CHAPTER 6

NEURAL-STATISTICAL MODEL OF TATA-BOX MOTIFS IN EUKARYOTES

Haiyan Wang
Max-Planck Institute for Molecular Genetics
whyinsa@yahoo.com

Xinzhong Li
Imperial College
xinzhong@doc.ic.ac.uk

Vladimir B. Bajić
Institute for Infocomm Research
bajicv@i2r.a-star.edu.sg

The TATA-box is one of the most important binding sites in eukaryotic Polymerase II promoters. It is also one of the most common motifs in these promoters. The TATA-box is responsible mainly for the proper localization of the transcription start site (TSS) by the biochemical mechanism of DNA transcription. It also has very regular distances from the TSS. Accurate computational recognition of the TATA-box can improve the accuracy of the determination of the TSS location by computer algorithms. The conventional recognition model of the TATA-box in DNA sequence analysis is based on the use of a position weight matrix (PWM). The PWM model of the TATA-box is widely used in promoter recognition programs. This chapter presents a different, nonlinear, recognition model of this motif, based on a combination of statistical and neural network modelling. The resulting TATA-box model uses "statistical filtering" and two LVQ neural networks. The model is derived for a sensitivity level that corresponds to approximately 67.8% correct recognition of TATA motifs. The system is tested on an independent data set used in the evaluation study by Fickett and Hatzigeorgiou, and it performs better in promoter recognition than three other methods, including the one based on the matching score of the TATA-box PWM of Bucher.

ORGANIZATION.

Section 1. We first give some background about promoter recognition via the recognition of the TATA-box.

Section 2. Then we provide a statistical analysis of TATA motifs and their surroundings.

Section 3. Next we describe the use of the LVQ ANNs in modelling TATA motifs.

Section 4. Finally, we present a new multistage LVQ ANN system that models the TATA motifs. The results of application of this system to promoter recognition are also presented.

1. Promoter Recognition via Recognition of TATA-Box

Extracting new biological knowledge in a computational manner from recorded biological sequence databases is one of the key issues of bioinformatics.[26] One of the currently most important general problems in bioinformatics is the annotation of uncharacterized biological sequences. It presumes finding and locating functionally active segments of biological sequences. Once these are found, their positions, functions, and other relevant information are recorded and stored in databases.

At present, the primary target of annotation of sequences originated from eukaryotes is the location of protein coding genes.[123, 127, 166, 247, 304, 587] However, correctly recognizing the starting and ending points of different genes is not a simple task, and methods for this purpose are not sufficiently accurate yet.[123, 127, 166, 247, 304, 334, 587] The starting end of genes can be more precisely determined through the location of promoters, since promoters are usually located before the respective gene, so that recognizing a promoter allows for a more precise determination of the gene's 5' end.[38, 249, 306]

Computational promoter finding has received more attention in the last decade.[38, 43, 44, 46–48, 116, 150, 168, 190, 211, 249, 262, 321, 330, 377, 383, 439, 446, 535, 549, 565, 626, 627, 629, 630, 674, 679, 681, 690, 691, 703, 753, 784, 796, 901, 931] Current knowledge of promoter functionality relies on extensive experimental work.[249, 656] In simplified terms, the promoter region of protein coding genes of eukaryotes—shortly called eukaryotic Pol II promoters—represents a section of DNA to which RNA Polymerase II enzyme and different transcription factors (TFs) bind, forming the so-called transcription preinitiation complex that makes the initiation of the DNA transcription possible. Different transcription factors bind to different subsections of the promoter region. These docking sites, called transcription factor binding sites, are recognized by transcription factors via the bio-chemical machinery of the cell. One of the most important transcription factor binding sites in eukaryotes is the so-called TATA-box.[75, 116, 173, 314, 624, 661, 773, 889]

The eukaryotic promoters are far more complex that the prokaryotic ones, and possess very individual structures which are not common to large groups of promoters. It is thus difficult to design a general algorithm for recognition of eukaryotic promoters. There exists a number of computer programs to aid in promoter identification; see recent overviews by Fickett and Hatzigeorgiou[249] and Prestridge.[681]

Although much progress has been achieved in recent years in developing such general type promoter-recognition software, the general conclusion of the study by Fickett and Hatzigeorgiou[249] is that the performances of the publicly available promoter recognition programs are not very satisfactory. That study has demonstrated that on the specific test set these programs recognize the existing promoters in the range between 13% to 54%, producing false positives in the range of about 1/460bp down to 1/5520bp. The general tendency is an increased number of false predictions as the number of correct predictions increases. A recently reported result[753] that aims at a low level of false recognition recognizes only 29% of the true promoters on the same test set, although it makes only 1 false prediction per 2547bp.

Although there are a number of transcription factor binding sites in eukaryotic Pol II promoters—such as Inr, CCAAT-box, and GC-box—that are shared among larger promoter groups, the most common transcription factor binding site among them seems to be the TATA-box. Some estimates are that it is present in 70%–80% of eukaryotic Pol II promoters.[681] This motivates using the recognition of TATA-box as a part of eukaryotic promoter prediction algorithms.[44, 116, 439, 549, 679, 784]

Recognition of this element is heavily dependent on finding good matches to the TATA-like motif.[116, 661] This is frequently done by using the position weight matrix (PWM) description[322, 795, 801, 803, 805] of the TATA-box.[116] The very first attempts of using the PWM models in biological sequence analysis are in modeling the transcription start site and translation initiation site in *E. coli*.[322, 805] However, the recognition performance based on PWM score only is not very good and too many false recognitions are produced if a high sensitivity level of recognition is required. It thus makes sense to search for a better characterization of the TATA motif, as it may improve recognition accuracy of both the TATA-box and the promoter.

One such possible approach is presented in this chapter. It should be mentioned that it is a general opinion[38] that the recognition of eukaryotic promoters cannot be made very accurate if it relies on the detection of the presence of only one of the transcription factor binding sites—even if that is the TATA-box—and that other promoter characteristics or binding sites should be used simultaneously in promoter recognition algorithms. Although we are aware of this, we still test the new neural-statistical model of the TATA motif in the recognition of eukaryotic Pol II promoters. Our result outperforms three other previously tested programs, including the one based on the original TATA-box PWM from Bucher.[116]

We develop in this chapter a statistical nonlinear characterization of TATA-box motifs of eukaryotic Pol II promoters. Development of this model is based on the representation of nucleotides by the electron-ion-interaction potential (EIIP),[857, 858] the PWM of the TATA motif,[116] domain characteristics of neighboring segments of the TATA-like motifs, positional information of the motif, and Artificial Neural Network (ANN) based modeling. The new model is obtained

- by finding regularities in the DNA segments around the TATA-box and in the distribution of the TATA-box motifs based on the PWM matching score,
- by developing a new system for TATA motif recognition that combines the LVQ ANNs[441, 443, 445] augmented with statistical analysis, and
- by using a genetic algorithm (GA)[254] to optimize the initial weights of the LVQ ANNs in an attempt to improve the prediction accuracy of the model.

2. Position Weight Matrix and Statistical Analysis of the TATA-Box and Its Neighborhood

In this section we develop some results based on the PWM of TATA motifs, and combine these with the properties of local neighborhoods of TATA-boxes. These are used together with information on motif position in subsequent sections to support TATA motif modelling by ANNs. The model development in this section is backed by the statistical and biological regularities of the TATA motif and local surrounding regions.

2.1. *TATA Motifs as One of the Targets in the Search for Eukaryotic Promoters*

Promoters perform a crucial role in the initiation of the DNA transcription process. They indicate and contain the starting point of transcription near its 3’ end. Eukaryotic Pol II promoters contain numerous different transcription factor binding sites that are not always present in each of the promoters, whose location with respect to the transcription start site (TSS) may change in different promoters, as may also their ordering. These facts result in a huge number of micro-organizational combinations of these functional regions, and consequently in a large number of possible eukaryotic promoter structures, which is also reflected in the high complexity of eukaryotic organisms. Hence, it is difficult to create a unique eukaryotic promoter model precisely.

The recognition of TATA-box is frequently used as a part of computer algorithms[44, 116, 439, 549, 211, 679, 784] that aim at recognizing eukaryotic Pol II promoters. Other binding sites are also used in many solutions.[150, 151, 259, 260, 439, 446, 549, 678–680, 683, 689, 705, 784] This provides the motivation for improving the quality of models of these different short DNA motifs, so that their recognition becomes more accurate. Our interest in this study is focussed on the TATA motif as it is a frequent binding site in eukaryotic Pol II promoters. The TATA-box is a hexamer sequence with a consensus given by TATAAA, though it is more accurate to describe it with a PWM.[116]

The TATA motif is found in a majority of protein coding genes, with the position of its first nucleotide generally falling in the range of –25 to –35 upstream from the TSS and centered about position –30; see Figure 1. In the transcription process, the TATA-box serves as the binding site to TFIID[656] and, when it is present in the promoter, it is responsible for the correct determination of the TSS by the transcription machinery.[656]

2.2. *Data Sources and Data Sets*

The data used for building a model of the TATA motif is collected from some public nucleotide databases. All extracted sequences were from vertebrate organisms. For the core promoter regions the data source was the Eukaryotic Promoter Database (EPD),[115, 662] which contains an annotated collection of experimentally mapped TSS and surrounding regions. We have extracted 878 vertebrate core promoter sequences. All promoter sequences are of length 43bp from position –45 to –3 relative to the TSS. The location of the TSS is assumed to be between the nucleotide at position –1 and the nucleotide at position +1. In this notation there is no nucleotide in position 0 and the first transcribed nucleotide is the one at position +1. The other two group of sequences—belonging to the exon and intron regions—are extracted from the GenBank database, and they are also taken from vertebrate organisms. All exon and intron sequences are divided into non-overlapping segments of length 43bp.

