BIOINFORMATICS ORIGINAL PAPER
Vol. 26 no. 24 2010, pages 3043–3050
doi:10.1093/bioinformatics/btq577

Sequence analysis
Advance Access publication October 17, 2010

High-quality annotation of promoter regions for 913 bacterial genomes

Vetriselvi Rangannan and Manju Bansal*
Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Associate Editor: Martin Bishop

ABSTRACT

Motivation: The number of bacterial genomes being sequenced is increasing very rapidly and hence, it is crucial to have procedures for rapid and reliable annotation of their functional elements such as promoter regions, which control the expression of each gene or each transcription unit of the genome. The present work addresses this requirement and presents a generic method applicable across organisms.

Results: Relative stability of the DNA double helical sequences has been used to discriminate promoter regions from non-promoter regions. Based on the difference in stability between neighboring regions, an algorithm has been implemented to predict promoter regions on a large scale over 913 microbial genome sequences. The average free energy values for the promoter regions as well as their downstream regions are found to differ, depending on their GC content. Threshold values to identify promoter regions have been derived using sequences flanking a subset of translation start sites from all microbial genomes and then used to predict promoters over the complete genome sequences. An average recall value of 72% (which indicates the percentage of protein and RNA coding genes with predicted promoter regions assigned to them) and precision of 56% is achieved over the 913 microbial genome dataset.

Availability: The binary executable for '*PromPredict*' algorithm (implemented in PERL and supported on Linux and MS Windows) and the predicted promoter data for all 913 microbial genomes are available at http://nucleix.mbu.iisc.ernet.in/prombase/.

Contact: mb@mbu.iisc.ernet.in

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

Received on June 29, 2010; revised on September 23, 2010; accepted on October 3, 2010

1 INTRODUCTION

Understanding the mechanism that regulates gene expression and identifying the key regulatory elements that aid gene expression is a major challenge in molecular biology. Recent advances in the high-throughput genome sequencing has led to an exponential growth in the number of completely sequenced microbial genomes and hence it is essential to have fast, efficient and reliable computational methods to identify, annotate and tabulate the coding regions and non-coding functional elements such as transcription factor binding sites (TFBSs), promoter and enhancer regions in these microbial genomes. The identification of the location and function of promoter regions is very challenging because the promoter regions in genomic sequences very often do not adhere to specific sequence patterns or motifs and are difficult to determine experimentally. Several motif finding algorithms have been developed based on various motif models and their performance has been assessed (Das and Dai, 2007). Whole-genome expression profiles have also led to characterization of bacterial and archaeal transcriptomes (Passalacqua *et al.*, 2009; Wurtzel *et al.*, 2010). These data can be used to validate the various promoter and DNA binding site prediction algorithms developed based on sequence motifs (Carlson *et al.*, 2007; Chakravarty *et al.*, 2007; Gordon *et al.*, 2003, 2006; Jacques *et al.*, 2006; Mann *et al.*, 2007; Reese, 2001; Solovyev and Shahmuradov, 2003; Studholme and Dixon, 2003) as well as those using structure-based properties of DNA (Abeel *et al.*, 2008b; Dekhtyar *et al.*, 2008; Du *et al.*, 2008; Gan *et al.*, 2009; Kanhere and Bansal, 2005b; Rawal *et al.*, 2006; Wang and Benham, 2006; Yadav *et al.*, 2008). Most of the methods for the identification of promoter regions are either specific to a particular genome or only aim to compare the properties of promoter sequences with other regions in a general manner. These protocols have not been applied over the entire set of microbial genomes, nor have the predictions been validated on a genomic scale. In particular, most methods are unable to identify promoters for RNA genes.

In our algorithm, we have made use of the earlier reported observation that promoter regions in genome sequences have several sequence-dependent structural properties such as low stability, high curvature and less bendability, as compared with the flanking genomic sequence (Kanhere and Bansal, 2005b). Among these, lower stability which is calculated as the sum of the free energy of constituent dinucleotide steps is found to be the most ubiquitous physicochemical property of promoter regions (Abeel *et al.*, 2008a; Holloway *et al.*, 2007; Kanhere and Bansal, 2005b). A scoring function was defined and an algorithm '*PromPredict*' was developed to predict promoter regions, which used threshold values specific to some select organisms namely *Escherichia coli*, *Bacillus subtilis* and *Mycobacterium tuberculosis* (Kanhere and Bansal, 2005a; Rangannan and Bansal, 2007). Here, we briefly summarize the modifications and improvements made on prediction methodology of '*PromPredict*', in order to generalize it for predicting promoter regions in all microbial genomes and possibly even eukaryotic genomes.

2 METHODS

2.1 Dataset

Of total, 913 microbial genome sequences along with their annotation information were downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/

*To whom correspondence should be addressed.

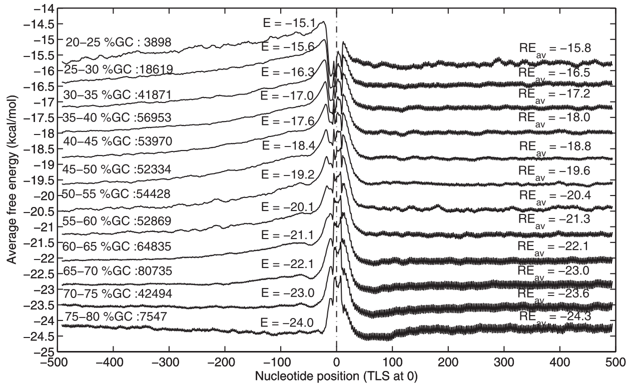


Fig. 1. AFE profiles for genomic sequences of 1001 nt length (spanning −500 to +500 nt w.r.t TLSs) with varying GC content, from TLS-based training dataset. Of total, 5 31 285 TLS (constituting only 5% of total TLS data) that are 500 nt apart and corresponding to protein genes have been combined and categorized based on %GC of flanking 500 nt sequences (refer Section 2). The %GC content and number of sequences in each %GC category is given above the plot. The values of E the AFE over −80 to +20 region and RE$_{av}$ which is the AFE over the +100 to +500 region with respect to TLS are also shown for all %GC classes which have more than 100 sequences.

genomes/Bacteria/) in March 2009 (only full-length chromosome sequences were considered). All together, the 913 microbial genome dataset consists of 26 71 868 genes that code for proteins. Out of these, gene translation start sites (TLSs) that are 500 nt apart, and constitute 5% (5 31 285) of the total dataset, have been combined to create the training dataset. In all, 1001 nt long genomic sequences (spanning −500 to +500 nt with respect each to TLS) have been extracted from the respective genomes for calculating the threshold values from the TLS-based training dataset.

