developing an ANN model for predicting the curvature of a DNA in terms of its retardation anomaly value, which is a measure of the electrophoretic anomaly of the curved DNA reflecting the additional friction of the DNA in the gel due to curvature (Marini et al., 1982). The relative electrophoretic mobility of most curved DNA fragments monotonously decreases with the fragment length. This is usually characterized as the ratio of the apparent to actual DNA length, and the ratio termed as the 'R_L factor' is found to increase with increasing fragment length. In an earlier study (Parbhane et al., 1998), an ANN-based prediction of the R_L factor using the EIIP code was successfully conducted and the results obtained thereby have been utilized here for comparison purposes.

The data (54 sequences) comprising circular and curved, and straight synthetic fragments and their experimental R_L values were taken from the study by Bolshoy et al. (1991). The choice of such data was based on the consideration that the data set pertains to the most exhaustive experimental gel retardation study of DNA sequences. The respective experiments were carried under standard gel conditions and hence the data is ideal for EBPN training. The sequences are of uneven length that varies between 10 and 42 base pairs. Each sequence forming the network input was coded separately using the dinucleotide-specific wedge, twist and random code values. Since a single wedge/twist/random code value describes a dinucleotide, a sequence say 21 base-pair long, can be coded using ten values. To complete the coding of the entire sequence, the 21st nucleotide was paired with the first one and coded accordingly. All the sequences with odd lengths were analogously coded. For CODE-4 strategy, the sequences were coded using four digit binary numbers as described earlier. It is necessary for the network training that all the input patterns are of the same length. Since the nucleotide sequences are of variable length, the shorter ones (length smaller than 42 base pairs) represented using the wedge/twist/random codes were uniformly padded with a small dummy number (0.01) until each short sequence becomes 21 ($=42/2$) units long. For CODE-4, similar padding was applied till each fragment was 168 ($=42\times4$) units long. This is an indirect way of informing the network that the sequence position valued 0.01 does not belong to either A, T, G or C. The resulting data can be viewed as a matrix of size (54×21) for the wedge/twist/random codes, and of size (54×168) for the CODE-4. Next, each column of the (54×21) matrix was normalized so that each column element upon normalization lies between 0.05 and 0.95. While performing normalization, the padded elements of a sequence were not processed. In order to differentiate between the circular and linear sequences, two additional inputs were considered at the end position of each coded sequence. Specifically, the circular fragments were described as (0.05, 0.90) and the linear ones by (0.90, 0.05). Such an addition of two inputs at the end position of each coded sequence resulted in the data matrix of size (54×23) for the three dinucleotide-based codes and a matrix of size (54×170) for the CODE-4. The experimental R_L values that formed the target output for each input pattern (coded sequence) were also normalized to lie in the [0.05, 0.95] range. Upon normalization, the data set of 54 coded sequences (inputs) and their R_L values (outputs) was divided into the training (40 patterns) and test (14 patterns) sets, respectively (see Table 1 in Parbhane et al., 1998). During network training, the training set is used for adjusting the network weights while the test set is used to evaluate the generalization performance of the network.

The optimal values of the EBPN's structural parameters, GDR parameters, and the RMSE values corresponding to the training and test sets for all the five coding strategies are listed in Table 2. A rigorous statistical analysis has been additionally performed for comparing: (i) the predictions of the five ANN models

Table 2
Details of optimal EBPN architectures and RMSE values corresponding to three case studies

Coding strategy	Case study I: DNA curvature prediction ($\eta = 0.15,\ \alpha = 0.10$)			Case study II: promoter strength prediction ($\eta = 0.3,\ \alpha = 0.15$)			Case study III: prokaryotic transcription terminator prediction ($\eta = 0.5,\ \alpha = 0.9$)		
	N_I:N_H:N_O[a]	RMSE		N_I:N_H:N_O	RMSE		N_I:N_H:N_O	Classification accuracy[b]	
		Training set	Test set		Training set	Test set		Training set	Test set
CODE-4	170:1:1	0.320	0.098	280:1:1	0.244	0.147	204:7:1	99.43	98.12
EIIP	44:1:1	0.055	0.051	70:1:1	0.237	0.146	51:7:1	96.59	95.62
Wedge	23:1:1	0.072	0.064	35:1:1	0.032	0.036	26:1:1	95.17	92.50
Twist	23:1:1	0.074	0.069	35:1:1	0.006	0.050	26:1:1	95.45	90.00
rnd di[c]	23:1:1	0.071	0.072	35:1:1	0.016	0.138	26:1:1	76.70	75.00

[a] N_I, number of input neurons; N_H, number of hidden neurons; N_O, number of output neurons; η, learning rate; α, momentum coefficient.
[b] Percentage of correctly classified sequences.
[c] rnd di denotes random dinucleotide coding scheme.

with the experimental R_L values, and (ii) the predictions of a combination of ANN models, wherein all possible model combinations have been considered. In here, apart from computing the correlation coefficient (r_{xy}) values, we have performed the Z-test (for large sample size, i.e. the number of points, $n > 30$) and the F-test. The procedures for the Z- and F-tests are described in Appendix A. The purpose of performing these tests, in essence, is to answer the query 'how significant are the differences between the means and variances of the R_L predictions made by two coding strategies, namely, x and y?' The r_{xy} values along with the results corresponding to the Z- and F-tests are tabulated in Table 3.

