Plant Tissue Cult. & Biotech. **18**(2): 123-130, 2008 (December)

PTC&B

Identification of Drosophila Promoter Using Positional Differential Matrix and Support Vector Machine from Sequence Data

Azizul Haque, Firoz Anwar[1], Taskeed Jabid[1], Syed Murtuza Baker[1], Haseena Khan[2], Mohammad Nurul Islam[3] and Mohammad Shoyaib*

Institute of Information Technology, University of Dhaka, Dhaka-1000, Bangladesh

Key words: Drosophila, Promoter, Sequence data

Abstract

Promoter region plays an important role in controlling gene expression of any living organism. It regulates gene transcription by providing space to the RNA polymerase and transcription factors to bind and interact with. Binding of appropriate transcription initiation complex is determined by the specific promoter sequence carrying gene specific motifs. The promoter recognition process is a part of the complex process where genes interact with each other over time and actually regulates the whole working process of a cell. Thus computational method for identifying promoter is a focal point for researchers. This paper presents an algorithm for identifying *Drosophila melanogaster* promoter using differential positional frequency matrix between promoter and non-promoter sequences which shows maximum 90.36% tenfold cross validation accuracy. The proposed method exhibits greater accuracy for detecting promoters. Also higher sensitivity and specificity results elucidate that the proposed method is less prone to false negatives and false positives compared to some other existing methods.

Introduction

To understand the transcriptional process it is necessary to identify and characterize the promoter as the motifs residing within these promoters actually work as switches to trigger the transcriptional process.

The promoter is a region on the genomic sequence, which resides upstream of the transcription start site (TSS). It plays a major role while the DNA is transcribed into messenger RNA (mRNA) and it also largely controls the biological activation of the gene (Pedersen et al. 1999). Due to this inherent

*Corresponding author. E-mail: <shoyaib@univdhaka.edu>. [1]Department of Computer Science and Engineering, East West University, 43 Mohakhali, Dhaka-1212, Bangladesh. [2]Department of Biochemistry and Molecular Biology, University of Dhaka, Dhaka-1000, Bangladesh. [3]Department of Botany, University of Dhaka, Dhaka-1000, Bangladesh.

relationship, identification of promoters will provide better understanding on the implication of promoter over gene annotation.

Due to the lack of straightforward procedure for detecting a promoter from sequence data using an open reading frame, it is difficult to locate the position, amount, and the strength of selective functional regulatory elements of the gene. Transcriptional regulation model is not just a simple activation or suppression by transcription factors, rather, also includes competitive binding of proteins (Small et al. 1991), co-operative binding (Burz et al. 1998), chromatin bending and other molecular interactions that are not always reflected in the nucleotide sequence.

Several computational methods have been proposed in the past few years. CONPRO (CONsensus PROmoter) can correctly detect promoters for approximately half of human gene (37 - 71%) of which around 85 - 90% are true promoters (Rongxiang and David 2002). PromoterExplorer (Xie et al. 2006), analyzed the different roles of various features on the sequence data. A combined local distribution of pentamers, positional CpG Island and digitized DNA sequence were used to construct a higher-dimensional input vector. A cascade AdaBoost-based learning procedure was then adopted to select the most 'informative' or 'discriminating' features. Recent research has presented another general method for characterizing a set of sequences by their recurrent motifs. They have demonstrated the use of prevalent features extraction and proposed a framework for identifying promoter (Sharan and Myers 2005).

Most of the available computational methods for core promoter prediction are based on solid machine learning techniques like probabilistic sequence models, Hidden Markov Model (HMM) or Support Vector Machines (SVM) and have shown good performance on fly predictions (Ohler 2006). These methods show different success rates with different datasets. Mostly they have identified promoters by analyzing various positional features. To identify unknown promoters a machine learning system has been trained with the analyzed features that can distinguish between a promoter and a non-promoter and then tested on test sequences.

In this paper, a new algorithm "DrosophilaPromoterIdentification" is proposed to identify the *Drosophila melanogaster* promoter from its gene sequences. Then the performance of the proposed algorithm for detecting *Drosophila melanogaster* promoters using different DNA sequences is evaluated. Consistent and promising results have been obtained, which proves that the proposed method can greatly improve the promoter identification performance and also outperforms some other existing methods.

Materials and Methods

The *Drosophila melanogaster* promoter sequences have been obtained from the Division of Biological Sciences of University of California, San Diego (UCSD). A comprehensive dataset of 340 different *Drosophila melanogaster* promoter sequences each of length 92 bps (positive dataset) is collected. Though this dataset contains relatively small number of promoters of *Drosophila melanogaster*, additional promoter sequences from other datasets have not been collected due to the integrity and authenticity of the dataset. These data contain TATA box, CAAT box, BRE and DPE. A recent study on the fruit fly shows that the core promoters mostly span from [–50, +50] position relative to the TSS of the DNA sequence (Ohler 2006). So, 340 gene sequences each of length 92 bps (–791 to -700) are taken from Eukaryotic Promoter Database (EPD), which surely reflects negative dataset (non-promoter). The length of both positive and negative datasets are intentionally kept the same so that the training of SVM does not become bias.

