BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Prediction model

Volume 6(6)

An ANN-GA model based promoter prediction in *Arabidopsis thaliana* using tilling microarray data

Hrishikesh Mishra, Nitya Singh, Krishna Misra, Tapobrata Lahiri*

Division of Applied Sciences and Indo-Russian Centre for Biotechnology, Indian Institute of Information Technology, Allahabad, India; Tapobrata Lahiri - Email: tlahiri@iiita.ac.in; Phone: 91-532-2922242; Fax: 91-532-2430006; *Corresponding author

Received February 10, 2011; Accepted May 09, 2011; Published June 06, 2011

Abstract:
Identification of promoter region is an important part of gene annotation. Identification of promoters in eukaryotes is important as promoters modulate various metabolic functions and cellular stress responses. In this work, a novel approach utilizing intensity values of tilling microarray data for a model eukaryotic plant *Arabidopsis thaliana*, was used to specify promoter region from non-promoter region. A feed-forward back propagation neural network model supported by genetic algorithm was employed to predict the class of data with a window size of 41. A dataset comprising of 2992 data vectors representing both promoter and non-promoter regions, chosen randomly from probe intensity vectors for whole genome of *Arabidopsis thaliana* generated through tilling microarray technique was used. The classifier model shows prediction accuracy of 69.73% and 65.36% on training and validation sets, respectively. Further, a concept of distance based class membership was used to validate reliability of classifier, which showed promising results. The study shows the usability of micro-array probe intensities to predict the promoter regions in eukaryotic genomes.

Keywords: core promoter, TATA box, artificial neural network, class membership.

Background:
With the continuously increasing number of various genome sequencing projects, predicting promoters has become one of the prime focuses for gene identification and annotation. Promoter-prediction is multi-informative, as along with delineating one end of gene and being key to gene regulation; it also gives a cue to functional aspect of gene **[1]**. In eukaryotes promoter guides the cell development and differentiation, tissue morphogenesis and specificity, hormonal communication, and cellular stress responses **[2]**. Moreover the knowledge of promoter may lend clues for function of completely anonymous proteins which cannot be retrieved from already predicted amino acid sequences **[3]**. In eukaryotes promoters are classically defined as the start site of transcription (TSS) **[1]**. According to a geneticist's view promoters are cis-acting elements deciding the site and rate of transcription, while according to biochemist's view; these are target sites of transcription factors **[4]**. Eukaryotic promoters are of three types viz., core promoters, proximal promoters and distal promoters that have different roles in gene regulation. Core promoters, also known as minimal promoters are located ~80 to 100 base pairs (bp) around TSS and are required for initiation of transcription. Proximal promoters located ~250 to 1000 upstream of core promoter, are position and species specific, involve transcription factor binding site and initiate basal transcription. Distal promoters located further upstream and also known as enhancers involve additional regulatory element binding sites at distal regions from transcription start site **[5, 6]**. For computational prediction, core promoter is more important as it is the first essential site required for the initiation of transcription **[1]**. Eukaryotic core promoters have TATA box as an integral part which lies 25-30 bp upstream of TSS **[7]**. So for retrieving the locations of promoter sequences, location of TATA box can be searched and region upstream to TSS with TATA box of known genes can be targeted. Major challenges associated with promoter prediction methods involve weak models used for promoter regions and algorithmic constraints **[5]**. Also, it has been reported that using existing promoter prediction methods miss 30-40% of true promoters and have a false positive rate of 45-60% **[8]**. Complex architecture of promoter sequences presents a computational problem yet to be solved satisfactorily **[3]**. Methods for promoter prediction are in their infancy and level of accuracy achieved is low **[9]**.