4. Case study II: prediction of promoter strength

A promoter is a start signal at the beginning of a gene or a gene cluster that directs RNA polymerase to initiate RNA synthesis. RNA polymerase measures the efficiency of transcription in terms of the promoter strength that refers to the relative rate of synthesis of the full-length RNA product from a given promoter. The transcription efficiency of a given promoter sequence is regulated by many factors such as: (i) nucleotide sequence of the -35 region; (ii) nucleotide sequence of the -10 region; (iii) spacing between the -35 and -10 regions and (iv) nucleotide sequence especially A + T content in the 5′-flanking region upstream from the -35 region (McClure, 1985). The additive rule states that the individual contributions of nucleotide sequence spacer length, deoxyribonucleic acid (DNA) conformation, and electrostatic binding within a promoter, collectively establish the total promoter strength. It can thus be noticed that a number of factors influence the strength of a promoter. Owing to the difficulties in the experimental evaluation of the stated contributing factors, it is advantageous to build a promoter strength prediction model that does not require explicit knowledge of the various factors influencing the transcription efficiency. With this objective, we have examined the efficacy of the wedge and twist codes vis-a-vis CODE-4, EIIP and random dinucleotide codes for the ANN-based prediction of the promoter strength.

For this study, an EBPN was trained using the experimental data by Deuschle et al. (1986), where in vivo promoter strengths of the various promoters transcribed by *E. coli* RNA polymerase have been determined. The data set comprising 14 promoter sequences and their corresponding strengths was divided into training (ten patterns) and test (four patterns) sets, respectively (refer to Table 5). In these data, all but one of the promoter sequences are 70 nucleotides long; the remaining one is 69 nucleotides long. For ANN modeling, the sequences were coded using the wedge, twist and random code values specified earlier. For coding the 69-nucleotide long promoter sequence, the last nucleotide was paired with the penultimate nucleotide, i.e. from the group of three nucleotides (AAG) at the sequence end, two dinucleotide pairs (AA and AG) were formed, and coded accordingly. To make all the input vectors of same size, the 69 base-pair long sequence was uniformly padded with 0.1 till it was 280 ($= 70 \times 4$) units long for the CODE-4 scheme and 70 ($= 70 \times 1$) units long for the EIIP code. The resulting data can be viewed as a matrix of size (14×35) for the wedge, twist and random dinucleotide codes and, ma-

Table 3
Statistical analysis of different combinations of sample sets comprising experimental and network predicted R_L values

No.	Sample set of R_L values	$\bar{X}$	$\bar{Y}$	s_x^2	s_y^2	r_{xy}	Z_c [a]	F_c [b]
1	x = expt, y = code4	1.05	1.06	0.021	0.015	0.877	−0.386	1.40
2	x = expt, y = eiip	1.05	1.05	0.021	0.019	0.954	−0.111	1.09
3	x = expt, y = wedge	1.05	1.06	0.021	0.016	0.931	−0.209	1.30
4	x = expt, y = twist	1.05	1.06	0.021	0.016	0.920	−0.305	1.31
5	x = expt, y = rnd di	1.05	1.05	0.021	0.016	0.890	−0.167	1.28
6	x = code4, y = eiip	1.06	1.05	0.015	0.019	0.901	0.274	0.78
7	x = code4, y = wedge	1.06	1.06	0.015	0.016	0.892	0.186	0.93
8	x = code4, y = twist	1.06	1.06	0.015	0.016	0.890	0.081	0.94
9	x = code4, y = rnd di	1.06	1.05	0.015	0.016	0.867	0.230	0.91
10	x = eiip, y = wedge	1.05	1.06	0.019	0.016	0.920	−0.095	1.20
11	x = eiip, y = twist	1.05	1.06	0.019	0.016	0.911	−0.193	1.21
12	x = eiip, y = rnd di	1.05	1.05	0.019	0.016	0.906	−0.052	1.18
13	x = wedge, y = twist	1.06	1.06	0.016	0.016	0.989	−0.103	1.00
14	x = wedge, y = rnd di	1.06	1.05	0.016	0.016	0.968	0.044	0.98
15	x = twist, y = rnd di	1.06	1.05	0.016	0.016	0.963	0.147	0.98

[a] $Z_\alpha = 2.33$ at $\alpha = 0.01$, and $Z_\alpha = 1.64$ at $\alpha = 0.05$.
[b] $F_{53,53,0.01} = 1.60$ for $n_x = n_y = 54$ at $\alpha = 0.01$, and $F_{53,53,0.05} = 1.39$ at $\alpha = 0.05$.

