Journal of Bioinformatics and Sequence Analysis Vol. 1 (1), pp. 011-016, May, 2009
Available online at http://www.academicjournals.org/JBSA
© 2009 Academic Journals

Full Length Research Paper

A computational technique for prediction and visualization of promoter regions in long human genomic sequences

Q. M. Alfred[1]*, K. Bishayee[1], P. Roy[2] and T. Ghosh[3]

[1]University Institute of Technology, University of Burdwan, West Bengal, India 713104
[2]Department of Biotechnology, University of Burdwan, West Bengal, India, 713104
[3]Burdwan Medical College and Hospital, West Bengal, India.

Accepted 29 April, 2009

This communication proposes a simple algorithm with high specificity and sensitivity for determining promoter regions in human genomic sequences. This method relies upon non-redundant and experimentally verified promoter data sets form Eukaryotic Promoter Database (EPD) as training parameters. This technique predicts and computationally satisfies the promoter regions in the NCBI annotated database around gene sequences.

Keywords: Promoter, CpG islands, transcription start site (TSS), discrete fourier transform/ fast fourier transform (DFT/FFT).

INTRODUCTION

Objective of human genome project is to correctly annotate the regulatory regions, transcription start and stop site, coding regions, exons and introns etc. Promoter is a fragment of DNA sequence centered on transcription-start-sites (TSS), is biologically responsible for the transcription from DNA to RNA sequence. Therefore reliable recognition of promoter region is essential for understanding the biological mechanism as well as helping the field of genetic engineering. As every gene is recognized by the features of promoter sequence and widely varies among species to species. Some promoter features are well reported in literatures, for example, TATA box which is sometimes located at -10 to -35 positions upstream of TSS (0 postion), CpG islands is another well known promoter feature mostly found in eukaryotic (human, mouse etc.) genomes but not in prokaryotes(*Escherichia coli* etc.). Till date, no feature is found which deterministically confirms the existence promoter sequence. The above mentioned feature in combine with some other features predicts the existence of promoter sequence in large genomes. Experimentally (biochemical method) finding a promoters form huge genome like human is almost impossible for researchers. Hence, prediction of promoters by computational method is a highly regarded area of interest. Several research groups have developed techniques and algorithms for in-silico (in computer) promoter recognition. Among them, weight matrix model (Prestridge, 1995; Bucher, 1990), Hidden Markov Model (HMM) (Burge and Karlin, 1997; Kulp and Haussler, 1996), feature (signal/context) based model (TATA, CpG etc.) (Pedersen et al. 1998; Ponger and Mouchiroud 2002; Wu and Xie, 2007; Zhang, 1998a; Fickett and Hatzigeorgiou 1997), neural network model (Brunak et al., 1991; Pedersen and Engelbrecht 1995), graph based model (Matsuda et al., 2002) etc. But each method has its inherent advantage and disadvantage. Most of the models suffer from computational complexities and specificities in promoter prediction.

Motivated by the importance and presence of good research authors have proposed a simple but novel approach in promoter region identification as well as potential TSS prediction. Present method conceptually differs from the above well known techniques

METHOD

This method is highlighted by the following steps:

*Corresponding author. E-mail: quazi_alfred@yahoo.co.in.