From the extracted promoter sequences, we randomly select 640 which make up the training set P_{tr} for promoters, while the promoter test set P_{tst} contains the remaining 238 promoter sequences. From the set of exon sequences we randomly select 8000 sequences and this set is denoted as S_{cds}. Analogously, from the set of intron sequences, another 8000 sequences are randomly selected to form the set S_{int}. We used the whole S_{cds} set as the negative training set, and the S_{int} set as a negative test set.

2.3. *Recognition Quality*

Since we are developing a new model of TATA-box motifs, we also have to test the quality of recognition of the TATA-like sequences based on this new model. Thus we have to use proper measures to express the success of such recognition.

There are several measures that can be used to assess the quality of motif recognition. The four customary numbers TP, TN, FP, and FN, denoting “true positive”, “true negative”, “false positive”, and “false negative” recognition respectively, are generally used in different combinations to describe the basic recognition quality. These four numbers are expressed as percentages in this chapter. However, it is often more convenient to use measures expressed by a single number that relate to the recognition quality. For a recent review on these, see Bajic.[50]

Any way, during model development, we use the correlation coefficient (CC) to measure the success of prediction, although this measure does not sufficiently penalize the large FP prediction. However, for the final assessment of the quality of the developed model we use the average score measure (ASM)[50] so as to be able to make the relevant comparison of the promoter prediction results obtained by the new TATA motif model and the results obtained by other algorithms.

2.4. *Statistical Analysis of TATA Motifs*

The DNA sequence is composed of 4 different nitrogenous bases denoted in the customary way as “A”, “T”, “C”, and “G”, relating to adenine, thymine, cytosine, and guanine respectively. Before any analysis of a DNA sequence is made, we convert it to a numeric sequence by replacing each of the nucleotides by its electron-ion-interaction potential (EIIP)[857, 858]. After that, the values of EIIP for A and T are increased by tenfold. The collection of the original values of EIIP for C and G, and the modified EIIP values for A and T, are denoted as the modified EIIP values (MEIIPVs).

Given an uncharacterized genomic DNA sequence, the goal is to recognize the TATA-box motif by moving along the sequence. If accurate recognition of the TATA-box can be made, then one can also make a recognition of the promoter. In this study, promoter recognition is made only on the basis of the detected presence of TATA motifs. The TATA-box is modeled based on the PWM determined in Bucher[116] from a set of 502 unrelated eukaryotic Pol II promoters. In order to better represent the core TATA motif, we also use some features of the short regions 6bp before and 6bp after the core motif.

Thus, due to the overlapping of the PWM length and the three hexamer regions, the data window analyzed in the search for the TATA motif is only 20bp wide. This information is finally combined with the position information of the TATA motif relative to the TSS. All these data serve as input information to the ANN system developed in the subsequent sections.

2.4.1. *PWM*

Computer models of specific biological signals normally form part of the tools for nucleotide sequence analyses. Since a large amount of data is available in databases, proper statistical analysis is possible in many cases and can help in developing predictive models of specific biological signals. A very convenient method for such modeling is based on the position weight matrix (PWM) of specific motifs.[322, 795, 801, 803, 805] This has found broad and very successful applications in many motif search computer programs.

PWM is a statistical motif descriptor. It is derived from the base-frequency matrices that represent the probabilities of a given nucleotide occurring at a given position in a motif. Since it basically relates to probabilities, the PWM attempts to describe in a statistical fashion some characteristics of the motif found in a collection of sequences. For the proper determination of the PWM it is important that enough data is available.

Bucher[116] has enhanced the basic algorithm for the determination of the PWM by introducing an optimization criterion based on a measure of local over-representation. It is possible to estimate in this way the cut-off matching score to the weight matrix, as well as its width and location of the preferred region of occurrence. Due to the way that it is determined, the PWM for DNA motifs are represented by a rectangular matrix that has 4 rows, where each row corresponds to one of the 4 bases, and its number of columns is equal to the length of the motif. The PWM of the TATA-box motif from Bucher[116] is of size 4×15, giving the motif length of 15 nucleotides, with the start of the core motif at column 2 of the PWM.

This PWM can be used to scan a sequence for the presence of the TATA motif. A window of length 15 nucleotides slides along the sequence. The matching score for the motif is calculated based on the nucleotides found and the PWM. The matching score for the window is given by

$$x = \sum_{i=i}^{15} w_{b_i,i} \tag{1}$$

where $w_{b_i,i}$ is the weight of base b_i at the ith position of the motif, and b_i is the ith base of the sequence in a window; b_i relates to A, C, G, or T. To determine which of the matching scores suggests the presence of the TATA-like motifs, we compare the matching score with the threshold τ. If $x > \tau$, where the cut-off value τ is usually determined either experimentally or statistically, then the data window suggests the presence of the TATA-box.

2.4.2. *Numerical Characterization of Segments Around TATA-Box*

DNA is an organized high capacity biochemical structure—a macromolecule—that stores information. It thus seems logical that the nucleotides in the DNA sequence are ordered following certain regularities. The mechanism of recognition of binding sites in promoters by transcription factors is not completely known. However, one may assume that a binding site and its immediate surrounding regions—the region before the binding site and the region after the binding site—have some properties that enable easy recognition of the site by the respective transcription factors.

Based on this assumption, we expect certain regularities in the case of the TATA-box, in the segments before and after the core TATA motif, which enable easier recognition of the motif. We also expect that these regularities for the functional TATA motifs are different from those relating to non-functional ones.

Consequently, we focus our attention on the distribution of the bases A, C, G, and T around TATA motifs. To describe the potential features of the segments around the TATA-box, we consider three hexamer segments of DNA associated with this motif in the direction from the 5' toward the 3' end of the sequence: The segment S_1 immediately before the TATA-box, the segment S_2 representing the TATA-box hexamer, and the segment S_3 representing the hexamer immediately after the core TATA-box. All three segments are 6bp in length each. However, in addition to these three hexamer we also use the PWM of the TATA motif to calculate the matching score. Since the core TATA motif is represented in the PWM starting from position 2, we need to consider in the modeling process a subregion of the total length 20bp relating to the TATA motif . We use certain characteristics of these three segments in the development of our TATA-box model.

Consider a sequence s of length 43 nucleotide positions. Assume it contains a TATA-box. Let AV_1, AV_2, and AV_3 denote the average of the MEIIPVs of the nucleotides contained in the segments S_1, S_2, and S_3 respectively. Let e_j, $j = -45, -44, ..., -3$, represent the MEIIPV of each of the nucleotides in the sequence s. Further assume that the 5' end—the starting nucleotide—of the TATA-box is at the position i relative to the TSS. Then AV_1, AV_2, and AV_3 are given by

$$AV_1 = \frac{1}{6} \times \sum_{k=1}^{6} e_{i-k}, \;\; AV_2 = \frac{1}{6} \times \sum_{k=0}^{5} e_{i+k}, \;\; AV_3 = \frac{1}{6} \times \sum_{k=6}^{11} e_{i+k} \qquad (2)$$

2.4.3. *Position Analysis of TATA Motifs*

In order to obtain initial information on the location of the TATA motif, we scan each sequence from P_{tr} using a window of 15 nucleotides. The first window starts at position –40 relative to the TSS, and the last one starts at position –17. We calculate according to Equation 1 the matching scores for each of the windows with the PWM for the TATA-box as determined by Bucher.[116] Then we assume that the position of the window with the maximum matching score for a sequence corresponds to the window which contains a hypothetical TATA motif. This computational determination of putative TATA motifs is not the best one since the highest score does not imply biological relevance. But in the absence of experimentally determined TATA motifs for all sequences in P_{tr}, we adopt this unified way of determining the putative motifs. It is also assumed that at most one of the TATA motifs can exist in any of the sequences from P_{tr}. With a selected threshold of $\tau = 0.45$, a total of 475 sequences from P_{tr} are extracted. This amounts to about 75% of the sequences in P_{tr} which can be regarded as those that contain the TATA-box. The extracted sequences form the set P_{TATA}.

The sequences from P_{TATA} are clustered into different groups G_i, according to the distance of the 5' end of the TATA-box hexamer from the TSS. In the PWM of Bucher[116] the first nucleotide of the TATA-box hexamer is in position 2, so that the detected TATA motifs by using the PWM begin from position –39 to –16, making in total 24 possible positions for the distribution of the motif. Let $G_i, i = -39$, -38, ..., -16, denote the group of promoter sequences containing the TATA-box hexamers which all start at position i. Thus, $P_{TATA} = \bigcup_{i=-39}^{-16} G_i$. The number of sequences $N(G_i)$ in different G_i's is not the same. The total number of the computationally found TATA-box motifs in P_{tr} with the threshold $\tau = 0.45$—*i.e.*, the number of sequences in the set P_{TATA}—is $N_{TATA} = \sum_{i=-39}^{-16} N(G_i) = 475$.