Table 4
Comparison of different coding strategies using leave-k-out cross-validation method

Coding strategy	Case study I ($k=9$, $\eta=0.15$, $\alpha=0.10$)			Case study II ($k=2$, $\eta=0.3$, $\alpha=0.15$)		
	N_I:N_H:N_O	Average RMSE		N_I:N_H:N_O	Average RMSE	
		Training	Test		Training	Test
CODE-4	170:1:1	0.161	0.046	280:1:1	0.148	0.107
EIIP	44:1:1	0.137	0.094	70:1:1	0.112	0.081
Wedge	23:1:1	0.112	0.091	35:1:1	0.013	0.033
Twist	23:1:1	0.102	0.074	35:1:1	0.006	0.045
rnd di	23:1:1	0.159	0.098	35:1:1	0.124	0.055

trices of sizes (14×280) and (14×70) for the CODE-4 and EIIP code, respectively. Each column element of the (14×35) and (14×70) matrices was normalized such that it lies between 0.05 and 0.95 upon normalization. The values of the experimental promoter strength that formed the target output for each input pattern were also normalized to lie in the [0.05, 0.95] range.

The five networks utilizing different coding schemes were rigorously trained and optimized as described earlier. The details of the optimized network architectures and the GDR parameters are listed in Table 2. The table also gives the RMSE values corresponding to the training and test sets for the five coding schemes.

As in case study I, a rigorous statistical analysis has been conducted by employing the Student's t-test (for small sample size, i.e. when the number of data points $n < 30$) and the F-test. The procedure for Student's t-test has been described in Appendix A.

5. Case study III: prokaryotic transcription terminator prediction

Terminators are sequences that primarily regulate the gene expression by providing stop signals at the end of transcription units and, thus, allowing adjacent genes and/or operons to be transcribed and regulated independently (von Hippel et al., 1984). Studies have shown that the 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 has been witnessed that many independent terminators do not comply with the consensus pattern of the dyad symmetry and T-stretch (Brendel and Trifonov, 1984) and, therefore, conditions for termination are not well defined. It is thus important to develop methods for identifying (classifying) the terminators comprising inconsistent consensus patterns. ANNs utilizing the CODE-4 and EIIP formalisms have been already found to be successful in this task (Nair et al., 1994). Our objective in the present case study is to examine the classification efficiency of the wedge and twist codes vis-a-vis CODE-4, EIIP and the random dinucleotide coding schemes. Towards this objective, three network models utilizing wedge, twist and random codes have been developed and their classification performance is compared with the CODE-4 and EIIP code results obtained by Nair et al. (1994).

The terminator sequences for the ANN simulations were taken from the compilation by Brendel et al. (1986). From a total of 128 terminators of length 51 nucleotides, 88 were chosen for training the network and the remaining 40 were used as the test data. A pseudo-random number generator was used for constructing the random sequences with equal compositions of A, T, G and C. These random sequences were combined with the terminator sequences in 1:3 ratio. The resulting 352 patterns formed the training set inputs; the test set inputs (160 patterns) were constructed analogously. Since the length of terminator sequences is an odd number (51 nucleotides), the last nucleotide was paired with the penultimate nucleotide of the same sequence and coded accordingly. Subsequently, the column elements of the resulting matrices of size (512×26) were normalized to lie between 0.05 and 0.95. In this case study, the target output equal to one represents a terminator sequence, and the target output of zero refers to a random (non-terminator) sequence.

The three networks utilizing the wedge, twist and random dinucleotide input coding schemes were rigorously optimized following the procedure described earlier. The details of the optimized network structures and the GDR parameters along with the percentage classification accuracy for all the coding schemes can be found in Table 2.

6. Results and discussion

6.1. Case study I

The statistical Z-test checks whether or not the mean values of two large samples drawn from respective populations are statistically different. In the present context, a sample refers to a set of the R_L values either determined experimentally or those predicted by each of the five ANN models. In essence, the Z-test verifies the validity of the null hypothesis (H_0) that the difference in the means (μ_x and μ_y) of two populations is statistically insignificant. It can be noted (see Table 3) from the Z-statistic values (Z_c) corresponding to the 15 different combinations of x and y samples that the absolute value of Z_c is less than both $Z_{0.01}$ ($=2.33$) and $Z_{0.05}$ ($=1.64$). Thus, we may accept the null hypothesis, H_0 (with 1 and 5% levels of significance), that the differences in the respective μ_x and μ_y values are insignificant.

