Bioinformatics, 36(8), 2020, 2375–2384
doi: 10.1093/bioinformatics/btz941
Advance Access Publication Date: 7 January 2020
Original Paper

Sequence analysis

A novel method *SEProm* for prokaryotic promoter prediction based on DNA structure and energetics

Akhilesh Mishra[1,2], Sahil Dhanda[1], Priyanka Siwach[1,3,*], Shruti Aggarwal[1] and B. Jayaram[1,2,4,*]

[1]Supercomputing Facility for Bioinformatics & Computational Biology, [2]Kusuma School of Biological Sciences, Indian Institute of Technology, New Delhi 110016, India, [3]Department of Biotechnology, Chaudhary Devi Lal University, Sirsa 125055, India and [4]Department of Chemistry, Indian Institute of Technology, New Delhi 110016, India

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on May 24, 2019; revised on November 8, 2019; editorial decision on December 10, 2019; accepted on January 2, 2020

Abstract

Motivation: Despite conservation in general architecture of promoters and protein–DNA interaction interface of RNA polymerases among various prokaryotes, identification of promoter regions in the whole genome sequences remains a daunting challenge. The available tools for promoter prediction do not seem to address the problem satisfactorily, apparently because the biochemical nature of promoter signals is yet to be understood fully. Using 28 structural and 3 energetic parameters, we found that prokaryotic promoter regions have a unique structural and energy state, quite distinct from that of coding regions and the information for this signature state is in-built in their sequences. We developed a novel promoter prediction tool from these 31 parameters using various statistical techniques.

Results: Here, we introduce SEProm, a novel tool that is developed by studying and utilizing the in-built structural and energy information of DNA sequences, which is applicable to all prokaryotes including archaea. Compared to five most recent, diverged and current best available tools, SEProm performs much better, predicting promoters with an '*F*-value' of 82.04 and 'Precision' of 81.08. The next best '*F*-value' was obtained with PromPredict (72.14) followed by BProm (68.37). On the basis of 'Precision' value, the next best 'Precision' was observed for Pepper (75.39) followed by PromPredict (72.01). SEProm maintained the lead even when comparison was done on two test organisms (not involved in training for SEProm).

Availability and implementation: The software is freely available with easy to follow instructions (www.scfbio-iitd.res.in/software/TSS_Predict.jsp).

Contact: bjayaram@chemistry.iitd.ac.in or psiwach29@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The pace at which whole genomes are sequenced is far ahead of that of genome sequence annotation, let alone a molecular understanding of the genome organization. Compared to experimental approaches, computational tools offer a faster option for sequence annotation; reliability being largely decided by their careful design which in turn is dependent on a fair understanding of the molecular mechanisms concerned.

One of the most sought after task in genome annotation is identification of promoter regions, not only for the validation of predicted genes and identification of novel genes but also to understand transcriptomic regulatory networks (with respect to promoter location and architecture). During initial years of sequence accumulation, promoter identification was regarded as some trivial task because of certain experimentally derived sequence information rules (Pribnow, 1975). However, with each passing year, it has emerged as one of the most daunting challenges in genome annotation. Genomic sequences in prokaryotes are highly adaptive within genomes and highly diversified across species to enable their survival in diverse and extreme conditions. This makes it difficult to detect conserved regulatory sites by sequence homology. Further, variability in the length of 5′ untranslated regions and presence of multiple transcriptional start sites (TSSs) does not make it obvious to look for promoters in the immediate upstream region of the annotated coding sequence. The complexity is further compounded by high gene densities, with neighbouring genes having generally very short intergenic spaces or in some cases have overlapping coding regions.

Recent reports on pervasive transcription, where transcription can initiate from any location, have further deepened the mystery (Wade and Grainger, 2014).

Many efforts have been made for developing efficient promoter prediction tools, based on different logics. Sequence based computational methods for promoter prediction have been moderately successful (Dekhtyar *et al.*, 2008; Jacques *et al.*, 2006), though machine learning on huge training data of promoter sequences have led to some good predictive but highly genome specific tools (de Jong *et al.*, 2012; de Silva *et al.*, 2011; Lai *et al.*, 2019; Shahmuradov *et al.*, 2016; Solovyev and Salamov, 2011; Umarov and Solovyev, 2017; Umesh *et al.*, 2014). Lately, attention has also been given to capture the structural and/or energetic signals of promoter regions. Some of the structural and/or energetic properties used for promoter prediction are bendability, curvature, inter-base pair (BP) properties, free energy, A-philicity and stress induced DNA duplex destabilization among others (Abeel *et al.*, 2008; Florquin *et al.*, 2005; Goñi *et al.*, 2007; Rangannan and Bansal, 2010; Wang and Banham, 2006). DNA shape determined by four distinct features—minor grove width, propeller twist, roll and helical twist—have been found to be important determinants in the identification of transcription factor binding sites (TFBs) and TSSs (Chiu *et al.*, 2015; Levo *et al.*, 2015; Zhou *et al.*, 2013, 2015). These studies undoubtedly establish that three dimensional structure of DNA, beyond the primary sequence, is a determinant of protein–DNA binding specificity. Despite the many insights resulting from such studies over the years, development of a promoter prediction tool giving high performance is still awaited. It is apparent that the extant conceptual frameworks are not adequate yet to understand the nature of promoter signals fully. There is a need to develop new ideas/models to explain the fine tuning of structural and energy state of promoter sequence with respect to interacting protein/transcription factors/ligands.

Rationale: target was to develop a new model for capturing the structural and energy state of promoters. For structural characterization, instead of taking DNA shape (the cumulative effect of various parameters), we decided to use all the individual parameters (28 in all) involved in spatial organization of the bases and BP steps—backbone organization, inter-BP arrangements, intra-BP arrangements and the relative positioning with respect to BP axis. Rationale was to have a broad horizon for finding some novel information. Structural analysis was guided by some breakthrough studies in the analysis of nucleic acid structures during last few decades (Beveridge *et al.*, 2004, 2012; Dixit *et al.*, 2005; Hassan and Calladine, 1995; Lavery *et al.*, 2009, 2010; Olson *et al.*, 1998; Pasi *et al.*, 2014; Yanagi *et al.*, 1991). For energy characterization, we relied on our own experience. We have been putting efforts to understand the language of DNA in terms of its energetics during the last 15 years and have obtained modest success in the process. In this series, we have reported that hydrogen bond, stacking and solvation energies show clear signatures of functional destinies of DNA sequences (Dutta *et al.*, 2006; Khandelwal *et al.*, 2012, 2014; Khandelwal and Bhyravabhotla, 2010; Khandelwal and Jayaram, 2012; Singh *et al.*, 2017; Singhal *et al.*, 2008). So with 31 parameters (20 structural and 3 energy), we characterized 16 519 primary prokaryotic promoter sequences (Mishra *et al.*, 2018). It was found that all parameters yield a signature signal at/near TSS and the information for this signature signal is in-built in promoter sequences. Power of an accurate model lies not only in its ability to explain but also to predict precisely. If this structural and energetic model of promoters is accurate, it should lead to a reliable promoter prediction tool with an uncommon ability to efficiently and accurately predict the prokaryotic promoters irrespective of genome/species. With this aim, we directed our efforts to develop a suitable promoter prediction algorithm from the 31 parameters by using various statistical techniques. Here, we present a novel method, *SEProm*, which is applicable to all prokaryotes including archaea and performs considerably well in comparison with the available promoter prediction programmes. The tool is freely downloadable with easy to follow instructions.