The success of prediction depends upon biological model, type and quality of training data utilized **[10]**. In general, most of the computational methods for promoter prediction are based on models searching for organization of promoters, promoter location **[11]** or for the hallmarks of promoters as CpG islands **[12]**, TATA boxes **[13]**, CAAT boxes **[14]**, specific transcription factor binding sites (TFBSs) **[13, 14]**, pentamer matrix **[15]** and oligonucleotides **[16]**. Various computational strategies applied for this purpose involve: neural networks **[12]**, linear and quadratic discriminant analyses **[17]**, interpolated Markov Model **[14]**, independent component analysis (ICA) **[18]**, and non-negative matrix factorization (NMF) **[19]**. All the methods have their own advantages and limitations. Selection of an appropriate combination of biological features and computational approach for accurate promoter prediction is still an open issue. In the present work we have dealt with this problem by utilizing tilling microarray intensities for nucleosome rich DNA of chromosome 1 of *Arabidopsis thaliana*. Although, microarray expression data have been used to establish correlation between DNA sequence and promoter-strength **[20]**, but using it for predicting the promoter location is a novel approach.

It has been shown that nucleosome is located at ~ +40±15 bp of TSS i.e., mostly downstream from typical TATA box position. It can be inferred that TATA box is situated within 5' half of nucleosomal DNA or right upstream of it. Further, nucleosomal distribution around the TSS has a strong correlation with promoter region **[21]**. Therefore, microarray experiment was designed to retrieve nucleosomal rich DNA regions to ensure the incorporation of promoter regions. Further, as analyzed from various biological features adopted for promoter prediction, length and composition of DNA sequence i.e., GC content plays an important role in distinguishing promoter region from non-promoter **[22]**. Difference in melting temperature due to different composition of promoter sequence may result into different pattern in hybridization intensity which can be detected. Our work explored conglomerative machine learning approach to utilize data obtained from high throughput and high resolution tilling microarray data for detection of promoter region.

Methodology:

Dataset:

Intensity values for mononucleosomal DNA regions were obtained from tilling microarray data downloaded from gene expression Omnibus (GEO) at NCBI having GEO accession: GEO25553, available at http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=gse25553. Also whole genomic DNA tilling microarray intensity values generated under the same experimental conditions were obtained from same geo data series, which were used for normalization of mononucleosomal DNA intensity data. Intensity data for mononucleosomal regions was vectorized to consider the context intensity values of neighboring probes as a window around centrally positioned intensity value. Several window sizes were tested and finally size of 41 was chosen by trial and error to include 41 consecutive probes, considering the average range of number of probes comprising different known promoter regions in training data.

Each window comprised the intensity for the probe in question (P) positioned at the centre of the window i.e., 21st position so that P was represented with 20 intensity values on both sides. A probe in question (P) refers to the tilling array position against which the nucleosome was to be detected. For extremities zero padding of required size was done. Thus finally the input feature for 779303 data yields 779303 row vectors each consisting of 41intensity values. Promoter locations for fifty genes from chromosome 1 of *A. thaliana* were identified considering the presence of TATA box in the proximity (17-25 base pairs upstream) of start codon for the genes concerned. Intensity vectors corresponding to these promoter regions were retrieved to prepare the initial data set. All the 1492 vectors representing 50 promoter regions and 1500 vectors representing non-promoter regions were retrieved. Thus we had 2992 data each comprising a vector of 41 consecutive probe intensity values.

Training set and Validation set:

Data was divided with a ratio of 30:20 for training (897 and 900 data for promoters and non-promoters respectively) and validation (595 and 600 for promoters and non-promoters respectively) data set respectively. Thus training and validation sets consisted of 1797 and 1195 vectors respectively. Mean was calculated for data of each column of training set. Training data was normalized by subtracting the column wise mean from each element of column concerned. The mean calculated from training data was used for normalizing the validation set also.