The F-test is meant for testing whether there exists a statistically significant difference between the variance values (σ_x^2 and σ_y^2) of two populations. When the F-test is used on the samples consisting of variance values of the experimental and CODE-4 predicted R_L values, it is seen (see Table 3, row 1, column 9) that the absolute value of F_c ($=1.40$) is less than $F_{53,53,0.01}$ ($=1.60$), but greater than $F_{53,53,0.05}$ ($=1.39$). Hence, we may accept: (i) the null hypothesis (H_0), that σ_x^2 is equal to σ_y^2 at 1% level of significance and, (ii) an alternative hypothesis (H_1), that σ_x^2 is greater than σ_y^2, at 5% level of significance. For the rest fourteen combinations of samples x and y, the absolute values of the F_c are smaller than both $F_{53,53,0.01}$ and $F_{53,53,0.05}$ and, therefore, we may accept H_0 at both 1 and 5% levels of significance.

From Tables 2 and 3, it can be noticed that the RMSE and r_{xy} values for the wedge, twist, random dinucleotide and EIIP codes are comparable, although the last one fares marginally better than the two new coding strategies. On the other hand, the RMSE values (0.32, 0.098) corresponding to the training and the test set of CODE-4 are the highest among five coding schemes. Also, the magnitude of the coefficient of correlation ($r_{xy}=0.87$) between the experimental and CODE-4 based network predicted R_L value is the lowest among all coding schemes. These trends suggest that the CODE-4 is the least efficient of the five input coding strategies for the ANN-based prediction of R_L. This is consistent with the F-test results where it was observed that the variances in respect of the experimental and CODE-4 based network predicted R_L values are different at 5% level of significance. The result indicates that the CODE-4 based model has not captured the variations in the experimental R_L values with statistically significant accuracy. It can be also noticed from the r_{xy} values listed in Table 3 (column 7, entries 3, 4 and 5) that the wedge and twist codes perform better, albeit marginally, than the random dinucleotide code. These wedge and twist code results essentially indicate that the codes possess good potential as sequence coding schemes since both the strategies resulted in relatively high r_{xy} values (≥ 0.92) and low RMSE values (≤ 0.069) for the test set. Also, the mean (1.06) and variance (0.016) values associated with the R_L predictions of the ANNs using these codes are statistically consistent with the mean (1.05) and variance values (0.021) of the experimental R_L values.

While coding the DNA sequences in this case study, the effect of overlapping dinucleotides was not taken into account though it is well known that the curvature of a sequence depends on the overlapping dinucleotides. Such a simplified coding approach though leaves out half of the relevant information contained in a sequence, was used still with a view of keeping the complexity of the coding procedure to a bare minimum. To check whether this simplification has any effect on the prediction accuracy of the trained network, we performed a control study for the networks utilizing the wedge and twist codes. In here, the first nucleotide was removed from each DNA sequence (Table 1 in Parbhane et al., 1998), and the remaining portion of the sequence was coded using wedge and twist codes. The resultant input patterns are different from those wherein the first nucleotide was retained during sequence coding. These input patterns were then used to repredict the R_L values for which the optimal weights obtained originally were utilized. It was observed that the repredicted R_L values match their desired (experimental) values with the same accuracy as obtained when first nucleotide was considered for the input coding. The correlation coefficient for the experimental and repredicted R_L values for the wedge and twist codes were found to be 0.93 and 0.924, respectively, which almost match those listed in Table 3 (0.931 and 0.92). It can thus be inferred from the results of control simulations that it is not essential in ANN-based R_L prediction studies to account explicitly for the overlapping dinucleotides.

While partitioning the available data (54 patterns), a care was exercised that the 14 examples in the test set are the true representatives of the 40 examples in the training set. It is however essential to verify whether the available data was adequate at all for effecting the said partition. Accordingly, 'cross-validation' simulations were performed using the leave-k-out methodology. In this approach, the entire set of available data is randomly divided into N subsets each comprising k patterns. Next, the network is trained N times using each subset in turn as the test set with the remaining $(N-1)$ subsets collectively representing the training set. Upon completing this exercise, the RMS errors corresponding to the training and test sets are averaged; the mean

Table 5
Listing of various promoters transcribed by *E. coli* RNA polymerase and their in vivo promoter strengths expressed in P_{bla} units

No.	Promoter	Promoter strength
1	P_{H207}	55 (4)
2	$P_{D/E20}$	56 (8)
3	P_{N25}	30 (5)
4	P_{G25}	19 (2)
5	P_{J5}	9 (1)
6	P_{A1}	76 (9)
7	P_{A2}	20 (4)
8	P_{A3}	22 (3)[a]
9	P_{L}[b]	53 (8)[a]
10	P_{lac}	5.7 (0.5)
11	P_{lacUV5}	3.3 (0.3)[a]
12	P_{tacI}	17 (2)[a]
13	P_{con}	4 (0.2)
14	P_{bla}	1

[a] Promoter sequences that were part of the test set.
[b] Promoter strength taken from Knaus and Bujard (1988).

RMSE in respect of the test set gives an estimate of the overall network performance that could be achieved if more data were available for the network training.