2 Materials and methods

2.1 Promoter sequence dataset

Initially, a total of 16 519 primary TSS positions were selected from global TSS mapping data of 12 organisms (*Methanolobus psychrophilus*, *Thermococcus kodakarensis*, *Halofrex volcanii*, *Mycobacterium tuberculosis* H37Rv, *Streptomyces coelicolor* A3, *Helicobacter pylori*, *Salmonella enteric* serovar *Typhimurium*, *Escherichia coli*, *Pseudomonas aeruginosa* PA14, *Bacillus amyloliquefaciens*, *Chlamydia pneumonia CWL029* and *Synechocystis* sp. PCC6803) (Albrecht *et al.*, 2011; Babski *et al.*, 2016; Cortes *et al.*, 2013; Hershberg, 2001; Jäger *et al.*, 2014; Jeong *et al.*, 2016; Kopf *et al.*, 2014; Kröger *et al.*, 2012; Li *et al.*, 2015; Liao *et al.*, 2015; Sharma *et al.*, 2010; Wurtzel *et al.*, 2012, respectively). On manual inspection, it was found that many primary TSSs for a given genome were located on consecutive positions. To remove redundancy in data, among the TSSs which were <5 bp apart, only one was selected based on the highest expression level (as reported in the paper concerned); the final number of primary TSSs selected in this way was 12 880. Sequences of 1001 nt length (spanning 500 nt upstream and downstream of the TSS positioned at 0), for the 12 880 selected TSSs positions, were extracted from respective genome sequences. As control dataset, a total of 6218 coding sequences (with length >1500 nt) from all the 12 organisms were taken and the central 1001 nt long region of each was extracted for final use. The complete set of promoter sequences and CDSs is available online (http://www.scfbio-iitd.res.in/software/SEProm_Data_TSS.jsp).

2.2 Parameters for promoter characterization

Complete structure (i.e. spatial arrangements of the constituent bases and BPs) of any DNA sequence can be completely defined by spatial arrangements of constituent BP steps defined by: backbone organization (alpha, beta, gamma, delta, epsilon, zeta, chi, phase and amplitude), inter-BP arrangements (shift, slide, rise, tilt, roll, twist, H-rise and H-twist), intra-BP arrangements (shear, stretch, stagger, buckle, propel and opening) and the relative positioning with respect to BP axis (X displacement, Y displacement, inclination, tip and axis-bend). We decided to use all the individual parameters instead of cumulative effect of these parameters (DNA shape) to capture the structural signals of promoters; besides having a broad horizon for extracting the information, amplification of signals was also the reason. Likewise, the complete energy state of DNA can be assessed by knowing the hydrogen bond energy, stacking energy and solvation energy of its constituent BP steps.

For extracting the standard values for all these parameters, we obtained crystal structures of 74 B-DNA (DNA only without any protein/ligands) molecules from NDB database (Supplementary Table S1). For these 74 DNA molecules, the values of 28 structural parameters (mentioned above) for unique 10 dinucleotide steps in the 5′–3′ direction were calculated using Curves+ (Lavery *et al.*, 2009) while in-house programmes were used for energy calculations (Singh *et al.*, 2017).

After calculating values for all the parameters for each B-DNA structure, all occurrences of unique 10 dinucleotide steps in the 5′–3′ direction were considered for each parameter and the average of all the occurrences was calculated to obtain a master table (Supplementary Table S2).

2.3 Structural and energy profiles of promoter sequences

Each sequence of the promoter dataset (12 880) and coding dataset (6281) were converted to 31 numeric sequences using the values of master table (Supplementary Table S2). The calculated dinucleotide values for each parameter were used for getting structural profile of 1001 nt long promoter and CDS sequences by performing moving average calculation on sliding window of 25 BPs covering 24 dinucleotide steps (Supplementary Figs S1 and S2). From the Supplementary Figure S1, it is clear that all the 31 parameters undergo a change in their values as the promoter DNA sequence proceeds to TSS, while there is no such change in coding sequence. At this

stage, it is difficult to assess which parameter contributes more than others. Since our promoter dataset represented promoters from a diverse range of prokaryotes, we thought of getting organism specific plots for each parameter so as to have more clarity about the behaviour of each parameter. From Supplementary Figures S2 (1–31), it is clear that though all parameters maintain same behaviour at TSS across organisms (i.e. increase or decrease in value), the strength and pattern of signal is somewhat different for each organism, meaning apparently that if we want to develop a universal tool with applicability to diverse prokaryotes, use of all parameters would be more profitable; however, relevant statistical tests studying discriminative power of various parameters to reach to a minimum set of parameters were not done.

The values were further made dimensionless using normalization; the values were normalized between 0 and 1 by subtracting the minimum value of the profile from each value and then dividing the value with the range of the profile (i.e. max. to min.). Using the normalized values for each parameter, the profiles of each sequence were obtained by performing the moving average calculation on a sliding window of 25 BP covering 24 dinucleotide steps; same exercise was followed for all the sequences. Then for each parameter, single profiles for promoters and CDS each were obtained by taking the average values for 12 880 promoter sequences and 6281 coding sequences, respectively. The profiles of each of the five categories (backbone, inter-BP, intra-BP, BP-axis and energy) were plotted in a single plot using Matlab software.

2.4 Training dataset

Of the 12 880 genomic sequences, 10 304 sequences (80%) were used as training dataset. Each sequence of training dataset was converted to 31 numeric sequences (one for each of the 31 parameters) by using values for 10 unique di-nucleotides steps from Supplementary Table S2. This was followed by normalization by performing moving average calculation on a sliding window of 25 BPs covering 24 dinucleotide steps. Hence for each sequence, 31 numeric sequences (each with a length of 975) were obtained.

2.5 Methodology

The very first step was to know whether the parameters were completely independent or were they correlated, and if correlated, to what degree? Correlation analysis was done using R software (Supplementary Table S3, in excel format). Varying degrees of correlation were observed for different pairs of parameters; only a few parameters were independent of some while being simultaneously correlated to others. For such a situation having multiple intercorrelated variables, principal component analysis (PCA) is best suited for extracting and visualizing the information which is expressed as a set of a few new variables called principal components. The numeric sequences of training dataset, obtained above, were subjected to PCA for the following selected regions of promoters.

(a)For the continuous stretch just upstream of TSS: two regions were selected viz. the '−75 to +5' (80 nt window) and the '−35 to +5' (40 nt window) with respect to TSS, as the positive dataset. The respective negative dataset was taken from 200 nt downstream of TSS i.e. +200 to +280 for the first window and +200 to +240 for the second window.

Values for the above four windows were extracted directly from the normalized list of numeric sequSEProm for prokaryotic promoter predictionences of training dataset. Values of each window for each parameter were converted to a single data point by taking average and in this way 31 data points were obtained for one window of each training sequence. For each window of each sequence, these data points were inserted into PCA analyses. Supplementary Figure S3 shows percentage of explained variance for the first 10 components; maximum variance (51.4%) is explained by PCA1. First five components are selected. Heat plots showing contribution of individual parameter in each component are also presented in Supplementary Figure S3. These values are then subjected to logistic regression analysis to get the regression equation, calculating the LOD score for that window (see Supplementary Material). The LOD score is converted to probability value. The threshold value for cut-off probability is selected by using values from 0.2 to 0.5 for sequence prediction; probability value of 0.5 is found to be satisfactory for optimum sensitivity and specificity. To get a Boolean value for both the lengths (40 and 80 nt), 'OR' operator is used. On the basis of 0.5 as probability threshold, probability results are classified as 0/1.

