IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Vol.-7, Special Issue-8, Apr 2019 E-ISSN: 2347-2693

Extensive Review on Computational Predictions of Genomic Regulatory Sequences

Sasikala S[1*], Ratha Jeyalakshmi T[2]

[1]Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli-12, Tamilnadu, India.
[2] Department of Computer Applications, Sri Sarada College for Women, Tirunelveli

Corresponding Author: ssk_sivapri27@yahoo.in

DOI: https://doi.org/10.26438/ijcse/v7si8.9194 | Available online at: www.ijcseonline.org

Abstract— This paper focuses an extensive study of the existing computational work related to prediction of gene regulatory sequences and the relevant factors based on different properties. A hybrid approach is studied which combines position correlation score function and increment of diversity to elucidate signal features and composition features of sequences to improve the accuracy of promoter classifiers. It is found that Markov Model of order K is used to extract features from k-mer frequency of the sequence. Also a Support Vector Machine (SVM) is applied with the transcription signals such as Gc box, TATA box, CAAT box, NIT box and CpG islands and modified Mahalanob discriminant to predict Eukaryotic and Prokaryotic promoters. It is studied that a new approach is implemented using Artificial Neural Network (ANN) with the properties namely curvature, stacking energy and Stress Induced duplex Destabilization (SIDD). To analyze sequence characteristics of prokaryotic and eukaryotic promoters, Convolutional Neural Networks (CNN) is found to be contributing a significant role. This paper analyses the use of a tool bTSSfinder for promoter predicting models. It is identified that an algorithm exist for promoter prediction based on evolutionarily conserved sequences by concentrating AT-rich elements and G-quadruplex sequences using various statistical measures such as recall, precision, specificity accuracy and F1- scores . Algorithms using machine learning based approach are studied to discover promoters in nucleotide sequence using entropy based feature. Some of the remarkable DNA structural features such as DNA bending stiffness, duplex free energy, duplex disrupt energy, stacking energy, DNA denaturation, protein deformation, nucleosome position, propeller twist are studied. Multifarious promoter prediction models which are found to be predicting promoters associated with PoI II sequence, sigma factors such as $\sigma^{70,}$ $\sigma^{66,}$ σ^{54} and transcription factor binding sites. This paper studied that bidirectional genes are co expressed and tends to be involved in the same biological functions with stronger expression correlation. Also studied the intergenic regions enriched of regulatory elements are essential for the transcription initiation. Though various models are found to be effective, still they need to uncover various characteristics. Because of the dynamicity of gene regulatory process, promoter prediction models still require improvement .Indeed this field has a wider exposure of detailed research work.

Keywords- Eukaryotic and Prokaryotic Promoters, Sigma Factor, TSS

I. INTRODUCTION

The basic unit of organisms is called cells. Cells contain nucleus which is a pack of Deoxyribo Nucleic Acid (DNA), RNA, histones and other compounds called as chromatin. Gene is defined as a region of DNA. Every gene is encoded with specific information. There are three regions in a gene which are promoter, coding region and a terminator region. Promoter is an important sequence of gene regulation process network, which is responsible for gene transcription initiation. Normally promoters reside near the transcription start site. But the region of promoter is not a fixed one .It may exist in any location of the DNA sequence. Identification of promoters is an important task in biology, given that they are central in understanding the process by which genes are regulated. Transcription factors which are directly bound with promoters control the transcription process.

Experimental methods related with the foresaid processes cause various inconveniences based on the aspects such as time, accuracy, handling volume of data whereas computational models can handle these inconveniences effectively that save effort and time sufficiently. This paper focuses a detailed study on different gene regulatory sequences and their impacts on various activities in gene regulation.

II. RELATED WORK

The performance of promoter prediction methodologies are evaluated based on various features such as sensitivity, specificity and correlation coefficient. High quality data sets are used for training set and test set. The factor known as Transcription Start Site (TSS) plays a vital role in gene regulation process.