For performing the above-described cross-validation simulations, the available data of 54 DNA sequences and their corresponding R_L values were partitioned into six subsets ($N = 6$, $k = 9$). The results of the cross-validation simulations in respect of the five coding schemes are presented in Table 4. A comparison of the test set RMSE values listed in Tables 2 and 4 indicates that the cross-validation results are better only in the case of CODE-4 scheme. This result suggests that the available data of 54 patterns was adequate for all the network models except the one using the CODE-4 coding scheme. The result is a natural consequence of the CODE-4 scheme producing largest (as compared with other codes) sized networks, thus needing more training data.

6.2. Case study II

In this case study also, a rigorous statistical analysis was performed on the promoter strengths predicted by the five ANN models. The results of the Student's t- and F-tests conducted thereby on the sample sets comprising experimental and ANN-predicted promoter strengths are tabulated in Table 6. It is noted from the various Table 6 entries that for all the 15 different combinations of x and y samples, the absolute values of t_c are less than $t_{2\alpha}$ ($= 1.315$) and $t_{2\alpha}$ ($= 1.706$), which correspond to 1% ($\alpha = 0.01$) and 5% ($\alpha = 0.05$) levels of significance, respectively. Thus, we may accept the null hypothesis (H_0) that the mean values (μ_x and μ_y) of the respective populations are statistically equal in all the 15 combinations of x–y sample sets at 1 and 5% levels of significance.

The F-statistic (F_c) values (see column 9) corresponding to the two combinations of x and y involving experimental promoter strengths and those predicted by the CODE-4 and EIIP based networks indicate that the respective F_c magnitudes (4.96 and 8.52) are greater than both $F_{13,13,0.01}$ ($= 2.42$) and $F_{13,13,0.05}$ ($= 3.59$). This result in essence suggests that the variance value (534.99) in respect of the experimental promoter strengths is greater (at 1 and 5% significance levels)

Table 6
Statistical analysis of different combinations of sample sets comprising experimental and network predicted promoter strength values

No.	Sample set of promoter strength values	$\bar{X}$	$\bar{Y}$	s_x^2	s_y^2	r_{xy}	t_c[a]	F_c[b]
1	x = expt, y = code4	26.50	26.20	534.991	107.783	0.631	0.043	4.96
2	x = expt, y = eiip	26.50	26.21	534.991	62.779	0.753	0.042	8.52
3	x = expt, y = wedge	26.50	27.12	534.991	475.569	0.994	−0.070	1.12
4	x = expt, y = twist	26.50	26.61	534.991	521.070	0.995	−0.012	1.03
5	x = expt, y = rnd di	26.50	29.23	534.991	509.097	0.969	−0.304	1.05
6	x = code4, y = eiip	26.20	26.21	107.783	62.779	0.326	−0.004	1.72
7	x = code4, y = wedge	26.20	27.12	107.783	475.569	0.604	−0.138	0.23
8	x = code4, y = twist	26.20	26.61	107.783	521.069	0.588	−0.059	0.21
9	x = code4, y = rnd di	26.20	29.23	107.783	509.098	0.590	−0.440	0.21
10	x = eiip, y = wedge	26.21	27.12	62.779	475.569	0.772	−0.141	0.13
11	x = eiip, y = twist	26.21	26.61	62.779	521.069	0.759	−0.058	0.12
12	x = eiip, y = rnd di	26.21	29.23	62.779	509.098	0.703	−0.454	0.12
13	x = wedge, y = twist	27.12	26.61	475.569	521.069	0.992	0.058	0.91
14	x = wedge, y = rnd di	27.12	29.23	475.569	509.098	0.978	−0.242	0.93
15	x = twist, y = rnd di	26.60	29.23	521.069	509.098	0.972	−0.294	1.02

[a] $t_{2\alpha} = 1.315$ at $\alpha = 0.01$, $t_{2\alpha} = 1.706$ at $\alpha = 0.05$.
[b] $F_{13,13,0.01} = 2.42$ for $n_x = n_y = 14$ at $\alpha = 0.01$, $F_{13,13,0.05} = 3.59$ at $\alpha = 0.05$.

than the variance values, 107.78 and 62.78, corresponding to the predictions of the CODE-4 and the EIIP code based networks. Since the absolute values of F_c for the remaining 13 combinations of x and y samples are always less than $F_{13,13,0.01}$ ($=2.42$) and $F_{13,13,0.05}$ ($=3.59$), we may accept the null hypothesis (H_0) that the respective variances are equal at both 1 and 5% levels of significance.

