Identification and annotation of promoter regions in microbial genome sequences on the basis of DNA stability

VETRISELVI RANGANNAN and MANJU BANSAL*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

**Corresponding author (Fax, 91-80-2360 0535; Email, mb@mbu.iisc.ernet.in)*

Analysis of various predicted structural properties of promoter regions in prokaryotic as well as eukaryotic genomes had earlier indicated that they have several common features, such as lower stability, higher curvature and less bendability, when compared with their neighboring regions. Based on the difference in stability between neighboring upstream and downstream regions in the vicinity of experimentally determined transcription start sites, a promoter prediction algorithm has been developed to identify prokaryotic promoter sequences in whole genomes. The average free energy (E) over known promoter sequences and the difference (D) between E and the average free energy over the entire genome (G) are used to search for promoters in the genomic sequences. Using these cutoff values to predict promoter regions across entire *Escherichia coli* genome, we achieved a reliability of 70% when the predicted promoters were cross verified against the 960 transcription start sites (TSSs) listed in the Ecocyc database. Annotation of the whole *E. coli* genome for promoter region could be carried out with 49% accuracy. The method is quite general and it can be used to annotate the promoter regions of other prokaryotic genomes.

[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]

1. Introduction

In recent years identification of promoter regions in genome sequences is one of the major challenges in bioinformatics and integrates comparative, structural and functional genomics. It remains important, not only to detect rarely expressed genes but also for regulatory analysis of genes whose full length transcripts are not known, which is currently true for most of the genes. Better understanding of the features associated with promoters will assist in recognizing promoter regions within genomic sequences as well as in identifying genes associated with rRNA, tRNA and other non-coding RNAs. The process of transcription begins with the RNA polymerase (RNAP) binding to the DNA at the promoter region, which is in the vicinity of transcription start site (TSS). Precisely, how RNAP locates this specific binding site in the large excess of DNA remains an active field of investigation. A promoter sequence is thought to comprise of sequence motifs positioned at specific sites relative to TSS (Harley and Reynolds 1987; Bucher 1990). These sequence motifs were identified based on the analysis of a large set of promoter sequences and they represent consensus sequences. The exact consensus sequence motifs are found in only a few promoters. Also these sequence motifs encompass 6–10 nucleotides and are degenerate; hence the probability of finding similar sequences in regions other than promoters is quite high and it is unlikely that these sequence motifs alone are responsible for binding of RNAP to promoter regions.

Many different algorithms for the identification of promoters, transcriptional start sites and transcription factor binding sites in DNA sequence have been proposed (Fickett and Hatzigeorgiou 1997; Pedersen *et al* 1999; Werner 1999; Ohler and Niemann 2001). Most of them

Keywords. DNA stability; free energy calculation; promoter; upstream and downstream region

Abbreviations used: nt, Nucleotides; RNAP, RNA polymerase; TSS, transcription start site are based on discovery of patterns (sequence motifs) in genome sequences. Some of them attempt to discriminate promoter from non-promoter regions in vertebrate DNA sequence based on hexamer frequency analysis (Hutchinson 1996) and to recognize primate promoter sequences by doing comparative density analysis of specific transcription factor binding sites (Prestridge 1995). A recent method (Reese 2001), which combines recognition of TATA box and *Inr* using the time delay neural network architecture, allows for variable spacing between features. The general picture emerging from the above mentioned methods is that, when a promoter prediction algorithm is used under conditions where they find reasonable percentage of known promoters (true positives), then the number of wrongly predicted promoters (false positives) is far too high.

