International Journal of Engineering Research and Development
e-ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 9, Issue 10 (January 2014), PP. 01-07

Determining the Promoter Region in the DNA using ke-Rem (ke-Rule Extraction Method)

Günay Karlı[1]
[1]*International Burch University, Faculty of Engineering and IT, IT Department, Sarajevo, Bosnia and Herzegovina.*

Abstract:- Biologists endeavor to reveal the secrets of life by looking into gene sequences. Concordantly, determining the promoter region in the DNA is an important step in the process of detecting genes. However the gene sequence data grow too huge recently. Conventional methods remain incapable to predict promoter so it becomes increasingly important to automate the identification of functional elements, such as coding region, genomic region or promoters. Thus many computer scientists are interested in the biological technology, and improve some data mining methods which take advantages of computer power to see into gene sequences. In this study, we employ ke-REM (ke-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 identifying promoter regions.

Keywords:- Promoter prediction, inductive learning, data mining, bioinformatics, DNA.

I. INTRODUCTION

Genome coding segments for transfer ribonucleic acids (tRNAs), messenger ribonucleic acids (mRNAs) and ribosomal ribonucleic acids (rRNAs) are known as genes [1]. Proteins have amino acids whose sequence is determined by mRNAs. Prokaryotic cells have a simple mechanism since all the genes tend to be converted into the corresponding mRNA and then finally into proteins [2]. Gene finding or genome analysis generally relates to that part of computational biology that involves identification of stretches of sequence algorithmically, and it is basically the genomic DNA which is functional biologically. This particularly incorporates protein-coding genes as well as various functional elements like regulatory regions and RNA genes. To clearly understand a species' genome, one of the most crucial and significant step is to understand gene finding.

For prokaryotes, it is relatively simple to predict the Computational Gene since all the genes are basically converted into the corresponding mRNA and finally into proteins. However, for eukaryotic cells, the process becomes more complicated since there is interruption of coding of the DNA sequence by random sequences commonly known as introns. There are a number of questions which biologists have currently attempted to answer and they include [3]:

- What section of a DNA sequence codes for a protein and what section is considered as junk DNA?
- How can a junk DNA be classified as intron, transposes, untranslated region, regulatory elements, dead genes, etc.?
- Dividing a genome that has been newly sequenced into the coding genes and non-coding regions.

In a DNA, determination of the promoter region is a crucial step in the process of detecting genes and this implies that the problem of determining a promoter is of major significance in biology [4] [5]. Biologists have tried to analyze the secrets of life by investigating the gene sequence. Nevertheless, the growth of data on the gene sequence has been growing rapidly. Therefore, most of computer scientists use biological technology to come up with methods generated by the power of a computer to see gene sequences [6].

Despite the fact that approaches for determining regions of coding in genome DNA sequences have been in existence since the nineteenth century, programs designed for combining coding sequences into mRNA sequences that could be translated were invented in early 20th century [7]. Biologists have since then had various programs at their disposal including GenViewer [8], GeneID [7], GenLang [9], GeneParser [10], FGENEH [11],

SORFIND [12], Xpound [13], GRAIL [14], VEIL [15], GenScan [16], etc. Two tools, which include GRAIL and GenScan, are the most widely used tools both in the industry and in academics [17].

Most of the above approaches are founded on motifs searching in regard to DNA sequence to establish whether it forms a promoter or not [18]. In this context, the search is done through the help of Markov models and position weight matrices [19][20]. Artificial intelligence has also been used to supplement statistical intelligence. More specifically artificial neural networks have produced acceptable values with the only disadvantage being high positive rates that are false [21][22]. However, this has not curtailed their application to solve other bioinformatics problems.

Computational method for predicting accurate promoter is still to be wholly developed. Currently, biologists use ke-REM. The purpose of this paper is to explicate ke-REM can be successful be used for promoter sequences.

