**Structural bioinformatics**

**Analysis of *E.coli* promoter recognition problem in dinucleotide feature space**

T. Sobha Rani*, S. Durga Bhavani and Raju S. Bapi

Computational Intelligence Lab, Department of Computer and Information Sciences, University of Hyderabad, Hyderabad 500046, India

Received on September 28, 2006; revised on December 13, 2006; accepted on December 30, 2006

Advance Access publication January 19, 2007

Associate Editor: Anna Tramontano

### ABSTRACT

**Motivation:** Patterns in the promoter sequences within a species are known to be conserved but there exist many exceptions to this rule which makes the promoter recognition a complex problem. Although many complex feature extraction schemes coupled with several classifiers have been proposed for promoter recognition in the current literature, the problem is still open.

**Results:** A dinucleotide global feature extraction method is proposed for the recognition of sigma-70 promoters in *Escherichia coli* in this article. The positive data set consists of sigma-70 promoters with known transcription starting points which are part of regulonDB and promec databases. Four different kinds of negative data sets are considered, two of them biological sets (Gordon *et al.*, 2003) and the other two synthetic data sets. Our results reveal that a single-layer perceptron using dinucleotide features is able to achieve an accuracy of 80% against a background of biological non-promoters and 96% for random data sets. A scheme for locating the promoter regions in a given genome sequence is proposed. A deeper analysis of the data set shows that there is a bifurcation of the data set into two distinct classes, a majority class and a minority class. Our results point out that majority class constituting the majority promoter and the majority non-promoter signal is linearly separable. Also the minority class is linearly separable. We further show that the feature extraction and classification methods proposed in the paper are generic enough to be applied to the more complex problem of eucaryotic promoter recognition. We present Drosophila promoter recognition as a case study.

**Availability:** http://202.41.85.117/htmfiles/faculty/tsr/tsr.html

**Contact:** tsrscs@uohyd.ernet.in

*To whom correspondence should be addressed.

1 INTRODUCTION

Promoter recognition unlike other problems of exon prediction and gene recognition, does not yield good results with methods of alignment or sequence similarity searches. In general, patterns (TATA box, CAAT box, Inr, etc.) in the promoter sequences within a species are known to be conserved. But there exist many exceptions to this rule such as the presence or absence of a particular region (TATA box) that makes the promoter recognition a difficult problem. Several techniques based upon local and global signal methods are proposed for promoter prediction. The local signal based methods locate any short consensus patterns including those corresponding to the the binding sites and hence there is a high probability of finding putative promoters which dominate actual promoters. Huerta and Collado-Vides present a thorough computational analysis of the *Escherichia coli* promoter recognition problem in their paper (Huerta and Collado-Vides, 2003). They give evidence to the hypothesis that there are promoters of different strengths and any alignment method assuming certain consensus signal could give rise to a number of putative promoters.

Machine learning methods can be used to address the issues mentioned above to properly classify the promoter sequences. Well studied *E.coli* species is used for this purpose.

2 E.COLI PROMOTER RECOGNITION

Promoter recognition problem is quite often posed as a binary classification problem in which a given sequence segment is predicted to be a promoter or a non-promoter (Mahadevan *et al.*, 1994). The data set of *E.coli* promoters is a standard data set obtained from Promec (Hershberg *et al.*, 2001) but the investigators of this problem consider different kinds of sets for the non-promoter background. Randomly generated data sets are chosen (Ma *et al.*, 2001) and more recently the biologically meaningful gene and intergenic segments are chosen to form the non-promoter background for promoter recognition.