(b)For the important 'Motif' regions in different combinations: it is well established that RNA polymerase interacts with promoters by making specific contacts at specific regions: the '−10' region, the '−35' region and the '−70' region. Reviews are available regarding the exact location of these motifs (Haugen *et al.*, 2008 and references within) which can be summarized as:

'The −10 region': it is characterized by the presence of one 'core −10 element', which is generally present from −7 to −13 positions with respect to TSS. In some cases, an 'extended -10 element' spanning from −14 to −17 position is also present along with the core element while some promoters are characterized by the presence of a 'discriminator element' present from +1 to −6 position. Generally, extended −10 element or discriminator element is present along with core '-10' element when '-35' element is absent.

'The −35 region': the location, as per the studies cited above, most likely pertains to '−29 to −35' or '−30 to −36' positions; though, as per many reports [cited in Haugen *et al.* (2008) mentioned above] the position of this hexameric motif is variable and can be found at '-28 to -34' or −31 to −37, while in some sequences, the designated hexameric motif may be altogether absent.

'The −70 region': it can be present as a single stretch somewhere between '−63 to −80' in some promoters while some may have two sub-sites: proximal (somewhere between −40 and −50) and distal (somewhere between −55 and −68).

The above three motif regions were combined in the following different ways and were considered as positive dataset:

Combination 1: '−7 to −14', '−30 to −37', '−63 to −80'

Combination 2: '−7 to −14', '−29 to −38', '−40 to −46', '−50 to −60'

Combination 3: '−7 to −17', '−28 to −37', '−40 to −50', '−55 to −66'

Combination 4: '−1 to −13', '−28 to −37', '−40 to −50', '−55 to −68'.

For respective negative dataset for each combination, equal size motif windows, with equal inter-window distance, were extracted from 200 nt downstream of TSS. Since for each combination there were many windows, to reduce the volume of the input data, instead of using the 31 parameters individually, we grouped these to obtain four derived vectors namely, structure increasing vector, structure decreasing, energy increasing and energy decreasing, based on whether the values increase or decrease at TSS (Supplementary Material). For each motif, values of these four vectors were subjected to PCA analysis and top four PCAs were selected. Equations for calculating the LOD score for each motif are available in Supplementary Material. The method for conversion to probability and optimization of threshold probability were same as described above. Probability value of 0.5 was found to be satisfactory for optimum sensitivity and specificity and probability results were classified as 0/1.To get the Boolean value for a given combination, condition for any two motifs for first combination while any three motifs for the rest three combinations, was set. Further, the final Boolean value was obtained by using condition of any one of the four combinations.

(c)Combining the above (a) and (b) to obtain the Final algorithm: the results of both the above algorithms were combined using 'AND' operator. To polish the resultant peaks, a sliding window of 5 is taken, and if 3 or more 1's are found in that window, it is considered as 1, else 0. This removes the minor errors of 0's or 1's. Also, a few constant values were taken, to remove some false peaks: minimum peak size: 35, minimum distance between the two peaks 20. On running the software, two output files are created, one with

'_tss.csv' extension, and the other with '_pos.csv'. The '_tss.csv' contains positions of each peak found in each sequence, and the '_pos.csv' file contains the proper display of peaks using 0/1's.The complete algorithm is depicted as a flowchart in Figure 1 and the tool so developed is named as *SEProm*.

2.6 Promoter prediction and evaluation

The performance of the algorithm was tested by running on the complete data (12 880). If the peak was observed within −450th to +150th position with respect to TSS, it was assigned as true positive (TP), rest of the positions correspond to true negatives (TN). Though some previous studies have analysed no more than 200 bp around TSS as the TP position (12, 13), we observed in our previous study (34) that signals for many of the selected parameters start around −450th/400th position and continue to +100th/150th position with respect to TSS, and so a wider range was taken as TP region. Measure of performance was made by calculating specificity and sensitivity. These statistical measures are briefly described in Supplementary Material. Area under the precision-recall curves are presented as Supplementary Figure S4.

2.7 Predicting the number and positions of TSSs associated with a promoter

The promoter dataset (12 880) of this study represents primary TSSs at 500th position of each sequence. On manual inspection of this dataset, it was observed that many sequences have a single TSS (at 500th position) on the entire length while on others, primary TSS at 500th position is flanked by many secondary TSSs on both sides. In the latter case, designation of primary or secondary TSS for a particular gene varies with the environmental conditions (primary is the one having the highest expression values under given conditions), as per the global TSSs annotation data reported originally for respective organisms. In other words, there are certain promoters which allow transcription to start from a fixed site no matter what the environmental condition is, while certain promoters let the environmental conditions decide the exact location of TSS. We wanted to know the difference between the signals of these two types of promoters, as it can go a long way in identifying the regulatory nature of that promoter. For this, we divided the total data (12 880 sequences) into two groups: Dataset-1 comprising sequences having single primary TSS at 500th position with no other TSS at the entire sequence length (9346 sequences) and Dataset-II where sequences have primary TSS at 500th position flanked by two or more secondary/alternative TSSs (3534 sequences). *SEProm* was run on these two datasets and observations were made with respect to the number of peaks, peak length and distance of last peak (moving from 5′ to 3′ direction) from the TSS located at 500th position.

2.8 Prediction of promoters specific to different sigma factors

Architecture of bacterial promoters is different for different sigma factors. We wanted to check the performance of SEProm in prediction of different sigma specific promoters. Rationale was—whatever motif architecture (−10/−35 elements or −12/−24 elements) a promoter may have, it will definitely undergo a change in structural and energy state to facilitate binding with sigma factors. And our tool is simply capturing any structural and energy change occurring at promoter compared to nearby sequence. It was anticipated that SEProm would be able to predict promoters despite differential sigma factor specificity, though it will not be able to categorize the predicted promoters.

For this, we obtained *E.coli* sigma specific promoters (Sigma24, Sigma28, Sigma32, Sigma38, Sigma54, and Sigma70) from RegulonDB database. Additionally sigma54 promoters were also obtained from the work of Zhao *et al.* (2010).

2.9 Comparison with the state of the art

A number of promoter prediction tools are available presently—CNNPromoter_b (Umarov and Solovyev, 2017), bTSSfinder (Shahmuradov *et al.*, 2016), Pepper (de Jong *et al.*, 2012), BPROM (Solovyev and Salamov, 2011), BacPP (de Silva *et al.*, 2011), Prompredict (Rangannan and Bansal, 2010), PromoterHunter (Klucar *et al.*, 2010), PATLOC (Mrazek and Xie, 2006), NNPP2 (Burden *et al.*, 2005) and Virtual Footprint (Munch *et al.*, 2005). For comparison, we selected five recent, diverged and best performing promoter prediction tools, viz. CNNPromoter_b, bTSSfinder, PromPredict, Pepper and BPROM.

CNNPromoter_b uses the Convolutional Neural Networks to analyse sequence characteristics of promoters and is trained on promoters of *E.coli* and *Mycoplasma pneumonia* bTSSfinder is the most recent tool developed especially for prediction of promoters in cyanobacteria and *E.coli* and is based on neural network learning using 19–20 features. PromPredict is a web-based tool to identify promoter regions in genomic DNA sequences on the basis of differences in the stability between neighbouring regions. Differences in free energy or stability of neighbouring regions are calculated and compared with the assigned cut-off (obtained from the energy difference between upstream and downstream regions in the vicinity of known TSS), to predict promoters in genomic DNA sequences. Pepper predicts promoters on the basis of position specific probability matrices and Hidden Markov Models of −35 and −10 consensus sequences and various sigma TFBs. BProm uses weight matrices of five conserved sequence motifs, distance between −10 and −35 elements and the ratio of densities of octa-nucleotides overrepresented in known bacterial transcription TFBs relative to their occurrence in the coding regions for discriminating between promoters and non-promoters. Prompredict gives output in two possible forms—a promoter region (length identified) as well as a single TSS position, while the other four tools give output in the form of a single position only.