2.2 Threshold calculation

The genomic sequences of 1001 nt length, for the TLS-based training dataset, were retrieved and grouped based on their GC content. The average free energy (AFE) or stability profile was computed for each group (Fig. 1). E, the AFE over −80 to +20 region and RE$_{av}$, the AFE over the downstream region from +100 to +500 with respect to TLS, are found to discriminate promoter regions from coding regions and were hence used to derive the threshold values E and D (where, D = E − RE$_{av}$) (Rangannan and Bansal, 2009). It may be worth mentioning that the AFE value for E and RE$_{av}$ for the genomic fragments from TLS-based training dataset with GC content 30–60%, have AFE values (Fig. 2) very similar to those derived earlier from a smaller TSS-based dataset (i.e. experimentally identified transcription start sites) from *E.coli, B.subtilis* and *M.tuberculosis* (Rangannan and Bansal, 2009). Hence, threshold values have been calculated for genomic DNA sequences with GC content varying over a much larger range (shown in Fig. 2 and hereafter referred to as TSS-TLS-based cutoff values) and applied for annotation of promoter regions in all bacterial genome sequences.

2.3 Scoring function

The stability of double-stranded DNA can be expressed in terms of the free energy of constituent dinucleotides. The AFE of a long continuous stretch of DNA was calculated as described in Kanhere and Bansal (2005a). The energy values corresponding to the 10 unique dinucleotide sequences are taken from the unified parameters obtained from melting studies on 108 oligonucleotides (Allawi and SantaLucia, 1997; SantaLucia, 1998).

The scoring function defined below has been used to calculate the relative stability (DE) between neighboring regions of 100 nt length with respect to every nucleotide position n. The average energy was assigned to the center position corresponding to '$n+50$'.

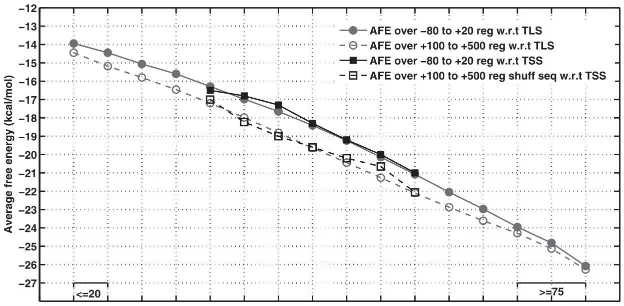
$$\mathrm{DE}_{(n+50)} = E1_{(n+50)} - E2_{(n+50)}$$

where,

$$E1_{(n+50)} = \frac{\sum_{n}^{n+100} \Delta G^0}{100}$$

$$E2_{(n+50)} = \frac{\sum_{n+150}^{n+250} \Delta G^0}{100}$$

Thus $E1_{(n+50)}$ and $E2_{(n+50)}$ represent the free energy averages for 100 nt fragments starting from nucleotides 'n' and '$n+150$', respectively. DE is the difference between $E1$ and $E2$. A stretch of DNA sequence is assigned as a promoter only if its AFE ($E1$) and the difference in free energy (DE) as compared with its neighboring downstream region are greater than the chosen cutoff values (E and D) for the corresponding %GC range, as defined in the TSS-TLS-based cutoff values table (shown in Fig. 2).



%GC	10–15	15–20	20–25	25–30	30–35	35–40	40–45	45–50	50–55	55–60	60–65	65–70	70–75	75–80	80–85	85–90
E (kcal/mol)	−14.5		−15.1	−15.6	−16.3	−16.8	−17.3	−18.3	−19.2	−20.0	−21.0	−22.0	−23.0	−24.0		
D (kcal/mol)	1.0		1.0	1.0	1.0	1.2	1.4	1.4	1.4	1.4	1.2	1.2	1.2	1.2		

Fig. 2. Threshold values of free energy used to predict promoters in genomic DNA with varying GC content. Of total, 1001 nt long genomic sequences spanning 500 nt on either side of 457, 282 and 40 TSSs of protein coding genes in *E.coli, B.subtilis* and *M.tuberculosis* (dataset from Rangannan and Bansal, 2009) and also 500 nt on either side of TLSs of protein coding genes from 913 microbial genomes were categorized based on their %GC content (dataset as in Fig. 1). The AFE value 'E', over the −80 to +20 nt region for the promoter sequences with varying GC composition is shown (as solid lines with filled square and circle markers) for both datasets. 'RE$_{av}$', the AFE values calculated over the +100 to +500 nt regions downstream of the TSSs and TLSs are also plotted (as dashed lines with hallow square and circle markers, respectively). The TSS-TLS-based threshold values assigned to the parameters E and D (the difference between E and RE$_{av}$) for identifying promoter regions within genomic DNA with varying %GC content (details given in Section 2) are tabulated below the plots.