Both promoters and non-promoters are composition of four nucleotides (i.e. A T C G). The proposed method was based on the differential frequency distribution of each of the nucleotides a particular position between the promoters and the non-promoters. In order to do so, at first all the compiled 340 *Drosophila melanogaster* promoter sequences were aligned. Then the frequency of A, C, G and T located in the first column/position of all 340 sequences were calculated. This process was continued for each column of the whole sequence. Same methodology was applied over the 340 aligned non-promoter sequences.

The process can be denoted using Eq. 1 and Eq. 2. Here, *PE* stands for promoter elements and *NPE* stands for non-promoter elements. The notation i represents the column/position and n represents the total number of column on the sequence.

$$\sum_{i=0}^{n} PE_A[i] \text{ Or } \sum_{i=0}^{n} PE_C[i] \text{ or } \sum_{i=0}^{n} PE_T[i] \text{ or } \sum_{i=0}^{n} PE_G[i] \quad \text{(Eq. 1)}$$

(for promoter)

$$\sum_{i=0}^{n} NPE_A[i] \text{ Or } \sum_{i=0}^{n} NPE_C[i] \text{ or } \sum_{i=0}^{n} NPE_T[i] \text{ or } \sum_{i=0}^{n} NPE_G[i] \quad \text{(Eq. 2)}$$

(for non-promoter)

After counting the frequency on each column a frequency matrix of every sequence character i.e, DFqMatrix [4, n] and NPFqMatrix [4, n] (where DFqMatrix is Drosophila promoter frequency matrix and NPFqMatrix is non promoter frequency matrix) was constructed. Then the difference between every

sequence element of each column from DFqMatrix and NPFqMatrix was calculated and the result was stored in another matrix called DiffMatrix [4, n].

$$\text{DiffMatrix [1, n]} = \sum_{i=0}^{n} PE_A[i] - \sum_{i=0}^{n} NPE_A[i] \qquad \text{(Eq. 3)}$$

$$\text{DiffMatrix [2, n]} = \sum_{i=0}^{n} PE_C[i] - \sum_{i=0}^{n} NPE_C[i] \qquad \text{(Eq. 4)}$$

$$\text{DiffMatrix [3, n]} = \sum_{i=0}^{n} PE_T[i] - \sum_{i=0}^{n} NPE_T[i] \qquad \text{(Eq. 5)}$$

$$\text{DiffMatrix [4, n]} = \sum_{i=0}^{n} PE_G[i] - \sum_{i=0}^{n} NPE_G[i] \qquad \text{(Eq. 6)}$$

For developing the learning model an inductive model was constructed from the DiffMatrix, which can be used for further implication on new data. For the learning purpose, SVM was used as it has proven to be a better tool compared to the other available tools for analyzing biological data (Kasabov and Pang 2004).

Cortes and Vapnik (1995) developed SVM at AT&T Bell laboratories. It was a statistical learning technique used as a classifier based on pattern recognition. It can also perform real valued function approximation tasks. Support Vector Machines can non-linearly map their n-dimensional input space into a high dimensional feature space (Cortes and Vapnik 1995, Vapnik 1982). For a typical learning task P ($\overline{x}$, y) = P (y | $\overline{x}$) P ($\overline{x}$), an inductive SVM learner aims to build a decision function

$f_L : \overline{x} \to \{-1,+1\}$ based on a training set S_{train}, which is

$$f_L = L(S_{train}) \qquad \text{(Eq. 7)}$$

where: $S_{train} = (\overline{x_1}, y_1), (\overline{x_2}, y_2), ...(\overline{x_n}, y_n)$.

Two criteria were widely used for evaluating the performance of promoter prediction program. They were sensitivity (*Sn*) and specificity (*Sp*). It can be defined as following:

$$Sn = \frac{TP}{TP + FN} \qquad \text{(Eq. 8)}$$

$$Sp = \frac{TP}{TP + FP} \qquad \text{(Eq. 9)}$$

where, *TP*, *FP* and *FN* denote the numbers of true positives, false positives and False negatives, respectively. In general, the larger value of *Sn* symbolizes less false negative and the smaller value of *Sp* represents more false positive. It was a trade-off to balance *Sn* and *Sp*.