Comparison was made on two data—first the complete promoter dataset (12 880 sequences) of this study and second, for fairness-to rule out any effect of training on SEProm performance, on the test dataset comprising promoters of Halobacterium salinarum NRC-1—an archaean, and Leptospira interrogans—an eubacteria, which were not used for training during this study. The test dataset was prepared by taking the map positions for primary TSSs from the genome wide TSS mapping data of these two organisms (Koide *et al.*, 2009; Zhukova *et al.*, 2017), followed by extraction of 500 nt from upstream as well as downstream positions with respect to each TSS from the whole genome sequence. All the programmes were run with default settings unless indicated otherwise. bTSSfinder was run for the identification of promoters for all the sigma factors. If the peak was observed within −450th to +150th position with respect to TSS, it was assigned as TP, rest of the positions correspond to TN. To rank the different programmes, we used the *F*-score, precision values and balanced accuracy values (relevant formulas for calculations are given in Supplementary Material).

3 Results

3.1 Structural and energy profiles of promoter and coding sequences

Numeric profiles of nine backbone angles, eight inter-BP, six intra-BP, five BP-axis and three energy parameters for the pooled primary promoter (12 880) and CDSs (6218) are presented as Figure 2. For all the parameters, the promoter sequences show unique intrinsic value at TSS and nearby regions resulting in a sharp/broad peak/cleft at/near TSSs. The profiles speak about the potential these parameters hold for their significant use in promoter identification in any given sequence. Correlation analysis using 'R' software revealed multiple inter-correlations among these 31 parameters (Fig. 3); correlation values are available in Supplementary Table S3 (as excel file); correlation ranged from strong to moderate to low. There were few parameters which exhibited no correlation with some of the parameters (Fig. 3). Applying different statistical techniques to these 31 parameters (discussed above in methodology), a promoter prediction tool, SEProm, was then developed.

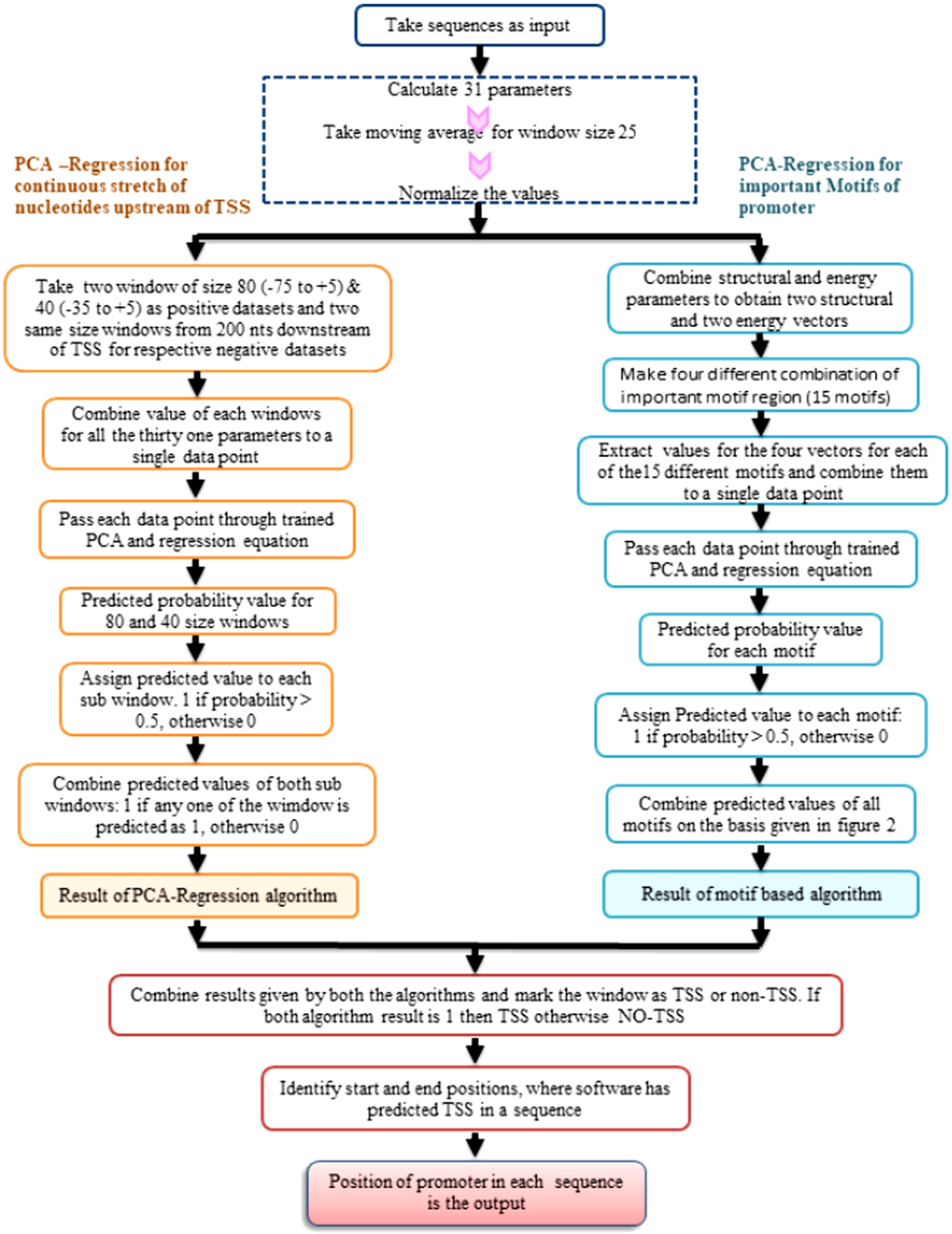
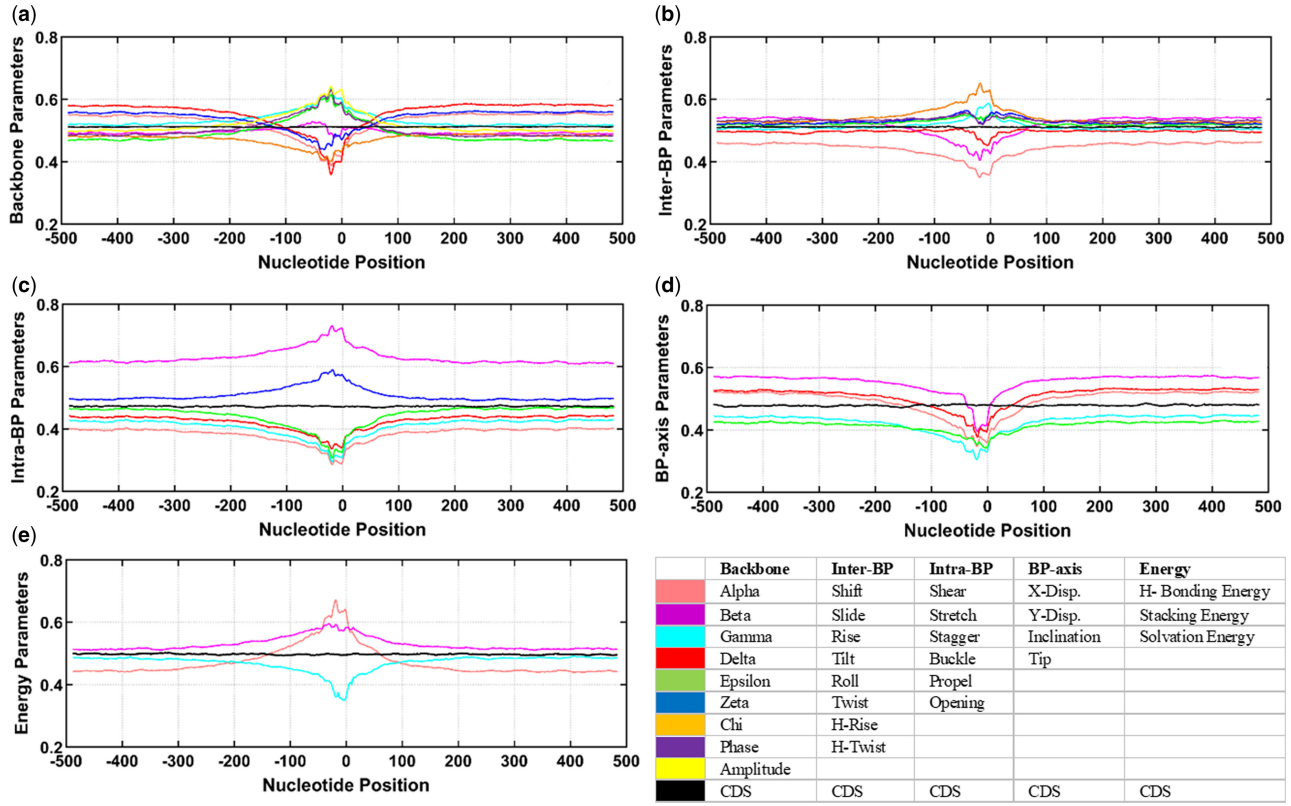