2.4 Method for classification of reliability level of predictions

The average DE value (which is the difference in AFE between neighboring 100 nt long regions) for each predicted promoter region is denoted as PP_DE$_{ave}$. Since no correlation was observed for PP_DE$_{ave}$ based on %GC (Supplementary Fig. S1A), in the current study we have chosen this as an unbiased parameter to define the reliability level for each prediction within whole genome. The average DE values for all predicted regions in a particular genome have also been calculated and denoted as WPP_DE$_{ave}$ (μ). For each predicted promoter region, its PP_DE$_{ave}$ has been compared with WPP_DE$_{ave}$ to assign a reliability level following the criteria described below (which has been illustrated clearly in Supplementary Fig. S1B, for

E.coli predicted promoter regions).

$$\mathrm{PP\_DE}_{\mathrm{ave}} \leq \mu - 1\sigma \rightarrow \text{ Low;}$$

$$\mu - 1\sigma > \mathrm{PP\_DE}_{\mathrm{ave}} \leq \mu \rightarrow \text{ Medium;}$$

$$\mu > \mathrm{PP\_DE}_{\mathrm{ave}} \leq \mu + 1\sigma \rightarrow \text{ High;}$$

$$\mu + 1\sigma > \mathrm{PP\_DE}_{\mathrm{ave}} \leq \mu + 2\sigma \rightarrow \text{ Very high;}$$

$$\mathrm{PP\_DE}_{\mathrm{ave}} > \mu + 2\sigma \rightarrow \text{ Highest;}$$

Low, medium, high, very high and highest are the five prediction reliability levels classified for the predicted promoter regions.

2.5 Evaluating parameters

Recall, Precision and their harmonic mean (F-score) are the parameters used to evaluate the promoter prediction results over the 913 microbial genomes and they are defined as follows:

$$\text{Recall} = \frac{\text{No. of genes with an identified TP}}{\text{Total number of genes}}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision+Recall}}$$

where, true positive (TP) is total number of TP predictions and FP is total number of false positive predictions. The 500 nt sequence upstream of gene translation start site (TLS) is generally considered as the proximal promoter region, where multiple TFBSs are located to trigger the initiation of gene expression (Solovyev and Shahmuradov, 2003). Hence, we assigned a predicted promoter region to be a TP, if it lies within the 500 nt upstream region of a gene (with respect to the TLS). If a predicted promoter region lies within the coding region of a gene which is in the same transcribing strand (i.e. gene and the prediction both have same transcription direction) then it is considered as false positive (FP). The predictions which do not satisfy these criteria are ignored.

3 IMPLEMENTATION AND AVAILABILITY

'*PromPredict*', an algorithm for promoter prediction based on DNA relative stability has been implemented in PERL. Based on the observation and analysis of predicted promoter regions in 913 microbial genomes, a database termed as '*PromBase*' is developed using MySQL and Apache with all major browsers supported. The web interface is managed by a collection of HTML, PERL cgi scripts. The binary executable for '*PromPredict*' algorithm and the predicted promoter data for all 913 microbial genomes are available at http://nucleix.mbu.iisc.ernet.in/prombase/. Apart from the promoter prediction data being displayed (both graphically as genome browser and in tabular form), *PromBase* provides information about several other sequence and structural features for all 913 bacterial genomes analyzed here.

4 RESULTS AND DISCUSSION

4.1 Sequence preference in the vicinity of TSS and TLS

Analysis of percentage occurrence of a tetranucleotide in promoter region (−150 to +50 w.r.t. TSS) as compared with its downstream region (+200 to +400 w.r.t. TSS) from *E.coli*, *B.subtilis* and *M.tuberculosis* genomes has shown that AT-rich tetramers are predominant in *E.coli* and *B.subtilis* but not in high GC containing *M.tuberculosis* (Supplementary Fig. S2). However, even in *E.coli* and *B.subtilis*, several AT-containing tetramers, other than TATA, showed higher percentage occurrence in promoter region. Nucleotide base composition analysis and %GC distribution at different regions w.r.t. TLS for all 913 microbial genomes can be viewed at *PromBase* (http://nucleix.mbu.iisc.ernet.in/prombase), a database displaying various features of promoter regions, in all 913 microbial genomes.

4.2 Correlation between %GC and the average free energy

The GC base composition of bacteria varies extensively between species (16.6–74.9%). Such a wide range of variation has been attributed to mutational bias (Cox and Yanofsky, 1967; Sueoka, 1962) and also correlated with environmental influences on the system (Chen and Zhang, 2003; Foerstner *et al.*, 2005) but its effect on transcriptional efficiency has not been explored. A large number of genomes (∼16%) have GC content within each of the ranges 35–40% and 65–70% for which very little transcription data are available. In addition, regions of a given genome have GC content differing significantly from the whole-genome GC composition. For the present study, we have combined the data from all 913 bacterial genomes to derive cutoff values and hence a wide variation is seen in GC content of the 1001 nt long sequences, extracted with respect to the training set of TLS data. The AFE plots for the 1001 nt long sequences (constituting the TLS-based training dataset), pooled according to their %GC content, are shown in Figure 1. It is interesting to note the shift in the entire free energy profile based on %GC content and the pronounced low stability peak in the vicinity of TLS in all cases.