A. What is promoter and importance of promoter prediction?

A promoter by definition is a DNA non-coding region responsible for initiating transcriptions a specific gene. They are usually located on the upstream and on the same strand of the DNA-towards the anti-sense strand's 3' region also referred to as the template strand. A promoter may be characterized by a nucleotide sequence of between 100 to 1000 base pair long [23]. A special enzyme, RNA polymerase, is needed for mRNA transcription. This enzyme needs to attach itself to the DNA near a gene for it to qualify to be called a promoter sequence. Sequences comprises of specific response and DNA sequences responsible for providing a fully secure primary binding site for the enzyme as well as for proteins referred to as transcription factors.

Eukaryotic and Prokaryotic promoters vary from each other. In prokaryotic organism ς70 sigma factor is able to identify specific promoter sequences, which in this case are 5'TATAAT3' and 5'TTGACA3' (-10 and -35 respectively) through the help of ς70 subunit of the polymerase enzyme[24]. Eukaryotic organism on the other hand is more complex requiring at least 7 different factors for the polymerase II enzyme to bind to the promoter.

The promoter intensity correlates with identity degree to the sequence but separated by the spacer length. Dense promoters are however founded closer to gene [25][26]. It has for long been thought that for transcription activity to be optimal, various promoter elements' combinations including -35 and -10 hexamers, must be in existence coupled with downstream and upstream regions [25]. According to this school of thought, RNAP works both regions of the two hexamers in sequence and promoters of A+ T-rich sequences upstream of the −35 hexamer [26][27] in several E. coli or Bacillus subtilis were identified as facilitating increased transcription in vitro when accessory proteins were absent [28]. Different upstream sequences show different effects on transcription increasing it from a mere 1.5 to 90 fold [29]. Those promoter sequences that are characterized by powerful binding affinity have a direct effect on mRNA transcription.

Regardless of whether a transcribed DNA sequence can be identified through biological testing or not, experiments are known to be time consuming and costly. The promoter prediction approach can however narrow promoter regions amongst huge DNA sequences. A subsequent experiment can be established and tested thus saving time and money [6].

II. METARIAL 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.

A. Data set

The research sought to incorporate the E. Cole promoter gene arrays of DNA in the testing the proficiency of ANN. Such data were collected from the UCI Repository [30]. 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 [5].

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 [30].

B. ke-REM (ke-Rule Extraction Method)

This section introduces the novel development referred to as ke-REM (ke-Rule Extraction Method) and addresses its ability in utilizing DNA promoter region predictions. As provided for above, an e.coli dataset consists of a total of 106 DNA sequences, each containing a length of 57 nucleotides. The computer science perspective expresses the dataset for e-coli as consisting of 106 instances containing 57 attributes bearing four values. The attributes for these instances can be expressed as nucleotides locations for the 57-element sequence. Each attribute accommodates 4 values, namely T-Thymine, A-Adenine, C-Cytosine and G-Guanine.

ke-REM constructs a rule-base by applying the data set attribute-value pairs. In an effort to generate a robust rule-base, attribute-value pairs with significant importance are used. The significant question at this point queries, "How are pairs with significant informational value determined?" the new ke-REM upgrade uses a "gain function" in computing the informational value for the set's pair. ke-REM considers the higher gain value as a higher informational value indicator. Therefore, the attribute-value with a higher value has a greater priority in the processing of rule-base for the prediction system. keREM (ke-Rule Extraction Method) was upgraded to have the ability to obtain IF-THEN rules from a given set of examples. It proactively discards encountered pitfalls commonly present in inductive learning algorithms. keREM applies the gain function value, to determine which attributes are of significant importance and are thus given a higher priority and as such, serve to further provide rules that are more commonly acceptable.

The following is a summary of the algorithm:

Step 1: In a particular training set, a person computes class distribution and probability distribution rate of every attribute-value.

Step 2: For every attribute in the data set, you compute the power of classification.

Step 3: for every attribute-value pairs, its Class-based Gain is calculated with the use of computed probability distributions, power of classification and class distribution rate.