Fig. 1. Flowchart of the algorithm implemented in the SEProm programme



	Backbone	Inter-BP	Intra-BP	BP-axis	Energy
	Alpha	Shift	Shear	X-Disp.	H- Bonding Energy
	Beta	Slide	Stretch	Y-Disp.	Stacking Energy
	Gamma	Rise	Stagger	Inclination	Solvation Energy
	Delta	Tilt	Buckle	Tip	
	Epsilon	Roll	Propel		
	Zeta	Twist	Opening		
	Chi	H-Rise			
	Phase	H-Twist			
	Amplitude				
	CDS	CDS	CDS	CDS	CDS

Fig. 2. Normalized values of 31 structural and energy parameters of 1001-nt long sequences having primary promoters (coloured lines) as well as sequences with no promoters i.e. CDSs (black line). Parameters are shown in five graph based on their respective group. (a) Nine parameters of backbone, (b) eight inter-BP parameters, (c) six intra-BP parameters, (d) five BP-axis parameters and (e) three energy parameters. In each graph, each colour line plot represents specific parameter based on legend table, while line plot in black colour represents average of parameters shown in that group for coding sequence. (Color version of this figure is available at *Bioinformatics* online.)

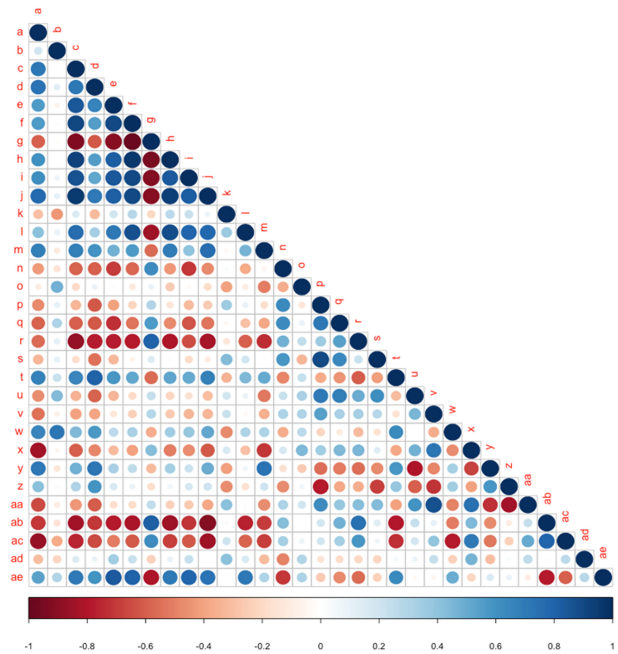


Fig. 3. Graphical representation of correlation values among 31 parameters (parameters names for codes used above is available in Supplementary Table S4)

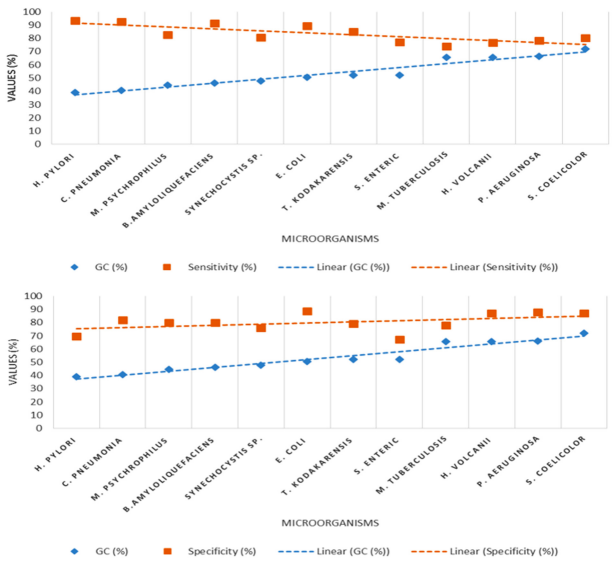


Fig. 4. Scatter plot, with best fit line, for correlation coefficient values between the %GC content of genome and sensitivity (the upper one) and specificity (the lower one)

3.2 SEProm: the tool

SEProm has been written and compiled in Node.js. Node.js is an open-source, cross-platform JavaScript run-time environment that executes JavaScript code outside of a browser and is supported by almost all the operating system including Linux, macOS, Microsoft Windows, SmartOS, FreeBSD and IBM AIX. User can freely download the Node.js source code or a pre-built installer for his/her platform from Node.js home site. After installing Node.js, the SEProm software can be downloaded and extracted in a directory.

Table 1. Characteristic features of peaks obtained with *SEProm* on two promoter datasets

	Characteristics features of sequences	Number of peaks	Percentage of sequences in total prediction (%)	Range of peak length (in bp)	Range of distance between TSS and last peak[a] (in bp)
Dataset I	Primary TSS at 500th position with no other TSS at entire length	Single peak	37	40–618	+55 to −208
		Multiple peaks (2–5)	63	41–559	+35 to −300
Dataset II	Primary TSS at 500th position, flanked by many (up to 10) secondary TSSs	Single peak	34	41–615	+17 to −150
		Multiple peaks (2–5)	66	41–545	+45 to −350

[a]Last nucleotide of peak while moving from 5′ to 3′.
Positions downstream to TSS are designated '+', while towards upstream are designated as '–'.