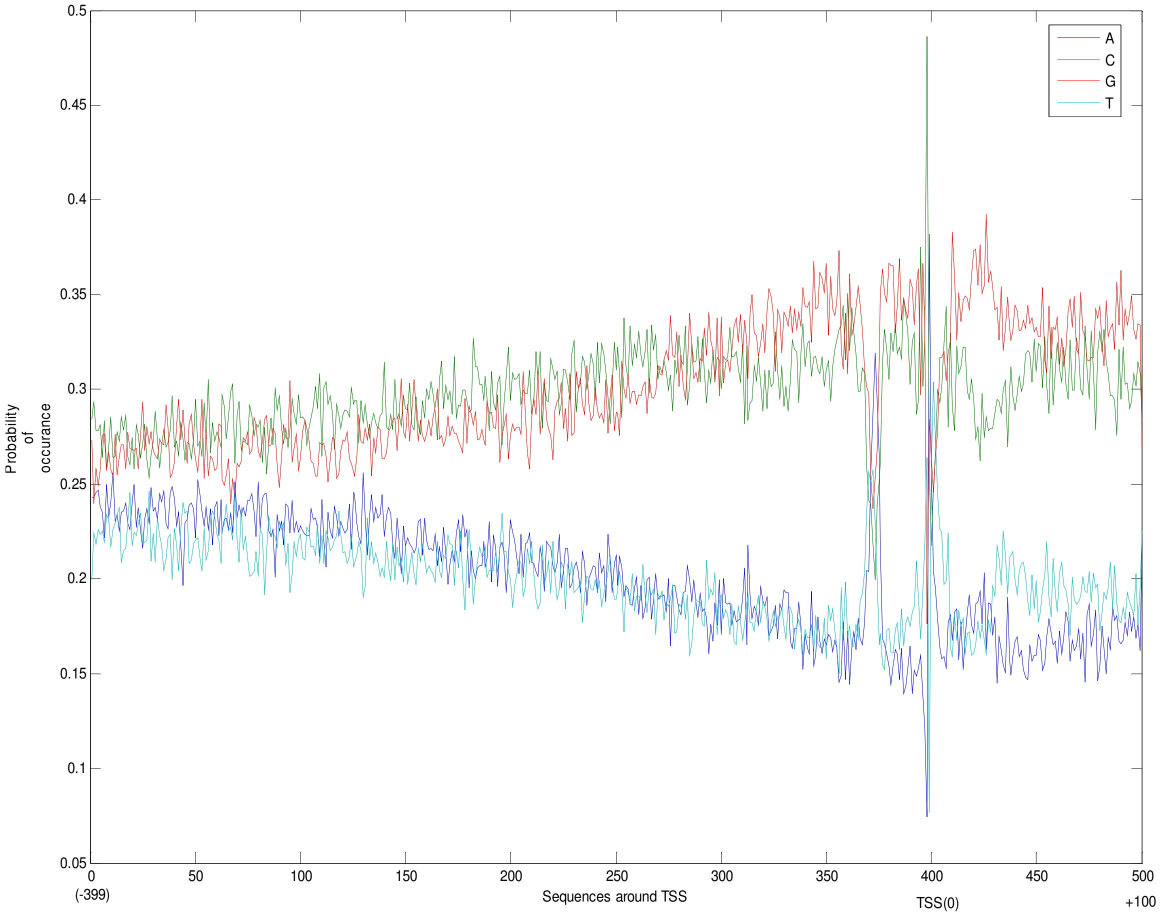

Figure 1. Plot of probability matrix around TSS of nucleotides (A, T, C, G).

Step-1:

First, more than 1500 Homo sapiens genes and their known promoter regions are gathered from SIB-EPD (Eukaryotic Promoter Database) database (Cavin et al., 1998). These data sets are selected as they are non-redundant, experimentally verified and filtered. Human promoter sequences in the region of -399 to +100(500 bp) around TSS are considered as testing data sets from these experimentally known genes.

From these data sets, a positional frequency matrix for four (4) nucleotides is derived at each 500 positions (4×500 matrix). As the number of promoter is very high, this frequency matrix may be approximated as positional probability matrix from the following relation.

$$P_{i<A,T,C,G>}=f_{i<A,T,C,G>}=\frac{n_{i\langle A,T,C,G)}}{N}$$

Where;
i =-399,-398……..0(TSS).. 1, 2…..100
N=no. of promoters, n_i=no. of nucleotide at i^{th} positions

Step-2:

This 4×500 matrix signifies discrete probability distribution of four nucleotides at each 500 individual positions, is graphically plotted in Figure 1. Unlike the weight matrix model, where a weight table is calculated in terms of background frequency ¼ (occurrence of any nucleotide), but here the probability at each position is exploited in the calculation of score.

Now, any unknown and long human genomic sequence is scanned by sliding window of length 500, which is then shifted by 1bp (may be shifted more for fast computation compromising error). Within each window, the scores are entered according to the occurrence of any of the four nucleotides with reference to the probability matrix. Total score is calculated by adding scores at each position within a window. This process is repeated by shifting the window by 1bp along the forward strand (5-3) direction.

