BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis
Volume 10(4)

A novel sequence and context based method for promoter recognition

Umesh P[1]*, Jitendra Kumar Dubey[2], Karthika RV[1], Betsy Sheena Cherian[3], Gopakumar Gopalakrishnan[2] & Achuthsankar Sukumaran Nair[1]

[1]Department of Computational Biology and Bioinformatics, University of Kerala, Thiruvananthapuram – 695581, Kerala, India; [2]Department of Computer Science and Engineering, National Institute of Technology, Calicut - 673601, Kerala, India; [3]Faculty of Science, Kuwait University, P. O. Box: 5969, SAFAT: 13060, Kuwait; P Umesh – Email: toumesh@gmail.com; Phone: +91 4712308759, Fax: +91 4712308759; *Corresponding author

Received March 12, 2014; Revised March 17, 2014; Accepted March 18, 2014; Published April 23, 2014

Abstract:
Identification of promoters in DNA sequence using computational techniques is a significant research area because of its direct association in transcription regulation. A wide range of algorithms are available for promoter prediction. Most of them are polymerase dependent and cannot handle eukaryotes and prokaryotes alike. This study proposes a polymerase independent algorithm, which can predict whether a given DNA fragment is a promoter or not, based on the sequence features and statistical elements. This algorithm considers all possible pentamers formed from the nucleotides A, C, G, and T along with CpG islands, TATA box, initiator elements, and downstream promoter elements. The highlight of the algorithm is that it is not polymerase specific and can predict for both eukaryotes and prokaryotes in the same computational manner even though the underlying biological mechanisms of promoter recognition differ greatly. The proposed Method, Promoter Prediction System - PPS-CBM achieved a sensitivity, specificity, and accuracy percentages of 75.08, 83.58 and 79.33 on *E. coli* data set and 86.67, 88.41 and 87.58 on human data set. We have developed a tool based on PPS-CBM, the proposed algorithm, with which multiple sequences of varying lengths can be tested simultaneously and the result is reported in a comprehensive tabular format. The tool also reports the strength of the prediction.

Availability: The tool and source code of PPS-CBM is available at http://keralabs.org.

Background:

High throughput sequencing technologies has contributed to the increase in the number of sequencing projects which in turn increase the number of completely sequenced genomes. This increase in sequenced genomes demands hike in genome annotation which includes gene identification and its functional annotation. One of the important parts in genome annotation is the identification of promoter, the specific regions on the upstream of gene where the RNA polymerase binds to initiate transcription which is one of the major steps in gene regulation.

The sequence, structure and composition of promoters of eukaryotes and prokaryotes differ in many aspects. Prokaryotic promoter prediction is a comparatively easier task due to their relatively simpler gene structure and their high sequence conservation. The complexity of eukaryotic promoter structure makes their identification a difficult task. Even though many computational methods and machine learning techniques are powerful in making highly accurate predictions by learning the underlying data patterns, they are not yet widely explored for the eukaryotic promoter prediction.

The available promoter prediction algorithms are of three types- signal based, property based, and hybrid approach. In signal-based method, DNA sequences are first mapped to signals and are analysed to capture the strong signature features for the promoter. In property based method, various physico-chemical properties are derived from training data set.

In the hybrid approach, a combination of signal based and physico-chemical property based approaches are used. Computational methods like soft computing techniques are largely based on the presence of core-promoters, TSS region and specific motifs like TATA box, INR, DPE, BRE **[1-3].** The performances of all these algorithms are highly dependent on the training dataset used for deriving the information.

Elucidating the sequence to be a promoter in terms of quantitative characterization, termed promoter strength, is one of the new themes in promoter prediction. Many experimental methods are available for measuring the promoter activity based on reporter protein activity determination such as luminescence, fluorescence and spectrometry of enzyme products **[4].** The main challenge is to fit the data obtained from the wet lab experiments into the computational models.