### 2.1 Local signal based methods

Promoter recognition has been attempted mostly by first identifying the putative binding sites and then classifying using the signal extracted at these binding sites. We name this class of methods as ‘Local signal-based methods’. Harley and Reynolds (1987) initiated the ongoing experimental study of *E.coli* promoters almost two decades ago, the results of which...
are still being used by the researchers working on this problem. Many computer assisted prediction methods developed for E.coli promoter recognition are based on this accepted model of a promoter region consisting of two binding sites with Transcription start site (TSS) at +1 position. Lawrence and Reilly (1990) used Expectation Maximization (EM) Algorithm to extract features for the identification of the binding sites Cardon and Stormo 1992 extended this work to account for variable spacing between binding sites Ma et al. used Bayesian Maximum A Posteriori based EM algorithm for locating the binding sites and used a feed forward neural network for promoter recognition. They reported an average precision of 96.3% (Ma et al., 2001). Ma et al. do not consider biological data sets for the background. Anuj et al. (2003) extract regions around binding sites using EM algorithm as above and use dinucleotide features and a feed forward neural network for achieving promoter recognition. Huerta and Collado-Vides (2003) do not treat the problem as a binary classification problem but as one of recognition. They proposed a position weight matrix model (PWM) which searches for conserved motifs using multiple sequence alignment methods and extract a number of weight matrices for residues around these motifs including the binding sites which are then used to locate promoters within a DNA sequence (Huerta and Collado-Vides, 2003). Ben-Gal et al. (2005) proposed a model called variable order Bayesian network model which generalizes the PWM models for the identification of sigma-70 binding sites in E.coli. Lin et al. (2004) using the same method as Ma et al. (2001) for feature extraction but with a slightly different encoding scheme giving rise to smaller input vectors and reported an average precision of 90.9%.

2.2 Global signal-based methods
On the other hand, binding sites alone do not constitute the entire signal for the promoter (Huerta and Collado-Vides, 2003) There is a growing understanding with respect to the promoter signal strength in the recent literature. It is proposed that along with the two binding sites at the −10 and −35 regions some specific elements beyond −35 actually contribute to the promoter strength. It is also seen that several non-consensus bases within −10 and −35 regions have a positive effect on the promoter strength (Huerta and Collado-Vides, 2003; Kiryu et al., 2005; Ross et al., 1998). Hence one needs to look at the other approaches for promoter classification that are based on global signal present in the promoter sequences. Gordon et al. (2003) consider promoter sequences consisting of 80 nucleotides and a non promoter data set from the DNA genic and intergenic segments. They extract a global signal using a matching function between sequences and develop a kernel function for a support vector machine which is then used for the classification achieving a precision of about 85%. A neural network based multi classifier system using different encodings for the representation of the entire length of 100 base pairs around TSS is designed by Ramawana and Palade (2005) for the problem of promoter recognition reporting a near 98% precision.

2.3 Our approach
We propose a simple global feature extraction scheme that extracts an average signal from the entire promoter sequence of length 80. This signal is used by a neural network to achieve a comparable performance for the different non-promoter data sets that are proposed in literature. In-depth analysis of the classified and misclassified sequences in promoter data set against the biological background throws up two distinct kinds of signals in promoter data set. For this analysis the machinery required is very elegant. We find that simple dinucleotide features form an adequate global signal to discriminate the promoter. In fact this technique can be extended automatically to the eucaryotic promoter recognition. We also report the preliminary results obtained with regard to Drosophila promoter recognition.

3 FEATURE EXTRACTION
Extraction of word features of different lengths is quite common in the sequence problems. As the conserved patterns in procaryotes are of length 6, it would be natural to experiment with words of length 3 i.e. trinucleotide features. We notice that very surprisingly simple dinucleotide features perform quite well on E.coli promoter recognition problem. This problem is addressed by looking at the global signal characterized by the frequency of occurrence of dinucleotides in the promoter region. We show in Section 7 that these same features perform well for eucaryotic promoter recognition.

To extract the global signal for a promoter, frequency of occurrence of dinucleotides is calculated. On the DNA alphabet {A,T,G,C}, the set of 16 possible pairs such as AA, AT, AG, AC, TA, etc. constitutes all the dinucleotide features. Let \( f_i \) denote the frequency of occurrence of the \( i \)-th feature and let \( |S| \) denote the length of the sequence. The feature values \( v_i \) are normalized frequency counts given in (1).

\[
v_i = \frac{f_i}{|S|-1}, \quad 1 \leq i \leq 16
\]

Here, the denominator denotes the number of two tuples possible in a sequence of length \( |S| \) and hence \( v_i \) denotes the proportional frequency of occurrence of \( i \)-th feature. Thus each promoter and non-promoter sequence of the data set is represented as a 16-dimensional feature vector \( (v_1, v_2, \ldots, v_{16}) \).