Variations are seen for %GC and AFE at different regions spanning TLS (−400 to −200, −150 to +50, −80 to +20 and +200 to +400 regions with respect to TLS) within the sequences that belong to the same %GC category (Supplementary Table S1). However, the absolute values of both E and $\mathrm{RE}_{\mathrm{av}}$ become larger, as the GC content increases, indicating their higher stability. This is surprising since it indicates that even in the core and proximal promoter regions the AT content decreases almost proportionately. The difference between E and $\mathrm{RE}_{\mathrm{av}}$ reduces slightly for extreme %GC categories, due to the overall AT richness of even the flanking sequences in very low GC containing genomes and small numbers of AT bases in the core promoter as well as in the downstream coding region in the genomes with very high GC content. With an increase in %GC of the 1001 nt long sequence, the difference in AFE values between upstream intergenic region (−400 to −200 w.r.t. TLS) and downstream coding region (+200 to +400 w.r.t TLS) also reduces linearly. Sequences with low %GC show a broad less stable region upstream of TLS while the sequences with high %GC show a narrow low stability peak with both upstream and downstream regions being almost equally GC rich. This is due to greater occurrence of AT-rich UP elements (specifically containing oligo A-tracts) in the promoter regions of the genomes with low %GC, as seen in Supplementary Figure S2 for *B.subtilis* (with 43.5% GC) as well as in promoter regions of other genomes such as *E.coli* (with 50.8% GC). On the other hand, highly GC-rich genomes lack A-tracts and other AT-rich tetranucleotides in the vicinity of TSS/TLS (as shown in Supplementary Fig. S2 for *M.tuberculosis* which has 65.6% GC content) and have smaller variation of tetranucleotide frequencies in their non-coding regions (Bohlin *et al.*, 2008; Davenport and Tummler, 2010).

The absence of A-tracts in upstream regions has led to suggestions that the genomes with high %GC may use different mechanisms to those mediated by the A-tracts for DNA packing in bacterial nucleoids (Tolstorukov *et al.*, 2005). The narrow low stability peak at TLS in Figure 1 for the fragments with high %GC also suggests that due to the paucity of AT bases in these genomic regions, the few that are present tend to be localized in close proximity of TSS to facilitate duplex opening and even upstream TFBSs in these genomes may be biased toward GC-rich motifs. Hence, the promoter prediction programs, such as NNPP (Reese, 2001) which are trained on AT-rich sequence motifs, fail to predict promoter regions in GC-rich *M.tuberculosis* genome (Supplementary Table S2).

4.3 Refinement of threshold value for TLS-based analysis

In our earlier analysis, we have shown that the promoter regions from *E.coli*, *B.subtilis* and *M.tuberculosis* (with whole-genome GC content of 50.8, 43.5 and 65.6%, respectively) are in general less stable than the flanking regions, but their AFE values vary depending on the GC composition of the genome (Rangannan and Bansal, 2007). Based on this observation, free energy threshold values were derived based on the %GC content of an individual genome to predict promoter regions over whole genome sequences of the above-mentioned three bacterial systems and found to be moderately sensitive in identifying promoter regions in the vicinity of experimentally validated TSSs (Rangannan and Bansal, 2007). We then enhanced the '*PromPredict*' algorithm so that it identifies potential promoter regions based on relative stability of DNA sequences by applying the threshold values derived by considering the GC content of genomic regions in the near vicinity of TSSs (referred to as TSS based cutoff values). Whole-genome annotation using the TSS-based thresholds resulted in %recall of 59% and 49% for *E.coli* and *B.subtilis* genomes, respectively (Rangannan and Bansal, 2009). For the GC-rich *M.tuberculosis* whole-genome annotation 45% recall was achieved. The free energy threshold values used to discriminate promoter sequences from coding sequences were obtained from an analysis of 1001 nt long sequences (spanning −500 to +500 nt region with respect to the experimentally identified TSSs) from the above-mentioned three systems and encompass the %GC range from ∼ 30% to 60% (Rangannan and Bansal, 2009). When these TSS based threshold values were applied to carryout whole-genome annotation of promoter regions over 913 microbial genomes, for the genomes with high GC composition (>65%), the number of predictions with respect to all gene TLSs were considerably less as compared with genomes with lower GC content. Consequently, for the genomes with high GC content, the %recall for promoter prediction was very low in several cases, while the %precision was low in case of genomes with low GC content (Supplementary Fig. S3). Hence, it was necessary to rationalize the cutoff values for the genomic fragments with GC content >60% as well as <35%. However, experimentally validated TSS information is not available for genomes with these extreme values of GC content while annotated TLS data are readily available. Since the relative distance between the experimentally determined TSS and corresponding TLS of closest regulated gene in the operons of *E.coli* and *B.subtilis* shows a maximum at about 20 nt (median distance of −62 and −39 nt, respectively as shown in Supplementary Fig. S4), the TSS and TLS in prokaryotic genomes seem to be in close proximity. Hence, we have re-designed the method for derivation of threshold values using the available TLS-based training data, in order to rationalize the predictions of promoter regions in genomes with very low and very high GC content.

The calculated stability profile for the fragments with varying %GC content from TLS-based training dataset (shown in Fig. 1) have features similar to those seen in case of promoter sequences retrieved with respect to the experimentally validated TSSs from *E.coli, B.subtilis* and *M.tuberculosis* (TSS dataset) (Rangannan and Bansal, 2009). Figure 1 also shows the AFE values for E and RE$_{av}$, calculated over −80 to +20 region and +100 to +500 region (with respect to TLS), respectively. Figure 2 illustrates the variation in E and RE$_{\rm av}$ values based on %GC for both TSS dataset from *E.coli*, *B.subtilis* and *M.tuberculosis* and TLS-based training data. From Figure 2, it is interesting to see that the values for both E and RE$_{av}$ from the promoter sequences retrieved with respect to the known TSSs for above-mentioned three organisms and those retrieved with respect to the TLS-based training dataset nearly merge, over most of the GC range for which TSS data are available. The refined threshold values using the TSS and TLS-based training dataset (given at bottom of Fig. 2 and termed as TSS-TLS-based thresholds) have now been assigned to predict promoter regions over the complete genomic sequences of all 913 prokaryotic genomes, with %GC content varying between 20% and 80%.

