IJERST

International Journal of Engineering Research and Science & Technology

ISSN : 2319-5991
Vol. 3, No. 1
February 2014

www.ijerst.com

Email: editorijerst@gmail.com or editor@ijerst.com

ISSN 2319-5991 www.ijerst.com
Vol. 3, No. 1, February 2014

Research Paper

PROMOTER PREDICTION USING IREM (INDUCTIVE RULE EXTRACTION METHOD)

G Karlı[1]*

Corresponding Author: **GünayKarlı ✉ gunay.karli@ibu.edu.ba*

Nowadays, the prediction of promoters has attracted many researchers' attention. Because, the number of DNA sequences has been growing fast since the beginning of the Human Genome Program (HGP) and so it becomes increasingly important to automate the identification of functional elements, such as coding region, genomic region or promoters. Promoters are regulatory regions that govern the expression of genes, and their prediction is reputed difficult, so that this issue is still open. In this study, we employIREM (Inductive Rule Extraction Method) classifier to predict promoters of DNA sequences, and evaluate their performances.The obtained results show that the classifier competes the existing techniques for identifyingpromoter regions.

Keywords: Promoter prediction, Inductive learning, Data mining, Bioinformatics

INTRODUCTION

Being the carriers of the message contained in the DNA, Proteins serve as one of the most important classes of biological molecules. Two processes are majorly involved in the synthesis of proteins, transcription being the first of these. In transcription, a single stranded RNA molecule, called messenger RNA is synthesized by using the complementary of the bases from one of the strands of DNA corresponding to a gene coding for a particular type of protein. Transcription begins with the binding of an enzyme called RNA polymerase. The resultant occurs in a certain location on the DNA molecule known as the Promoter Region. The region determines which of the two strands of DNA will be transcript and in what direction (Gabriela and Bocicor, 2012).

The difficulty in promoter identification is an indispensable aspect of biology (Óscar and Santiago, 2011; Gabriela and Bocicor, 2012). Efforts are placed to investigate the secrets of life by going into gene sequences. Yet, the genetic sequence data grows too huge to render the efforts unfruitful as indicated by numerous recent experiments.

This has led to computer scientists getting into the biological technology, and giving numerous techniques which take advantages of digitalized mechanisms to see into gene sequences (Huang, 2003).

Methods for predicting coding regions in genomic DNA sequences existed long since the 1980s, though; the programs for assembling

[1] International Burch University, Department of Information Technologies, Sarajevo, BiH.

coding sequences into translatable mRNA sequences were unavailable until the early 1990s as noted byGuigo and Burset (Guigo and Burset, 1996). Several programs that are available for biologists, such as GenViewer (Milanesi, Kolchanov and Rogozin, 1993), GeneID (Guigo, Knudsen, Drake and Smith, 1995), GenLang (Dong and Stormo, 1994), GeneParser (Stormo and Snyder, 1993), FGENEH (Solovyev, Salamov, and Lawrence, 1994), SORFIND (Hayden and Hutchinson, 1992), Xpound (Skolnick and Thomas, 1992), GRAIL (Xu Mural and Uberbacher, 1994), VEIL (Henderson, Salzberg, and Fasman, 1997), GenScan (Karlin and Burge, 1997), etc. From such techniques GeneScan and GRAIL, (Hrishikesh *et al*., 2011) are most widely used in academia and industry (Hrishikesh, Nitya, and Krishna, 2011).

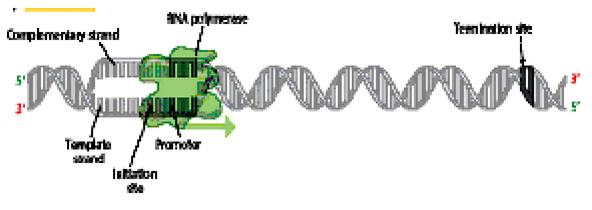
Gordon *et al.* (2006) note that these methods are based on searching motifs in a DNA sequence to decide whether a promoter exists or not.Position weight matrices and Markov models (Liu, 2002) (Luo and Yang, 2006) (Premalatha and Aravindan, 2009) are used besides statistical strategies. Additionally, artificial intelligence is also applied. ANN shave has proven acceptable profound results, but the high false positive rate (Abeel T S, 2008) (Zhang, 2009) has affected the specificity and as a result neural network has been applied. Approaches that are of this kind in promoter prediction are still a controversial. We have dealt with this problem by utilizing IREM in this research, obtaining results by evaluating the classification model proposed confirms that applying IREM for promoter sequence recognition is promising.

