Experimental Investigation of Enhancer-Promoter Interactions out of Genomic Big Data based on Machine Learning

DESISLAVA IVANOVA[1], PLAMENKA BOROVSKA[1], VESKA GANCHEVA[2]
Department of Informatics[1],
Department of Programming and Computer Technologies[2],
Technical University of Sofia
1000 Sofia, Bul. “Kliment Ohridski” 8
BULGARIA
d_ivanova@tu-sofia.bg, pborovska@tu-sofia.bg, vgan@tu-sofia.bg

Abstract: This paper reviews the existing methods for detection of enhancer-promoter interactions. It presents the experimental investigation for detection of enhancer-promoter interactions from genomic big data based on machine learning. The authors are spent time to explain the importance of promoters and enhancers and their impacts on gene expression. The main purpose of the paper is to propose a pipeline for detection of enhancer-promoter interactions. It is realized by using Decision Tree and Support Vector Machine classifiers. The experimental framework is based on Apache Spark environment that allows streaming and real time analysis of big data. Machine learning library of Apache Spark (MLlib) is implemented in python programming language for processing genomic big data. To perform the results, the enhancer-promoter interactions GM12878 and K562 datasets are used. Finally, the experimental results are presented and discussed.

Key-Words: Genomic Big Data, Enhancer-Promoter Interactions, Machine learning, Experimental Investigation, Spark Apache, MLlib

1 Introduction

Nowadays there is sharp need of efficient algorithms and computational power for analyzing the growing amounts of data across most of the industries. Genomics is one of the main domains in the field of big data. The challenge is not only to extract meaningful information from the big data, but to gain knowledge, to discover previously unknown insight, to look for patterns and to make sense of the data. Many different approaches, including statistical and graph theoretical methods, data mining and machine learning methods, have been applied in the past, however with partly unsatisfactory success especially in terms of performance [1, 2].

Many advances in powerful computational tools in recent years have been developed by separate communities with different philosophies: machine learning researchers tend to believe in the power of their methods to identify the relevant patterns - mostly automatic, without human intervention, however, the dangers of modeling artifacts grow when end user comprehension and control are diminished. Consequently, it is a grand challenge to work towards enabling effective human control over powerful machine intelligence by the integration and combination of machine learning methods and advanced analytics methods to support insight and decision making. Effectively tackling these challenges is possible by bringing together the better of two worlds: a synergistic combination of traditional theories, methods and approaches and Knowledge Discovery from Data (KDD). Such approaches need an interdisciplinary methodology [3, 4]. The great example for that is the processing and analysis of genomic big data that is part of the experimental investigation in this paper.

The living organism is a complex system depends on synchronous actions of different groups of genes. The determination of “active and deactivate” gene is a challenge in genomics era. The first and key step in gene expression is promoter recognition by RNA polymerase enzyme. Promoter is a key region that is involved in differential transcription regulation of genes and protein coding. It is near the starting point of the protein biosynthesis and thus plays a vital role. Coordination of gene expression is achieved to a large extent by different transcriptional control mechanisms characteristic for each gene and controlling timing, rate, and level of its transcription. Promoter regions may contain many short motifs that serve as recognition sites for proteins providing initiation of transcription as well as specific regulation of gene expression [5, 6].

The boundaries of promoters are not very clear, but most important transcriptional signals known today are generally located within the segment relative to the transcription start site (TSS), which often already includes proximal enhancers.

Gene promoters have been responsible for the integration of different mutations favorable for the environmental conditions. A major question in evolutionary biology is how important tinkering with promoter sequences is to evolutionary change, for example, the changes that have occurred in the human lineage after separating from other primates.

Some evolutionary biologists, for example Allan Wilson, have proposed that evolution in promoter or regulatory regions may be more important than changes in coding sequences over such time frames.

A key reason for the importance of promoters is the potential to incorporate endocrine and environmental signals into changes in gene expression. A great variety of changes in the extracellular or intracellular environment may have impacts on gene expression, depending on the exact configuration of a given promoter. The combination and arrangement of specific DNA sequences that constitute the promoter defines the exact groups of proteins that can be bound to the promoter, at a given time point [7, 8].

Promoters represent critical elements that can work in concert with other regulatory regions (enhancers, silencers, boundary elements/insulators) to direct the level of transcription of a given gene. A promoter is induced in response to changes in abundance or conformation of regulatory proteins in a cell, which enable activating transcription factors to recruit RNA polymerase.

A comprehensive study of the promoter content information is presented by Schultzaberger [9]. Very recent work on predicting enhancer-promoter interactions based on functional genomic features is existed [10].

This paper is focuses on the experimental investigation results for detection of enhancer-promoter interactions out of genomic Big Data based on machine learning.