In supervised classification approach, a major portion of the data set is used for training the classifier and the rest which is not exposed to the classifier is used as the test data set. We denote the set of promoters as positive data set and the set of non-promoters as the negative data set. The frequency values of dinucleotide features are computed for each positive and negative sequence. This collection of vectors is divided approximately in the ratio of 70:30 into training and test sets. Each of the training and test sets has an equal number of positive and negative sequences. A neural network classifier is then trained using the 16 dinucleotide features vectors. The test set is used to evaluate the performance of the classifier.
4 E.COLI PROMOTER CLASSIFICATION

4.1 Data set

Positive data set is built by taking 669 sigma-70 promoter sequences of length 80 with 60 base pairs upstream of the Transcription Start Site (TSS) and the rest downstream as is proposed in the literature from RegulonDB and Promec databases (Gordon et al., 2003). There is no standard negative data set available. We consider all the different types of negative data sets proposed in the literature. The negative data set of Gordon et al. is chosen in a biologically meaningful way by taking sequence fragments outside the promoter region, 709 sequence fragments from the coding region and 709 sequence segments from intergenic portions. In all we consider four negative data sets, two built from Gordon et al. (2003) and the other two are generated randomly with 60 and 50% A+T composition.

The data set is divided into training and test data sets. Except in the case of non-promoters comprising of both gene and intergene portions, in all other cases the positive and negative data sets are taken to be of same size for training. In the case of non-promoter data set consisting of both gene and intergene portions, the proportion of positive data set to the negative data set is taken as 1:2. Each promoter and the non-promoter sequence of the data set is encoded as a dinucleotide feature vector by the method explained in section 3. The graph of the dinucleotide frequency averages against each of the dinucleotide features for both the promoter data set and randomly generated 50% A+T rich negative data set is given in Figure 1. The plot shows the clear discrimination of the average signal of the promoter from the non-promoter for more than 50% of the features.

4.2 Classification results

A feedforward neural network with three layers, namely, an input layer, one hidden and an output layer is used for promoter classification. The number of nodes in the input layer is 16 corresponding to the dinucleotide feature vector. After some trial-and-error experimentation, a hidden layer consisting of 48 hidden nodes is found to give an optimal classification performance. The output layer has one node to give a binary decision as to whether the given input sequence is a promoter or non-promoter. These simulations are done using Stuttgart Neural Network Simulator SNNS.

This neural network is trained on the training set and then the classification performance is evaluated on the test set. All the classification experiments are carried out using a 5-fold cross validation procedure (Mitchell, 1997).

The classification results are evaluated using the performance measures such as Precision, Specificity and Sensitivity. Specificity is the proportion of the negative test sequences that are correctly classified and sensitivity is the proportion of the positive test sequences that are correctly classified. Precision is the proportion of the correctly classified sequences of the entire test data set. The classification results for various negative data sets are presented in Table 1 in which the column under data set gives the proportion of positive and negative data sets.

The classification precision achieved by the neural network is around 78% with biologically meaningful non-promoter data sets. With randomly generated negative data sets the classifier achieves a much higher precision of about 96% just as reported in the literature. For the sake of comparison we have used a multi-layer perceptron in the case of synthetic data sets also. Otherwise same results are obtained without a hidden layer. Hence the promoters against a synthetic background are distinguishable to a high degree of accuracy. Though the use of synthetic data sets as a plausible negative data set is very much in question.

In the case of promoter and non-promoter sequences, negative set consisting of both gene and intergenic portions, in the ratio of 1:1, precision turns out to be 77.1%, specificity 75.69% and sensitivity 80.47%. It can be seen that even though a much better recognition of promoters is achieved, false positives increase compared with the case when the training data set is in the ratio of 1:2. Hence only the 1:2 case is reported in the Table 1.