Step -3:

Assuming background probability 0.25, the cut-off score is chosen as 125 for 500 positions.

During scanning any genomic sequence, when the total score in the sliding window exceeds 125, is selected as hit segment. Now, the scores and the positions of these hits are sorted out for further processing.

In each hit segment, probabilities during 371-420 positions are Fourier transformed (DFT), which will be used later.

Step-4:

Maximum occurrence value among four nucleotides are interpolated for all 500 positions in the probability matrix to generate a

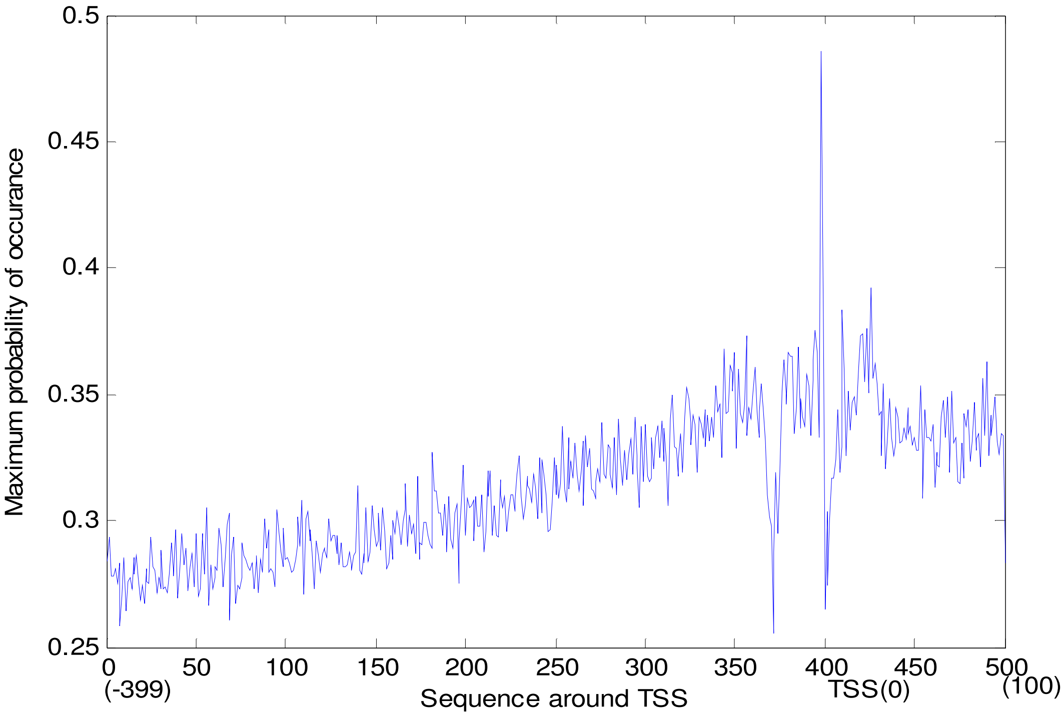

Figure 2. Plot of maximum probability of occurrence.

1×500 matrix, is plotted in Figure2. Then for 50 positions (-371 to +420) around transcription start site (TSS) are transformed (DFT) into frequency domain, shown in Figure 3. After DFT (FFT), the coefficients are obtained at 50 discrete frequency points. Among these, 0th coefficient (DC value) indicates sum of all max-values or global shape of the template but form 1st coefficient to the rest carries detailed feature of this template shape(shown in Figure 2).

Step-5:

It is found that, in eukaryotic genome the score is high (>130) in CpG rich regions whether the potential promoter exists or not. Therefore, to avoid false prediction rate (false positive), the probability score form position -29(371) to +20(420) duration is Fourier transformed for all hit positions (where score>125) at time of scanning a genomic sequences. First 20(1:20) DFT coefficients of each hit sequence (500 lengths) are compared with the template DFT (1:20) (Figure 3). When this yields minimum difference value, confirms the pattern matching with template, shown in Figure 2. Here, the objective is to consider those sequences as promoters which shows minimum error with the reference pattern of promoters

RESULTS

This algorithm is applied on the five moderately long genomic contigs of Homo sapiens chromosome 22 from NCBI's GenBank (Benson et al., 1998) of total length 3.65 Mbp and 41 TSSs in the forward strands. Table 1 shows the overview of these genomic sequences.