2 Enhancer-Promoter Interactions Detection from Genomic Big Data based on Machine Learning

2.1 State of the art

With respect to promoters' detection, the scientists need to know the TSS, because the promoters will be in that region. They are usually found by applying a conserved motif to these sites and are logged in the genetic sequence databases.

The scientists align the promoter sequence with the genome sequence, thus finding the TSS. Normally for that purpose, they used a pairwise sequence alignment algorithm that is designed to decrease the time needed to align millions of mouse genomic reads and expressed sequence tags against the human genome sequence and can be used to find promoters, which lie in tightly constrained positions relative to the TSS [11]. The other option is to use machine learning, in order to train a predictor, based on respective feature representations [12, 13].

2.2 The proposed approach

In this paper, the experimental investigation for detection of enhancer-promoter interactions from Genomic Big Data is proposed based on machine learning algorithms: Decision Tree (DT) and Support Vector Machine (SVM).

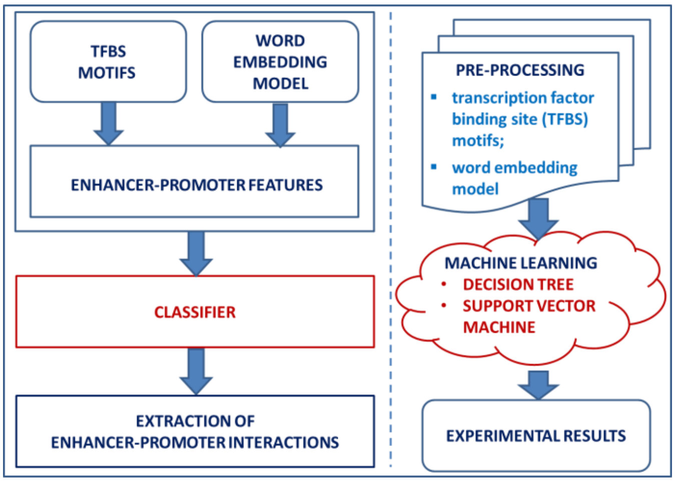

Fig.1 Pipeline of experimental investigation for detection of enhancer-promoter interactions

The pipeline presents the detection of enhancer promoter interactions out of genomic big data based on machine learning. It is consisted of three stages: *1) Pre-processing:* this stage includes data cleaning and feature detection; *2) Classifier* to detect the enhancer promoter interactions; and *3) Presenting the experimental results.*

The first stage incorporates two approaches for feature extraction directly from the DNA sequences of enhancer and promoter elements, Fig 1. The first approach search for patterns of known transcription factor binding site (TFBS) motifs in the sequences involved in enhancer-promoter interactions datasets. Second, the word embedding model is used. Word embedding model is the collective name for a set of language modeling and feature learning techniques where words or phrases from the vocabulary are mapped to vectors of real numbers. This model is used to directly embed the sequences of enhancer and promoter regions into a new feature space. These two approaches are previously utilized by PEP framework [7].

The next stage is the data classification based on the extracted features. In the proposed approach, Decision Tree and Support Vector Machine are used to implement the classifiers.

Decision Tree is very important data machine learning technique for classification of data. Decision tree induction is the learning of decision trees from class labeled training tuples [14]. A decision tree is a flow chart like tree structure, where each internal node denote a test on an attribute, each branch represent an outcome of the test, and each leaf node hold a class label. The topmost node in a tree is the root node. Decision tree can handle high dimensional data. Decision tree algorithm is simple and fast. These tree classifiers have good accuracy. Decision tree induction algorithms have been used for classification in many application areas such as medicine, genomics and molecular biology.

Support vector machine are basically linear classifiers. SVM is widely accepted classifier, considered very effective for pattern recognition, machine learning and bioinformatics [15]. In SVM, a separator hyper plane between two classes is chosen to minimize the functional gap between two classes, the training data on the marginal sides of this optimal hyper plane called support vector. The learning process is the determination of those support vectors. For non-linearly separable data, SVM maps the input vector from input space to some normally higher dimension feature space given by kernel function. The kernel function is an important step is successful design of a SVM in specific classification task.

The performance parameters are given in the final stage of the proposed pipeline.

3 Experiments and Result Analysis

3.1 Experimental Framework

The experimental framework is based on Apache Spark environment that allows streaming and real time analysis of big data. Machine learning library of Apache Spark (MLlib) is implemented in python programming language for processing genomic data [16]. For the experimental investigation, the enhancer-promoter interactions data from TargetFinder project are used [17]. The dataset includes six cell lines (GM12878, HeLa-S3, HUVEC, IMR90, K562, and NHEK). The data for each cell line consist of enhancer-promoter pairs which are annotated as positive (interacting) or negative (non-interacting) using high-resolution genome-wide measurements of chromatin contacts in each cell line.