Our first conclusion then is that with simple features and using a low-end supervised learning algorithm, the classification results are comparable to that reported in the literature (Gordon et al., 2003; Ma et al., 2001) We do not carry out any run-time comparisons with the other approaches presented in the literature. The emphasis here is that a simple scheme without any prior knowledge of the data set like the number

![Fig. 1. Dinucleotide frequency averages for promoter data set and 50% A+T rich negative data set.](Image)

Table 1. Promoter classification for different negative data sets

<table>
<thead>
<tr>
<th>Negative data set</th>
<th>Data set</th>
<th>Precision</th>
<th>Specificity</th>
<th>Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequences from gene segments</td>
<td>1:1</td>
<td>78.4</td>
<td>79</td>
<td>80</td>
</tr>
<tr>
<td>Sequences from gene segments and intergenic portions</td>
<td>1:2</td>
<td>76</td>
<td>88.01</td>
<td>63.6</td>
</tr>
<tr>
<td>60% AT rich</td>
<td>1:1</td>
<td>95.5</td>
<td>98.18</td>
<td>93</td>
</tr>
<tr>
<td>50% AT rich</td>
<td>1:1</td>
<td>97.3</td>
<td>99.3</td>
<td>95.8</td>
</tr>
</tbody>
</table>
of binding sites, the range of spacer lengths between the binding sites, etc. achieves a rate of recognition comparable to the several methods proposed in the literature. It is to be noted that the combination of coding and non-coding sequences together as negative data set has not been tried by Gordon et al. (2003). We feel that this represents a better realistic background than considering them individually. To the best of our knowledge such a scheme achieving this kind of precision in recognition has not been reported so far in the literature. This scheme of feature extraction and classification will be extended to the problem of Drosophila promoter recognition in Section 7.

During our experiments we notice that promoter classification with an accuracy of near 96% is achievable by a single layer perceptron for synthetic negative data sets, potentially indicating the linear separability of the promoter data sets. The issue of linear separability has not been raised previously in the promoter classification literature and hence we investigate further. In the subsequent section we report that a single layer perceptron can achieve adequate classification accuracy with the biological negative data also.

5 ANALYSIS OF CORRECTLY CLASSIFIED AND MISCLASSIFIED PROMOTER SEQUENCES

One cause of concern during the training phase of the neural network is that irrespective of network architecture, such as the number of hidden layers and the number hidden nodes in each hidden layer, the network could not achieve a training performance beyond 85%. In order to carry out a deeper analysis of classification of promoters, we select a set of sequences randomly from both promoters and non-promoters (consisting of both gene and intergene portions) in the ratio of 1:2, respectively for training. That is, a set of 454 sequences are taken from promoter data set as positive set, and 454 sequences are taken from each gene and intergene sequence sets. The rest of the data set is used as test data. Feature extraction and classification are done as described in Section 3 and Section 4.2. The set of misclassified sequences is given a closer attention in this section.

To analyse this problem further, the sequences that the neural network finds difficult to learn are isolated from both positive and negative data sets. Promoter sequences that are classified correctly, i.e. true positives (TP), misclassified (FN) and the non-promoter sequences that are correctly classified (TN) and misclassified (FP) are treated as four classes. A 4-way classifier is designed to check the assumption that there may exist four different signals. But, the results of this classification show that TP comprising of a majority of promoter sequences and a minor portion of non-promoter sequences (FP) are grouped as one class. Similarly TN and FN are grouped as another class. We can clearly illustrate these results by plotting the dinucleotide frequency averages of these four sets against the negative data set consisting of gene and intergene portions. Figure 2 depicts the euclidean distances between TP and TN as well as FP and FN. For completeness of discussion we give dinucleotide frequency average plots for other negative data sets as well.