Table 2. Performance of SEProm in prediction *E.coli* promoters with different sigma specificity

Sigma factor	No. of sequences	TP	FP	TN	FN	Sensitivity	Specificity	Precision	*F*-score
24	537	447	122	415	90	83.240	77.281	78.558	80.831
28	156	131	35	121	25	83.974	77.564	78.915	81.366
32	320	261	53	267	59	81.562	83.437	83.121	82.334
38	243	208	42	201	35	85.596	82.716	83.200	84.384
54	139	116	22	117	23	83.453	84.172	**84.057**	83.754
70	1982	1718	348	1634	264	86.680	82.441	83.155	**84.881**
—	—	—	—	—	Average	84.084	81.268	81.834	82.925

Note: The highest values of Precision and F-score are highlighted as bold values.

SEProm can simply be called from Node.js command prompt by providing the first argument as sequence file (on path of sequence file) and second argument is path of output directory. As output user will get two files, first presenting the position of promoters predicted and second as promoter predicted with sliding window of 100 on each provided sequence; '0' means no promoter while '1' denotes promoter predicted. Time taken by the programme depends on DNA sequence and the speed of the system on which it is run. It takes usually 40 s for 100 sequences each of 1000 nt length. The software can be downloaded from the following link—www.scfbio-iitd.res.in/chemgenome/TSS_Predict.jsp. README.txt file along with input and output files are provided at the link—www.scfbio-iitd.res.in/software/TSS_Predict.jsp.

3.3 Evaluation of SEProm performance

We tested SEProm on the complete promoter dataset of each of the 12 organisms (12 880 sequences) by measuring sensitivity and specificity for prediction (Fig. 4) (Supplementary Table S5). Significantly good performance was observed for all the organisms with an average sensitivity of 83.38% and specificity of 80.19%; highest sensitivity (93.43%) was achieved for promoters of *H.pylori* though with a compromised specificity (69.57%) while promoters of *E.coli* were predicted with outstanding levels of both sensitivity (89.39%) as well as specificity (88.68%). The performance of the software was observed to be independent of phylogeny of the organism (Supplementary Table S5). However, moderate correlation was observed between performance and %GC content. Sensitivity and specificity exhibited correlation to %GC content with coefficient values of −0.763 and 0.513, respectively (Fig. 4).

Recent studies have reported the presence of multiple/alternative TSSs for a single gene, at the BP level. Growing evidence has emerged in support of a dynamic usage of alternative TSSs for the fine regulation of gene expression in prokaryotes (Li *et al.*, 2015). The change of TSS usage is guided by change in developmental or environmental conditions (Zhang *et al.*, 2017), indicating thereby that the promoter region with many alternative TSSs has the potential to interact with a wide range of regulatory proteins/ligands. On the other hand, promoters of genes with single TSS need to interact with basic/default regulatory proteins/ligands only. Since the structural and energetic state of any given DNA region is fundamental to its interaction with proteins/ligands (Rohs *et al.*, 2009), we hypothesized that different categories of promoters (based on the nature of regulation as reflected by the number of TSSs) may give different structural and energetic signals. So now the big question in front of us was—can our tool capture these differential signals? If yes, then can these signals guide us to crack the number and positions of all alternative TSSs associated with a promoter?

To find answer to these questions, we divided the total promoter dataset in two categories—Dataset-1 comprising sequences having single primary TSS at 500th (9346 sequences) and Dataset-II where sequences have primary TSS at 500th position flanked by many alternative/secondary TSSs (3534 sequences), as discussed above in the methodology section. It was expected that sequences of each dataset should give nearly uniform signals, characteristic for respective datasets. When SEProm was run on sequences of these two datasets, results were not as anticipated. Nearly one third of the predicted sequences in both the datasets yielded single peak while for rest of the sequences, multiple peaks were observed (Table 1). No correlation seems to occur between number of peaks and the number of TSSs as otherwise sequences of dataset I (each having a single TSS) would have given single peak and sequences of dataset II (all having many TSSs in the vicinity of a primary TSS) would have shown some uniform peak pattern with respect to the number of TSSs. Occurrence of multiple peaks in dataset I seem possible—there might be certain sequences which have alternative TSSs yet to be identified with right experimental conditions. But observation of single peaks in the sequences of dataset II could not be explained with the existing understanding. So, it does not seem possible to know whether predicted promoter is having a single TSS or multiple/alternative TSSs on the basis of profiles obtained.

One consistent observation was made with respect to minimum peak length, it was always 40/41 BPs irrespective of the category it belongs to. This observation resonates with the presence of a minimal core promoter region, nearly 35–40 bp long, identified for prokaryotic promoters by many groups. When a promoter (from any of the two datasets) yielded single peak, maximum peak length observed was nearly 615. On the other hand, when multiple peaks