DNA sequence-dependent three-dimensional structure is also important for transcriptional regulation, both at the level of single binding sites and at the level of entire promoter regions. Regulatory sequences such as promoter regions not only contain specific sequence elements that serve as targets for interacting proteins, but also exhibit distinct structural properties which are a reflection of the sequence. In case of bacteria as well as in eukaryotes, various properties, such as curvature of regions upstream of TSS, differ from that of downstream regions (Kanhere and Bansal 2003). Some of these properties were studied independently on genomic sequences (Vollenweider *et al* 1979; Margalit *et al* 1988). A comprehensive analysis of promoter regions in prokaryotic and eukaryotic genomes reveals that they are expected to have several common structural features, such as lower stability, higher curvature and lesser bendability as compared with their neighboring regions (Kanhere and Bansal 2005a). In general, these similarities and differences in the structural features of DNA sequences, particularly differences in the stability of DNA provide much better criteria for identifying promoter region from non-promoter region, than sequences alone (Kanhere and Bansal 2005a, Wang and Benham 2006). Based on the analysis of difference in stability between neighboring regions (upstream and downstream regions), a promoter prediction algorithm had been developed in our laboratory to identify prokaryotic promoter regions and tested for known *E. coli* promoters (Kanhere and Bansal 2005b). The applicability of promoter prediction program has now been enhanced in order to predict the promoter regions across the entire *E. coli* genome as well as on a large data set of *Bacillus subtilis* promoter sequences. We achieve good reliability level, when the free energy differences between the known *E. coli* promoter sequences and entire *E. coli* genome are used as thresholds to search for promoters in *E. coli* genome sequence. Here, we illustrate the procedure and the results on annotation of the whole prokaryotic genomes for promoter regions.

2. Materials and Methods

2.1 *Sequence data*

The *E. coli* and *B. subtilis* promoter sequences were obtained from public domain databases. Whole genome sequences were downloaded from the NCBI site.

2.1.1 Escherichia coli *data set*: Whole genome sequence of *E. coli*, which consists of 4,639,675 nt was downloaded from NCBI (accession No: NC_000913). The transcription start sites for *E. coli* were acquired from Ecocyc database (Version 9.1, updated on 12th May, 2005) (Keseler *et al* 2005). This Ecocyc database provides a compilation of 1044 TSSs and details about 4474 annotated genes of *E. coli*. Using the transcription start site details from Ecocyc, 1001 nt long promoter sequences were extracted from *E. coli* genome sequence database (NCBI accession No: NC_000913). We have analysed the *E. coli* promoter sequences in the vicinity of experimentally identified TSS to check for promoters in close proximity of TSSs.

The data is divided into two sets. The first set contains 611 promoter sequences, which are at least 100 nt apart, that are culled from the 1044 TSSs identified in Ecocyc. These sequences are 101 nt length (spanning from –80 to +20 with respect to TSS). This first set of annotated promoter sequences are used to derive the cutoff values (*see* § 2.4) that are used to search for promoter regions in the whole genome sequence. Out of the 1044 TSS from Ecocyc, only the 615 TSSs with experimental citation were considered to derive the second dataset, of 1001 nt long sequences (covering the region from –500 to +500 with respect to TSS) and with the TSS being at least 500 nt apart. This second set thus comprises of 251 experimentally determined promoter sequences. The cutoff values derived from the first set of sequences are applied, as a test case, to predict promoter regions, in the second set of longer promoter sequences.

2.1.2 Bacillus subtilis *data set*: Whole genome sequence of *B. subtilis*, which consists of 4,214,630 nt was downloaded from the NCBI site (accession No: NC_000964). The TSS for B. subtilis promoters were obtained from DBTBS database (Makita et al 2004) and 879 unique transcription start sites were compiled from DBTBS. The required length of sequences around the TSS were extracted from the *B. subtilis* genome sequence (NCBI accession No: NC_000964). The two data sets of promoter sequences (of length 101 nt and 1001 nt) for B. subtilis were derived in the same manner as in case of *E. coli*. The first set contains 339 promoter sequences of 101 nt length and the second set contains 283 promoter sequences of 1001 nt length. In order to avoid multiple TSS in the 1001 nt long and 101 nt long sequences, TSS that were less then 500 nt apart and 100 nt apart respectively, were excluded.