2.1 Eukaryotic and Prokaryotic Promoters

The Prokaryotic and Eukaryotic promoters use different DNA sequences to regulate gene expression. In prokaryotes, the transcription of most of the genes is regulated by sigma 70 (σ^{70}) promoters. In Eukaryotes, all protein-coding genes and certain small nuclear RNAs are regulated by PoI II promoters. Hao Lin et al. [9] have proposed a hybrid approach which combines position correlation score function and increment of diversity to elucidate signal features and composition features of sequences with modified Mahalanobis discriminant to predict Eukaryotic and Prokaryotic promoters.

C.Premalatha et al. [8] have represented a way to improve the accuracy of promoter classifiers. Markov Model of order K is used to extract features from k-mer frequency of the sequence. Also a Support Vector Machine (SVM) is applied with the transcription signals such as GC box, TATA box, CAAT box, NIT box and CpG islands. The classifier models, 4-mer classifier and 5-mer classifier are proposed. The authors have used only limited set of features. Wenying He et al. [18] have predicted a model 70ProPred to improve the accuracy of sigma 70 promoters in Prokaryote. It achieved sufficient accuracy level.

Charles Bland et al. [6] have suggested a new approach using Artificial Neural Network (ANN) includes properties namely curvature, stacking energy and Stress Induced duplex Destabilization (SIDD). Threshold energy level of SIDD compared against the suggested ANN approach. Additionally they showed that the DNA structural properties play an important role in promoter prediction. The method suggested in this article may be validated for overfitting.

To analyze sequence characteristics of prokaryotic and eukaryotic promoters Convolutional Neural Networks contributed a considerable role [16]. The authors built prediction models using CNN for five different organisms. They are human, mouse, plant (Arabidopsis) and E. coli, B. Subtilis. bacteria. Though this model reached a reasonable accuracy level still requires improvement.

For five classes of sigma factors in Cyano bacteria (σ^{A}, σ^{C}, σ^{F}, σ^{G} and σ^{H}) and five classes of sigma factors in E.*Coli*. Bacteria (σ^{70}, σ^{38}, σ^{32}, σ^{28}, σ^{24}) Ilham Ayub Shahmuradov et al. [17] have introduced a tool bTSSfinder for promoter predicting models .And it is found that this tool gained considerable accuracy levels.

Marco Di Salvo et al. [19] have presented an algorithm for promoter prediction based on evolutionarily conserved sequences by concentrating AT-rich elements and G-quadruplex sequences. Various statistical measures, recall, precision, specificity accuracy and F1- scores followed to get better results..

The properties of DNA structures have crucial role in various genomic functionalities. The second order structural information encoded in promoter regions could be identified using RNA polymerase. Yangalan Gan et al. [10] have proposed a way to select reliable features which are closely connected with the promoter sequences. The remarkable structural features analyzed are DNA bending stiffness, duplex free energy, duplex disrupt energy, stacking energy, DNA denaturation, protein deformation , nucleosome position, propeller twist. Also the authors have identified the structural profile with two peaks. The two peak values used to determine the transcription process.

Zaw ZawHtikea et al. [11] have presented a machine learning approach named as weightily averaged one dependence estimators. To lower the error rate and to increase the efficiency optimistic nucleotide sequences are selected using entropy based measure.

Ravi Gupta et al. [7] presented a computational model to identify promoters associated with PoI II enriched region. Also it is discriminated with other sequence. This paper identified the models based on the approaches, Bagging and Random forest which gave remarkable result.

Guofeng Meng et al. [5] have suggested computational evaluation on recovering TF binding sites from differentially expressed genes using the approaches, over representation motif analysis using oPOSSUM and de novo prediction with weeder tool.

Bidirectional gene pair is a two adjacent genes which are in opposite strands of DNA with transcription starting site (TSS) which are not more than 1000 base pairs apart and the intergenic region between two TSSs is commonly termed as bidirectional promoter. Wang et al. [4] have analyzed whole genome of Arabidopsis thaliana for the existence of bidirectional promoters. The analysis showed that bidirectional genes are co expressed and tend to be involved in the same biological functions with stronger expression correlation. Also identified the intergenic regions enriched of regulatory elements are essential for the transcription initiation.