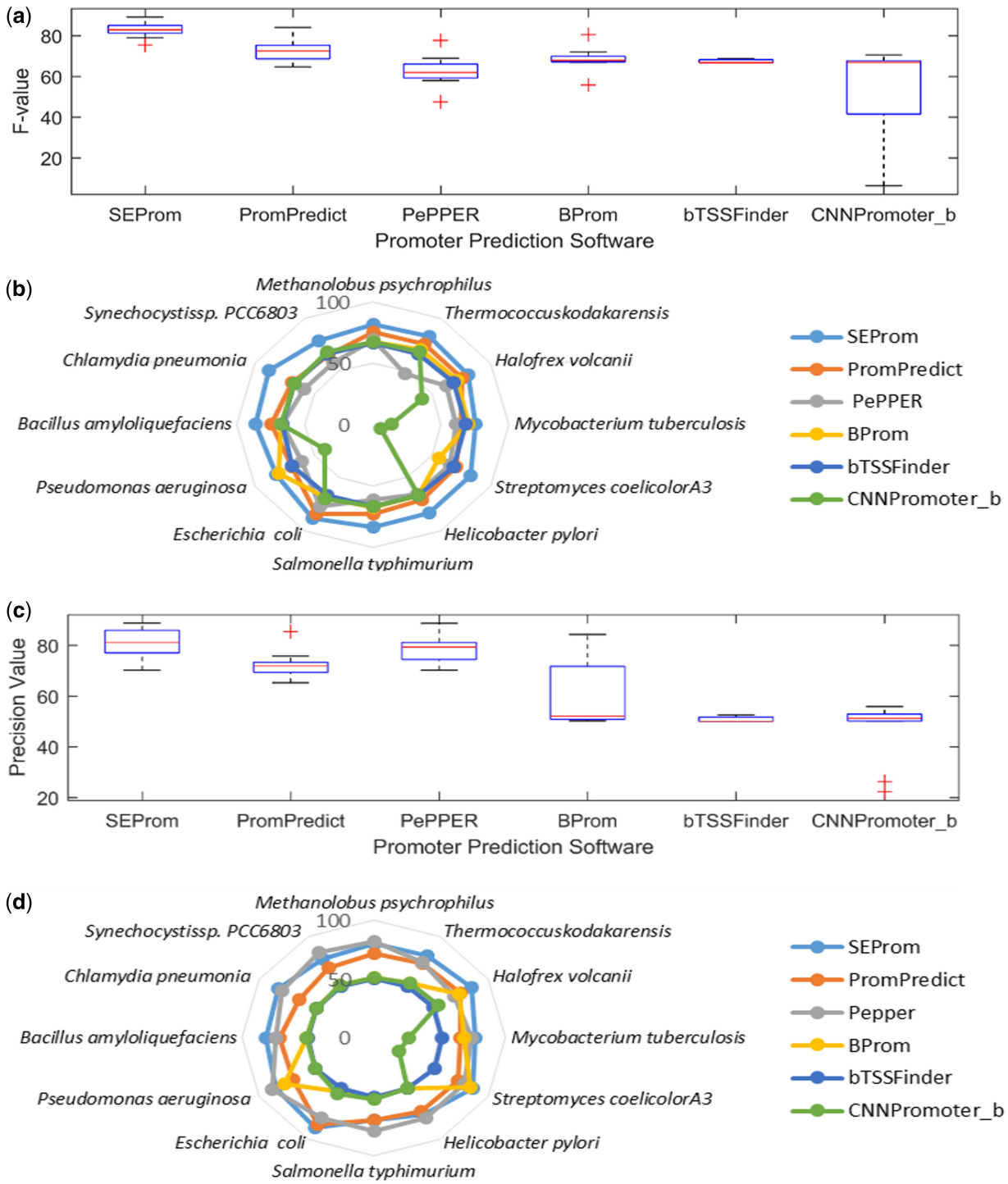


Fig. 5. Comparison of SEProm with 5 other tools in prediction of promoters for 12 organisms used in the study; (a) average *F*-value, (b) *F*-value for individual organisms, (c) average precision value and (d) precision value for individual organisms

were observed for a promoter (from any of the two datasets), longest peak length observed was around 550. Regarding location of the peaks with respect to TSS, large variations were observed in the distance between TSS and the last nucleotide of last peak while moving from 5′ to 3′ direction. Tracing the exact position of TSS seemed really very tricky. Our viewpoint is manifested more loudly in the next section where a comparison of our tool is made with some other tools which predict exact positions of TSSs.

3.4 Prediction of promoters with different sigma specificity

As clear from Table 2, SEProm could predict *E.coli* promoters with different sigma specificity, consistently good precision and *F*-score. Maximum precision was observed for Sigma54 promoters, while highest *F*-score was obtained for Sigma70 promoters. Since SEProm is capturing the structural and energy changes only and is not designed to differentiate among different types of such signals,

Table 3. Comparison of SEProm with five popular promoter prediction programmes on two test datasets (promoters of *Halobacterium salinarum* and *Leptospira interrogans*)

Name of organism	Tool	TP	FP	TN	FN	Sensitivity	Specificity	*F*-score	Precision	Balanced accuracy
Halobacterium salinarum (an archaean) (1355 sequences)	*SEProm*	838	162	1193	517	**61.84**	**88.04**	**71.16**	**83.8**	**74.94**
	PromPredict	842	342	1013	513	62.14	74.76	66.32	71.11	68.45
	BProm	815	238	1117	540	60.14	82.43	67.69	77.39	71.28
	bTSSfinder	1340	1213	142	15	98.89	10.47	68.57	52.48	54.68
	CNNPromoter_b	1335	266	334	1021	1089	19.63	75.33	44.33	27.21
Laptospira interrogans (an eubacterium) (2866 sequences)	*SEProm*	2599	1204	1662	267	**90.68**	**57.96**	**77.93**	**68.32**	**74.32**
	PromPredict	1860	1116	1750	1006	64.89	61.06	63.67	62.5	62.97
	BProm	2865	2851	15	1	99.96	0.523	66.76	50.12	50.24
	bTSSfinder	2865	2865	1	1	99.96	0.034	66.65	50	49.99
	CNNPromoter_b	2866	2864	2860	6	2	99.93	0.20	50.03	66.68

Pepper tool could not be used as its server was not responding. The highest values are highlighted as bold values.

categorization of different sigma specific promoters could not be made out on the basis of observations made.

3.5 Comparison of SEProm with other TSS prediction tools

Out of the five tools selected for comparison, four tools viz. Pepper, BProm, bTSSfinder and CNN give output in the form of a single position (called as predicted TSS-TSSpr). For the fifth tool (Prompredict), we opted for getting output in the form of promoter region rather than a single TSS (since both options were available), as our tool (SEProm) also gives output in the form of a promoter region. Performances were compared by calculating the *F*-values and precision values of all the six tools (Fig. 5, Supplementary Table S6).

As clear from Figure 5, SEProm outperforms all the five tools in promoter prediction of the 12 organisms studied. The average *F*-value and precision value of *SEProm* are 82.04 and 81.08, respectively, which are significantly higher than the respective values with other five tools. Further, SEProm's performance is almost uniform for all the organisms proving its universal applicability. Pepper showed an overall better precision, second to SEProm while PromPredict stood second for F-value. bTSSfinder, trained and developed especially for *E.coli* and *cyanobacteri*a, gave poor results even for these two; *F*-value of 66.72 and 66.60 is obtained for *E.coli* and *Synechocystis* species, respectively, with corresponding precision values of 50.08 and 50. On the other hand, *SEProm* predicted the promoters of *E.coli* and *Synechocystis* sp. with *F*-value of 89.07 and 78.88 and precision value of 88.76 and 77.10, respectively. The results clearly indicate that structure and energy, when taken together (SEProm), has more potential to precisely capture the promoter signals as compared to the energy alone (PromPredict) or the direct sequence information interpreted by any approach (Pepper, BProm, bTSSfinder and CNNPromoterb). Further, the not-so-good performance of the four tools viz. Pepper, BProm, bTSSfinder and CNNPromoterb (each giving a single position as output), despite a large region (−450 to +150) as TP, indicates that TSS prediction is less reliable than the promoter prediction. This resonates with the observation we made in the previous section.

For fairness, we also assessed all tools on two test datasets—promoters of Halobacterium salinarum NRC-1 and Leptospira interrogans; these organisms were not used for training purpose during this study (Table 3). It is clear from the table that SEProm maintains the lead. Our tool performs with a good balance of sensitivity and specificity and works considerably well for these two test organisms as well. The results emphasize the point that universal signatures for promoters can be truly captured by its structural and energy state. The conversion values for structural and energy parameters used in this study were calculated for unique di-nucleotides steps. We anticipate that the sensitivity and specificity of capturing the signals can be enhanced considerably by taking conversion values for tetranucleotide steps, since it takes into consideration the nearest neighbour effects more effectively.

4 Concluding remarks

Promoter prediction in prokaryotes is still a difficult task in genome analysis. Available tools are unable to handle the challenge satisfactorily probably because of poorly understood nature of promoters. We first found that prokaryotic promoters are characterized with a universally similar structural and energetic state, quite distinct from the coding sequences. Using 28 structural and 3 energy parameters, we developed an efficient promoter prediction tool which was named SEProm. We believe that SEProm is the first tool that can recognize promoters universally across diverse prokaryotes with high sensitivity as well as equally high specificity.

Acknowledgements

Funding from the Department of Biotechnology, Govt. of India is gratefully acknowledged. A.M. is a UGC Fellow. P.S. acknowledges Chaudhary Devi lal University Sirsa for granting her sabbatical leave.

Funding

This work was supported by the Department of Biotechnology, Ministry of Science and Technology, Government of India.

Conflict of Interest: none declared.

References

Abeel,T. *et al.* (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.

Albrecht,M. *et al.* (2011) The transcriptional landscape of *Chlamydia pneumoniae*. *Genome Biol.*, **12**, R98.

Babski,J. *et al.* (2016) Genome-wide identification of transcriptional start sites in the halo archaeon *Haloferax volcanii* based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*, **17**, 629.

Beveridge,D.L. *et al.* (2004) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides-I. Research design and results on d(CpG) steps. *Biophys. J.*, **87**, 3799–3813.