The distribution of the putative TATA motifs according to the position of their starting nucleotide with respect to the TSS is shown in Figure 1. This distribution can be approximately modelled by

$$Y = 0.192 \times e^{-0.125 \times (i+30)^2}$$

where Y is the probability that the TATA hexamer starts at position $i = -3$, ..., -16. This model distribution is also depicted in Figure 1 as the dotted curve. Note that the model distribution is based on the P_{tr} promoter set and, in principle, may change if the set is changed. However, it resembles well the previously observed positional distribution of TATA-like motifs in eukaryotic Pol II promoters.

Figure 1 indicates that the starting point of the TATA-box hexamer in the promoter set P_{tr} concentrates around the position –30. The middle point of the TATA-box hexamer is by convention considered to be represented by the second T in the

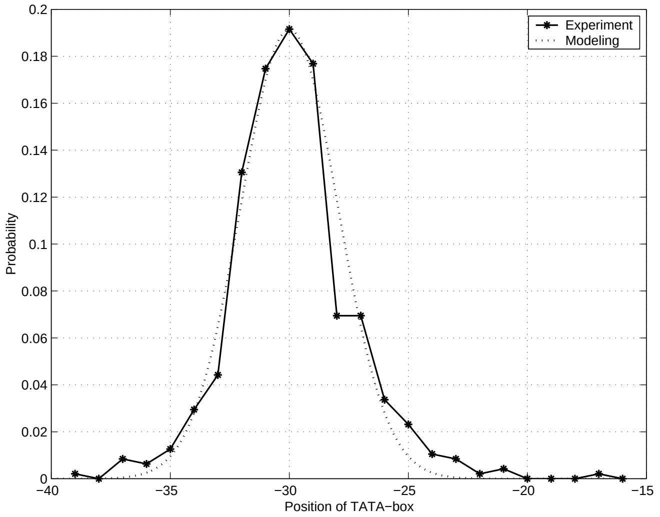

Fig. 1. Experimental and modelled distributions of TATA-box motifs found in R_{tr} based on PWM matching score greater than the threshold $\tau = 0.45$.

TATAAA consensus. This implies that the middle point of the motif is concentrated around position –28. This is close to experimental results, which suggests the location of the TATA-box to be about –25bp to –30bp upstream of the TSS. The distribution of the TATA-box can be regarded as one of the features of the motif, as it expresses a very regular statistical property.

2.4.4. *Characteristics of Segments S_1, S_2, and S_3*

We want to find some regularities that stem from the positional distributions of the TATA motif. For this purpose we use the P_{TATA} set of promoter sequences, and a part of the sequences from S_{cds} that are selected as follows. For each of the sequences from S_{cds} the maximum matching score is obtained with the PWM for TATA-box motifs. Then if such score is greater than $\tau = 0.45$, the sequence passes the filter, otherwise it is rejected. We obtain 1040 filtered sequences from S_{cds} in this way to make the set S^f_{cds}. Further, we find for each sequence from $P_{TATA} \cup S^f_{cds}$, based on the TATA PWM, the maximum matching score among the scores that correspond to the positioning of the 5' end of the TATA hexamer at –39, ..., –16, relative to the TSS. Each such position is considered to determine a putative TATA motif. Then we calculate the respective AV_1, AV_2, and AV_3 values of the segments S_1, S_2, and S_3 using MEIIPVs.

The following observations follow from the analysis of AV_i values:

- there are ranges of AV_i values produced only by sequences from P_{TATA};
- analogously, there are ranges of AV_i values produced only by sequences from S^f_{cds}; and
- some ranges of AV_i values contain a majority of data produced by either P_{TATA} or by sequences from S^f_{cds}.

These observations are used in analyzing the distributions of values of two additional parameters $D_1 = AV_2 - AV_1$, and $D_2 = AV_2 - AV_3$, that describe the contrasts between the neighboring segments S_1 vs. S_2, and S_2 vs. S_3 respectively. The observations about the ranges of numerical parameters AV_1, AV_2, AV_3, as well as the derived parameters D_1 and D_2, suggest that restricting the values of these parameters can serve as a sort of filtering of TATA-containing sequences, and thus could be used in possible recognition of the TATA motifs and consequently the promoters.

In order to illustrate the possible effects of such filtering on TATA motif (and promoter) recognition, we slide a window of 20 nucleotides in length along each sequence in P_{tr}, P_{tst}, S_{cds}, and S_{int}. For each position of the window, the AV_1, AV_2, and AV_3, as well as D_1 and D_2, are calculated based on the content from positions 1 to 18 within the window. The PWM matching score for TATA motif is calculated based on content from position 6 to 20 of the window.

If the AV_2, D_1, and D_2 values fall into pre-selected data ranges, and the PWM score is higher than the selected threshold τ, the sequence is considered as a TATA-box containing sequence (and thus a promoter). The position of S_2 for which this is observed relates to the position of the found TATA motif. Such filtering possibilities, by restricting the values of D_1, D_2, and AV_2, are illustrated in Figure 2, where the results of promoter recognition via the TATA motif recognition are given by using some ranges of D_1, D_2, and AV_2 values, together with the threshold $\tau = 0.45$ of the PWM matching score, as a filtering criteria.

For example, Figure 2 indicates that $FP = 4.8\%$ is obtained on the set of S_{int} and $FP = 2.6\%$ on the set of S_{cds}, when $TP = 58.1\%$ on P_{tr} and $TP = 45.4\%$ on P_{tst}. The results in Figure 2 are not all that good by themselves. Nevertheless, they illustrate the positive effect of restricting the values of numerical parameters D_1, D_2, and AV_2. This is utilized in combination with the other information in the construction of our TATA motif model.

Case	Value Intervals		$P_{tr}(\%)$	$P_{tst}(\%)$	$S_{cds}(\%)$	$S_{int}(\%)$
	D_1	D_2				
1	$[0.3, 1.3]$	$[0.39, 1.2]$	58.1	45.4	2.6	4.8
2	$[0.5, 1.3]$	$[0.43, 1.2]$	43.1	33.2	1	1.5
3	$[0.5, 1.3]$	$[0.58, 1.2]$	41.4	31.5	0.51	0.55
4	$[0.7, 1.3]$	$[0.6, 1.2]$	22.4	15.1	0.3	0.19
5	$[0.95, 1.3]$	$[-0.1, 1.2]$	13.1	6.7	0.025	0
6	$[0.97, 1.3]$	$[-0.1, 1.2]$	11.1	5.8	0	0
$AV_2 \in [1.07, 1.097] \cup [1.275, 1.4]$ for all cases						

Fig. 2. Recognition of promoters based on attempted recognition of the TATA motif by constraining numerical parameters that characterize S_1, S_2, and S_3 and using threshold $\tau = 0.45$

2.5. *Concluding Remarks*

Up to now we have described some statistical regularities of TATA-like motifs in eukaryotic Pol II promoters. These properties are based on the PWM and are associated with certain data ranges obtained by the analysis of numerical representation of the TATA-box hexamer and its immediate neighboring hexamers. Finally, strong regularities are also detected in the positional distribution of the motif. These regularities reflect different aspects of the statistical, and partly biological, features of the TATA-containing eukaryotic Pol II promoter sequences.

Based on these regularities one can derive a set of information data for each promoter sequence as follows. For a promoter sequence from P_{tr} of length 43bp spanning positions –45 to –3, we slide a window of length 15 nucleotides from the 5' end toward the 3' end of the sequence. We calculate 24 matching scores p_i against the TATA-box PWM, associated with the starting positions of the window $i = -40$, -39, ..., -17, relative to the TSS. For the sequence, the window that achieves the maximum score is considered to be most likely the one which contains the TATA-box. If the position of the first nucleotide in the window with the maximum matching score p_i is i, then the position of the first nucleotide in the TATA-box hexamer is $j = i + 1$.

The following 8 data are then generated and associated with the detected

TATA-box hexamer

$$\left.\begin{array}{ll} x_1 = \max_i p_i & x_2 = \dfrac{j+40}{24} \\ x_3 = 0.192 \times e^{-0.125\times(j+30)^2} & x_4 = AV_1 \\ x_5 = AV_2 & x_6 = AV_3 \\ x_7 = AV_2 - AV_1 & x_8 = AV_2 - AV_3 \end{array}\right\} \quad (3)$$

These data reflect the statistical and distributional information of the TATA-box sequence and will be used in the training of ANNs for TATA motif and promoter recognition.

3. LVQ ANN for TATA-Box Recognition

There are obviously many approaches that can be used to recognize the presence of a TATA-box in genomic uncharacterized sequences. These may be based on the consensus sequence,[801] PWM, other statistical methods including Hidden Markov Models, Factor Analysis, *etc.*, or even simple feature extraction. However, a well-known approach to pattern recognition based on multidimensional feature space is the use of artificial neural networks (ANNs). ANNs often perform pattern recognition with a high recognition rate. Moreover, the use of ANNs represents a complementary approach to conventional methods based on explicitly defined feature extraction, so that systems for pattern recognition based on ANNs can be combined with other methods in a "mixture of experts" approach.

Many ANN structures and learning algorithms can be successfully employed in TATA motif recognition. However, we concentrate here on using the so-called Learning Vector Quantization (LVQ) ANNs,[441,443,445] for this task. The LVQ ANNs have shown to be very successful in pattern recognition.[204,572] Thus we expect that them to perform well on this problem.

In this section we explain:

- how data preprocessing is done to enable easier training of LVQ ANNs,
- how LVQ ANNs can be used for the TATA motif recognition, and
- how initial parameters of the LVQ ANN can be optimized by a genetic algorithm for more efficient subsequent network training.

These explanations provide the background for the design of the final model of the TATA-box motif to be presented in the next section.