Step 4: One rule of selection is that you can select any value whose probability distributions one for n=1. The next step is to convert the attribute-values into rules and then you mark the classified examples.

Step 5: Move to step 8.

Step 6: Starting from the first example that is unclassified, you form combinations with the n values by using the attribute-values that has a bigger gain.

Step 7: You apply each combination in all examples. Using the values that are made up of n combinations, those that match only with on class are converted into a rule. You mark the classified examples.

Step 8: In the training set, when all examples are classified, you move to step 11.

Step 9: perform the expression n=n+1

Step 10: go to step 6 if n<N

Step 11: Select the most general rule if there is over one rule that represents the same examples.

Step 12: End.

III. RESULTS AND DISCUSSION

The aim of this section is to experimentally analyze our approach for promoter sequences recognition with the use of ke-REM and compare it with the existing approaches. Ke-REM has an important feature which allows it to compute class-based gain for every attribute-value in a particular training set. First, in this context, the class distribution and probability distribution rate of every nucleotide that forms DNA sequence is computed on the basis of promoter and non-promoter classes. For every attribute in the data set, you contribute power of classification. However, there is no class information in the results for the attribute-value pairs. Therefore, using class distribution

rate, probability distribution and the power of classification, you compute class-based gain for every value in the DNA sequence data set. Using this method, rules that the algorithm produces were formed by attribute-value which has a maximum information value.

There are two phases which fulfill the experiments of promoter prediction and they reflect the principles behind a supervised learning algorithm which include testing and training. Classification model is built during training, and during testing, the model is applied for classification of unseen examples that are DNA sequence and they consist of promoter and non-promoter region. You evaluate the performance of ke-REM with the use of a standard 5-fold cross-validation. Dataset is partitioned randomly in the 5-fold cross-validation into five subsets. Every subset contains an equal ratio of both the promoter and non-promoter region. For five times, ke-REM is trained using 4 subsets for each time for training and remaining the 5thh subset for testing. 5 models are generated in this way during cross-validation. You obtain the final prediction performance by averaging the results that are achieved from every model.

We determined prediction performance of the algorithm by measuring the threshold-dependent parameters sensitivity (SE), specificity (SP), accuracy (ACC) and Matthew's Correlation coefficient (MCC). The following equations were used to calculate ACC, SE, SP and MCC.

$$SE=TP/(TP+FN) \quad (1)$$

$$SP=TN/(TN+FP) \quad (2)$$

$$ACC= (TP+TN) / (TP+TN+FP+FN) \quad (3)$$

$$MCC= ((TP*TN)-(FN*FP))/SQRT((TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)) \quad (4)$$

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.

Table 1: Performance of ke-REM in term of SE, SP, ACC and MCC

Threshold-Dependent Parameters	1. Model	2. Model	3. Model	4. Model	5. Model	Average
ACC	0.75	0.90	0.80	0.90	0.69	0.8085
SE	0.80	0.90	0.80	1.00	0.54	0.8077
SP	0.70	0.90	0.80	0.80	0.85	0.8092
MCC	0.50	0.80	0.60	0.82	0.40	0.6246

In the literature, the way learning algorithms perform can be evaluated by applying cross-validation with the use of a "leave-one-out" methodology. Leave-one-out cross validation (LOOCV) is considered as a special type of k-fold cross-validation and k represents the number of instance in the data. This implies that in all iteration, close to every data apart from the single observation are utilized for training and then you test the model on the single observation. An estimate that is accurate that is obtained with the use LOOCV is considered as almost unbiased [31]. This is commonly used when the data that is available is rare, particularly in bioinformatics when there is only dozens of data sample available.