What is Promoter and Importance of Promoter Prediction?

The Blueprint; a common name for genetic coding is believed to possess the instructions needed by cells in environmental adaptability. This is further collaborating with research that cites that in addition to the instructions carried by the blueprint, there exists a synthesis point for most of the molecule including the RNA and proteins. The instructions contained in the blueprint are designed in such a manner that they are only readable in two ways as mentioned earlier in this paper i.e. transcription and translation (Clancy, 2008).

The messenger RNA; usually a single strand RiboNucleic Acid molecule, is amalgamated usually from one of the strands contained in DNA through a complementary of the origins. This strand does relate to a gene. Transcription will usually commence with the RNA polymerase being engulfed by a lone point in the DNA molecule known as the Promoter (See Figure 1). (Clancy, 2008).This process is facilitated majorly by the Promoter having much consideration of the responsibility it has on the transcription from a DNA strand to an RNA strand. This is further identified as the sequential upstream from the Transcriptional Start Site (TSS). This is well illustrated in the Figure 1. The entire process commences with the binding of the RNA polymerase to a promoter array in the DNA molecule up to a point where the coding is realized. Coding occurs during the upstream movement around the promoter usually starting at 3’ end of the DNA molecule to the 5’ point in a DNA molecule (Huang, 2003).

Figure 1: Position of The Promoter in aDNA Sequence



In rare cases, then there may be the existence of the upstream of the TSS of DNA array that possess transcriptional characteristics, thus the presence of the promoter may not be a necessity. When promoter prediction is incorporated, a researcher is at ease to constrict down the entirely colossal DNA sequences. This paper acknowledges that through biological interventions, it is made easier to verify the DNA sequence that may either be transcribed or no. This though comes along with economic constraints.

In the recent past, the integration of computerization in the promoter identification and prediction has raised debate. And the results obtained by evaluating the classification model proposed in this paper confirm that applying IREM for promoter sequences recognition is promising.

METARIALS AND METHODS

There are two core classes of the promoter prediction, namely '+' and '–'. These classes will denote the existence of promoter prediction in the DNA sequence, having the '+' denoting for a positive indication of promoter location in the DNA sequence and the '–' denoting the absence of promoter locations in the DNA sequence. This research paper proposes to deal with a supervised learning technique in the prediction of promoter regions in the DNA sequence.

Data Set

The research sought to incorporate the E. Cole promoter gene arrays of DNA in the testing the proficiency of IREM. Such data were collected from the UCI Repository (Frank and Asuncion, 2010); this contains a set of 106 promoter and non-promoter instances. The research paper notes that such data is viable in the comparisons of ANN with the models existing in the literature; additionally such information involving the use of the data set is publicly available (Gabriela and Bocicor, 2012).

The 106 DNA arrays are composed of 57 nucleotides each. 53 of the DNA sequences in the data set had a '+' denoting, indicating the presence of promoter location in the DNA array. The research then sought to align the (+) parameter instances separately allowing for transcription. The following data characterize the (+) instances as observed from the experiment. One is that for every occurrence the (+) represents for the promoter positive presence, a name was also given in each instance and a classification of the DNA array was made composing of A, T, G and C stand for Adenine, Thymine, Guanine, Cytosine (Frank & Asuncion, 2010).

IREM (INDUCTIVE RULE EXTRACTION METHOD)

Usage of this technique in predicting promoter region in DNA is introduced in this section which is a newly developed IREM (Inductive Rule Extraction Method). E. coli data set is composed of 106 DNA sequences, each having a length of 57 nucleotides as described above in this paper. When engaging a digitalized mindset, E. coli data set consists of 106 instances with 57 attributes having 4 values. These attributes are defined as the position of nucleotides in the 57-element sequence with each attribute taking 4 values which are A-Adenine, C-Cytosine, G-Guanine and T-Thymine. IREM builds its rule-base using attribute-value pairs in the set.