Model development using Neuro-GA approach:

The classification of data was done using a feed-forward backpropagation network model, where normalized probe intensity vectors were fed as input. Several architectures for neural network were tried for training. Finally a three layer neural network giving maximum efficiency was chosen as classification model. While the input layer comprised forty one inputs, first and second hidden layers consist of twenty and ten nodes respectively. The output layer contains two nodes to represent two classes of our interest i.e., promoter and non-promoter. Target output vectors were created as [-1 1] for promoter region and [1 -1] for non promoter regions. Tan sigmoid transfer function was used in both hidden and output layers. To enhance the speed of learning Levenberg-Marquardt algorithm was applied. Mean square error of decisions was used as performance function for the network. The weights and biases of trained ANN were arranged into a vector comprising 1072 variables and optimized using a strategy following a published protocol using Genetic Algorithm **[23]**. Initial population was created by randomly adding or subtracting uniformly distributed random numbers between 0 to 10% of the value of each element of the combined vector obtained by the combination of weight-matrices and bias-vectors. Fitness scaling was done using the rank of each individual. Scattered crossover and uniform mutation with rate of 0.01 was used to generate new generations of population. Percentage efficiency in correctly classifying validation data using current weights and biases was used as fitness function. Optimized weights and biases were used for promoter-prediction.

Cross validation using efficiency measures and distance based membership:

Sensitivity, specificity and accuracy were used as measures of efficiency of classification as the classification of probe intensity values into promoter and non-promoter classes is a binary classification **[24]**. Further a new statistic 'distance based membership' of correctly predicted class by neural network in the original class was calculated. Euclidian distances were measured between the predicted output and original target output as given in **Supplementary material**.

Discussion:

Accuracy, sensitivity and specificity of trained network on training data set were found to be 69.73%, 73.91% and 65.56% respectively. Accuracy, sensitivity and specificity of trained network on validation data set were found to be 65.36%, 71.26% and 59.50% respectively, as tabularized in **Table 1 (see Supplementary material)**. Our method has given false positive rate of 40.50% on validation data set as compared to existing methods which have a false positive rate of 45-60% as discussed previously. Similarly false negative rate was found to be 28.74% which is lower than currently available methods known to miss 30-40% of true promoters. Results obtained from this pilot study in the direction of a new approach for predicting eukaryotic promoters appear to be promising. Nucleosomal DNA experiment intensity values were calibrated by using whole genomic DNA intensities generated by hybridizing MNase treated genomic DNA (~150 bp) to similar Arabidopsis tilling array chip. So finally the calibrated data had intensity values for nucleosome rich regions along with embedded genomic DNA sequence intensities. This facilitated statistical analysis which used the concept of context to incorporate the effect of genomic DNA intensity values. This might have helped neural network classifier to detect the hidden pattern in data for promoter regions.

Moreover the concept of distance based class membership was introduced that gives an idea about the confidence with which the neural network predicted the class for input data. Values obtained for distance based class membership were found to be 64.14% and 64.94% for training and validation sets respectively. It indicates that the classification decision obtained through our classifier was reliable. The study paves a way to utilize high throughput microarray data for fast prediction of promoter locations. None the less, the study is also helpful to establish a link between promoter location, gene expression and nucleosomal dynamics in the eukaryotic genome. Study of nucleosomal positioning and dynamics in genome is of great importance and interest as it governs the gene expression or suppression events which are at the heart of cell metabolism.

Conclusion:

The results obtained from our study which involves robust classification of tilling microarray data provide new insights for the promoter prediction problem. The error involved in this method is possibly because of low resolution in the tilling microarray system. Moreover, the apparent error may be due to the identification of unknown promoter sites for further investigation. The distance based class membership statistics used to validate the classifier accuracy makes the results more reliable and may be used by the machine learning community to validate the outcome of their classifier.

Acknowledgement:

We gratefully acknowledge the helpful discussion about microarray data used in this work by research team of Dr Sameer Sawant, National Botanical Research Institute (NBRI), Lucknow, India. We are also grateful to Ministry of Human Resource and Development (MHRD) and Department of Science and Technology (DST) for their joint financial support to Division of Applied Sciences and Indo-Russian Centre for Biotechnology at IIIT to continue this work.