Beveridge,D.L. *et al.* (2012) The ABCs of molecular dynamics simulations on B-DNA, circa 2012. *J. Biosci.*, **37**, 379–397.

Burden,S. *et al.* (2005) Improving promoter prediction Improving promoter prediction for the NNPP2.2 algorithm: a case study using *Escherichia coli* DNA sequences. *Bioinformatics*, **21**, 601–607.

Chiu,T.P. *et al.* (2015) GBshape: a genome browser database for DNA shape annotations. *Nucleic Acids Res.*, **43**, D103–D109.

Cortes,T. *et al.* (2013) Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in *Mycobacterium tuberculosis*. *Cell Rep.*, **5**, 1121–1131.

de Silva S. *et al*. (2011) BacPP: bacterial promoter prediction—a tool for accurate sigma-factor specific assignment in enterobacteria. *J. Theor. Biol*., **287**, 92–99.

de Jong,A. *et al*. (2012) PePPER: a webserver for prediction of prokaryote promoter elements and regulons. *BMC Genomics*, **13**, 299.

Dekhtyar,M. *et al*. (2008) Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. *BMC Bioinformatics*, **9**, 233.

Dixit,S.B. *et al*. (2005) Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. II: sequence context effects on the dynamical structures of the 10 unique dinucleotide steps. *Biophys. J*., **89**, 3721–3740.

Dutta,S. *et al*. (2006) A physicochemical model for analyzing DNA sequences. *J. Chem. Inf. Model*., **37**, 78–85.

Florquin,K. *et al*. (2005) Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res*., **33**, 4255–4264.

Goñi,J. *et al*. (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol*., **8**, R263.

Hassan,M.E. and Calladine,C. (1995) The assessment of the geometry of dinucleotide steps in double-helical DNA: a new local calculation scheme. *J. Mol. Biol*., **251**, 648–664.

Haugen,S.P. *et al*. (2008) Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat. Rev. Microbiol*., **6**, 507–519.

Hershberg,R. (2001) PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res*., **29**, 277–270.

Jacques,P. *et al*. (2006) Detection of prokaryotic promoters from the genomic distribution of hexa-nucleotide pairs. *BMC Bioinformatics*, **7**, 423.

Jäger,D. *et al*. (2014) Primary transcriptome map of the hyperthermophilic archaeon *Thermococcus kodakarensis*. *BMC Genomics*, **15**, 684.

Jeong,Y. *et al*. (2016) The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). *Nat. Commun*., **7**, 11605.

Khandelwal,G. and Bhyravabhotla,J. (2010) A phenomenological model for predicting melting temperatures of DNA sequences. *PLoS One*, **5**, e12433.

Khandelwal,G. and Jayaram,B. (2012) DNA–water interactions distinguish messenger RNA genes from transfer RNA genes. *J. Am. Chem. Soc*., **134**, 8814–8816.

Khandelwal,G. *et al*. (2012) DNA-energetics-based analyses suggest additional genes in prokaryotes. *J. Biosci*., **37**, 433–444.

Khandelwal,G. *et al*. (2014) A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes. *Biophys. J*., **106**, 2465–2473.

Klucar,L. *et al*. (2010) phiSITE: database of gene regulation in bacteriophages. *Nucleic Acids Res*., **38**, D366–D370.

Koide,T. *et al*. (2009) Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol*., **5**, 285.

Kopf,M. *et al*. (2014) Comparative analysis of the primary transcriptome of Synechocystis sp. PCC 6803. *DNA Res*., **21**, 527–539.

Kröger,C. *et al*. (2012) The transcriptional landscape and small RNAs of *Salmonella enteric* serovar *Typhimurium*. *Proc. Natl. Acad. Sci. USA*, **109**, E1277–E1286.

Lai,H.-Y. *et al*. (2019) iProEO: a computational predictor for predicting promoter. *Mol. Ther. Nucleic Acids*, **17**, 337–346.

Lavery,R. *et al*. (2009) Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res*., **37**, 5917–5929.

Lavery,R. *et al*. (2010) A systematic molecular dynamics study of nearest-neighbor effects on base pair and base pair step conformations and fluctuations in B-DNA. *Nucleic Acids Res*., **38**, 299–313.

Levo,M. *et al*. (2015) Unraveling determinants of transcription factor binding outside the core binding site. *Genome Res*., 185033–185114.

Li,J. *et al*. (2015) Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon *Methanolobus psychrophilus*. *Sci. Rep*., **5**, 9202.

Liao,Y. *et al*. (2015) The global transcriptional landscape of *Bacillus amyloliquefaciens* XH7 and high-throughput screening of strong promoters based on RNA-seq data. *Gene*, **571**, 252–262.

Mishra,A. *et al*. (2018) Towards a universal structural and energetic model for prokaryotic promoters. *Biophys. J*., **115**, 1180–1189.

Mrazek,J. and Xie,S. (2006) Pattern locator: a new tool for finding local sequence patterns in genomic DNA sequences. *Bioinformatics*, **22**, 3099–3100.

Munch,R. *et al*. (2005) Virtual footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. *Bioinformatics*, **21**, 4187–4189.

Olson,W.K. *et al*. (1998) DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA*, **95**, 11163–11168.

Pasi,M. *et al*. (2014) μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res*., **42**, 12272–12283.

Pribnow,D. (1975) Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA*, **72**, 784–788.

Rangannan,V. and Bansal,M. (2010) High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics*, **26**, 3043–3050.

Rohs,R. *et al*. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.

Shahmuradov,I. *et al*. (2016) bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and *Escherichia coli*. *Bioinformatics*, **33**, 334–340.

Sharma,C.M. *et al*. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, **464**, 250–255.

Singh,A. *et al*. (2017) Physico-chemical fingerprinting of RNA genes. *Nucleic Acids Res*., **45**, e47.

Singhal,P. *et al*. (2008) Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J*., **94**, 4173–4183.

Solovyev,V. and Salamov,A. (2011) *Metagenomics and Its Applications in Agriculture, Biomedicine, and Environmental Studies*. Nova Science Publisher's, Hauppauge, NY.

Umarov,R. and Solovyev,V. (2017) Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS One*, **12**, e0171410.

Umesh,P. *et al*. (2014) A novel sequence and context based method for promoter recognition. *Bioinformation*, **10**, 175–179.

Wade,J. and Grainger,D. (2014) Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol*., **12**, 647–653.

Wang,H. and Benham,C. (2006) Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. *BMC Bioinformatics*, **7**, 248.

Wurtzel,O. *et al*. (2012) The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog*., **8**, e1002945.

Yanagi,K. *et al*. (1991) Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol*., **217**, 201–214.

Zhang,P. *et al*. (2017) Relatively frequent switching of transcription start sites during cerebellar development. *BMC Genomics*, **18**, 461.

Zhao,K. *et al*. (2010) Promoter and regulon analysis of nitrogen assimilation factor, σ^{54}, reveal alternative strategy for *E.coli* MG1655 flagellar biosynthesis. *Nucleic Acids Res*., **38**, 1273–1283.

Zhou,T. *et al*. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res*., **41**, W56–W62.

Zhou,T. *et al*. (2015) Quantitative modeling of transcription factor binding specificities using DNA shape. *Proc. Natl. Acad. Sci. USA*, **112**, 4654–4659.

Zhukova,A. *et al*. (2017) Genome-wide transcriptional start site mapping and sRNA identification in the pathogen *Leptospira interrogans*. *Front. Cell. Infect. Microbiol*., **7**, 10.