The classification on SVM generated by a two-step procedure: First, the sample input vectors were mapped into a higher dimensional space. Then, the

SVM finds a hyperplane in this high-dimensional space with the largest margin separating classes of data. Here first the proposed algorithm was used to construct hyperplane in a very high dimensional "mapped" space and then reviewed for identifying corresponding classifying surface in the original space. SVM was trained in a supervised manner on a collection of promoter and non-promoter sequences. The training of this system was made on the compiled dataset. Fig. 1 describes the whole methodology in a block diagram.

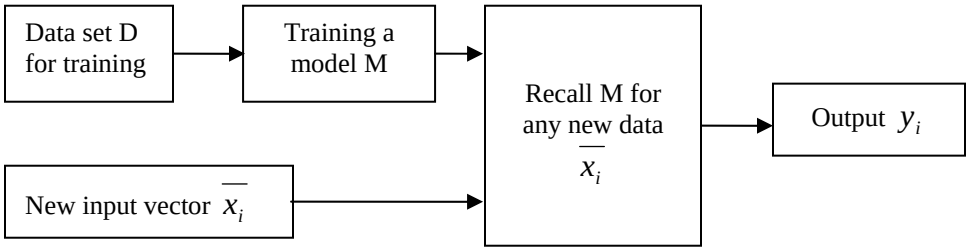

Fig. 1. A block diagram of an inductive reasoning system. A global model M was created based on data samples from D and then recalled for a new vector.

The promoter identification method proposed in this paper can be summarized into the following steps:

(i) Find column frequency for each nucleotide {A, C, T, and G} of the sequence.

(ii) Calculate i^{th} column frequency for each nucleotide combination and continue the process until the last column for Drosophila promoter sequence as well as the chosen non-promoter sequence.

(iii) Subtract the nucleotide combination frequency of nonpromoter from promoter and store the difference in DiffMatrix (difference matrix).

(iv) Train the SVM according to the DiffMatrix value. (If the first element of both promoter and nonpromoter sequence is A then the 1st feature of the SVM input data will be the value of A in the difference Matrix 1st column and so on)

Using the model created during the training of SVM test on set of known data: Two kinds of test can be conducted. Cross validation, where the whole set of data was divided (also known as folded) into *n* fragments and *n-1* fragments were used for learning and to create the model. Then this model was tested on the remaining fragment. The other method allows the whole dataset for learning and then applies this model on data, which are mutually exclusive from the training dataset.

Results and Discussion

Three different kinds of folding for the cross validation were applied and the result shows improvements on every occasion. A threefold, seven-fold and tenfold cross validation exhibit 87.69, 88.13 and 90.36%, respectively. The cross validation result presented on Fig. 2 shows significant performance of both promoter and non-promoter data mapping. The cross validations were applied to the whole dataset of 340 positive dataset and 340 negative dataset. That means 340 promoter sequences and 340 gene sequences were taken for the cross validation of the proposed system and it showed good accuracy. In tenfold cross validation an accuracy of 90.36% has been achieved which is quite high accuracy considering the classification of other database.

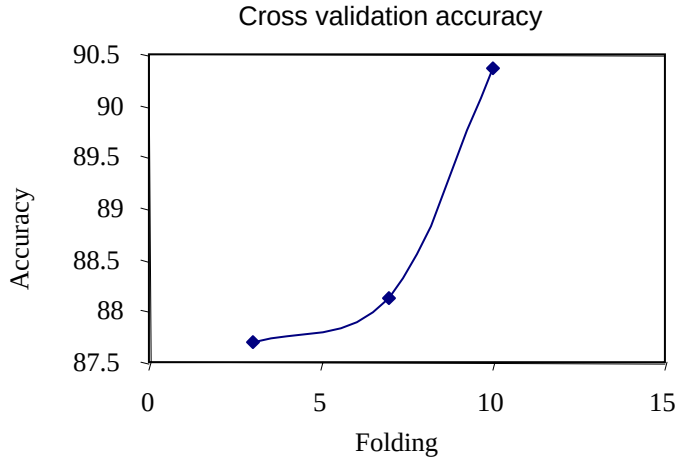

Fig. 2. Cross validation accuracy using 3, 7 and tenfold using SVM.

For the random check of the accuracy of the system a model was developed taking 325 promoter sequences and 325 non-promoter sequences. The remaining 30 sequences from both promoters and non-promoters were then tested through the system to analyze the accuracy. From the random input of 15 promoter and 15 non-promoter data features average result showed an accuracy of 90.667%, which was consistent compared to other promoter identification methods. The sensitivity (*Sn*) and specificity (*Sp*) analysis also showed that during prediction of promoters and non-promoters the proposed method was less prone to false negative (*FN*) and false positive (*FP*) compared to other methods. The results of the experiments are presented in Table 1. The high sensitivity and specificity also shows significantly higher accuracy rate, as it showed in the cross validation.