For each dinucleotide feature \( i \), the average of frequency values \( v_{ij} \) over an entire data set \( D \), is computed as \( a_{i,D} \), where \( |D| \) denotes the size of the data set.

\[
a_{i,D} = \frac{1}{|D|} \sum_{j=1}^{C_{ij}} v_{ij}
\]

The average euclidean distance between two data sets \( D_1 \) and \( D_2 \) is given by

\[
d(D_1, D_2) = \sqrt{\sum_{i=1}^{16} (a_{i,D_1} - a_{i,D_2})^2}
\]

For example, the euclidean distance between TP and FP is denoted as \( d(\text{TP,FP}) \), etc. Distances between these data sets TPs, TNs, FPs, FNs are presented in Table 2.

Both the Figures 2 and 3 corresponding to the two biological negative data sets show that the correctly classified promoter sequences and the misclassified non-promoter sequences are close and also misclassified promoter and correctly classified non-promoter sequences are close. It is interesting to note that \( d(\text{TP,FP}) \) and \( d(\text{TN,TN}) \) are much smaller than \( d(\text{TP,TN}) \) and \( d(\text{FP,FN}) \). These results clearly demonstrate that there is a small confusion set in both the promoter and non-promoter data sets. That is, there exist promoter like non-promoters and non-promoter like promoters. This insight gives us the motivation to dissolve the confusion set by constructing two data sets one which captures ‘majority’ promoter class and

![Fig. 2. Separation between promoter data set and negative dataset consisting of segments from gene and intergene portions of the DNA.](image)

<table>
<thead>
<tr>
<th>Distance</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>d(TP,FP)</td>
<td>0.015962</td>
</tr>
<tr>
<td>d(TN,FN)</td>
<td>0.030809</td>
</tr>
<tr>
<td>d(TP,TN)</td>
<td>0.112933</td>
</tr>
<tr>
<td>d(FP,FN)</td>
<td>0.081533</td>
</tr>
<tr>
<td>d(TP,FN)</td>
<td>0.087181</td>
</tr>
<tr>
<td>d(TN,FP)</td>
<td>0.112983</td>
</tr>
</tbody>
</table>
the other that reflects the ‘minority’ signal. This is taken up in the next section.

A perceptive reader will notice that a preliminary evidence for linear separability of promoter data set against a synthetic negative background is already pointed out in an earlier Section 4.2. Figure 4 depicts the dinucleotide frequency averages for promoter and 60% A+T rich negative data set having euclidean distance as small as 0.031599. The analysis in the next section shows that while biological data sets suffer from a confusion set, the promoters are easier to discriminate against a synthetic negative background. This possibly explains why the results of classification accuracy in the literature are much higher for synthetic negative data sets (Ma et al., 2001).

In fact this motivates us into regrouping the data set by moving away the confusion set and investigating the possibility of single layer perceptron for classification against biological negative data.

5.1 A single layer perceptron for promoter recognition

The data sets are regrouped into two sets: sequences that are correctly classified by the neural network are called Major Set (Maj). Similarly the sequences that are misclassified are grouped as Minor Set (Min). Using these two data sets, two separate new neural networks $NN_{Maj}$ and $NN_{Min}$ are constructed. $NN_{Maj}$ is trained only on Maj and similarly $NN_{Min}$ is also trained on only Min. These neural networks achieve 100% training performance with no hidden layer. That is the Majority class constituting the majority promoter and the majority non-promoters is linearly separable. Similarly Min is linearly separable.

At this point it is interesting to evaluate the overall classification performance of Maj and Min classifiers with the original test data without the major and minor segregation. Results of one I-fold are shown in Table 3. $NN_{Maj}$ achieves 80% precision while $NN_{Min}$ recognizes the rest of the 20% (up to a max of 3% error) correctly. Moreover, the sequences that are the TPs and TNs of $NN_{Maj}$ become the false negatives and the false positives, respectively of $NN_{Min}$. In this way, $NN_{Min}$ behaves in a kind of complementary fashion to $NN_{Maj}$ on the test set. To summarize, our results demonstrate that by segregating the promoter data into major and minor classes, it is possible to construct linearly separable classifiers.