Combinatorial synthesis **[5]** and artificial engineering of promoters is a fast growing area in synthetic biology. The characterization of promoters will accelerate the bottom up approach for designing the genetic construct for useful applications. The massive parallel sequencing method like ChIP-seq can be used for the identification of promoters in different cellular conditions **[6].** This will give more resolution to the mechanism and function of promoters. This can be extended as a tool for automated synthesis of promoter sequence and can serve as a module in tools and programming languages for synthetic biology **[7].** In this study, our aim was to develop a model for predicting promoter sequence and to predict its strength in both prokaryotes and eukaryotes.

Methodology:

Dataset

Promoters of *E. coli* and humans were downloaded from RegulonDB **[8]** and Eukaryotic Promoter Database (EPD) **[9]** respectively. For E. coli, non-promoter data sets were prepared from both the intergenic regions of *E. coli* genome downloaded from Ecogene **[10]** and the gene sequences downloaded from RegulonDB. Human negative set was obtained from gene sequences downloaded from NCBI-Gene database.

Context structure features

The presence of n-mers is considered as the context structure feature for both *E. coli* and human promoter prediction **[11, 12].** In this study, all possible 5 mers (45) are taken into account. Every test sequence is assigned a "ContextScore" which is calculated from the count of occurrence of *5-mers* in the training data as well as test data.

feature_value(X)=c/l, where X is a 5-mer present 'c' number of times. 'l'=total number of *5-mers* in the sequence. The mean (m1(X)) and standard deviation d1(X) of all the feature values for feature X for a set of N positive sequences is: **(Please see supplementary for equation 1 & 2).**

Z-test is used as one of the measures to classify a test sequence as promoter or non-promoter. Based on the Z-test, "score1", "score2" and the ContextScore is calculated as follows:

if $|x-m_1(X)|/d_1(X) <3$ then score1=1 else score1=0
if $|x-m_0(X)|/d_0(X) >3$ then score2=1 else score2=0
ContextScore = (score1 + score2)/(2*(length(D))) **(3)**

Prediction of promoters in human sequences

For identifying a human promoter sequence, the presence of TATA box is searched and its presence makes TATA_score 1 or else score is 0. The initiation element (Inr) in the test sequences were then searched and if it is present, the presence of Down Stream Promoter Element (DPE) is also confirmed. If both INR and DPE are present, Inr_score =1, else Inr_score = 0. CpG island association is another sequence signal feature considered for human promoter prediction. GC percentage and observed/expected ratio are used to confirm the presence of CpG islands.

Observed/Expected (o/e) = p(CG)/{p(C) * p(G)} **(4)**

Where, n(C) is the number of C's in a test sequence, n(G) is the number of G's in a test sequence, L is the length of the test sequence, and n(CG) is the number of CG's in a test sequence. Finally if GCp > 0.5 and o/e > 0.4, then D is CpG related and GC_score = 1, else GC_score = 0.

Final score for the test sequence is calculated as follows: FinalScore = (ContextScore + TATA_score + Inr_score + GC_score)/4 **(5)**

Prediction of promoters in E. coli sequences

For E.coli sequences we searched for TATAAT, TTGACA and TSS. To calculate TATAAT_score, TTGACA_score and TSS_score of a test sequence, sub-sequences separated by 20 bases are searched for a match of minimum three positions within the sequence TATAAT and then TTGACA. TATAAT_score and TTGACA_score are then calculated as the sum of the individual position scores calculated based on the probability of occurrence of individual symbols at their specific positions in the consensus sequence, as in **Table 1 (see supplementary material)** divided by six **[13].** If stop codons are not present, then TATAAT_score is set to zero. If the latter sub-sequence only matches less than three positions with the test sequence, then TTGACA_score is set to zero. If TATAAT_score> 0 and TTGACA_score>0, then the Transcription Start Site score, TSS_score = 1 for the test sequence.