3.1. *Data Preprocessing: Phase 1*

For successful training of ANNs, a high quality training data set is extremely important. The method of training data selection, data coding, as well as data information content, always play a significant role. Although training an ANN can be relatively fast for moderate sized training data sets given current capabilities of computers and learning algorithms, the training with a large number of training data is still usually slow. This speed also depends on the size of the feature vectors that represent input patterns. This is a typical situation when dealing with the DNA sequences because, depending on the problem in question, these sequences may have feature vectors of high dimension and the number of sequences in the training set can be huge.

It is generally ineffective to feed a large amount of raw sequence data to an ANN and to expect it to be able to capture accurately the essential transformation from the input feature space to the target output space. In order to make the learning of the ANNs more efficient, as well as to enhance their generalization capabilities, we normally perform a number of preprocessing transformations of raw data to make it more suitable for ANN processing, as well as to emphasize more important information contained in the raw data. However, the coding of the raw DNA data may prevent proper data preprocessing, as is the case of binary coding of DNA sequences.

The data we select for the training and test sets, as well as the numerical representation of DNA sequences, have been described earlier. We have also explained, based on the preliminary statistical analysis in the last section, how we generate a set of 8 numerical data to represent a TATA-box motif. These are given in Equation 3. In principle, a vector $X = \langle x_1, x_2, ..., x_8 \rangle$ may be used as an input to the LVQ ANN. However, we proceed differently below.

Due to the statistical nature of the matching score for TATA-like motifs obtained by the PWM, and the distribution range of D_1 and D_2 values, we first use these to perform the primary filtering of the training feature vectors. This procedure considerably reduces the quantity of negative data that represent non-TATA motifs, although it sacrifices to an extent the positive data. Since our ultimate goal is to obtain a system with high recognition accuracy of TATA-box motifs, we change the level of filter threshold and examine its effect. The idea is to find the most suitable threshold—one that makes the greatest difference between the correct and false recognition—and to leave most of the further recognition task to the ANN.

At first, we set a threshold τ for the matching score p_i. Only those feature vectors whose first element x_1 is greater than τ are selected as potentially representing the TATA-box motif, and are used for further analysis. We also use the distributions of x_7 and x_8 as constraining conditions for the initial filtering of input data. These conditions are based on the statistical analysis of the last section.

The set of sequences that satisfies these first filtering conditions is defined by

$$S_{F1} = \{X \mid x_1 > \tau,\ x_7 \in [0.3, 1.3],\ x_8 \in [0.3, 1.2]\} \tag{4}$$

The initial training set T_{tr} is formed by taking the union of P_{tr} and S_{cds} and labeling the sequences with their class labels. The sequences from this initial training set is then prefiltered based on the conditions in Equation 4.

The efficiency of such initial filtering is depicted in Figure 3, showing the effects of different threshold values τ combined with the value range filtering for x_7 and x_8. The recognition quality represented by TP and FP varies with the change of the threshold, while the ranges for x_7 and x_8 are fixed as mentioned above. The curve "TP" is given for the set P_{tr}, while the curve "FP" is given for S_{cds}. The discrimination factor, $TP - FP$, is above 50% when $\tau \in [0.34, 0.46]$. This filter is denoted as SF_1.

Another prefiltering that we use in the stage of training LVQ ANNs is, when values ranges for x_7 and x_8 are changed, so that the sequences satisfying these new filtering conditions are given by

$$S_{F2} = \{X \mid x_1 > \tau,\ x_7 \in [0.1, 1.3],\ x_8 \in [0.1, 1.2]\}$$

This filter is denoted as SF_2 and its efficiency agains threshold τ is presented in Figure 4.

As can be seen, both filters have a maximum discrimination for τ of around 0.42. However, the filter SF_2 allows for more feature vectors to pass the filtering condition. Also, the maximum discrimination of SF_2 is about 63%, while that of SF_1 is 59.5%.

Since these filters operate only at the initial stage of the process of the TATA-box recognition, we select $\tau = 0.4$, which is a bit smaller than the τ with the best discrimination ability for any of the filters, to allow more promoter sequences to pass the filter and to make further "filtering" by the ANN system. The value of $\tau = 0.4$ allows about 67% of TP and 8% FP for SF_1, as well as 77% of TP and 17% of FP for SF_2. Note that the filter SF_1 is used in the final TATA motif model, while the filter SF_2 is used in the process of training LVQ ANNs. So, these two "statistical filters" are never used simultaneously.

Feature vectors X that pass such initial filtering are subjected to further processing, as explained in the next subsection on Phase 2 of data preprocessing. This converts vectors X into the transformed feature vectors Y that are input to the LVQ ANN. The overall structure of the preprocessing phases and the LVQ ANN that makes use of these data is presented in Figure 5 for both the training phase and the testing phase.

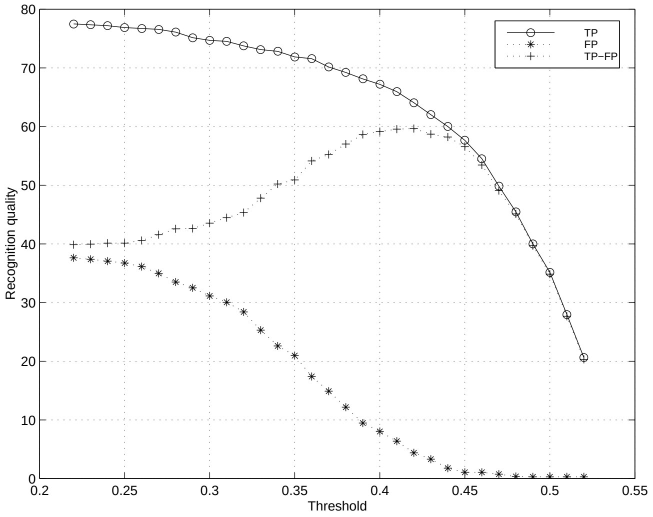

Fig. 3. The efficiency of the combined threshold and data range restriction filtering—*i.e.*, "statiscal filter"—with $x_1 > \tau$, $x_7 \in [0.3, 1.3]$, and $x_8 \in [0.3, 1.2]$. This filter is named SF_1.

3.2. *Data Preprocessing: Phase 2 — Principal Component Analysis*

After sequences are initially filtered by the statistical filter SF_2, we are faced with the question of whether the filtered data perform efficiently as input data to the LVQ ANN. We need only the relevant information to be fed to the LVQ ANN and that non-crucial information be neglected. One of the well-known statistical techniques that ranks the information contained in a data set according to their information content is the so-called Principal Component Analysis (PCA).[154, 326]

This technique is very useful for the preprocessing of the raw data. It helps eliminate components with insignificant information content, and in this way it performs dimension reduction of the data. This can significantly enhance the ability of ANNs to learn only the most significant features from the data set. PCA has been widely used in data analysis, signal processing, *etc.*[154, 326] PCA is a method based on linear analysis aimed at finding the direction in the input space where most of the energy of the input lies. In other words, PCA performs specific feature extraction. The training data set Y_{tr} that we use for the LVQ ANN is obtained by first filtering the sequences from $P_{tr} \cup S_{cds}$ by the statistical filter SF_2, after which normalization and PCA are applied as follows. Let us assume that after statistical filtering the promoter sequences that pass the filter is the set P_f. Then the

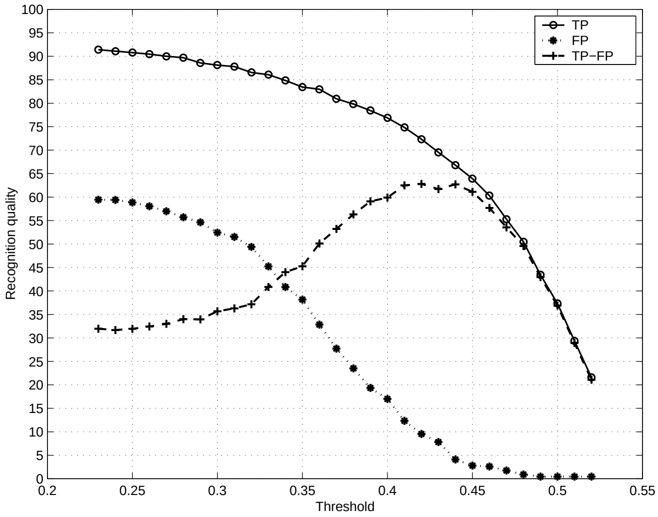

Fig. 4. The efficiency of the "statiscal filter" with $x_1 > \tau$, $x_7 \in [0.1, 1.3]$, and $x_8 \in [0.1, 1.2]$. This filter is named SF_2.

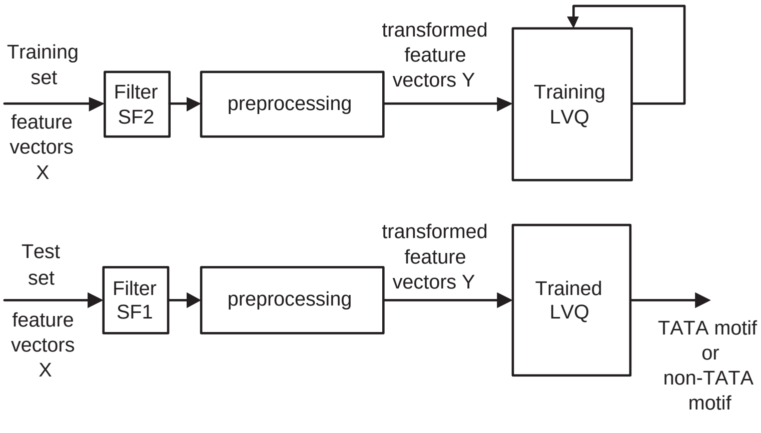

Fig. 5. Statistical filtering, preprocessing, and LVQ ANN.

following steps are made:

- **Step 1: Normalization.** We generate for each sequence in P_f a feature vector

X as explained before. All these vectors constitute the set X_{pf}. This set is normalized to have zero mean and a standard deviation of one, resulting in the set X_{pfN}.