It is to be noted that not only the numbers in each of the boxes of Table 3 match, but also the exact sequences in each of these boxes match. For example, the set of 101 FNs of $NN_{Min}$ is a subset of the 107 TPs of $NN_{Maj}$. Thus we see that both promoter and the non-promoter sequences have two distinct patterns, one being recognized by $NN_{Maj}$ and the other by $NN_{Min}$. But for a confusion of five to seven sequences the $NN_{Maj}$ and $NN_{Min}$ behave in a complementary fashion. A small portion (14%) of the non-promoter data set is similar to a majority (70%) of the promoter data set and also 86% of the non-promoter data set (TN) is closer to 30% of the promoter data set (FN).

Huerta and Collado-Vides (2003) name promoters which have signal close to the consensus pattern at the binding sites as strong promoters and others as weak promoters. Our analysis also gives a clear indication of two kinds of promoters. A majority of the promoters being easily distinguishable from the background we call as strong promoters, and small portion having a signal closer to the non-promoters as weak promoters. It is to be emphasized that in this process we have successfully built a neural network $NN_{Maj}$ which is a single layer perceptron achieving promoter recognition performance.
of 80% which is comparable to the powerful classifiers that are presented in the literature (Gordon et al., 2003).

6 PROMOTER RECOGNITION IN GENOME SEQUENCE

One of the main goals of promoter recognition is to locate promoter regions in the genome sequence. In this section we propose a scheme for locating promoters in a given DNA sequence segment of *E.coli* genome of length *N* in a particular direction (say, 5' to 3'). We do not address the issue of locating TSS in the promoter region. Going along with the scheme for classification by the neural network *NNMaj*, we consider a moving window of length 80 extracting segments from the start of the DNA sequence considered, that is, 1–80, 2–81, 3–82 and so on. These are represented as the 16-length dinucleotide feature vectors which are fed to the neural network classifier. Each of the segments gets classified as promoter (P) or non-promoter (NP). If a segment *m* = *(m + 79)* is classified as a promoter, then the nucleotide *m* is annotated as *P* and if it is classified as non-promoter then *m* is annotated as *NP*. This process of annotation is continued for the entire sequence to get a sequence of *P*'s and *NP*’s. We propose that if a contiguous segment of length more than a certain threshold has all *P*’s then we annotate that region as promoter region otherwise as non-promoter region.

This algorithm is implemented on a segment of length 10 000 base pairs taken from the Section 1 of the complete genome of *E.coli* K12 MG1655 of NCBI database as a preliminary exercise. We find that all the promoter regions that are located around the TSS at 106, 139, 5084, 8209 and 9279 (NCBI) are identified by this simple annotation scheme. For example, all the nucleotides from 1 to 132 are annotated as *P* and the annotation reported in the NCBI data base confirms 71–99 and 104–132 as promoter regions. We also obtain short lengths like 2214–2219 as *P* but if we set an appropriate threshold these short spurious promoters can be rejected. With a threshold set at say, 20 base pairs, all the true positives of the 10 000 length sequence are recognized by our scheme and four false positives are obtained which are not reported in the NCBI site. It is to be emphasized that just a segment of length 80 is not sufficient to say if it contains a promoter region. Since a moving window scheme is being applied one needs a minimum length of 140 if a contiguous segment of *P*’s of length 20 is to be found even starting at the 60th place in the sequence. The rate of performance in any case is tied up with the classification performance of the classifier and also it is possible that the false positives so obtained are in fact unknown promoters. The number of times the classifier needs to be run on the sequence is *N* = 80.

Satisfactory results are obtained when the threshold is set at 30. Also experimentation is carried out with different kinds of windows including that of consecutive segments of length 80, that is, 1–80, 81–160 and so on. This windowing scheme obtains a false negative by identifying only four out of the five promoter regions correctly. Hence it is reasonable to adopt an intermediate windowing scheme and also thus reducing the time complexity, by leaving 30 out instead of leaving one out for which classifier needs to only run for *N*/30 times and this scheme does achieve satisfactory results as reported above. We considered window centered at 1 whereas it is more common to consider windows actually centered in the middle and the central nucleotide being annotated accordingly. The promoter region is not really symmetric with respect to the TSS which is located at around 60th position and the two binding sites at either end being located approximately at 25 and 50 positions, respectively. Hence the central nucleotide may not actually belong to either of the binding sites. The scheme of annotation certainly needs to be fine-tuned by considering larger length of genome sequences and an appropriate threshold chosen after experimentation on these larger data sets.