Transcription Start Site score, TSS_score = 1 for the test sequence.
AT content in a test sequence is calculated using the following measures.
p (A) = number of As in the test sequence/length (test sequence); p (T) = number of Ts in the test sequence/length (test sequence); p (AT) = number of ATs in the test sequence/length (test sequence); AT_content = p (A) + p (T)
Observed/Expected, oeAT = p(AT)/(p(A)*p(T))

Final score for the test sequence is calculated as follows:
FinalScore=(score+AT_Score+TSS_score)/4 **(6)**

FinalScore and thresholds were selected by a random trial and error method **Table 2 (see supplementary material)** based on the training set. **Figure 1** illustrates Promoter Prediction System Using Context based Method for Human and *E. coli* Promoters. The test results obtained are tabulated as in **Table 3 (see supplementary material)**. A comparison with similar tools is given in supplementary file 1. The result shows that the prediction performance of the method is higher compared to other commonly used tools.

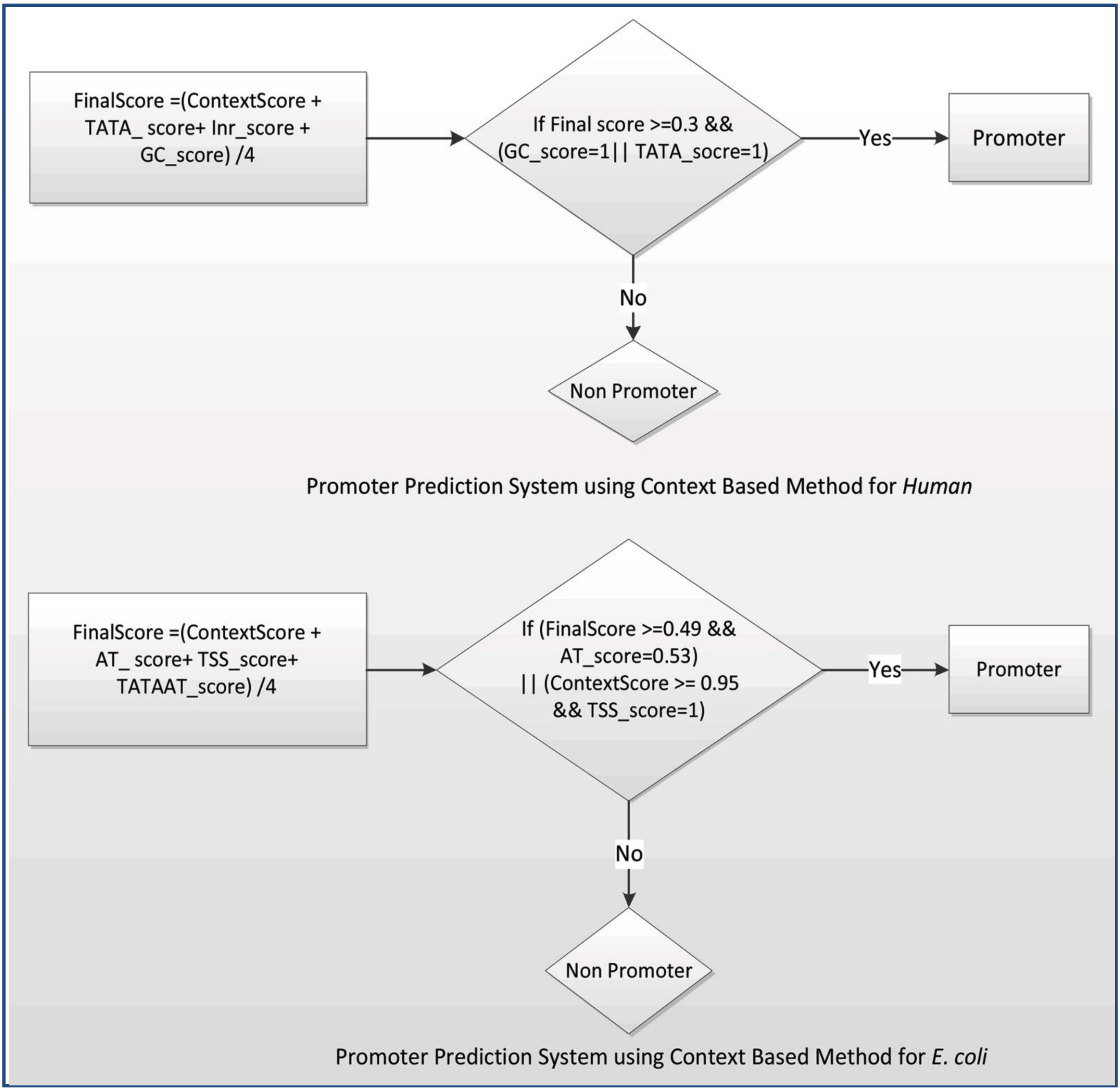
FinalScore =(ContextScore + TATA_ score+ Inr_score + GC_score) /4
If Final score >=0.3 && (GC_score=1|| TATA_socre=1)
Yes
Promoter
No
Non Promoter
Promoter Prediction System using Context Based Method for Human
FinalScore =(ContextScore + AT_ score+ TSS_score+ TATAAT_score) /4
If (FinalScore >=0.49 && AT_score=0.53) || (ContextScore >= 0.95 && TSS_score=1)
Yes
Promoter
No
Non Promoter
Promoter Prediction System using Context Based Method for E. coli