Earlier Xiomeng Li et al. (2008) and Lu et al. (2008) has compared the performance between four well known promoter prediction techniques in comprehensive manner. Table 2 comparatively shows their performance in terms of specificity and sensitivity when checked with Chromosome-22 sequence annotated by Sanger institute (http://www.sanger.ac.uk/HGP/Chr22).

Among them, DragonGSF and HPR-PCA (Lu et al., 2008) are preferable for predicting promoter region for long genomic sequences. Table 3 details the performance comparison of this algorithm (designated as PR-DFT) along with DragonGSF and HPR-PCA.

To illustrate this algorithm with an example, Figure 4 graphically represents the score of matching when scanning the contig NT_037887 in chromosome-16. The peaks relative to the surroundings predicts potential promoter regions. Figure 5 confirms the result after matching the DFT coefficients to the reference shown in Figure 4. Lower the value of their difference (score) higher the probability of prediction.

From the figure it is obvious that it predicts eight (8) promoter (gene) regions in *NT_037887* between 140000 to 175000bp.

According to NCBI annotation this region consists of seven (7) promoters. The predicted TSS also satisfies the annotated site with acceptable accuracy.

This algorithm is implemented in Matlab® environment with SUN Ultra-40M2 workstations.

Data sets

Data sets used here as training data is taken from SIB-EPD promoter database. NCBI annotated human data-

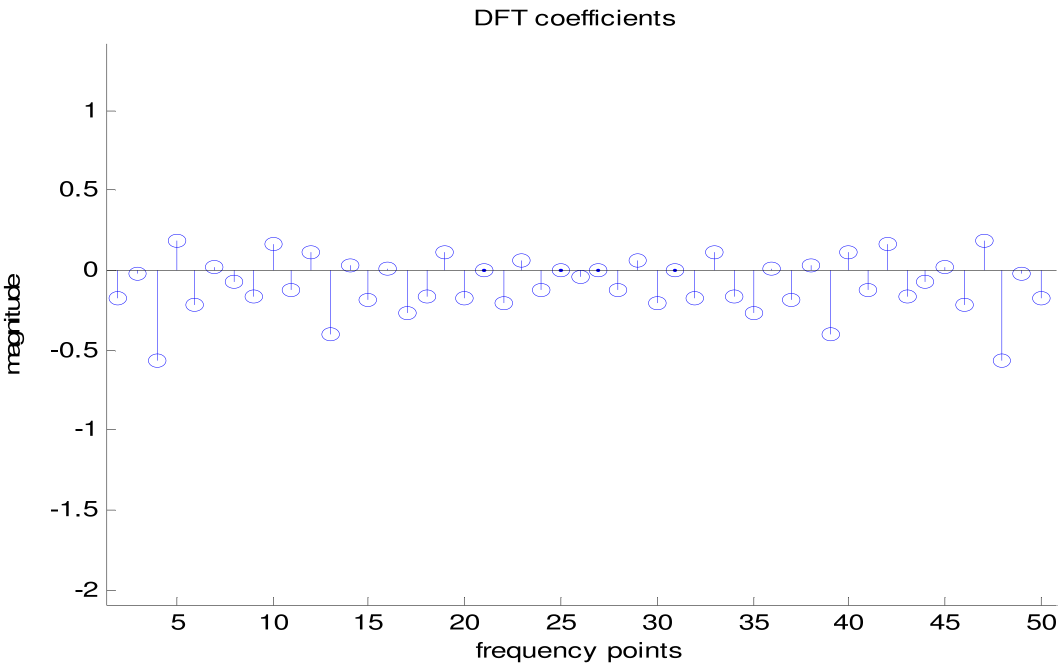

Figure 3. DFT coefficients for 50 frequency points (point 1 is not shown having magnitude 17).

Table 1. Description of large genomic sequence used as test set.