As can be noticed from Table 2, the RMSE values for the test sets of the wedge and twist coded networks are the lowest and the second lowest, respectively. Also, the r_{xy} magnitudes (refer to Table 6, column 7, entries 3 and 4) for the predictions made by the wedge and twist code based networks are very high ($\cong 1$). These results suggest that the networks utilizing the two codes have near-accurately approximated the relationship between a DNA sequence and its promoter strength. In comparison, the prediction performance of CODE-4 ($r_{xy}=0.63$) and EIIP ($r_{xy}=0.75$) strategies is very poor. This conclusion is consistent with the F-test results where it was observed that the sample variances of the experimental, and CODE-4 and EIIP based network predicted promoter strength values are different at both 1 and 5% levels of significance. The result indicates that the CODE-4 and EIIP based models have not captured the variations in the experimental promoter strength values with statistically significant accuracy. Among the three dinucleotide coding schemes, the random dinucleotide coding approach ($r_{xy}=0.96$) performs only marginally worse than the other two (wedge and twist) schemes. A plausible explanation for such a behavior is: since the random code — unlike wedge and twist codes — does not explicitly take into account any DNA sequence dependent property or characteristic (such as the curvature), it fails to predict with comparable accuracies.

A graphical comparison of the experimental and the network predicted promoter strengths (P_{bla} units) for the training and test sets of the wedge code is shown in Fig. 2(a and b), respectively, wherein for clarity the promoter strengths are arranged in the descending order of their magnitudes. A similar comparison for the twist code is depicted in Fig. 2(c and d).

The cross-validation test was performed for this case study also wherein the available data of 14 patterns was partitioned into seven ($N=7$) subsets each comprising two ($k=2$) patterns. The results of the cross-validation simulations using the 'leave-2-out' scheme are given in Table 4. A comparison of the cross-validation results with those in Table 2 for the test set indicates that the RMSE values corresponding to the cross-validation simulations are lower for all the codes. This suggests that the prediction performance of all the five networks can improve further if more data are available for training the networks. It can however be inferred from the approximately equal RMSE values for the wedge (0.036 and 0.033) and twist (0.05 and 0.045) codes (see Tables 2 and 4) that such an improvement, though possible, can only be marginal. In essence, the results of this case study indicate that the dinucleotide coding schemes fare better than the mononucleotide based schemes (CODE-4 and EIIP). The results corresponding to the wedge and twist codes are important in the sense that even under extreme paucity of the training data, the two new coding strategies have performed significantly better than the existing ones.

6.3. Case study III

In this case study, which examines the performance of wedge and twist codes for classification applications, the accuracy of classification is defined as the percentage of correctly classified input patterns; for a given input sequence, the network output in [0.5, 1.0] range signifies a terminator, otherwise it is regarded as a random sequence. The network utilizing the random dinucleotide code was found to possess poorest classification accuracy as it could correctly classify only 120 (75%) of the 160 test patterns and 270 (76.7%) of the 352 training patterns. On the other hand, the wedge and twist code based networks could correctly classify 148 (92.5%) and 140 (90%) test patterns, and 335 (95.17%) and 336 (95.45%) training patterns, respectively. Although the classification accuracy of the wedge and twist codes for the test patterns is reasonably good, it is lower than that obtained using the EIIP (95.62%) and CODE-4 (98.12%) schemes. In the classification study by Nair et al. (1994), a similar observation has been made where it was found that the CODE-4 strategy fares better than the EIIP code. The higher classification accuracy of CODE-4 was attributed to the larger EBP network size, which means larger parameter space as compared with the EIIP code. This explanation also holds when the classification accuracies corresponding to the CODE-4 and EIIP schemes are compared with those of the wedge and twist codes. It can thus be observed from Table 2 that as the size of the network's input space decreases (CODE-4 > EIIP code > wedge/twist codes), the classification accuracy for the test patterns decreases accordingly (98.12 > 95.62 > 92.5/90%). Notwithstanding this observation, it is important to note that the performance of the wedge and twist coding schemes is still acceptable since on an average 91.25% of the test patterns have been correctly classified.

7. Concluding remarks

In this paper, two input coding strategies namely, wedge code and twist code have been introduced for representing dinucleotides in the ANN-based modeling