Figure 1: Data flow diagram of Promoter Prediction System Using Context based Method for Human and *E. coli* Promoters.

Comparison with PHI

263 promoters with known PHI (Promoter Homology Index), the index to quantify promoter strength **[14],** are used for the study and have been compared with the computationally identified promoter strength. Also, the correlation between the predicted final score and PHI index are taken note of.

Results & Discussion:

The method employs the context structure feature which relies on the count of n-mers along with sequence signal features. Most of the existing systems include a selection of *5-mers* as feature elements. *5-mers* are significant since it is the smallest n-mer which can encapsulate the behavior of consensus sequence. The PPS-CBM reported in this paper stands out for its comprehensive use of the complete set of possible *5-mers* (1024). We have taken this approach as the biological significance of such a large number of *5-mers* cannot be easily analyzed. No literature has yet reported a list of highly impacting *5-mers* and their predictive powers. Using such a large number of 5mers naturally puts a heavy demand on the computational requirement. However this is not a major concern as this occurs only at the time of training, not during individual predictions.

It would be ideal to go from the comprehensive set to either (1) biologically significant *5-mers* or (2) statistically significant 5-mers, which are confirmed to have predictive power. Further, PPS-CBM also uses most of the consensus features - TATA box, Initiator Elements, and Downstream Promoter Elements for human promoter prediction. The TATA- box is a core promoter element in eukaryotes **[15].** If TATA boxes are absent in promoter sequences, Initiator elements **[16]** and Down Stream Promoter Elements (DPE) **[17]** are used to identify promoters. Another feature used is the presence of CpG islands. Since CpG Islands are found near the promoter region in mammalian genes **[18],** it is a significant feature to identify human promoter. TATAAT box and TTGACA box are the important consensus features used in case of *E. coli* promoters, for accomplishing the task of prediction. Promoter region is less stable compared to the coding region which shows the abundance of Adenine and Thymine. The AT percentage, a feature used for predicting *E. coli* promoters, thus makes a significant contribution to the prediction model. The final score calculated is correlated with Promoter Homology Index (PHI) values of *E. coli* sequences and a high correlation of 0.60 is obtained.

Conclusion:

We reiterate that PPS-CBM is the first one that is noticed to have a unified predictor for both eukaryotes and prokaryotes and in spite of this generality; the supplementary file 1 clearly shows an enhancement in the prediction performance. Another advantage of this model is that it gives a clue to the promoter strength which is not a common feature found in other existing tools. PPS-CBM reports a measure which can be correlated with the strength of the promoter. Based on the scores calculated, the promoters can be classified into weak, strong and very strong categories.