Table 1: Positive and negative samples in cell lines

Cell Line	Positive Pairs	Negative Pairs
GM12878	2,113	42,200
K562	1,977	39,500

Cell-line specific active enhancers and promoters are identified using annotations from the ROADMAP EPIGENOMICS and ENCODE Projects [18, 19]. In this paper for the experiments, GM12878 and K562 cell lines are used, Table 1.

3.2 Experimental Results and Analysis

The software that implements the proposed approach is written in python programming language and used Decision Tree and Support Vector Machine classifiers.

The software starts with data preparation which targets the GM12878 cell line followed by K562.

[Chromosome, start, end, name]

In dataset, we have a list of chromosomes whose sequences we need to obtain. We iterate the list where total is the length of the list and load each file from the learning data. Then it is opened the file with pairs which contain the start and end of all promoters and enhancers. Based on those start and end points, it is extracted the corresponding sequences from the chromosome and write them to the supervised file.

Second, it is red the starting and ending point of the known enhancers and promoters for the cell line. We use the start and end line, in order to filter the sequence of the chromosome and write only the theoretical location of our promoters and enhancers in the respected files. Finally, it is taken the previously extracted promoter and enhancer sequences and split them in even intervals. Information of enhancer-promoter pairs file is consisted of:

[bin,enhancer_chrom,enhancer_distance_to_promoter,enhancer_end,enhancer_name,enhancer_start,label,promoter_chrom,promoter_end,promoter_name,promoter_start,window_end,window_start,window_chrom,window_name,interactions_in_window,active_promoters_in_window]

The software used the unlabeled train enhancer-promoter files for training the word embedding model. Moreover the supervised enhancer-promoter data is used for training the decision tree and support vector machine classifiers. The generated paired enhancer-promoter samples indicate whether a sample is positive (interacting enhancer-promoter pair) or negative (non-interaction enhancer-promoter pair).

Finally, the code performed the cross validation and the result files with performance parameters are written in the main folder. In order to achieve the final results, it is needed a sequence for each chromosome that will be checked for matching the enhancers and promoters.

Table 2: Performance Results

Cell type		GM12878	K562
Accuracy	Decision Tree - Accuracy	**93%**	**91%**
	Support Vector Machine - Accuracy	**95%**	**92%**

The performance results of Decision Tree and Support Vector Machine classifiers are calculated using the formula:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \qquad (1)$$

where: TP (true positive), FN (false negative), FP (false positive), TN (true negative).

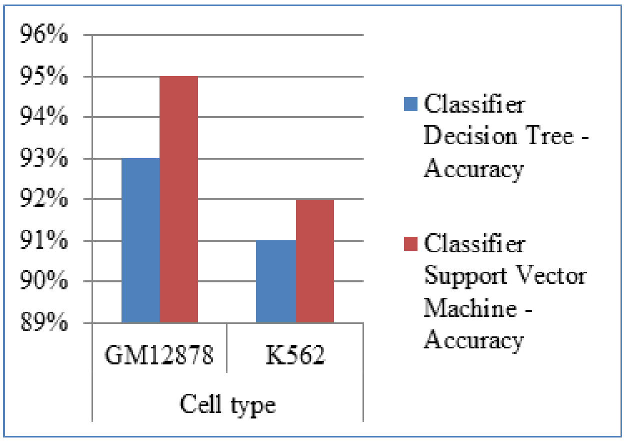

Fig.2 Performance evaluation of proposed approach for detection of enhancer-promoter interactions from GM12878 and K562 cell types

The enhancer - promoter interactions is one of the most challenging phenomena in genomics field. In this paper, we have investigated experimentally the detection of enhancer - promoter interactions from genomics big data based on machine learning, implementing the Decision Tree and Support Vector Machine classifiers. We demonstrated that achieved performance parameters of 91 – 95 % accuracy are competitive.

The future work is to compare the performance results with the state of the art method *TargetFinder* [17] that uses numerous functional genomic signals instead of sequence features and *SPEID* method [20] using the deep learning models to predict enhancer-promoter interactions based on sequence-based features.

4 CONCLUSION

This paper reviewed the existing methods for detection of enhancer-promoter interactions. It is presented the experimental investigation for detection of enhancer-promoter interactions out of genomic big data based on machine learning. The paper is proposed a pipeline for detection of enhancer-promoter interactions using Decision Tree and Support Vector Machine classifiers. The experimental framework is based on Apache Spark environment that allows streaming and real time analysis of big data. Machine learning library of Apache Spark (MLlib) is implemented in python programming language for processing genomic big data. The experimental results for detection of enhancer-promoter interactions have been performed with GM12878 and K562. Finally, the experimental results are presented and discussed.