Table 2: The errors of some machine learning algorithms on promoter data		
System	Errors	**Classifier**
REX-1	0/106	**Inductive L.A**
ke-REM	**3/106**	**Class-based gain**
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-	13/106	**A nearest neighbours algorithm**
ID3	**19/106**	**Quinlan's decision builder**

Unlike the classifiers that have been applied here for promoter prediction (Table 2), ke-REM that has been introduced in this document tends to outperform the present classifier for promoter prediction. When the classification error is considered, it becomes better compared to ID3, KB, NN, O'Neil and BP.

IV. CONCLUSIONS

Within the bioinformatics field, the promoter prediction is considered as a crucial problem from a computational perspective. Newly developed ke-REM (ke-Rule Extraction Method) in this document has been proposed as effective in tackling the problem. For rule-base with efficient rule to be generated, you employ attribute-value pairs with higher importance. To calculate information value of the pair in the set, ke-REM uses its own "gain function" to calculate information value. Because value of the higher gain tends to indicate higher information value, a greater priority is given to attribute-value with higher value when it comes to production of rule-base of the predicting system. According the results given above, it is possible to conclude that when ke-REM for promoter prediction is employed it leads to results that are promising. Moreover, additional improvement increases the accuracy of the results that have been obtained.

REFERENCES

[1]. D. Mount, Bioinformatics: Sequence and Genome Analysis, New York: Cold Spring Lab Press, 2013.
[2]. K. Davies, The Sequence, US: Weidenfeld & Nicholson, 2001.
[3]. P. Jayaram and K. Bhushan, Bioinformatics For Better Tomorrow, New Delhi: Indian Institute of Technology, 2000.
[4]. B. Óscar and B. Santiago, "Cnn-promoter, new consensus promoter prediction program," Revista EIA, no. 15, pp. 153-164, 2011.
[5]. C. Gabriela and M.-I. Bocicor, "Promoter Sequences Prediction Using Relational Association Rule Mining," Evolutionary Bioinformatics, vol. 8, pp. 181-196, 2012.
[6]. J.-W. Huang, Promoter Prediction in DNA Sequences, Kaohsiung,: National Sun Yat-sen University, 2003.
[7]. R. Guigo, S. Knudsen, N. Drake and T. Smith, "Prediction of gene structure," Journal of Molecular Biology, no. 226, pp. 141-157, 1995.
[8]. L. Milanesi, N. Kolchanov and I. Rogozin, "GenViewer: A computing tool for protein coding regions prediction in nucleotide sequences," in the 2nd International Congress on Bioinformatics, Supercomputing and Complex Genome Analysis,, 573-587, 1993.
[9]. S. Dong and G. Stormo, "Gene structure prediction by linguistic methods," Genomics, no. 23, pp. 540-551, 1994.
[10]. E. Stormo and E. Snyder, "Identification of coding regions in genomic DNA sequences: An application of dynamic programming and neural networks," Nucleic Acids Research, no. 21, pp. 607-613, 1993.
[11]. V. Solovyev, A. Salamov and C. Lawrence, "Prediction of internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames," Nucleic Acids Research, no. 22, pp. 5156-5163, 1994.