References:

[1] Pederson AG *et al. Comput Chem*. 1999 **23**: 191 [PMID: 10404615]
[2] Ioshikhes I *et al. Proc Natl Acad Sci U S A*. 1999 **96**(6): 2891 [PMID: 10077607]
[3] Solovyev VV & Shahmuradov IA. *Nucleic Acids Res*. 2003 **31**(13): 3540 [PMID: 12824362]
[4] Cavin Périer R *et al. Nucleic Acids Res*. 1998 **26**(1): 353 [PMID: 9399872]

[5] Carvalho AM *et al. IEEE/ACM Trans Comput Biol Bioinform*. 2006 **3**(2): 126 [PMID: 17048399]
[6] Zhang MQ. *BMC Bioinformatics*. 2007 **8** Suppl(6): S3 [PMID: 17903284]
[7] Smale ST & Kadonaga JT. *Annu Rev Biochem*. 2003 **72**: 449 [PMID: 12651739]
[8] Duret L & Bucher P. *Curr Opin Struct Biol*. 1997 **7**(3): 399 [PMID: 9204283]
[9] Abeel T *et al. Genome Res*. 2008 **18**(2): 310 [PMID: 18096745]
[10] Werner T. *Mamm Genome*. 1999 **10**(2): 168 [PMID: 9922398]
[11] Vanet A *et al. Res Microbiol*. 1999 **150**: 779 [PMID: 10673015]
[12] Bajic VB & Seah SH. *Genome Res*. 2003 **13**(8): 1923 [PMID: 12869582]
[13] Knudsen S. *Bioinformatics* 1999 **15**(5): 356 [PMID: 10366655]
[14] Ohler U *et al. Genome Biol*. 2002 **3**(12): RESEARCH0087 [PMID: 12537576]
[15] Bajic VB *et al. IEEE Intell Syst Mag*. 2002 17: 64
[16] Scherf M *et al. J Mol Biol*. 2000 **297**(3): 599 [PMID: 10731414]
[17] Davuluri RV *et al. Nat Genet*. 2001 **29**(4): 412 [PMID: 11726928]
[18] Chen L *et al. BMC Bioinformatics* 2008 **9**: 416 [PMID: 18837990]
[19] Hutchins LN *et al. Bioinformatics* 2008 **24**(23): 2684 [PMID: 18852176]
[20] Kiryu H *et al. Bioinformatics* 2005 **21**(7): 1062 [PMID: 15513998]
[21] Ioshikhes I *et al. Proc Natl Acad Sci U S A*. 1999 **96**(6): 2891 [PMID: 10077607]
[22] Choi CH *et al. Nucleic Acids Res*. 2004 **32**(4): 1584 [PMID: 15004245]
[23] Lahiri T *et al. Online Journal of Bioinformatics* 2009 **10**(1):29
[24] Baldi P *et al. Bioinformatics*. 2000 **16**(5): 412 [PMID: 10871264]

Edited by P Kangueane
Citation: Mishra *et al.* Bioinformation 6(6): 240-243 (2011)

BIOINFORMATION

open access

Supplementary material:

Euclidian distances:

$$d=\sqrt{((e_1-o_1)^2 + (e_2-o_2)^2)}$$

where e_1, e_2 are expected outputs and o_1, o_2 are predicted outputs of two output nodes respectively. For example, maximum possible distance between the predicted and original outputs for the promoter can be considered when the predicted outputs is (1 -1) instead of (-1 1) and vice versa for a non-promoter. Therefore, value of maximum possible distance (max_d) between predicted and expected output comes out to be $2\sqrt{2}$. Class membership was thus defined as: $m=1-d/max_d$. Using this distance their actual membership for the class assigned to them was calculated. Mean value obtained for class memberships for both the classes was used as measure of robustness of classification.

Table 1: Detailed results of classification on training and validation sets (TP= number of true positives, TN= number of true negatives, FP= number of false positives, FN=number of false negatives)

	Size of data	TP	TN	FP	FN	Sensitivity	Specificity	Accuracy
Training Set	1797	663	590	310	234	73.91%	65.56%	69.73%
Validation Set	1195	424	357	243	171	71.26%	59.50%	65.36%