- **Step 2: PCA.** Application of the PCA method to X_{pfN} transforms the normalized feature data so that the transformed feature vectors represent points in the space spanned by the eigenvectors of the covariance matrix of the input data. In addition, the size of the feature vectors may be reduced by retaining only those components that contribute more than a specified fraction of the total variation in the data set. The PCA transformation matrix M_{PCA} is generated to transform X_{pfN} to the new set Y_p. We set the required minimum contribution of components in the transformed data to be 0.01, which means that after the PCA transformation only components which contribute more than 1% to the total variation of the data set are retained. This also result in the reduction of the size of the transformed input feature vectors Y in Y_p to 6 elements instead of the original 8.
- **Step 3: Training set.** By applying statistical filtering SF_2 on the set S_{cds} we obtain filtered non-promoter sequences which make the set N_f. Then, for each sequence in N_f we calculate the feature vector X. All feature vectors for sequences in N_f make the set X_{Nf}. Using the mean of X_{pf}, the standard deviation of X_{pf}, and M_{PCA} obtained in Steps 1 and 2 for the promoter sequences, we transform feature vectors in X_{Nf} to new feature vectors that make the set Y_N. Thus, the training set for the LVQ ANN is represented by the set Y_{tr}, formed by taking the union of Y_p and Y_N labelled with their associated class labels. For the testing set we proceed analogously and obtain the set Y_{tst} of transformed feature vectors and their labels.

3.3. *Learning Vector Quantization ANN*

The input-output transformation performed by ANN in classification problems is discrete. The solution of a classification problem by an ANN is normally achieved in two phases. In a supervised ANN, there is initially the training stage during which the network attempts to learn the relationship between the input patterns and the desired output class by means of adjusting its parameters, and possibly its structure. Such learning procedures are in most cases iterative and attempt to tune the network parameters and structure until the discrepancy measure between the responses produced by the network and the desired responses is sufficiently small. After that phase, one considers the network training completed, and the values of network parameters remain fixed after that. In the testing phase, a test set of new examples, which are not contained in the training set, are presented to the network and compared to the desired network responses. In this way the test of the quality of the trained ANN can be evaluated and its generalization ability assessed.

An ANN system for TATA-box recognition can be considered as a classifier system. For classification problems, many types of ANN can be selected, such as feedforward ANNs, radial basis ANNs, decision based ANNs, *etc*. We use a Learning Vector Quantization (LVQ) ANN for this purpose. LVQ-competitive networks are used for supervised classification problems.[441, 443, 445] Each codebook vector—*i.e.*, input vector—is assigned to one of several target classes. Each class may have many codebook vectors. A pattern is classified by finding the codebook vector nearest to it and assigning the pattern to the class corresponding to that codebook vector. Thus, the LVQ ANN performs a type of nearest-neighbor classification. The standard vector quantization can be used for supervised classification and such methods can provide universally consistent classifiers[204] even in the cases when the codebook vectors are obtained by unsupervised methods. The LVQ ANNs attempt to improve this approach using the adaptation of the codebook vectors in a supervised way.

3.4. *The Structure of an LVQ Classifier*

An LVQ network has two layers. The first one is a competitive layer, while the second one is a linear layer. The role of the competitive layer is to learn to classify input feature vectors X into C_{hidden} clusters in the C_{hidden} space. After the patterns are clustered into C_{hidden} clusters by the competitive layer, the linear layer transforms the competitive layers' clusters into final target classes C_{tar} in the C_{tar} space. One neuron per class/cluster is used in both layers. For this reason the competitive layer can learn to classify up to C_{hidden} clusters. In conjunction with the linear layer, this results in C_{tar} target classes. The role of the linear layer is to convert the competitive layer intermediate classification into target classes. Thus, since the outputs of the competitive layer are vectors with only one element equal to 1, while all others are equal to 0, the weights of the linear layer can be fixed to appropriate 1s and 0s after the ANN initialization and need not be adaptively changed during the training process.

Let us assume that the input patterns are represented by N-dimensional feature vectors that have to be classified into K target classes, by using M intermediate classes at the output of the hidden layer. Then the LVQ ANN has N nodes in the input layer, M nodes in the hidden layer, and K nodes in the output layer. Let the weight matrix $W_{M\times N}$ describe the connection from the input layer to the hidden layer, while the matrix $V_{K\times M}$ describes the connection from the hidden layer to the output layer. For any input pattern, all the nodes in the hidden layer output 0s, except one node that outputs 1. Thus the output vectors of the hidden layer are composed of all 0s and only one element which is equal to 1. Also, only one node in the output layer produces 1 at its output, while all other nodes in the output layer produce 0s. The output vector $\langle 1, 0, ...0\rangle$ represents the first class, $\langle 0, 1, 0...0\rangle$ represents the second class, and so on. A structure of a trained LVQ ANN with two output classes is shown in Figure 6.

Because $V_{K\times M}$ is constant in LVQ ANN, we need only to tune $W_{M\times N}$ based on the training data. The purpose of the learning process is to place the codebook vectors in the input space in a way to describe the boundaries of the classes by taking into account data from the training set. Thus the LVQ algorithm attempts an optimal placement of the codebook vectors in the input space in order to optimally describe class boundaries. This is achieved via the adaptive iteration process. Class boundaries are segments of hyperplanes placed at the mid-distance of two neighboring codebook vectors that belong to different classes.

We know that for a classification problem the number of input and output nodes are determined by the nature of data and the problem setting. However, there are no clear and precise methods for determining the number of hidden nodes and learning rates, especially in the absence of prior knowledge about the class probability densities. Too few hidden nodes may not be enough for good class separation. Too many lead to a very long training phase and does not always produce good generalization. Thus, a compromise has to be achieved through experimentation so as to achieve the best result.

3.5. *LVQ ANN Training*

The training data are obtained as before, using statistical filtering, normalization, and PCA preprocessing. Each of the input vectors is obtained so as to correspond to the hypothetical TATA-like motif found in the original promoter sequence.

At the beginning of LVQ ANN training the weights should be initialized to small values either randomly or by some heuristic procedure. A proportion of neurons in the hidden layer, corresponding to the classes of feature vectors representing the TATA motif and those not representing this motif, is given as $4/6$, so that 40% of the neurons in the competitive layer relate to subclusters of Class 1 (TATA motif class) and 60% of the neurons in the competitive layer relate to subclusters for Class 2 (non-TATA motif class). Thus, the weight matrix V from the hidden subclasses layer to the output target classes is also defined. The number of nodes in the hidden layer is set to 100, and the number of learning epochs is set to 1000. The LVQ ANN is trained by the $lvq2$ learning method.[445]

The structure of the trained LVQ ANN is depicted in Figure 6. The input vec-

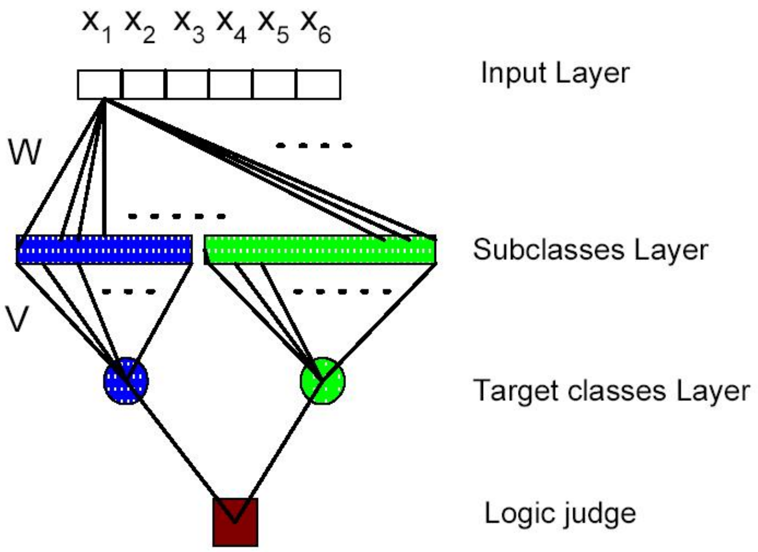

Fig. 6. A structure of a trained LVQ ANN.

tors are classified into two categories in the output target class layer, and the outputs are finally produced by the "logic judge." When the network is initialized, the neurons which relate to the different classes in the competitive layer are randomly distributed. However, through the process of learning they become properly organized to reflect the distribution of the classes in the input data, possibly forming clusters that relate to data subclasses within each of the output classes.

3.6. *Initial Values for Network Parameters*

An LVQ ANN usually performs efficiently if the initial weight vectors are close to their optimized final values. If the initialization of weights of the LVQ ANN is random, then the initial weights are far from the optimized values. When different initial values of the weights are attempted, and the ANN trained produce very different results in the recognition quality of the TATA motif. Thus it is decided to make a systematic search for the good initial weights that can ultimately produce an accurate recognition system. In the ANN field some heuristic procedures are frequently used for weight initialization, such as k-means for Kohonen's SOM ANN, *etc*. A good approach to the problem of optimizing initial weights in an LVQ ANN is to attempt to optimize these globally. This leads to a solution that employs a genetic algorithm for the determination of the optimized initial weights.