[12]. G. Hayden and M. Hutchinson, "The prediction of exons through an analysis of spliceable open reading frames," Nucleic Acids Research, no. 20, pp. 3453-3462, 1992.
[13]. A. Skolnick and M. Thomas, "A probabilistic model for detecting coding regions in DNA sequences," IMA J. Math. Appl. Med. Biol., no. 11, pp. 149-160, 1992.
[14]. Y. Xu, R. Mural and E. Uberbacher, "Constructing gene models from accurately predicted exons: An application of dynamic programming," Comput. Appl. Biosci, no. 10, pp. 613-623, 1994.
[15]. J. Henderson, S. Salzberg and K. Fasman, "Finding genes in DNA with a hidden Markov model," Journal of Computational Biology, vol. 2, no. 4, pp. 127-141, 1997.
[16]. C. Karlin and C. Burge, "Prediction of complete gene structures in human genomic DNA," J. Mol. Biol, no. 268:, pp. 78-94, 1997.
[17]. M. Hrishikesh, S. Nitya and M. Krishna, "An ANN-GA model based promoter prediction in Arabidopsis thaliana using tilling microarray data," Bioinformation, vol. 6, no. 6, p. 240–243, 2011.
[18]. J. Gordon, M. Towsey and J. Hogan, "Improved prediction of bacterial transcription start sites," Bioinformatics, vol. 22, no. 2, pp. 142-148, 2006.
[19]. R. Liu and D. States, "Consensus promoter identification in the human genome utilizing expressed gene markers and gene modelling," Genome Research, no. 12, pp. 462-469, 2002.
[20]. C. Premalatha and C. a. K. K. Aravindan, "On improving the performance of promoter prediction classifier for eukaryotes using fuzzy based distribution balanced stratified method.," in Proceedings of the International Conference on Advance in Computing, Control, and Telecommunication Technologies IEEE,, ACT, 2009.
[21]. T. Abeel, Y. Saeys, E. Bonnet and P. Rouzé, "Generic eukaryotic core promoter prediction using structural features of DNA," Genome Research, vol. 18, no. 2, pp. 310-323, 2008.
[22]. Y.-J. Zhang, "A novel promoter prediction method inspiring by biological immune principles.," Global Congress on Intelligent Systems, no. 569-573, pp. 569-573, 2009.
[23]. S. S. Roded, S. Karni and Y. Felder, Promoter Sequence Analysis, Lecture 11,, 2007.
[24]. A. Dombroski, W. Walter and M. Record, "Polypeptides containing highly conserved regions of transcription initiation factors 70 exhibit specificity of binding to promoter DNA," Cell, p. 501–512, 1992 .
[25]. H. Bujard, M. Brenner and W. Kammerer, Structure-function relationship of Escherichia coli promoters, New York: Elsevier, 1997.
[26]. D. Grana, T. Gardella and M. M. Susskind, "The effects of mutations in the ant promoter of phage P22," Genetics, p. 319–327, 1998.
[27]. M. Craig, M. L. Suh and T. Record, "HOz and DNase I probing of Es70RNA polymerase-l PR promoter open complexes: Mg21binding and its structural consequences at the transcription start site," Bio-chemistry, p. 15624–15632, 1995.
[28]. D. Frisby and P. Zuber, "Analysis of the upstream activating sequence and site of carbon and nitrogen source repression in the promoter of anearly-induced sporulation gene of Bacillus subtilis.," Bacteriol, p. 7557–7564, 1991.
[29]. R. Wilma, A. Sarah and J. Salomon, "Escherichia Colipromoters With UP Elements Of Different Strengths: Modular Structure Of Bacterial Promoters," Journal Of Bacteriology, p. 5375–5383, 1998.
[30]. A. Frank and A. Asuncion, "UCI machine learning repository," 2010. [Online]. Available: http://archive.ics.uci.edu/ml/.
[31]. B. Efron, "Estimating the error rate of a prediction rule: improvement on cross-validation," J. Am. Stat. Assoc., no. 78, p. 316–331, 1983.
[32]. Essentials of Cell Biology, Nature Education, 2010.
[33]. Q. Luo and W. a. L. P. Yang, "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, 2006.
[34]. S. Clancy, "Nature Education," DNA transcription, vol. 1, no. 1, p. 41, 2008.

[35]. M. Guigo and R. Burset, "Evaluation of gene structure prediction programs," Genomics, vol. 3, no. 34, pp. 353-367, 1996.

[36]. M. Wang, M. Yin and T. Jason, "GeneScout: a data mining system for predicting vertebrate genes in genomic DNA sequences," Information Sciences, vol. 163, no. Special issue, pp. 201-218, 2013.

[37]. T. S. Y. B. E. R. P. a. V. d. P. Y. Abeel, "Generic eukaryotic core promoter prediction using structural features of DNA," Genome Research, vol. 18, no. 2, pp. 310-323, 2008.