2.2 Sigma Factors

Sigma 70(σ^{70}) promoters regulate the transcription of most genes. Wenying He et al. [18] have presented a model 70ProPred to improve the accuracy of sigma 70 promoters in

Prokaryote. The model used two significant features such as Position-Specific Trinucleotide Propensity based on single stranded characteristic(PSTNPss) and electro-ion potential values for trinucleotides(PSeEIIP).

Qian-Zhong Li et al. [2] have suggested position correlation scoring matrix (PCSM) algorithm for predicting sigma 70 promoters and its performance is evaluated using 10- cross validation test. Conservation analysis is made based on the bounding coding regions such as tandem (TAN), divergent (DIV) or convergent (CON) in intergenic regions. To evaluate the prediction rate, the measures sensitivity, specificity, correlation coefficient and accuracy are taken into account.

An organism- specific model for Chamydia trachoma (CT) is introduced by Ronna R Mallios et al. [3] . This article uncovered different structural parameters such as DNA stability, curvature twist and stress-induced DNA duplex destabilization and σ^{66} promoter using hidden Markov model. Binary Logistic Regression is used to select optimal training data. Using model based genome-wide predictions a new model is revealed by including RNAP sigma factor and DNA binding. It has taken the genome which was small. It may further be improved for large genome and may be analyzed the similarities of different prediction models for other organism.

RNA polymerase helps to identify the promoters for σ^{54} resides in the conserved regions -24 and -12 nucleotide upstream from the transcriptional start site (TSS). Humberto Barrios et al. [1] have presented an updated compilation of σ^{54} dependent promoters and the derivation of an extended consensus sequence which is extended from the region - 8 to -31 relative to TSS.

Lin H et al. [13] Samples of DNA sequences are formulated using a novel feature vector known as pseudo K tuple nucleotide composition. The performance is analyzed using jackknife cross validation test. Statistical analysis also made for the representation of distance distribution between translation initiation site and transcription start site using gamma distribution.

2.3 Transcription Factor

Transcription factor (TF) finding is an important task among various activities in gene regulation. It needs sequence of steps to computationally perform the task of TF finding. Principal component analysis may lead the prediction to achieve high throughput. Smitha et al. [12] contributed to do the task. The authors have applied Random forest classifier to discover the TF. To analyze the efficiency, threshold curve, cost/benefit curve and ROC curve are used.

Tianyin Zhou et al. [15] presented a shape augmented model - an innovative way of taking DNA shape into account using Protein Binding Microarray (PBM). Support Vector regression is used to train the TF binding specificity.

Gusmao EG et al. [14] conceived a model hidden Markov model which is applied to combine DNas I Hyper sensitivity(DHS) sequence and histone sequence to locate chromatine open region and TF binding sites. The open chromatin region contains few footprint profiles, an indication of active TF binding sites. The measures specificity and sensitivity are considered for evaluation of this model.

III. CONCLUSION

Experimental methods for promoter identification and gene regulated process provide accuracy but suffer from being time consuming and lack of handling volume of data. In order to overcome these difficulties, computational methods for promoter prediction and algorithms to discover gene regulatory elements have been proposed. The previous studies include some analysis of patterns commonly found in promoter regions, such as -10 and -35 motifs .Though multifarious models are found to be effective, still they need to uncover various characteristics. Because of the dynamicity of gene regulatory process, promoter prediction models still require improvement. Indeed this field has a wider exposure of detailed research work.

REFERNCES

[1] Humberto Barrios, Brenda Valderrama & Enrique Morett , "Compilation and analysis of σ^{54} dependent promoter sequences" , Nucleic Acids Research 1999 , Vol. 27 , No.22 4305-4313

[2]Qian-Zhong Li, Hao Lin , "The recognition and prediction of sigma70 promoters in Escherichia coli K-12 " , Journal of theoretical Biology 242 (2006) 135 - 141

[3]Ronna R Mallios, David M Ojcius and David H Ardel , "An iterative strategy combining biophysical criteria and duration hidden Markov models for structural predictions of Chlamydia trachomatis σ66 promoters" , BMC Bioinformatics 2009 , 10:27