Reference:

[1] Zhang MQ, *Genome research*. 1998 **8**: 319 [PMID:9521935]
[2] Ohler U *et al. Bioinformatics* 1999 **15**: 362 [PMID:10366656]
[3] Kai Song, *Nucleic Acids Res.* 2011 **40**: 963 [PMID: 21954440]
[4] Skrlj N *et al. Anal Biochem*. 2010 **396**: 83 [PMID: 19720040]
[5] Murphy KF *et al. PNAS*. 2006 **104**: 12726 [PMID: 17652177]
[6] Gupta R *et al. BMC Bioinformatics* 2010 **11**: S65 [PMID: 20122241]
[7] Umesh P *et al. Syst Synth Biol*. 2010 **4**: 265 [PMID: 22132053]
[8] Huerta AM *et al. Nucleic Acids Res*. 1998 **26**: 55 [PMID: 9399800]
[9] Praz V *et al. Nucleic Acids Res*. 2002 **30**: 322 [PMID: 11752326]
[10] Rudd KE, *Nucleic Acids Res*. 2000 **28**: 60 [PMID: 10592181]
[11] Pedersen AG *et al. Comput Chem*. 1999 **23**: 191 [PMID: 10404615]
[12] Breathnach R & Chambon P, *Annu Rev Biochem*. 1981 **50**: 349 [PMID: 6791577]
[13] Hawley D K & William RM, *Nucleic acids research*. 1983 **11:** 2237 [PMID: 6344016]
[14] Harley CB & Robert PR, *Nucleic Acids Res.* 1987 **15:** 23430 [PMID: 3550697]
[15] Wobbe CR & Struhl K, *Mol Cell Biol*. 1990 **10:** 3859 [PMID: 2196437]
[16] Javahery *et al. Mol CellBiol*. 1994b **14:** 116 [PMCID: PMC358362]
[17] Burke TW & Kadonaga JT, *Genes Dev*. 1997 **11**: 3020 [PMID: 9367984]
[18] Pedersen AG *et al. Comput chem*. 1999 **23:** 191 [PMID: 10404615]

Edited by P Kangueane
Citation: **Umesh *et al.*** Bioinformation 10(4): 175-179 (2014)

Supplementary material:

Methodology

Context structure features:

The mean (m1(X)) and standard deviation d1(X) of all the feature values for feature X for a set of N positive sequences is:

$$m_1(X) = \frac{\sum_N feature\_value(x)}{N} \quad \textbf{(1)}$$

$$d_1(X) = Stdev(feature\_value(x)) \quad \textbf{(2)}$$

Similarly, m0(X) and d0(X) are calculated for a set of N negative sequences.

Table 1: Probability of occurrence of individual symbols at consensus sequence TATAAT and TTGACA

Consensus sequence	Position	Sub sequence character	If character, Score	Else Score
TATAAT	**1**	T	0.77	0.23
	2	A	0.76	0.24
	3	T	0.60	0.40
	4	A	0.61	0.39
	5	A	0.56	0.44
	6	T	0.82	0.18
TTGACA	**1**	T	0.69	0.31
	2	T	0.79	0.21
	3	G	0.61	0.39
	4	A	0.56	0.44
	5	C	0.54	0.46
	6	A	0.54	0.46

Table 2: Calculation of final score in Human and *E. Coli* test sequence

Organism	ContextScore	GC content	Observed/Expected GC	FinalScore= (ContextScore+TATA_score+ Inr_score+ GC_score)/4
Human	0.95	0.5	0.4	>0.3
E coli	0.95	0.53	0.51	>0.49

Table 3: Results of new promoter prediction algorithm for Human and *E. coli* Sequence. For datasets, see supplementary file. (Where TP- True positive, TN-True Negatives, FP- False Positive, FN-False Negative, MCC-Matthews Correlation coefficient)

TP	TN	FP	FN	Sensitivity (%)	Specificity (%)	Accuracy	Correlation Coefficient (MCC)
1040	1061	139	160	86.67	88.41	87.58	*0.75*
901	1003	197	299	75.08	83.58	79.33	0.59