4.4 Promoter prediction analysis

4.4.1 Evaluation of prediction results Analysis of prediction results for 913 microbial genomes on applying TSS-TLS-based cutoff values (Fig. 3) gives an average 71.7% ($\sigma = 8.3$) recall value with the %precision being 55.8% ($\sigma = 6.0$) and F-score of 62.2 ($\sigma = 4.2$). When compared to the results from TSS based cutoff values (data shown in Supplementary Fig. S3), overall %recall has improved considerably without significantly affecting the %precision. Interestingly, 70% ($\sigma = 15.6$) recall has also been achieved for RNA genes from 901 microbial genomes (12 out of 913 microbial genomes do not have annotation for RNA genes). The F-score, which is the harmonic mean of recall and precision, is higher by ∼3% and almost constant irrespective of the %GC of the genome. The %recall and %precision obtained for different clades of microbial genomes are shown in Supplementary Figure S5 as box *whisker* plots. Significantly lower % recall is seen for *Thermotogae* phylum (seven genomes with median value = 48.8) while *Bacteriodetes* (16 genomes) have highest average recall (median value = 81.3). *Actinobacteria* and *Spirochaetes* phyla (consisting of 59 genomes with high %GC and 23 genomes with very low %GC, respectively) have slightly lower precision (median value ∼49) as compared to others. But overall there is only a small spread in average precision and F-score values among the various phyla. Though we have considered 500 nt region upstream of gene TLS to assign TP, it is worth mentioning that majority (∼70%) of our TP predictions are within 150 nt upstream of TLS in all 913 microbial genomes (Supplementary Fig. S6). Comparison of the %recall versus %precision plots for annotation of promoter regions in 913 microbial genomes, using TSS based and TSS-TLS-based cutoff values, indicate that overall the current method of predicting promoter regions using the TSS-TLS-based cutoff values works

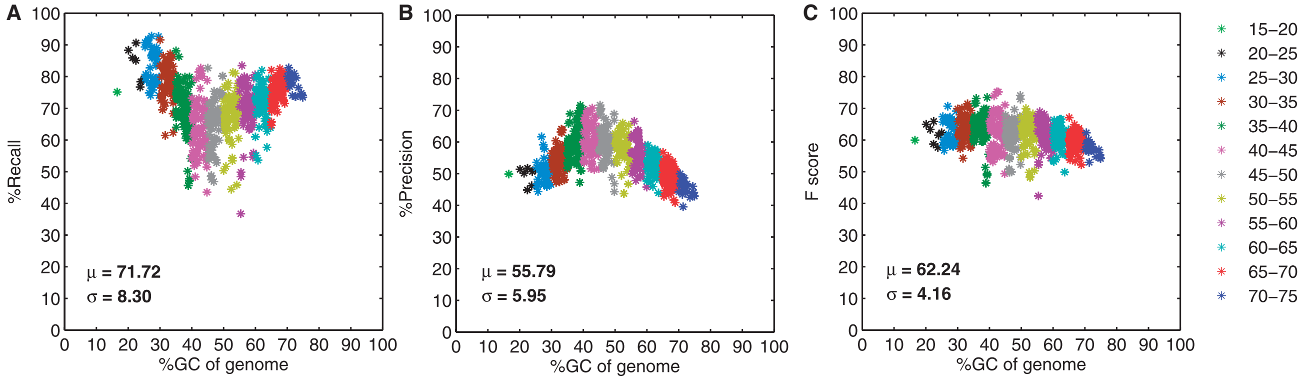


Fig. 3. Distribution of (**A**) %Recall (**B**) %Precision and (**C**) *F*-score obtained for whole-genome annotation of promoter regions in 913 bacterial genomes, using the TSS-TLS-based cutoff values, is shown against the %GC content of the microbial genomes. The color code used for representing genomes falling within each bin, corresponding to a 5% range of GC content, is shown on the right side. The overall mean and SD values for each of the parameters are given inside each plot.

better than the TSS based cutoff values, irrespective of the size of the genome and its GC content (Supplementary Fig. S7).

The characteristic features of TP and FP predictions were analyzed in detail in relation to their reliability level, GC content of the predictions, length of the predictions and for the occurrence of the universal prokaryotic promoter consensus sequence (refer Supplementary Figs S8–S11). In general, a higher percentage of FP promoter signals belong to low reliability category (see Supplementary Fig. S8 and Section 4.4.2 for details), contain higher %GC (Supplementary Fig. S9) and are shorter (Supplementary Fig. S10) as compared with the TP predictions. Also a smaller number of FP predictions contain the consensus promoter sequence motifs as compared with TP predictions (Supplementary Fig. S11). These characteristic features of FP predictions indicate that they probably correspond to short stretches comprising of AT-rich codons within the coding region, or weak signals indicative of internal promoters or antisense RNA promoters (Mendoza-Vargas *et al.*, 2009). Thus, the DNA stability-based promoter prediction method can be fine tuned as per user requirement of high precision/reliability or identification of a weak promoter.

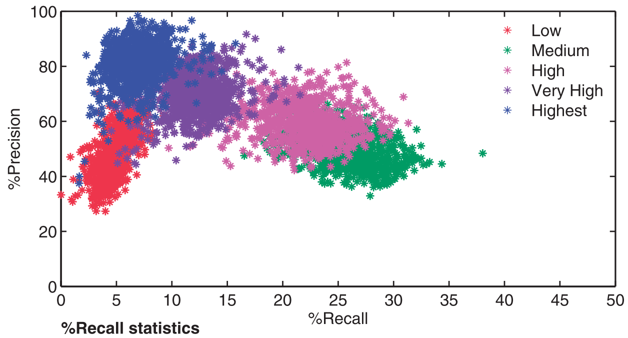
Analysis of 1977 TFBSs for *E.coli* from Regulon DB (version 6.4, Last updated on 10th August 2009) (Gama-Castro *et al.*, 2008) reveals that most of them are AT rich and a majority (~88%) lie within 300 nt region upstream of the translation start site, of the first transcribed gene in the downstream transcription unit (refer Supplementary Fig. S12). About 47% of experimentally determined TFBSs in *E.coli* (42.3% activator sites, 51.2% repressor sites and 52.1% dual regulator sites) are found to overlap with predicted promoter regions identified by our method. All of these overlap with TP predictions and a steady increase is seen in percentage occurrence of TFBS in predictions from low to high reliability classes.