5 ACKNOWLEDGMENTS

The paper is supported by the Project “Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research” funded by the Bulgarian National Science Foundation, Bulgarian Ministry of Education and Science, Competition for financial support of Fundamental Research (2016) under the thematic priority: Technical Science, contract № ДН07/24 - 15.12.2016.

References:

[1] Aaron M., Hanna M., Banks E., Sivachenko A., Cibulskis K., Kernytsky A., Garimella K., The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, 2010, 20(9):1297–1303. doi: 10.1101/gr.107524.110.

[2] Akalin A., Kormaksson M., Li S., Garrett-Bakelman FE., Figueroa ME., Melnick .A, Mason CE, Methylkit: a comprehensive R package for the analysis of genome-wide DNA methylation profile, *Genome Biol.*, *2012*, 13:R87. 10.1186/gb-2012-13-10-r87 doi: 10.1186/gb-2012-13-10-r87.

[3] K. Gasztonyi, Data Protection Officials Adopt Internet of Things Declaration and Big Data Resolution, *International Conference of Data Protection and Privacy Commissioners in Mauritius*, 2014.

[4] General Electric Company, Big Data, Analytics & Artificial Intelligence: The Future of Health Care is Here, *white paper*, 2016.

[5] B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, et al., The NIH roadmap epigenomics mapping consortium, *Nature biotechnology*, 28(10):1045–1048, 2010.

[6] M. J. Fullwood and Y. Ruan, Chip-based methods for the identification of long-range chromatin interactions, *Journal of Cellular Biochemistry*, 107(1):30–39, 2009, May 1;107(1):30-9. doi: 10.1002/jcb.22116.

[7] S. Whalen, Rebecca M. Truty, Katherine S. Pollard, Enhancer-promoter interactions are encoded by complex genomic signatures on looping chromatin, *Nat Genet.* 2016 May; 48(5): 488–496, published online 2016 Apr 4. doi: 10.1038/ng.3539.

[8] Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res.* 2012;22:1748–1759, doi: 10.11 01/gr.136127.111.

[9] Shultzaberger, R.K., Chen, Z., Lewis K.A., Schneider, T.D., Anatomy of Escherichia coli σ70 promoters, 2007, *Nucleic Acids Research*, Vol.35, No.3, pp. 771–788.

[10] Yang Y., Zhang R., Singh S., Ma J., Exploiting sequence-based features for predicting enhancer-promoter interactions, 2017, *Bioinformatics*, Jul 15, 33(14):i252-i260, doi: 10.1093/bioinformatics/btx257.

[11] Yamamoto YY., Yoshitsugu T, Sakurai T, Seki M, Shinozaki K, Obokata J., Plant J., Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis, 2009 Oct, 60(2):350-62. doi: 10.1111/j.1365-313 X.2009.03958.x. Epub 2009 Jun 29.

[12] Feng Liu, Hao Li, Chao Ren, Xiaochen Bo, Wenjie Shu, PEDLA: predicting enhancers with a deep learning-based algorithmic framework, Scientific Reports volume 6, Article number: 28517 (2016), doi:10.1038/ srep28517.

[13] Jianlin He, Ming-an Sun, Zhong Wang, Qianfei Wang, Qing Li, Hehuang Xie, Characterization and machine learning prediction of allele-specific DNA methylation, Genomics, Volume 106, Issue 6, December 2015, pp. 331-339, https://doi.org/10.1016/j.ygeno.2015.09.007.

[14] Last, M., Maimon, O. and Minkov, E., Improving Stability of Decision Trees, *International Journal of Pattern Recognition and Artificial Intelligence*, 16: 2,145-159, 2002.

[15] E. Osuna, R. Freund, F. Girosi, An improved training algorithm for support vector machines, In J. Principe, L. Gile, N. Morgan, and E. Wilson, editors, *Neural Networks for Signal Processing VII Proceedings of the 1997 IEEE Workshop*, pages 276 – 285, New York, IEEE.

[16] *Apache Spark:* fast and general engine for big data processing, with built-in modules for streaming: https://spark.apache.org

[17] TargetFinder project: https://github.com/ carringtonlab/TargetFinder.

[18] Roadmap Epigenomics Project launched by NIH Roadmap Epigenomics Mapping Consortium, website: http://www.roadmapepigenomics.org

[19] ENCODE (Encyclopedia of DNA Elements) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI), website: https://www.encodeproject.org/

[20] Shashank Singh, Yang Yang, Barnabas Poczos, Jian Ma, Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep *Neural Networks, Nov. 2, 2016,* doi: http://dx.doi.org/10.1101/085241, It is made available under a CC-BY-NC-ND International license 4.0.