After the weights are reasonably well tuned, we apply the $lvq2$ training algorithm to fine tune the decision boundaries in the LVQ ANN.

3.6.1. *Genetic Algorithm*

A genetic algorithm (GA) generally comprises a search and optimization method developed by mimicking evolutionary principles and chromosomal processing in natural genetics.[254] GAs are a part of evolutionary computing, and become a promising field associated with AI and global optimization problems. GAs are iterative optimization procedures that may succeed where other optimization approaches fail. They use a collection—a population—of possible solutions to the problem in each iteration. Normally, in the absence of knowledge from the problem domain, a GA begins its search with a randomly assigned initial population. Then, different "genetic" operators, such as $crossover$, $mutation$, *etc.*, are applied to update the population and thus complete one iteration. The iteration cycles are repeated until a sufficiently good solution, expressed by the fitness criterion, is achieved.

For determination of the initial weights in an LVQ ANN we define a chromosome—*i.e.*, an individual—as the whole set of the weights in the LVQ ANN. The mutation operator is defined as in the conventional GA. From generation to generation the population evolves, improving slowly the fitness criterion. The best solution in each iteration is copied without changes to the next generation in order to preserve the best chromosome. Consequently, the best solution can survive to the end of the process.

3.6.2. *Searching for Good Initial Weights*

To find the good initial weights of the LVQ ANN we proceed as follows. From each sequence in the set $P_{tr} \cup S_{cds}$, we determine a vector of 8 parameters related to a hypothetical TATA-box motif. These vectors are filtered by SF_2—*i.e.*, the threshold $\tau = 0.4$, and bounds for data regions for D_1 and D_2 selected as $D_1 = [0.1, 1.3]$, $D_2 = [0.1, 1.2]$. Then these vectors are normalized and PCA transformed, resulting in the set Y_{tr} that contains feature vectors of size 6 and their labels. In total 492 data vectors that originate from P_{tr} pass the filter SF_2, while 1359 data vectors that originate from S_{cds} pass the filter SF_2.

Then an LVQ ANN is constructed with 6 nodes in the input layer and 100 neurons in the competitive layer. Of these 100 neurons, 40 are related to Class 1 (TATA motif class) and 60 are related to Class 2 (non-TATA motif class). The ratio $4/6$ of the hidden layer neurons does not reflect properly the input data distribution $492/1359$. However, if the $492/1359$ proportion is used, then the network

τ	P_{tr} $(TP\%)$	S_{cds} $(FP\%)$	$CC_{training}$	P_{tst} $(TP\%)$	S_{int} $(FP\%)$	CC_{test}
0.4	67.5	2.4	0.6586	58.8235	2.875	0.4525

Fig. 7. Recognition using LVQ ANN1 after the initial weights are determined by GA

performance appears to be very poor for the reason that there is an insufficient number of neurons in the competitive layer associated with the TATA related feature vectors. On the other hand, if we want to increase sufficiently the number of neurons for Class 1 to achieve a good TP level, and at the same time keep the proportion of neurons relating to Class 1 and Class 2 correct, the total number of neurons in the competitive layer becomes very big and the training time for GA appears to be extremely long. Thus we decide to retain the ratio of dedicated neurons in the competitive layer as indicated. After the LVQ ANN is specified, we use GA to search for the good initial weights for the LVQ ANN. The GA is run for 21 epochs. After 21 epochs, the fitness has already reached about 87%. Thus it can be regarded as corresponding to relatively good initial weights.

Note that we filter feature vectors by SF_2 before the GA is applied to search for the optimized initial weights of the LVQ ANN. However, in order to test how good the determined initial weights are, we are not going to apply the complete SF_2 filtering, but only filtering by the threshold value of $\tau = 0.4$, and no restriction to the values of numerical parameters x_7 and x_8 is applied. This allows more feature vectors—in both the training and test sets—to be processed by the LVQ ANN than if the SF_2 filtering is applied. We apply the normalization and PCA transformation of feature vectors for all sequences from $P_{tr} \cup S_{cds}$ and $P_{tst} \cup S_{int}$, that pass the threshold filtering, using the preprocessing parameters as determined previously. Then such transformed feature vectors are presented to the LVQ ANN and their ability to recognize the TATA motif (and promoters) is evaluated. The results of recognition are presented in Figure 7. Then, the LVQ ANN is further trained with the $lvq2$ algorithm to provide fine tuning of the decision boundaries for the network. The learning rate $LR = 0.01$ and 5500 epochs are used for training. The recognition quality has improved as can be seen from results presented in Figure 8. However, the results are not sufficiently good for constructing a system for the recognition of TATA-like sequences in long stretches of DNA. Such a system is generated in what follows.

τ	P_{tr} $(TP\%)$	S_{cds} $(FP\%)$	$CC_{training}$	P_{tst} $(TP\%)$	S_{int} $(FP\%)$	CC_{test}
0.4	66.8750	1.2125	0.7198	58.8235	2.55	0.4712

Fig. 8. Recognition results with LVQ ANN1, whose initial weights are determined by GA and then fine-tuned by the $lvq2$ algorithm.

4. Final Model of TATA-Box Motif

4.1. *Structure of Final System for TATA-Box Recognition*

The system to be used for the recognition of the TATA-box is depicted in Figure 9. It consists of a statistical filter SF_1, a block for data normalization and PCA transformation, and a block with two LVQ ANNs (LVQB). The whole system is denoted as a multi-stage LVQ ANN (MLVQ) system. The feature vectors X that contain 8 numerical components are fed to the input of the statistical filter SF_1.

The feature vectors that pass the filter enter the normalization and PCA transformation block, at the output of which we get transformed feature vectors Y that contain only 6 numerical components. These transformed feature vectors enter the LVQB. At the output of the MLVQ system the input feature vectors X are classified into those that represent a TATA motif or those which do not.

The LVQB depicted in Figure 10 consists of two LVQ ANNs that process data sequentially. The input of the LVQ ANN1 are transformed feature vectors Y. If the feature vector Y is classified as the one representing the TATA motif, it is passed as input to the LVQ ANN2 that makes the final assessment whether Y represents the TATA motif or not.

4.2. *Training of the ANN Part of the Model*

The training of the LVQ ANNs within the MLVQ system is as follows.

- **The first stage LVQ ANN1.** This ANN is selected with 100 neurons in the hidden layer, and with the 50% of the neurons in the hidden subclasses layer corresponding to Class 1 (TATA motif class) and 50% of the neurons to relate to Class 2 (non-TATA motif class). The training set for this ANN includes feature vectors X of all sequences from P_{tr}, those of 640 sequences from S_{cds}, and the respective class labels. First we apply GA to search for the optimized initial weights of the network as explained in the last section. The

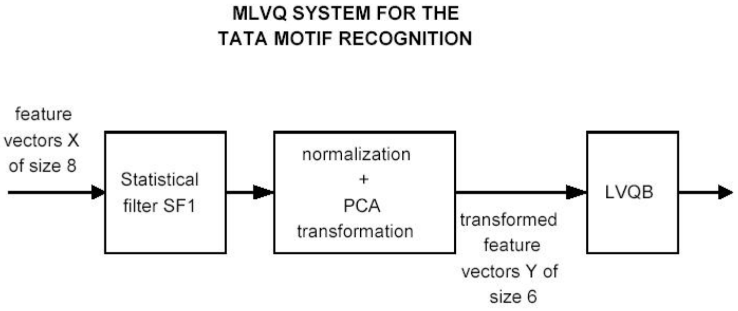
MLVQ SYSTEM FOR THE
TATA MOTIF RECOGNITION
feature vectors X of size 8
Statistical filter SF1
normalization + PCA transformation
transformed feature vectors Y of size 6
LVQB

Fig. 9. The system for the TATA motif recognition.

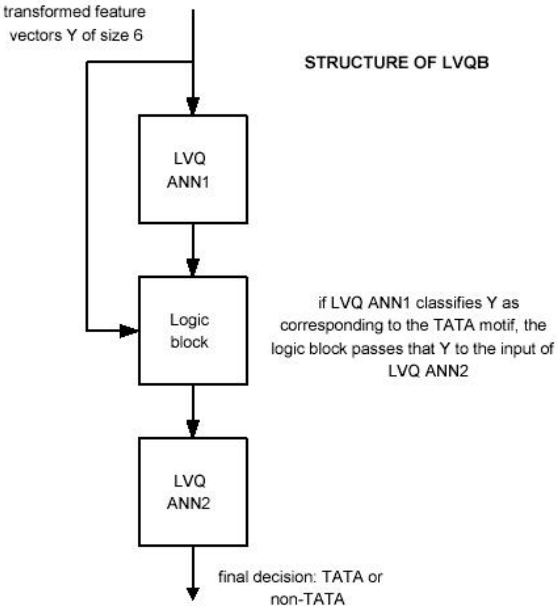
transformed feature vectors Y of size 6
STRUCTURE OF LVQB
LVQ ANN1
Logic block
if LVQ ANN1 classifies Y as corresponding to the TATA motif, the logic block passes that Y to the input of LVQ ANN2
LVQ ANN2
final decision: TATA or non-TATA

Fig. 10. The structure of LVQB. The transformed input feature vectors Y are classified at the output of the LVQ ANN as representing the TATA motif or not.

only reason for the restriction of the of the number of sequences from S_{cds} to 640 is the speed of GA search. After 51 generations, the fitness reaches 85.03%. In order to see its performance in attempted recognition of promoter

P_{tr}	S_{cds}	$CC_{training}$	P_{tst}	S_{int}	CC_{test}
$(TP\%)$	$(FP\%)$		$(TP\%)$	$(FP\%)$	
79.3750	11.9500	0.4707	74.3697	14.4125	0.27

Fig. 11. Recognition of promoters by means of LVQ ANN1 using initial weights obtained by GA.