Some of the publicly available promoter identification tools were tested using some random promoter and non-promoter sequences. The test results were summarized in Table 2. The result conforms the performance of the proposed method.

Table 1. Results generated by SVM for random dataset.

Prediction of promoter and non-promoter	Accuracy	Sensitivity	Specificity
15 promoter and 15 non-promoter (average over 5 sets of randomly selected data) sequences	90.67%	0.9094	0.9067

The ProScan (Prestridge 1995) mainly performs better for polymerase III promoters like primates or mammals. Similarly for Promoter 2.0 (Knudsen 1999), the algorithm was designed to be able to discriminate between vertebrate promoter and non-promoter sequences. This might be the reason why both the tools could not provide satisfactory result for *Drosophila melanogaster*.

Table 2. Comparison of accuracy against some existing methods.

Program used (%)	NNPP threshold (0.8)	SoftBerry (TSSP)	ProScan Vers. 1.7	Dragon Pro-moter Finder Vers. 1.4	Promoter 2.0 Pred. Server	Proposed method
Sensitivity	68	88	0	12	0	90.94
Specificity	76	90	100	100	78	90.67

For Dragon Promoter Finder (Bajic et al. 2002) the smaller sensitivity value indicates that it was more prone to false negative. The proposed method exhibits better accuracy compared to other two methods NNPP (Reese et al. 1996) and SoftBerry(TSSP).

The proposed method developed for identifying a *Drosophila melanogaster* promoter from a DNA sequence depends on statistical data analysis. TATA box, TSS, DPE, CpG Island, CAAT box or BRE element was not considered in this paper. Rather the differential frequency distribution between promoter and non-promoter sequences were exploited to successfully identify *Drosophila melanogaster* promoters. The experimental result exhibits that consistent and promising performance can be achieved using this approach. Also higher value of sensitivity and specificity indicates the proposed method is less prone to false negative and false positive. Further development of the method can be investigate by incorporating the TATA box, TSS, DPE, CpG Island, CAAT box or BRE element as additional feature to improve the result. Also the method can be tested for identifying promoters of other organisms.

References

Bajic VB, Seah SH, Chong A, Zhang G, Koh JLY and **Brusic V** (2002) Dragon promoter finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics **18**: 198-199.

Burz DS, Rivera-Pomar R, Jackle H and **Hanes SD** (1998) Cooperative DNA binding by Bicoid provides a mechanism for threshold-dependent gene activation in the Drosophila embryo. EMBO J. **17**: 5998 -6009.

Cortes C and **Vapnik V** (1995) Support vector network. Machine learning **20**: 273-297.

Hughes A (2003) The Central Dogma and Basic Transcription. Connexions. Version 1.5.

Kasabov N and **Pang S** (2004) Transductive Support Vector Machines and Applications in Bioinformatics for Promoter Recognition. Neural Information Processing - Letters and Reviews. **3**(2): 31-38.

Knudsen S (1999) Promoter 2.0: for recognition of Pol II promoter sequences. Biotechnologies **15**: 356–361.

Ohler U (2006) Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. Nucleic Acids Research. **34(20):** 5943-5950(8).

Pedersen AG, Baldi P, Chauvin Y and **Brunak S** (1999) The biology of eukaryotic promoter prediction - a review. Computers & Chemistry. **23**: 191-207.

Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. J. Mol. Biol. **249**: 923-932.

Reese M, Harris NL and Eeckman FH (1996) Large scale sequencing specific neural networks for

promoter and splice site recognition. Biocomputing Proceedings of Pacific Symposium. Edited by Hunter L, Klein T, World Scientific Co.

Rongxiang Liu and **David J** (2002) Consensus Promoter Identification in the Human Genome Utilizing Expressed Gene Markers and Gene Modeling. USAGenome Res. **12**: 462-469.

Sharan R and **Myers EW** (2005) A motif-based framework for recognizing sequence families. Bioinformatics. **21**(1): i387-i393(1).

Small S, Kraut R, Hoey T, Warrior R, and **Levine M** (1991) Transcriptional regulation of a pair-rule stripe in Drosophila. Genes Dev. **5**: 827-839.

Vapnik V (1982) Estimation of dependences based on empirical data. Springer-Verlag, New York.

Xie X, Wu S, Lam KM and **Yan H** (2006) PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. Bioinformatics. **22**(22): 2722-2728.

From Web sites:

Drosophila Core Promoter Database (DCPD) - Created by Alan K. Kutach and Scott L. in the laboratory of James T. Kadonaga. http://www-biology.ucsd.edu/labs/kadonaga/

EPD-The Eukaryotic Promoter Database, Current Release 91. http://www.epd.isb-sib.ch/