When generating a rule-base with powerful rule, attribute-value pairs with higher importance are employed. Consequently, a significant quest develops an "how the pairs with a higher information value can be determined". IREM applies its own "cost function" to calculate the information value of the pair in the set. It deals with regard to the values of the lower cost as an indicator of higher information value. With this, attribute-value with the higher value is given a greater priority in the process of producing rule-base of the predicting system as described by the following algorithm;

Step1. In a given E. coli data set (training set), distribution of probability of the each attribute-value pairs is computed.

Step2. The class-based entropy is computed for each attribute and value.

Step3. By using computed probability distributions and class-based entropy, the cost of the pairs is calculated.

Step4. Any value of which class-based entropy equals zero for $n = 1$ can be selected as a rule. The pairs are converted into rules. The classified examples (instances) are marked.

Step5. Go to step8.

Step6. Beginning of the first unclassified example, combinations with no values are formed by taking the value of the attributes whose cost is smaller.

Step7. Each combination is applied to all of the examples in the set. From the values composed of n combinations, those matching with only in class are converted into a rule. The classified examples are marked.

Step8. If all of the examples in the training set are classified then go to step11.

Step9. Perform $n = n + 1$ expression.

Step10. If $n < N$ the go to step6

Step11 If there is more than one rule representing the same examples; the most general one is selected.

Step12. End.

COST FUNCTION

Being able to compute the cost of each attribute-value in a given training set using class-based entropy is an indispensable feature of the IREM algorithm. This expression best evaluate the costs.

P(VC)	:	Probability distribution of each value in the data set.
|VC|	:	The number of value in any class
|V|	:	The number of value in data set
TV,C	:	Complement of probability of value V in class C
E(V)	:	Entropy of value V in data set
E(X)	:	Entropy of attribute in data set
E(V|C)	:	Class-based entropy of value V in class C
M(V|C)	:	Cost of value V in class C

X = {x1, x2, x3,…,xy} : attribute set of data set. y, the number of attributes in data set

V = {v1, v2, v3, …, vn} : value set of data set. n, the number of values in data set

C = {c1, c2, c3, …, cm} : m, the number of classes in data set

Class-based entropy of value V in class C, E(V|C) is calculated as follows.

$$P(VC) = |VC| / |V| \quad ...(1)$$

$$TV,C = 1 - P(VC) \quad …(2)$$

$$E(V|C) = TV,C * E(V) \quad …(3)$$

Cost of value V in class C, M(V|C) is calculated as follows.

$$M(V|C) = \begin{cases} 0, P(VC) = 1 \\ -1, P(VC) = 0 \\ TV, C{*}E(V) + TV, C{*}\ E(X), 0,<P(VC)<1 \end{cases} \quad …(4)$$

approach for promoter sequence recognition using IREM, in addition to providing a comparison past approaches. Ability to computing class-based entropy of each attribute-value in a given training set is the most important features of IREM algorithm. Probability distributions of each nucleotide forming DNA sequence were computed in terms of promoter and non-promoter classes.

Secondly, entropy of training set was evaluated, though the entropy does not contain class information for the attribute-value pairs. With this, the integration of entropy of the training set and the probability distributions of the attribute-value, class-based entropy was computed for each value in the DNA sequence data set. In this way, rules produced by the algorithm were formed by attribute-value whose information value was a successful maximum.

The section attempts to evaluate the approach to promoter sequences posed by the experiment with much consideration to other approaches highlighted in the literature. Such endeavors are usually conducted in two phases that showcase a precise learning algorithm, training and testing. Training will involve establishing a classification model; testing entails the implementation of the classification model previously established.

A standard 5-fold cross-validation was integrated into the evaluation of the IREM performance by having the dataset being randomly portioned into 5 subsets. This classification ensures an equal ratio of (+) and (–) promoter locations in the DNA array.