[4] Wang Q1, Wan L, Li D, Zhu L, Qian M, Deng M , "Searching for bidirectional promoters in Arabidopsis thaliana" , BMC Bioinformatics. 2009 Jan 30;10 Suppl 1:S29

[5] Guofeng Meng Axel Mosig and Martin Vingron, "A computational evaluation of over-representation of regulatory motifs in the promoter regions of differentially expressed genes", *BMC Bioinformatics* 2010 11:267

[6] Charles Bland, Abigail S Newsome, Aleksandra A Markovets , "Promoter prediction in E. coli based on SIDD profiles and Artificial Neural Networks" 7th Annual MCBIOS Conference Bioinformatics: Systems, Biology, Informatics and Computation Jonesboro, AR, USA. 19-20 February 2010 https://doi.org/10.1186/1471-2105-11-S6-S17

[7]Ravi Gupta, Priyankara Wikramasinghe, Anirban Bhattacharyya, Francisco A Perez, Sharmistha Pal and Ramana V Davuluri, "Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data" , BMC Bioinformatics 2010, 11(Suppl 1):S65

[8] C. Premalatha, Chandrabose Aravindan, Kamarajan Kannan , " Promoter prediction in eukaryotes using soft computing techniques" , 2011 IEEE Conference - Recent Advances in Intelligent Computational Systems

[9]Hao Lin Qian-Zhong Li , "Eukaryotic and prokaryotic promoter prediction using hybrid approach", Theory Biosci. (2011) 130: 91–100

[10] Yanglan Gan, Jihong Guan and Shuigeng Zhou , "A comparison study on feature selection of DNA structural properties for promoter prediction", BMC Bioinformatics 2012, 13:4

[11] Zaw ZawHtikea , Shoon LeiWin , " Recognition of Promoters in DNA Sequences Using Weightily Averaged One-dependence Estimators" , Procedia Computer Science, *2013 , Vol 23, Pages 60-67*

[12] Smitha C S , Saritha R , "Computational Transcription Factor Binding Prediction Using Random Forests" , International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT) 2014

[13]]Lin H, Deng EZ, Ding H, Chen W, Chou KC , " iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition" , Nucleic Acids Res. 2014 Dec 1;42(21):12961-72.

[14] Gusmao EG, Dieterich C, Zenke M, Costa IG , ".Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications ", Bioinformatics. 2014 Nov 15;30 (22): 3143-51

[15] Tianyin Zhoua, Ning Shenb,c, Lin Yanga, Namiko Abed, John Hortonc,e, Richard S. Mannd,f, Harmen J. Bussemakerf,g, Raluca Gordânc,e and Remo Rohsa , " Quantitative modeling of transcription factor binding specificities using DNA shape" , Proceedings of National Academy of Sciences , 2015 , 112(15), 4654–4659.

[16] Ramzan Kh. Umarov and Victor V. Solovyev , "Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks" , PLoS One . 2017; 12(2): e0171410.

[17] Ilham Ayub Shahmuradov , Rozaimi Mohamad Razali, Salim Bougouffa, Aleksandar Radovanovic and Vladimir B. Bajic , "bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli " , Bioinformatics, 33(3), 2017, 334–340 doi: 10.1093/bioinformatics/btw629

[18] Wenying He , Cangzhi Jia, Yucong Duan and Quan Zou , "70ProPred: a predictor for discovering σ^{70} promoters based on combining multiple features " , 11th International Conference on Systems Biology (ISB 2017) Shenzhen, China. 18-21 August 2017

[19] Marco Di Salvo, Eva Pinatel, Adelfia Tala, Marco Fondi , Clelia Peano and Pietro Alifano , "G4PromFinder: an algorithm for predicting transcription promoters in GC-rich bacterial genomes based on AT-rich elements and G-quadruplex motifs" BMC Bioinformatics (2018) 19:36

AUTHORS PROFILE

Mrs.S.Sasikala is a PhD Scholar, registered in Manonmaniam Sundaranar University, Tirunelveli. She is working as Assistant Professor in Computer Science department, Rani Anna Govt. College For Women,Tirunelveli. She has more than 20 years of teaching experience.