In our earlier analysis, we have already shown that promoter prediction by applying the TSS-based cutoff values to *E.coli* genome sequence outperforms the other methods of promoter prediction based on DNA sequence and structural properties (Rangannan and Bansal, 2009). We have now compared the whole-genome annotation of promoter regions in *E.coli, B.subtilis* and *M.tuberculosis* using the TSS-TLS-based cutoff values, with results from sequence-based promoter prediction programs PPP (http://bioinformatics.biol.rug.nl/websoftware/ppp/ppp_start.php) and NNPP (Reese, 2001) (refer Supplementary Table S2). Though the sequence-based methods (especially NNPP) are able to identify promoter regions for a higher number of genes in case of *E.coli* and *B.subtilis*, they also identify a larger member of false positives (which is reflected in the lower precision values for these genomes, as shown in Supplementary Table S2). They also predict a much smaller number of promoters for the GC-rich genome of *M.tuberculosis,* when compared with our method and hence both have very low recall values for this genome.

A similar anti-correlation is observed between the %GC of the genome and the percentage of protein coding genes containing the full consensus promoter sequence (with −35 and −10 motifs along with the spacer region; refer Supplementary Fig. S13). A mean recall value of 22.6% is obtained over all 913 genomes. Interestingly, even if the genes with −10 motif alone are included, then %recall only improves to 42.3%. These results indicate that in general the consensus sequence derived for σ^{70} promoters rarely occurs in upstream regions of genes in GC-rich genomes, and hence it may be difficult to identify promoters using these motifs, while our method based on DNA relative stability is sensitive to a promoter signal irrespective of the genome GC content (refer for example, to %recall values for genomes belonging *to Actinobacteria, Betaproteobacteria, Deltaproteobacteria* and *Deinococcus/Thermus*, all clades with GC-rich genomes, in Supplementary Figs S14 and S5A). However, as reported earlier (Sinoquet *et al.*, 2008) promoter regions in the AT-rich *firmicute* clade show the highest presence of these motifs (~72%) and this observation is in agreement with our analysis which gives ~74% mean recall value for this clade.

4.4.2 Classification of predictions based on their reliability level

Every predicted promoter region has been classified by us under the category of low, medium, high, very high or highest prediction reliability (refer Section 2). The %recall versus %precision plot for different reliability classes, along with the mean and SD values for each case is shown in Figure 4. The prediction classes with very high and highest reliability (shown in violet and blue color, respectively) together constitute only a small proportion



%Recall statistics

Prediction Reliability	μ	σ	Median	Min	Max
Low	4.8	1.2	4.9	0	9.5
Medium	25.9	2.7	25.8	16.5	38
High	22.1	3.1	22.2	9.7	31.3
Very High	12.1	2.4	12.0	4.6	27.6
Highest	6.8	2.0	6.6	1.6	15.7

%Precision statistics

Prediction Reliability	μ	σ	Median	Min	Max
Low	47.0	8.8	45.7	27.4	80.7
Medium	49.8	6.1	49.6	32.9	66.2
High	59.5	6.7	59.2	42.3	81.4
Very High	69.8	7.4	69.6	44.3	91.7
Highest	80.9	7.4	81.4	37.5	98.4

Fig. 4. Recall versus precision plots for all predicted promoter regions in 913 microbial genomes, after their classification into five groups, based on their reliability level (as described in Section 2). When calculating %recall, if a particular gene has more than one TP with different reliability level, then only TP with highest prediction reliability is considered. The color code used to discriminate the predictions under different reliability level is given in top right corner of the figure. The tables at bottom give the statistics of these parameters for each prediction class.

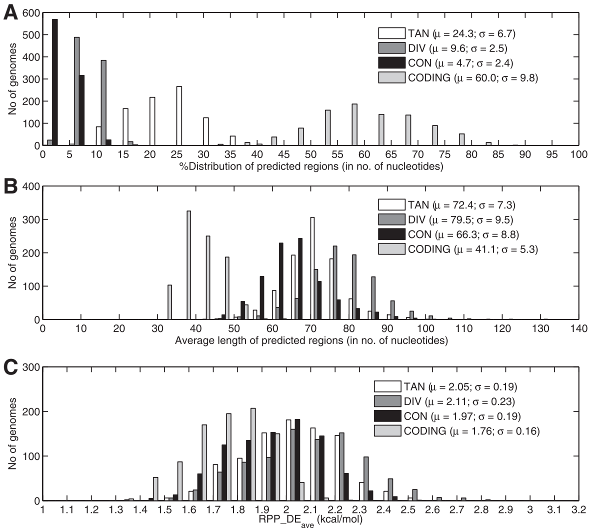


Fig. 5. Analysis of predicted promoters within CODING and various intergenic regions of microbial genomes. (**A**) Predicted promoter distribution in different regions. For this analysis, we have considered all the genes irrespective of their transcription direction to count the predicted promoter occurrence within CODING region and hence the %prediction within CODING is higher than the predictions assigned as false positive (refer Section 2 for false positive definition). (**B**) Predicted promoter region average length distribution in different regions. (**C**) Distribution of RPP_DE$_{\text{ave}}$, the average DE values for all predicted promoters within different genomic regions of microbial genomes. Since similar distribution is observed for predictions in reverse strand, prediction distribution in forward strand alone is shown. The color code used and the mean values for the respective distribution, along with SD values, are also indicated.