P_{tr}	S_{cds}	$CC_{training}$	P_{tst}	S_{int}	CC_{test}
$(TP\%)$	$(FP\%)$		$(TP\%)$	$(FP\%)$	
76.4063	5.75	0.5918	69.3277	8.1125	0.3436

Fig. 12. Recognition of promoters by the completely trained LVQ ANN1

sequences, based on recognition of the TATA motif, we use LVQ ANN1 with the weights obtained by GA to test directly both the set $P_{tr} \cup S_{cds}$ and the set $P_{tst} \cup S_{int}$. The results are captured in Figure 11. Also, we apply the $lvq2$ algorithm to further train LVQ ANN1 to fine tune the class boundaries. The learning rate is $LR = 0.01$ and the number of epochs is 500. This trains the LVQ ANN1 network. We test its ability to recognize promoter sequences by recognizing hypothetical TATA motifs. The results are given in Figure 12. It should be noted that no statistical filtering is applied. Thus, only the ability of LVQ ANN1 to recognize the class of feature vectors is evaluated. One can notice that the recognition performance has improved after $lvq2$ is applied. This is what we expected. However, the overall performance is not good enough. It will be improved after the second stage filtering by LVQ ANN2.

- **The second stage LVQ ANN2.** After filtering of feature vectors that correspond to $P_{tr} \cup S_{cds}$ by the LVQ ANN1, 489 feature vectors corresponding to sequences from P_{tr} ($TP = 76.4063\%$) and 460 feature vectors corresponding to sequences from S_{cds} ($FP = 5.75\%$) remain. These, together with the associated class labels serve as the training set for LVQ ANN2. The number of neurons in the hidden layer for LVQ ANN2 is also 100, and the distribution of neurons in the hidden layer is 50% for Class 1 and 50% for Class 2, as

P_{tr} $(TP\%)$	S_{cds} $(FP\%)$	$CC_{training}$	P_{tst} $(TP\%)$	S_{int} $(FP\%)$	CC_{test}
56.25	0.8000	0.6722	48.3193	1.4375	0.4767

Fig. 13. Promoter recognition results with LVQB, where LVQ ANN2 was tuned only by GA.

P_{tr} $(TP\%)$	S_{cds} $(FP\%)$	$CC_{training}$	P_{tst} $(TP\%)$	S_{int} $(FP\%)$	CC_{test}
58.1250	0.7625	0.6885	51.6807	0.6885	0.5377

Fig. 14. Promoter recognition results with the completely tuned LVQB.

the positive and negative data for training are approximately of the same size. GA is used to search for the initial weights for LVQ ANN2. After 101 generations, the set of optimized initial weights is obtained, producing the fitness of 79.66%. The ability of LVQB (LVQ ANN1 and LVQ ANN2 together) to recognize promoters by recognizing the presence of the TATA motif is now tested. The results are given in Figure 13. After the optimized initial weights for LVQ ANN2 are obtained, the LVQ ANN2 is further fine tuned using the $lvq2$ algorithm with the same parameters as for LVQ ANN1. Then, the LVQB is considered trained, and the results of its recognition accuracy testing are given in Figure 14. Compared with recognition results of LVQ ANN1, FP for S_{cds} has reduced 7.5 times, FP for S_{int} reduced 11.7 times, while TP for the training set falls about 31% and for the test set about 35%. Thus, the obtained improvement in accuracy is rather significant, which is also reflected through the increase in the CC values for training and for the test set results.

4.3. *Performance of Complete System*

The complete system for TATA motif recognition includes the statistical filter SF_1, preprocessing block (normalization and PCA transformation of feature vectors), and the tuned LVQB. The only global parameter that can influence the recog-

τ	P_{tr}	S_{cds}	$CC_{training}$	P_{tst}	S_{int}	CC_{test}
	$(TP\%)$	$(FP\%)$		$(TP\%)$	$(FP\%)$	
0.3600	57.9688	0.7625	0.6874	50.8403	1.0875	0.5312
0.3700	57.9688	0.7625	0.6874	50.8403	1.0875	0.5312
0.3800	57.9688	0.7625	0.6874	50.8403	1.0875	0.5312
0.3900	57.9688	0.7625	0.6874	50.8403	1.0875	0.5312
0.4000	57.9688	0.7125	0.6911	50.8403	1.0000	0.5411
0.4100	**57.6563**	**0.6625**	**0.6925**	50.8403	1.0000	0.5411
0.4200	57.5000	0.6625	0.6914	50.8403	0.9250	0.5500
0.4300	57.0313	0.6125	0.6918	**50.8403**	**0.9125**	**0.5515**
0.4400	56.2500	0.6000	0.6871	49.1597	0.9125	0.5384
0.4500	54.6875	0.5875	0.6767	47.4790	0.7625	0.5442
0.4600	52.6563	0.5875	0.6617	43.6975	0.6750	0.5254
0.4700	49.6875	0.5875	0.6393	39.4958	0.6625	0.4913
0.4800	46.2500	0.5000	0.6197	35.7143	0.4750	0.4867
0.4900	41.2500	0.4625	0.5825	30.6723	0.3000	0.4716
0.5000	36.2500	0.4625	0.5396	23.9496	0.1375	0.4408

Fig. 15. Recognition results of the complete tuned system.

nition accuracy and that is also simple for evaluation is the threshold level τ. From previous experimental results we expect the best recognition accuracy of the system for τ at around 0.42. The results of promoter recognition by means of hypothetical TATA motif recognition for the complete system with the threshold of the statistical filter variable, are presented in Figure 15.

It can be observed that a rather good recognition quality appears at the threshold $\tau = 0.43$ for the test set, providing $CC = 0.5515$. Also, the best recognition quality for the training set is achieved for $\tau = 0.41$, producing $CC = 0.6925$. Both of these threshold values are close to the expected best values of around $\tau = 0.42$. Note that the recognition level of $TP = 50.84\%$ on the test set corresponds to the correctly recognized promoter sequences based on the recognition of the hypothetical TATA-box. Since it was estimated in Section 2 that about 75% of promoter sequences used in this study contain a hypothetical TATA-box, the level of correct recognition of TATA motifs is $TP = 67.79\%$.

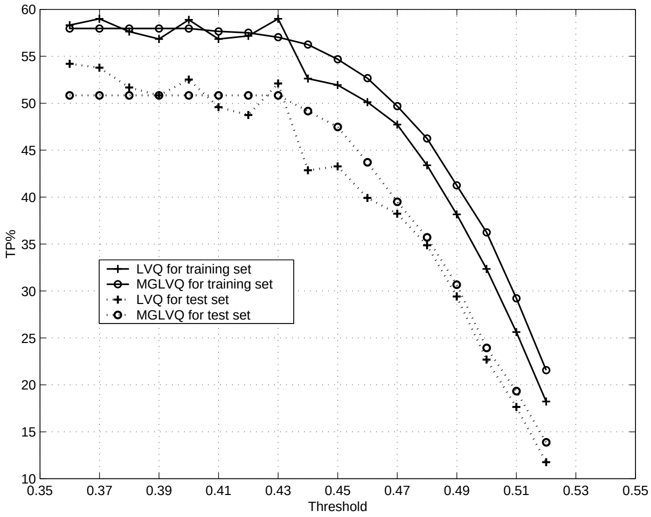

Fig. 16. Comparison of $TP\%$, as a function of τ, between the single LVQ system and the MLVQ system.

4.3.1. *Comparison of MLVQ and System with Single LVQ ANN*

It is interesting to compare the performance of the complete MLVQ system and the one consisting of the statistical filter SF_1, preprocessing unit, and a single LVQ ANN. The comparison results are depicted in Figure 16, Figure 17, and Figure 18. Figure 16 shows that the TP level of both systems are roughly similar as τ changes. However, a noticeable difference appears in the level of FP as indicated in Figure 17. For most of the τ values the FP for the test set of the MLVQ system is considerably lower than for the single LVQ system, implying a much better generalization ability of the MLVQ system. On the other hand, the single LVQ system performs better on the training set, indicating that it has been overtrained. These observations are also visible from the curves of CC changes as given in Figure 18. Although compared to the MLVQ system, the single LVQ system performs better on the training set, its performance is much weaker on the test set.