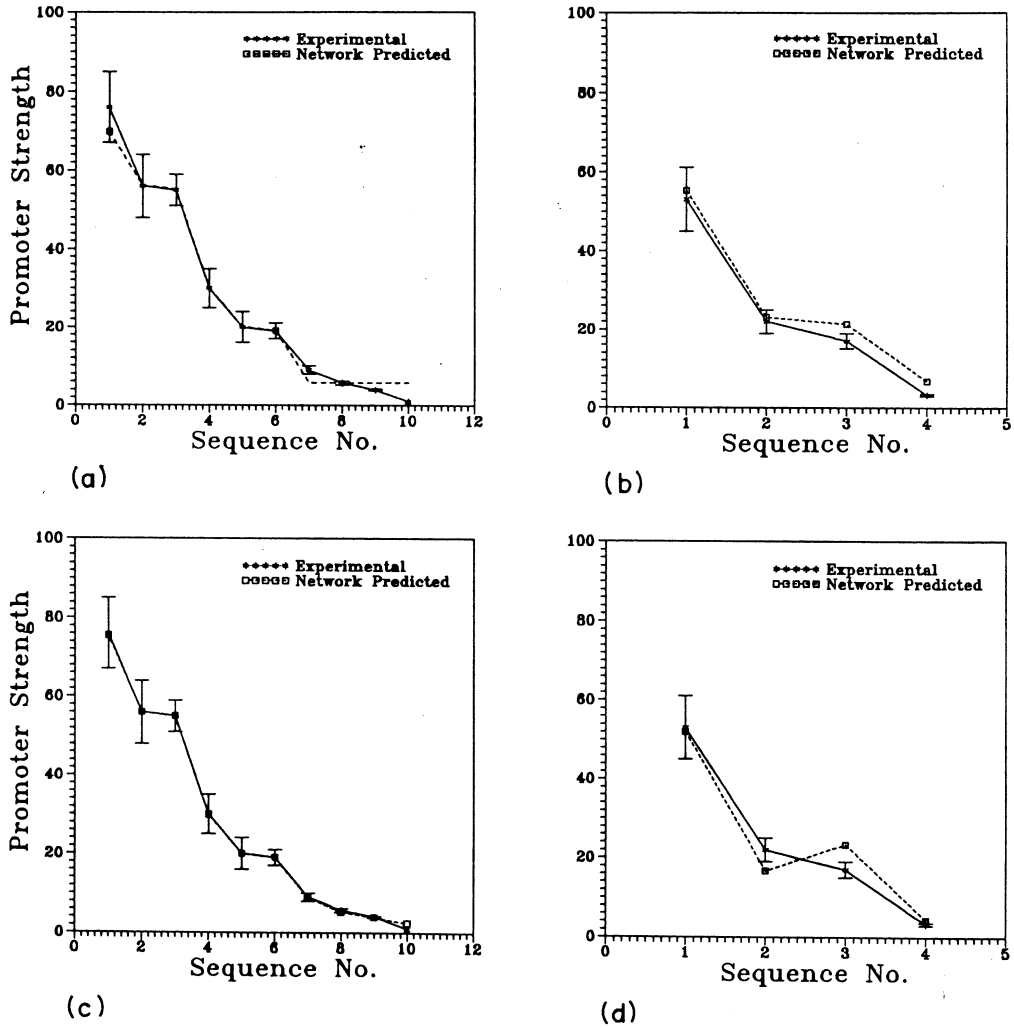

Fig. 2. Graphical comparison of the experimental and network predicted promoter strength (P_{bla} units) values using: (a) wedge code for the training data set; (b) wedge code for the test data set; (c) twist code for the training data set, and (d) twist code for the test data set.

of DNA sequences. These codes make use of the helical parameters namely, the wedge angle, twist angle, and the direction of deflection angle of a DNA. The principal advantage of the new coding strategies over the commonly used mononucleotide-based coding schemes such as CODE-4 and EIIP, is that they reduce the network's input dimensionality to one-eighth as compared with the CODE-4 strategy, and to one-half as compared with the EIIP scheme. Consequently, a smaller network that can be trained faster results. Such a network i.e. possessing less adaptable parameters (weights), in general possesses better generalization capability than the network with more parameters. The efficiency of the proposed strategies vis-a-vis other input coding schemes namely, CODE-4, EIIP and random dinucleotide code, has been evaluated by conducting three case studies involving ANN-based mapping and classification applications. In all the case studies, both the proposed coding strategies have been found to perform equally well. Also, the proposed codes have been found to perform better than the conventional strategies especially when the training data was limited (case studies I and II). In these studies, although the CODE-4 scheme that results into large input dimensionality did not perform well, the proposed codes with smaller input dimensionality have lead to some significant results. This feature of the proposed schemes is important since for many real systems the available data are often limited and generation of additional data can be an involved and costly task. It has been also observed that the networks using the wedge and twist codes fare better (i.e. yield higher

correlation coefficient magnitudes and classification accuracy) than the networks using the random dinucleotide code. Such a superior performance may be attributed to the DNA shape related property i.e. the helical parameters of a DNA used by the wedge and twist codes. Since the proposed codes are sufficiently general, they can also be used for representing DNA sequences in 'non-ANN-based' mapping and classification applications. The present work has also opened up a new gateway for tri- and tetra-nucleotide based DNA coding strategies.

Acknowledgements

RVP is supported by Senior Research Fellowship from Council of Scientific and Industrial Research (CSIR), New Delhi.

Appendix A. The computational procedures for evaluating *Z*-, *F*- and the Student's *t*-statistics

A.1. Z-test (for large sample, i.e. when the number of data points, n > 30)

This test, also known as the Normal test, checks whether the difference between two population means is statistically significant. In this test, Z-statistic (Z_c) is computed to test the null hypothesis (H_0): the means μ_x and μ_y of two populations are equal (i.e. $\mu_x = \mu_y$), against an alternative hypothesis, H_1: $\mu_x > \mu_y$. The Z_c is evaluated as:

$$Z_c = \frac{\bar{X} - \bar{Y}}{\sqrt{s_x^2/n_x + s_y^2/n_y}} \tag{A1}$$

where $\bar{X}$ and $\bar{Y}$ are the means of population samples x and y, respectively; s_x^2 and s_y^2 refer to the variances of x and y, respectively and n_x, n_y denote the respective sample sizes. The decision rule for the Z-test at α% level of significance is given as:

If $|Z_c| \geq Z_\alpha$, then reject H_0; otherwise accept H_0.