(~15% of total predictions) and hence make a modest contribution to the TP gene count, as reflected in their %recall values of 12.1 and 6.8%, respectively. However, their precision values are very high (an average of 70 and 81%, respectively), since these predictions are most often identified as TP and very rarely occur within the coding regions (Supplementary Fig. S8). The predictions with low reliability level (red dots in Fig. 4) are also small in number (~12% of total predictions) but these are often identified as FP and very rarely as TP (Supplementary Fig. S8). Hence both %recall and %precision values for the low reliability prediction class are small, as seen in Figure 4. The predictions classified under medium and high level have moderate recall and precision.

Thus, the predictions falling in the four reliability categories, varying from highest to medium, follow a trend from high precision/low-recall toward low-precision/high-recall, but the lowest reliability group is an exception. As mentioned above, the predicted promoter regions categorized as least reliable are quite small in number and are also very weak in terms of strength, being only marginally less stable than the flanking regions (as defined in Section 2). Hence, the probability of their being good candidate promoters or alternate promoters is expected to be quite low. Excluding these weak potential promoter regions from the analysis brings down the average recall value for all 913 genomes to 67% ($\sigma = 8\%$) but improves the average precision to 57% ($\sigma = 6\%$) as compared with the overall values of 71.7 and 55.8%, respectively (shown in Fig. 3). This indicates that the method adopted to classify the predicted region under different reliability levels works very well. Using this reliability level classification scheme, users can give high priority to the predictions with high reliability level, when looking for a promoter region for a specific gene in a particular organism.

4.4.3 Promoter prediction analysis in different regions of genome

The distribution of all predicted promoter regions within tandem (TAN), divergent (DIV) and convergent (CON) intergenic regions as well as within CODING regions (those labeled as CDS in the features table of GenBank files) was examined in all 913 microbial genomes and is shown in Figure 5. A salient feature of microbial genomes is dense packing of genetic elements. On an average, 86.4% ($\sigma = 5.3$) of DNA is transcribed as protein or RNA in these genomes and the number of overlapping genes is also very high in prokaryotes (Palleja *et al.*, 2008, 2009). Hence, the percentage occurrence of any sequence or structural motif within the CODING region is also high. Here, the predicted promoter distribution within CODING region is considered irrespective of their transcription direction and hence the %prediction within CODING in Figure 5 is higher than the predictions assigned as false positive. Overall, about 40% of predictions occur in the intergenic regions, which in general constitute <15% of the prokaryotic genome. The percentage of predictions occurring within TAN intergenic regions is high (24.3%) compared with the other intergenic regions (as seen in Fig. 5A), which correlates with the larger number of TAN IR in microbial genomes (Molina and van Nimwegen, 2008).

The predicted promoter region length distribution in different regions of microbial genome is shown in Figure 5B while Figure 5C shows the distribution of RPP_DE$_{\text{ave}}$, which is the average DE value for all predicted promoters within different genomic regions (TAN, DIV, CON IR and CODING regions) of microbial genomes. It is worth noting that the predicted promoter regions that occur in CODING region are short in length and exhibit low RPP_DE$_{\mathrm{ave}}$ value as compared with the predictions in intergenic regions, which indicates that predictions within CODING regions are in general weak signals. In some cases, these may correspond to internal promoters (Rangannan and Bansal, 2009).

The predicted promoter regions that overlap with divergent (DIV) intergenic region constitute about 10% of total predictions are longer and possess higher value for RPP_DE$_{\mathrm{ave}}$ than the others. This may be due to the long and AT-rich nature of the DIV intergenic region, which is essential to embed multiple upstream regulatory signals for two genes (Rogozin *et al.*, 2002).

About 5% of the predictions overlap with CON intergenic regions. It has been reported that the rho-independent intrinsic termination signal motif is a GC-rich dyad symmetry element, followed by an oligo(T) sequence (d'Aubenton Carafa *et al.*, 1990; Gusarov and Nudler, 1999). Hence, the CON intergenic region shows a low stability peak flanked by high stability region, giving a false positive signal (Rangannan and Bansal, 2009). The predictions within CON region are overall shorter and weaker as compared with the predictions within other intergenic regions (Fig. 5B and C) but a similar trend was observed for the percentage prediction distribution at different reliability levels in all intergenic regions (Supplementary Fig. S15A, B and C). However, as mentioned earlier (in Section 4.4.1) the predictions in the CODING region occur more frequently in the lower reliability level categories (Supplementary Fig. S15D).

5 CONCLUSION

AFE of DNA sequence is a well-defined property that can be used to distinguish promoter regions in a DNA sequence. The AFE profiles for promoter sequences show a less stable region upstream of the TLSs (which are generally in close proximity of the TSSs) when compared with the flanking genomic sequences. The AFE values in the near vicinity of TSS/TLS as well as in the flanking regions vary depending on the GC composition of the whole region, but the relative stability is maintained. These features have been used to derive threshold values for identifying promoter regions over whole-genome sequences. Promoter prediction using TSS-TLS-based cutoff values yields high %recall (an average of 72%), without significantly affecting the precision (average of 55.8%) when applied to annotation of 913 microbial genomes. The comparison of recall and precision parameters obtained when '*PromPredict*' is applied to *E.coli, B.subtilis* and *M.tuberculosis* with those obtained using other sequence-based methods (NNPP and PPP) indicate that overall our method performs better. The algorithm thus appears to be robust and applicable to all prokaryotic genomes regardless of their base composition. The '*PromPredict*' algorithm and TSS-TLS-based cutoff values can also be applied to other organisms and preliminary studies to predict promoter regions in plants as well as other eukaryotes show very promising results. Further this method can be combined with sequence motif-based methods and used along with structural properties such as curvature and bendability to improve the identification of promoter regions in genomes.

Funding: Microsoft Corporation and Department of Biotechnology, INDIA (in part).

Conflict of Interest: none declared.

REFERENCES

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

Abeel,T. *et al.* (2008b) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics*, **24**, i24–i31.