Contig	Description	Length	Number of TSS
NT_028395.3	Homo Sapiens	647850	9
NT_011521.4	Chromosome 22	830225	11
NT_011525.7	Genomic sequence	1384186	8
NT_019197.5		320440	5
NT_011526.6		464629	9
Total		3647330	41

Table 2. Performance of four prediction system (Source: Xiomeng Li et. al)

Systerm	TP	FP	S_e	S_P
DragonGSF	269	69	0.6844	0.7959
FirstEF	331	501	0.8422	0.3978
Eponin	199	79	0.5064	0.7158
HPR-PCA	301	65	0.7659	0.8224

database used for computational verification of the algorithm.

DISCUSSION

This communication presents a simple technique by visualizing promoter sequence. Identification of promoter regions demands some decision making by visualizing and comparing both the score plot as well as DFT difference plot.

Generally, peaks (high score) with respect to background (low score) can be decided as promoter sequence. But when there are high peaks along with background can be considered as high CG rich regions.

ACKNOWLEDGEMENT

This work is financially supported by AICTE, India under RPS grant.

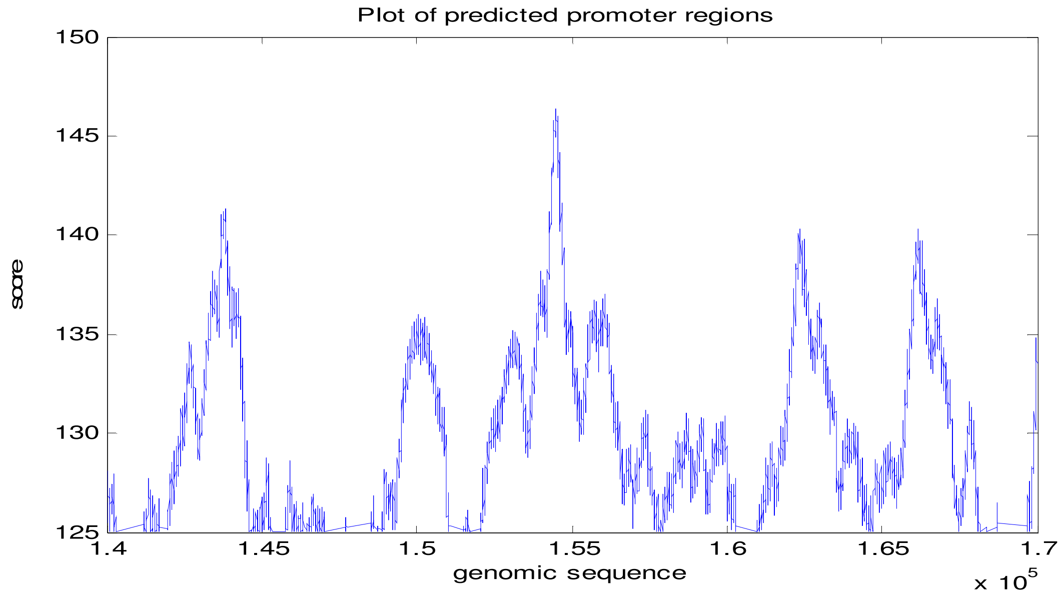

Figure 4. Plot of predicted promoter regions in NT_037887 (contig-1 14000-170000) of Chromosome-16.