The training occurred on the IREM for a series 5 times engaging only 4 subsets for each training while as retaining the remaining 5 for testing. As a result, 5 models were established during the cross-validation. Additionally, a final prediction performance was carried out on the subsets evaluating the average results from the experiment..

The performance of the promoter predictions was evaluated using the threshold parameters; accuracy (ACC), Mathew's Correlation Coefficient (MCC), sensitivity (SE) and specificity (SP). A couple of equations were integrated to affirm to the results. These were;

$$SE = TP/(TP+FN) \quad …(5)$$

$$SP = TN/(TN+FP) \quad …(6)$$

$$ACC = (TP+TN)/(TP+TN+FP+FN) \quad …(7)$$

$$MCC = ((TP*TN) – (FN*FP))/SQRT((TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)) \quad ...(8)$$

TP is true positive (promoter predicted as promoter)

FN is false negative (promoter predicted as non-promoter)

TN is true negative (non- promoter predicted as non- promoter)

FP is false positive (non- promoter predicted promoter).

The detailed performance of module in term of SE, SP, ACC and MCC is shown in Table 1.

An application of the cross-validation using a "leave-one-out" methodology was integrated for evaluating the performance of the learning algorithms in the literature. Leave-one-out cross-validation (LOOCV) is a special case of k-fold cross-validation where k equals the number of instances in the data. The research also integrated LOOCV with which an accuracy estimate was obtained to be almost unbiased(Efron, 1983). When the available data are very rare, especially in Bioinformatics where only dozens of data samples are available then this application is a vital integration.

IREM introduced in this paper outperforms existing classifier for promoter prediction: it is better than BP, ID3, KB, NN and O'Neil with a dire necessity to the consideration of occurrence of the error (see Table 2).

Table 1: Performance of IREM in Term of SE, SP, ACC and MCC

	Model 1	Model 2	Model 3	Model 4	Model 5	Average
ACC	0.75	0.95	0.80	0.95	0.69	0.8285
SE	0.80	1.00	0.80	1.00	0.54	0.8277
SP	0.70	0.90	0.80	0.90	0.85	0.8292
MCC	0.50	0.90	0.60	0.90	0.40	0.6632

Table 2: The Errors of Some Machine Learning Algorithms on Promoter Data Set

System	Errors	Classifier
REX-1	0/106	Inductive L.A
IREM	2/106	Class-based entropy
KBANN	4/106	A hybrid ML system
BP	8/106	Standard backpropagation with one layer
O'Neill	12/106	Ad hoc tech. from the bio. lit.
Near-Neigh	13/106	A nearest neighbours algorithm
ID3	19/106	Quinlan's decision builder

CONCLUSION

Taking into account of the results derived from the above experiment that employing IREM for promoter prediction has a promising result and posing room for improvements, undoubtedly, increasing the accuracy of the obtained results from a computational perspective, and proving promoter prediction as an indispensable aspect in Bioinformatics.

The newly developed IREM has been proposed for solving the problem. Attribute-value pairs with higher importance were employed in a bid for generating rule-base with an efficient rule. The values of the lowest cost are considered to indicate higher information value, thus, given the attribute-value with higher value a greater priority in the process of producing rule-base of the predicting system. IREM applies its own "cost function" to calculate the information value of the pair in the set.