Allawi,H.T. and SantaLucia,J. Jr (1997) Thermodynamics and NMR of internal G.T mismatches in DNA. *Biochemistry*, **36**, 10581–10594.

Bohlin,J. *et al.* (2008) Investigations of oligonucleotide usage variance within and between prokaryotes. *PLoS Comput. Biol.*, **4**, e1000057.

Carlson,J.M. *et al.* (2007) SCOPE: a web server for practical de novo motif discovery. *Nucleic Acids Res.*, **35**, W259–W264.

Chakravarty,A. *et al.* (2007) A novel ensemble learning method for de novo computational identification of DNA binding sites. *BMC Bioinformatics*, **8**, 249.

Chen,L.L. and Zhang,C.T. (2003) Seven GC-rich microbial genomes adopt similar codon usage patterns regardless of their phylogenetic lineages. *Biochem. Biophys. Res. Commun.*, **306**, 310–317.

Cox,E.C. and Yanofsky,C. (1967) Altered base ratios in the DNA of an Escherichia coli mutator strain. *Proc. Natl Acad. Sci. USA*, **58**, 1895–1902.

d'Aubenton Carafa,Y. *et al.* (1990) Prediction of rho-independent Escherichia coli transcription terminators. A statistical analysis of their RNA stem-loop structures. *J. Mol. Biol.*, **216**, 835–858.

Das,M.K. and Dai,H.K. (2007) A survey of DNA motif finding algorithms. *BMC Bioinformatics*, **8** (Suppl. 7), S21.

Davenport,C.F. and Tummler,B. (2010) Abundant oligonucleotides common to most bacteria. *PLoS One*, **5**, e9841.

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

Du,Z. *et al.* (2008) Genome-wide analysis reveals regulatory role of G4 DNA in gene transcription. *Genome Res.*, **18**, 233–241.

Foerstner,K.U. *et al.* (2005) Environments shape the nucleotide composition of genomes. *EMBO Rep.*, **6**, 1208–1213.

Gama-Castro,S. *et al.* (2008) RegulonDB (version 6.0): gene regulation model of Escherichia coli K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res.*, **36**, D120–D124.

Gan,Y. *et al.* (2009) A pattern-based nearest neighbor search approach for promoter prediction using DNA structural profiles. *Bioinformatics*, **5**, 2006–2012.

Gordon,L. *et al.* (2003) Sequence alignment kernel for recognition of promoter regions. *Bioinformatics*, **19**, 1964–1971.

Gordon,J.J. *et al.* (2006) Improved prediction of bacterial transcription start sites. *Bioinformatics*, **22**, 142–148.

Gusarov,I. and Nudler,E. (1999) The mechanism of intrinsic transcription termination. *Mol. Cell*, **3**, 495–504.

Holloway,D.T. *et al.* (2007) Machine learning for regulatory analysis and transcription factor target prediction in yeast. *Syst. Synth. Biol.*, **1**, 25–46.

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

Kanhere,A. and Bansal,M. (2005a) A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, **6**, 1.

Kanhere,A. and Bansal,M. (2005b) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic Acids Res.*, **33**, 3165–3175.

Mann,S. *et al.* (2007) A pHMM-ANN based discriminative approach to promoter identification in prokaryote genomic contexts. *Nucleic Acids Res.*, **35**, e12.

Mendoza-Vargas,A. *et al.* (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS One*, **4**, e7526.

Molina,N. and van Nimwegen,E. (2008) Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.*, **18**, 148–160.

Palleja,A. *et al.* (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? *BMC Genomics*, **9**, 335.

Palleja,A. *et al.* (2009) PairWise Neighbours database: overlaps and spacers among prokaryote genomes. *BMC Genomics*, **10**, 281.

Passalacqua,K.D. *et al.* (2009) Structure and complexity of a bacterial transcriptome. *J. Bacteriol.*, **191**, 3203–3211.

Rangannan,V. and Bansal,M. (2007) Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability. *J. Biosci.*, **32**, 851–862.

Rangannan,V. and Bansal,M. (2009) Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. *Mol. Biosyst.*, **5**, 1758–1769.

Rawal,P. *et al.* (2006) Genome-wide prediction of G4 DNA as regulatory motifs: role in Escherichia coli global regulation. *Genome Res.*, **16**, 644–655.

Reese,M.G. (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput. Chem.*, **26**, 51–56.

Rogozin,I.B. *et al.* (2002) Congruent evolution of different classes of non-coding DNA in prokaryotic genomes. *Nucleic Acids Res.*, **30**, 4264–4271.

SantaLucia,J. Jr (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc. Natl Acad. Sci. USA*, **95**, 1460–1465.

Sinoquet,C. *et al.* (2008) Large-scale computational and statistical analyses of high transcription potentialities in 32 prokaryotic genomes. *Nucleic Acids Res.*, **36**, 3332–3340.

Solovyev,V.V. and Shahmuradov,I.A. (2003) PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res.*, **31**, 3540–3545.

Studholme,D.J. and Dixon,R. (2003) Domain architectures of sigma54-dependent transcriptional activators. *J. Bacteriol.*, **185**, 1757–1767.

Sueoka,N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proc. Natl Acad. Sci. USA*, **48**, 582–592.

Tolstorukov,M.Y. *et al.* (2005) A-tract clusters may facilitate DNA packaging in bacterial nucleoid. *Nucleic Acids Res.*, **33**, 3907–3918.

Wang,H. and Benham,C.J. (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.* (2010) A single-base resolution map of an archaeal transcriptome. *Genome Res.*, **20**, 133–141.

Yadav,V.K. *et al.* (2008) QuadBase: genome-wide database of G4 DNA–occurrence and conservation in human, chimpanzee, mouse and rat promoters and 146 microbes. *Nucleic Acids Res.*, **36**, D381–D385.