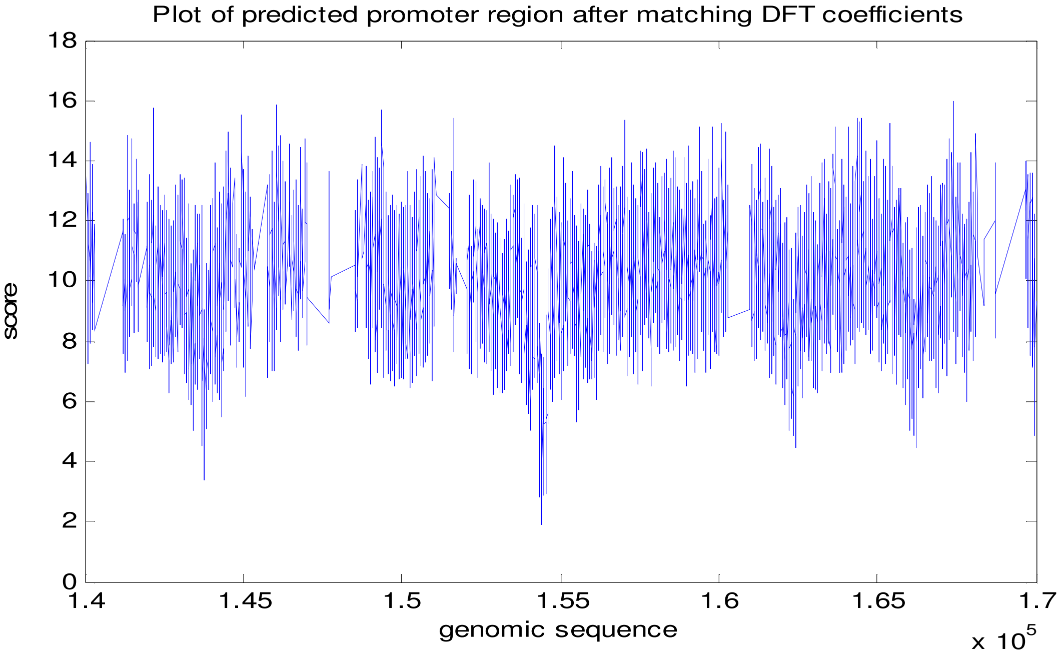

Figure 5. Plot of difference between DFT coefficients for Figure 4.

REFERENCES

Bucher P (1990). "Weight matrix descriptions of four eukaryotic RNA Polymerase II promoter elements derived from 502 unrelated promoter sequences. "Mol. Biol. 212: 563-589. Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, "*Genebank*", Nucleic Acids Res. 26(1):1-7.

Brunak S, Engelbrecht J, Knudsen S (1991). "Prediction of human mRNA donor and acceptor sites from the DNA sequence" J. Mol. Biol. 220: 49–65.

Burge C, Karlin S (1997). "Prediction of complete gene structures in human – 84- genomic DNA." J. Mol. Biol. 268: 78-94.

Cavin PR, Junier T, Bucher P (1998). "Eukaryotic Promoter Database EPD", Nucleic Acids Res.26, 353-357.

Fickett JW Hatzigeorgiou AG (1997)."Eukaryotic Promoter Recognition." Genome Res. 7: 861-878.

http://www.sanger.ac.uk/HGP/Chr22

Kulp D, Haussler D (1996), "A generalized hidden Markov model for the recognition of human genes in DNA." Proc Int Cong Intell Syst. Mol. Biol. 4: 134-142.

Li XM, Zeng J, Yan H (2008). "PCA-HPR: A New Method of Human Promoter Recognition Based on Principle Component Analysis." Bioinformation 2(9): 373-378

Liu YM, Li XM, Yan H (2008). Codon Relation Analysis for Promoter Recognition Using Indepent Component Analysis", J. Inf Comput. Sci. 5(1): 33-39

Matsuda T, Motoda H, Washio T (2002) "Graph-based induction and applications," Advanced Engineering Informatics", 16:135-143.

Pedersen AG, Baldi P (1998). "DNA Structure in Human RNA polymerase II Promoters.", J. Mol. Biol. 281: 663-673.

Pedersen AG, Baldi P (1999*).* "The Biology of Eukaryotic Promoter Prediction —a Review.", Comput. Chem. 23(3): 191-207

Pedersen AG, Engelbrecht J (1995). "Investigations Escherichia coli promoter sequences with artificial neural networks: new signals discovered upstream of the transcriptional start point" , Proceedings, third international conference on intelligent systems for molecular biology (ISMB-95) 292–299.

Ponger L, Mouchiroud D (2002). "CpG ProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics 18: 631-634.

Prestridge DS (1995). "Predicting Pol II promoter Sequences using Transcription -87- Factor Binding Sites." J. Mol. Biol. 249: 923-932.

Wu S, Xie X (2007). "Eukaryotic promoter prediction based on relative entropy and positional information.", Phys. Rev. E 75: 041908 1-7.

Zhang MQ (1998a). "A discrimination study of human core-promoters in silico.", Proc. Pacific Symp. Biocomputing 1998.