7 DROSOPHILA PROMOTER RECOGNITION

In order to explore the extendability of the proposed approach of dinucleotide feature and a multi layer perceptron classifier, we investigate Drosophila promoter recognition problem. Eucaryotic promoter recognition is very different from the problem of procraryotic promoter recognition. Eucaryotic promoters within a species and across the species have highly divergent promoter sequences (Werner, 1999). They may or may not contain conserved patterns such as the TATA box. In fact, in the case of highly regulated promoters, the extraneous regulation factors like enhancers, etc. occur adjacent to promoters, thus making the discrimination difficult. In this context, as a first step to get a base estimate for eucaryotic promoter recognition, we use dinucleotide features to extract a global signal from a promoter sequence of 241 base pairs. Using the same classification scheme as for *E.coli*, a neural network with backpropagation learning algorithm is used to train and classify a given Drosophila promoter.

7.1 Data set

The promoter data set of Drosophila is obtained from Eucaryotic promoter database (EPD). Positive data set contains sequences of length 241 bp with 200 base pairs upstream of the TSS and the rest downstream. The choice of length is made so as to include the downstream promoter element which usually occurs around +30 in case of TATA less promoters and an upstream element which occurs at −200 (Werner, 1999). We constructed a biologically meaningful negative data set by taking sequence fragments outside the promoter region, 300 sequence fragments from the coding region. Two more data sets are generated randomly with 60 and 50% A+T composition.

7.2 Classification results

Dinucleotide features of the data sets are extracted as given in Section 3. This feature extraction scheme has the advantage that it is independent of the length of the sequence. A neural network with these 16 dinucleotides as input vector with a single hidden layer of 48 nodes and an output layer with a single node is constructed. Five-fold cross-validation is done on each of these sets and the results are given in Table 4. Here also it can be observed that in the case of randomly generated
data set the performance of the classifier is as good as 98.8%. When a biologically meaningful data set is considered the performance of the system is about 75%.

It should be noted that this is only a preliminary investigation in the direction of eucaryotic promoter recognition. We need to do a deeper analysis to implement our method with the data sets proposed by Ohler et al. (2002), etc.

8 CONCLUSIONS

We show that a multi-layer perceptron with one hidden layer using dinucleotide features can achieve a binary classification of precision of around 80% for biologically meaningful non-promoter data set as background. Further, it achieves a precision for promoter recognition of 96% against the background of synthetic non-promoter data sets. Further, a windowing scheme to recognize promoters in a genome sequence is proposed.

The major thrust of the article is to estimate the recognition accuracy of a simple classifier for the promoter recognition problem which uses the dinucleotide features as input. The classifier achieves a precision of ~80%, which is comparable to the performance of the high powered classifiers using heavy feature extraction schemes that are presented in the literature. A similar scheme of feature extraction and classification is explored in the case of Drosophila promoter recognition problem, we report an average performance of 75% for recognition in the case of biological negative data set.

One of the other claims of this article is that, in fact, a single layer perceptron can be constructed which achieves as good a performance as the multi-layer perceptron. It is shown that there are two kinds of promoters: a major set and a minor set. A major set of promoters and a minor set of non-promoters have a similar signal in dinucleotide feature space. Similarly a minor set of promoters and a major set of non-promoters have similar signal. This analysis of the data set helps us in building another neural network called $NN_{Maj}$ which is a single layer perceptron that achieves 100% recognition on the major set. That is the majority class constituting the majority promoter and the majority non-promoter data sets is linearly separable. If the test data is not labelled as belonging to a major or a minor set, $NN_{Maj}$ achieves an overall precision of 80% for the promoter classification problem. This analysis demonstrates the usefulness of sub-classifying promoters before building a classifier.

ACKNOWLEDGEMENT

We would like to thank Gordon et al. (2003) for sharing their data sets (biological negative data sets) with us.

Conflict of Interest: none declared.

REFERENCES