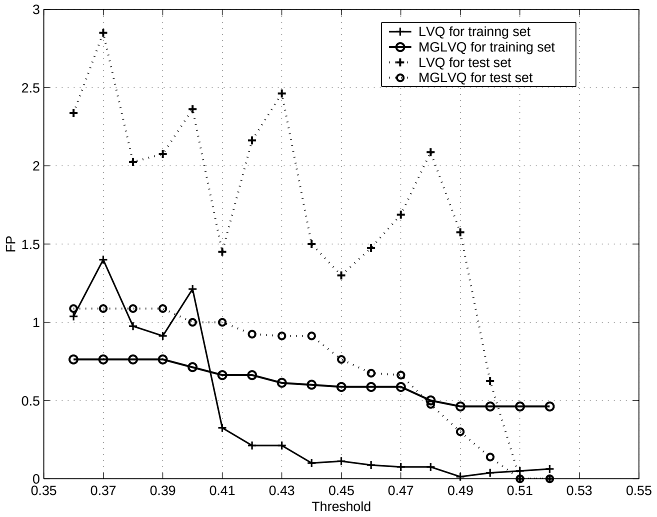

Fig. 17. Comparison of $FP\%$, as a function of τ, between the single LVQ system and the MLVQ system.

4.4. *The Final Test*

The final test of the neural-statistical model of the TATA-box implemented in the MLVQ system is made in an attempt to evaluate how well the system performs on longer DNA sequences. As the test set, we use a collection of 18 mammalian promoter sequences selected and used for the evaluation of promoter recognition programs in Fickett and Hatzigeorgiou.[249] The total length of these sequences is 33120bp, and the set contained 24 TSS. By applying the MLVQ system to this set we can assess the ability of the system to recognize TSS locations by means of recognizing potential TATA-box motifs.

The threshold is used with the values of $\tau = 0.41$ and $\tau = 0.43$. The former one corresponds to the highest CC value obtained previously on the training set, while the latter corresponds to the highest CC value obtained on the test set. We slide along each sequence from its 5' end toward the 3' end a window of 43 nucleotides in length. For 24 successive positions within the window we find the hypothetical TATA-like motif and calculate the 8 parameters associated with them as described previously. This data is fed to the MLVQ system and predictions of the location of the TATA-box are made. The predictied location of the TSS is counted as being 30bp downstream of the predicted TATA-box location. The same

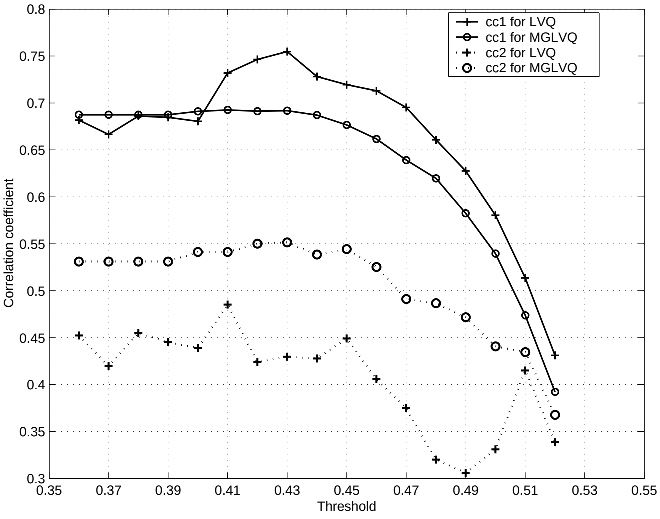

Fig. 18. Comparison of the correlation coefficient CC, as a function of τ, between the single LVQ system and the MLVQ system.

is done with the reverse complement of each of the sequences.

The same criterion, as used in Fickett and Hatzigeorgiou,[249] is applied to determine if a predicted TSS is correct: It is counted as correct if it falls within 200bp upstream or up to 100bp downstream of the actual TSS. Figure 19 shows the test result, where the "+" after a position indicates the correct prediction. The position indicated in bold is the one that counts as correct according to the adopted criteria. Positions with r in front are made on the reverse complement strand.

As can be observed with the threshold $\tau = 0.41$, the MLVQ system identifies 8 (33%) of the known promoters and makes 47 false positives (1/705bp). When the threshold is $\tau = 0.43$, the system finds 7 TSS (29%), whilst making 41 false positive predictions (1/808bp). Figure 20 shows the recognition quality of some other programs tested in Fickett and Hatzigeorgiou.[249] On this test set the MGLVQ system performs better than Audic, Autogene, and the PWM of the TATA-box. The results for comparison are given in Figure 20, where the last column shows the ranking results of different programs based on the average score measure from Bajic.[50] The average score measure is obtained based on 9 different measures of prediction quality and thus can be considered a balanced overall measure of prediction success for different programs. The ranking is in ascend-

No.	Seq. name + Length(bp)	(number) + [TSS location]	$\tau = 0.41$	$\tau = 0.43$
1	L47615(3321)	(1)[2078,2108]	2510, r2456, r1304	2510, r2456, r1304
2	U54701(5663)	(2)[935], [2002]	119, 228, 3118, 5369, r5619, r4272	119, 228, 3118, 5369, r5619
3	Chu(2003)	(2)[1483,1554], [1756,1783]	246, **1487(+)**	246, **1487(+)**
4	U10577(1649)	(2)[1169,1171], [r1040,r1045]	354	354
5	U30245(1093)	(1)[850,961]	194, 446, **690(+)**, r818, r709	446, **690(+)**, r818, r709
6	U69634(1515)	(1)[1450]	299, 934, r645	299, 934, r645
7	U29912(565)	(1)[143,166]	None	None
8	U29927(2562)	(2)[738,803], [1553,1717]	332, **1532(+)**, 2114, r1658	332, **1532(+)**, 2114
9	Y10100(1066)	(1)[1018,1033]	159, 781, r666, r349, r181	159, 781, r349, r181
10	Nomoto(2191)	(1)[1793,1812]	486, 867, 1191, 1375, **1643(+)**	486, 867, 1191, 1375, **1643(+)**
11	U75286(1984)	(1)[1416,1480]	r863	None
12	U52432(1604)	(1)[1512,1523]	217, **1380(+)**, r1451, r138	217, r1451, r138
13	U80601(632)	(1)[317,400]	94	94
14	X94563(2693)	(1)[1163,1200]	489, 2527, r2176, r227	489, 2527, r227
15	Z49978(1352)	(3)[855], [1020], [1150]	**1011(+),** **1024(+)**, r1160, r1122, r405	**1011(+)**, r1160, r1122, r405
16	U49855(682)	(1)[28,51]	r162	r162
17	X75410(918)	(1)[815,835,836]	473, **685(+),** **732(+)**, **786(+)**, r521	473, **685(+),** **732(+)**, **786(+)**, r521
18	U24240(1728)	(1)[1480]	r1529, r1071	r1529, r1071

Fig. 19. Results obtained by MLVQ system on the Fickett and Hatzigeorgiou data set.

ing order with the best performed program on position 1. One can observe that the MLVQ system performs much better than if the recognition is based on the PWM of the TATA-box from Bucher.[116] Since the goal is to develop a more accu-

#	Prog. name	$TP\#$	$TP\%$	$FP\#$	FP per bp	Rank
1	Audic[38]	5	24%	33	1/1004bp	9
2	Autogene[446]	7	29%	51	1/649bp	10
3	Promoter2.0	10	42%	43	1/770bp	4
4	NNPP	13	54%	72	1/460bp	6
5	PromoterFind[377]	7	29%	29	1/1142bp	5
6	PromoterScan	3	13%	6	1/5520bp	2
7	PWM of TATA-box[116]	6	25%	47	1/705bp	11
8	TSSG[784]	7	29%	25	1/1325bp	1
9	TSSW[784]	10	42%	42	1/789bp	3
10	MGLVQ ($\tau = 0.41$)	8	33%	47	1/705bp	8
11	MGLVQ ($\tau = 0.43$)	7	29%	41	1/808bp	7

Fig. 20. Results of other promoter finding programs tested on the Ficket and Hatizigoergiou data set.

rate model of the TATA motif, this improved accuracy is indirectly demonstrated through the better performance in promoter recognition.

One should note that the Audic program uses Markov chains to characterize the promoter region, while the part of the Autogene system that recognizes promoters (the FunSiteP program) bases its logic on the different densities of the transcription factor binding sites in the promoter and non-promoter regions. It it thus surprising in a way, that the MLVQ system that uses only one transcription factor binding site, the TATA-box motif, performs better than these programs. However, it would be interesting to evaluate the performance of MLVQ on a larger promoter set, since the one from Fickett and Hatzigeorgiou[249] is not very big and is not statistically structured to properly represent the eukaryotic Pol II promoters.

5. Summary

The chapter presents an application of ANN to the recognition of TATA-like motifs in eukaryotic Pol II promoters and, based on that, the recognition of promoter sequences. A short region of DNA of 20bp containing the hypothetical TATA-like motif is represented by 8 numerical parameters obtained as a combination of the TATA-box PWM matching score and a statistical analysis. Five of these numerical parameters are derived from the modified EIIP values of nucleotides in the TATA-box hexamer and the neighboring hexamers. These parameters show some regularities that were used in the design of the recognition system. The MLVQ system for TATA motif recognition contains a statistical filter, a normalization and PCA transformation block, and two LVQ ANNs. The initial weights of the ANNs in MGLVQ have been obtained by GA. These helped to obtain finally trained system with improved recognition quality. It is demonstrated on an independent data set of Fickett and Hatzigeorgiou[249] that the new neural-statistical model of the TATA-box (contained in the MLVQ system) achieves improved recognition accuracy as compared with the recognition based only on the use of the matching score of the TATA PWM. Moreover, the MLVQ system performed better than the other two promoter-finding programs that use other promoter characteristics in the search for promoters. Since our system is aimed at improved modeling of the TATA-like motifs and not for promoter recognition, we can conclude that a much better model of the TATA-box has been developed by the combined statistical filtering and ANN system.