REFERENCES

1. Abeel T, Saeys Y, Bonnet E and Rouzé P (2008), "Generic eukaryotic core promoter prediction using structural features of DNA", *Genome Research,* Vol. 18, No. 2, pp. 310-323.
2. Clancy S (2008), "Nature Education", *DNA transcription,* Vol. 1, No. 1, p. 41.
3. Dong S and Stormo G (1994), "Gene structure prediction by linguistic methods", *Genomics,* Vol. 23, pp. 540-551.
4. Efron B (1983), "Estimating the error rate of a prediction rule: Improvement on cross-validation", *J. Am. Stat. Assoc.,* Vol. 78, pp. 316-331.
5. Frank A and Asuncion A (2010), *UCI machine learning repository*, Retrieved from http://archive.ics.uci.edu/ml/
6. Gabriela C and Bocicor M-I (2012), "Promoter Sequences Prediction Using Relational Association Rule Mining", *Evolutionary Bioinformatics,* Vol. 8, pp. 181-196.
7. Gordon J, Towsey M and Hogan J (2006), "Improved prediction of bacterial transcription start sites", *Bioinformatics,* Vol. 22, No. 2, pp. 142-148.
8. Guigo M and Burset R (1996), "Evaluation of gene structure prediction programs", *Genomics,* Vol. 3, No. 34, pp. 353-367.
9. Guigo R, Knudsen S, Drake N and Smith T (1995), "Prediction of gene structure", *Journal of Molecular Biology*, Vol. 226, pp. 141-157.
10. Hayden G and Hutchinson M (1992), "The prediction of exons through an analysis of spliceable open reading frames", *Nucleic Acids Research,* Vol. 20, pp. 3453-3462.
11. Henderson J, Salzberg S and Fasman K (1997), "Finding genes in DNA with a hidden Markov model", *Journal of Computational Biology,* Vol. 2, No. 4, pp. 127-141.
12. Hrishikesh M, Nitya S and Krishna M (2011), "An ANN-GA model based promoter prediction in Arabidopsis thaliana using tilling microarray data", *Bioinformation,* Vol. 6, No. 6, pp. 240-243.
13. Huang J-W (2003), *Promoter Prediction in DNA Sequences,* Kaohsiung, National Sun Yat-sen University.

14. Karlin C and Burge C (1997), "Prediction of complete gene structures in human genomic DNA", *J. Mol. Biol.,* Vol. 268, pp. 78-94.

15. Liu R A (2002), "Consensus promoter identification in the human genome utilizing expressed gene markers and gene modelling", *Genome Research,* Vol. 12, pp. 462-469.

16. Luo Q and Yang W A (2006), "Promoter recognition based on the interpolated Markov chains optimized via simulated annealing and genetic algorithm", *Recognition Letters Pattern,* Vol. 9, No. 27, pp. 1031-1036.

17. Milanesi L, Kolchanov N and Rogozin I (1993), "GenViewer: A computing tool for protein coding regions prediction in nucleotide sequences", *The 2nd International Congress on Bioinformatics, Supercomputing and Complex Genome Analysis,* pp. 573-587.

18. Óscar B and Santiago B (2011), "Cnn-promoter, new consensus promoter prediction program", *Revista EIA,* Vol. 15, pp. 153-164.

19. Premalatha C and Aravindan C A (2009), "On improving the performance of promoter prediction classifier for eukaryotes using fuzzy based distribution balanced stratified method", *Proceedings of the International Conference on Advance in Computing, Control, and Telecommunication Technologies IEEE,.* ACT.

20. Skolnick A and Thomas M (1992), "A probabilistic model for detecting coding regions in DNA sequences", *IMA J. Math. Appl. Med. Biol.,* Vol. 11, pp. 149-160.

21. Solovyev V, Salamov A and Lawrence C (1994), "Prediction of internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames", *Nucleic Acids Research,* Vol. 22, pp. 5156-5163.

22. Stormo E and Snyder E (1993), "Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks", *Nucleic Acids Research,* Vol. 21, pp. 607-613.

23. Wang M, Yin M and Jason T (2013), "GeneScout: A data mining system for predicting vertebrate genes in genomic DNA sequences", *Information Sciences,* Vol. 163, Special Issue, pp. 201-218.

24. Xu Y, Mural R and Uberbacher E (1994), "Constructing gene models from accurately predicted exons: An application of dynamic programming", *Comput. Appl. Biosci.,* Vol. 10, pp. 613-623.

25. Zhang Y-J (2009), "A novel promoter prediction method inspiring by biological immune principles", *Global Congress on Intelligent Systems,* pp. 569-573.

IJERST

International Journal of Engineering Research and Science & Technology
Hyderabad, INDIA. Ph: +91-09441351700, 09059645577
E-mail: editorijlerst@gmail.com or editor@ijerst.com
Website: www.ijerst.com

ISSN 2319-5991
01
9 772319 599001