A.2. F-test

Similar to the Z-test for two means, the F-test is performed to check the validity of hypothesis involving two population variances (σ_x^2 and σ_y^2). The F-statistic (F_c) is computed as given below to validate the null hypothesis (H_0): $\sigma_x^2 = \sigma_y^2$, against an alternative hypothesis (H_1): $\sigma_x^2 > \sigma_y^2$:

$$F_c = \frac{s_x^2/n_x}{s_y^2/n_y} \tag{A2}$$

The decision rule for the F-test at α% level of significance and for $(n_x - 1)$, $(n_y - 1)$ degrees of freedom is:

If $|F_c| \geq F_{(n_x-1),(n_y-1),\alpha}$, then reject H_0; otherwise accept H_0.

A.3. Student's t-test (for small sample size, i.e., n ≤ 30)

In an event when the sample size is small ($n \leq 30$), Student's t-test is performed to check the validity of the null hypothesis (H_0): $\mu_x = \mu_y$, against an alternative hypothesis (H_1): $\mu_x > \mu_y$. The corresponding t-statistic (t_c) is evaluated as:

$$t_c = \frac{\bar{X} - \bar{Y}}{s\sqrt{1/n_x + 1/n_y}} \tag{A3}$$

where:

$$s = \sqrt{\frac{n_x s_x^2 + n_y s_y^2}{n_x + n_y - 2}} \tag{A4}$$

Note that the test statistic t_c follows Student's t distribution with $(n_x + n_y - 2)$ degrees of freedom. The decision rule for the t-test at α% level of significance is:

If $|t_c| \geq t_{2\alpha}$, then reject H_0; otherwise accept H_0.

References

Bisant, D., Maizel, J., 1995. Nucleic Acids Res. 23, 1632.
Bolshoy, A., McNamara, P., Harrington, R., Trifonov, E., 1991. Proc. Natl. Acad. Sci. USA 88, 2312.
Brendel, V., Hamm, H., Trifonov, E., 1986. J. Biomol. Struct. Dyn. 3, 705.
Brendel, V., Trifonov, E., 1984. Nucleic Acids Res. 12, 4411.
Demeler, B., Zhou, G., 1991. Nucleic Acids Res. 19, 1593.
Deuschle, U., Kammerer, W., Gentz, R., Bujard, H., 1986. EMBO J. 5, 2987.
Dickerson, R.E., et al., 1989a. J. Mol. Biol. 205, 787.
Dickerson, R.E., et al., 1989b. EMBO J. 8, 1.
Dlakic, M., Harrington, R.E., 1996. Proc. Natl. Acad. Sci. USA 93, 3847.
Freeman, J., Skapura, D., 1992. Neural Networks Algorithms, Applications, and Programming Techniques. Addison-Wesley, Reading, MA.
Kabsch, W., Sander, C., Trifonov, E.N., 1982. Nucleic Acids Res. 10, 1097.
Knaus, R., Bujard, H., 1988. EMBO J. 7, 2919.
Mahadevan, I., Ghosh, I., 1994. Nucleic Acids Res. 22, 2158.
Marini, J.C., Levene, S.D., Crothers, D.M., Englund, P.T., 1982. Proc. Natl. Acad. Sci. USA 19, 7664.
McClure, W.R., 1985. Annu. Rev. Biochem. 54, 171.
Nair, M., Tambe, S., Kulkarni, B., 1994. FEBS Lett. 346, 273.
Nair, M., Tambe, S., Kulkarni, B., 1995. Comp. Appl. Biosci. 11, 293.
Nair, M., 1996. In: Tambe, S.S., Kulkarni, B.D., Deshpande, P.B. (Eds.), Elements of Artificial Neural Networks with Selected Applications in Chemical Engineering and Chemi-

cal and Biological Sciences, Simulation and Advanced Controls, Louisville, KY, pp. 395–437.
Parbhane, R., Tambe, S., Kulkarni, B., 1998. Bioinformatics 14, 131.
Rumelhart, D., Hinton, G., Williams, R., 1986. Nature 323, 533.
Rumelhart, D., McClelland, J., 1986. Parallel and Distributed Processing: Explorations in the Microstructure of Cognition. MIT, Cambridge, MA.
Shpigelman, E., Trifonov, E., Bolshoy, A., 1993. Comp. Appl. Biosci. 9, 435.
von Hippel, P.H., Bear, D.G., Morgan, W.D., McSwiggen, J.A., 1984. Annu. Rev. Biochem. 53, 389.