2.2 *Free energy (stability) calculation*

The stability of a double stranded DNA molecule can be expressed in terms of the free energy of its constituent base paired dinucleotides. The standard free energy change (ΔG^{o}_{37}) corresponding to the melting transition of an ‘n’ nucleotide (or ‘n–1’ dinucleotides) long DNA molecule, from double strand to single strand is calculated as follows (SantaLucia 1998):

$$\Delta G^{o} = -(\Delta G^{o}_{ini} + \Delta G^{o}_{sym}) + \sum_{i=1}^{n-1} \Delta G^{o}_{i,i+1}$$

where,

ΔG^{o}_{ini} is the initiation free energy for dinucleotide of type ij.

ΔG^{o}_{sym} equals +0.43 kcal/mol and is applicable if the duplex is self-complementary.

$\Delta G^{o}_{i,j}$ is the standard free energy change for the dinucleotide of type ij.

The two terms ΔG^{o}_{ini} and ΔG^{o}_{sym}, which are more relevant for oligonucleotides, are not considered since our analysis involves long continuous stretches of DNA molecules. In the present study free energy over this long continuous stretch of DNA sequence was calculated by dividing the sequence into overlapping windows of 15 base pairs (or 14 dinucleotide steps). For each window, the free energy is calculated as given in the above equation and the free energy value is assigned to the central base pair of the window. 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).

2.3 *Details of promoter prediction methodology*

Difference in free energy or stability of neighboring regions are calculated and compared with the assigned cutoff values (obtained from the energy difference between upstream and downstream regions in the vicinity of known TSS, as described in § 2.4) to predict promoters in genomic DNA sequences. A slightly modified version of the scoring function D(n) defined below has been used. It looks for differences in free energy of the neighboring regions (upstream and downstream regions) with respect to every nucleotide position n:

$$\mathrm{D}(n) = \mathrm{E1}(n) - \mathrm{E2}(n)$$

where,

$$\mathrm{E1}(n) = \frac{\sum_{n}^{n+99} \Delta G^{o}}{100}$$

$$\mathrm{E2}(n) = \frac{\sum_{n+100}^{n+199} \Delta G^{o}}{100}$$

$$\mathrm{E1'}(n) = \frac{\sum_{n}^{n+49} \Delta G^{o}}{50}$$

(E1’ is used in place of E1 in the refinement cycle for false negatives).

Thus, E1(n) and E2(n) represent the free energy averages in a 100 nt window starting from nucleotide ‘n’ and the neighboring 100 nt region starting from nucleotide ‘n+100’, respectively. E1′(n) represents the free energy average in the 50 nt region starting from nucleotide ‘n’. E1′ is calculated instead of E1 in the refinement cycle for false negatives obtained in the first iteration. The E1 value represents the average free energy over the upstream regions of 100 nt length. D value represents the free energy difference in the two neighboring regions. A stretch of DNA is assigned as promoter only if the average free energy of that 100 nt region (E1) and the difference (D) in free energy as compared to its neighboring region is greater than the chosen E-cutoff and D-cutoff respectively (*see* § 2.4). The protocol followed to calculate the true positive, false positive and false negative signals, which define the sensitivity and precision of the method, for known promoter sequences is presented in the form of a flowchart in figure 1.

In our earlier analysis, use of a 50 nt window for E1 as well as E2 calculations, resulted in slightly lower precision than when a 100 nt window was used for E2 calculation and hence a 100 nt window was chosen for E2 calculation (Kanhere and Bansal 2005b). In our present study we have followed a two cycle procedure. In the first iteration, D is calculated using equal sized windows, of 100 nt length, for both E1 and E2, since this gives a high precision (*see* § 3.2). However the larger window sizes lead to more false negatives (lower sensitivity), so a second iteration has been carried out on these sequences, using the smaller window size of 50 nt for E1 (referred to as E1′) while retaining a 100 nt window size for E2. Incorporation of this refinement cycle in the present analysis has simultaneously enhanced the sensitivity and precision, as compared to our earlier method, which itself had been shown to work better than other promoter prediction programs (Reese 2001; Staden 1984).

2.4 *Derivation of threshold values and prediction of promoter regions over whole genome*

E-cutoff is the average free energy over the known promoter regions of 101 nt length (ranging from –80 to +20 with

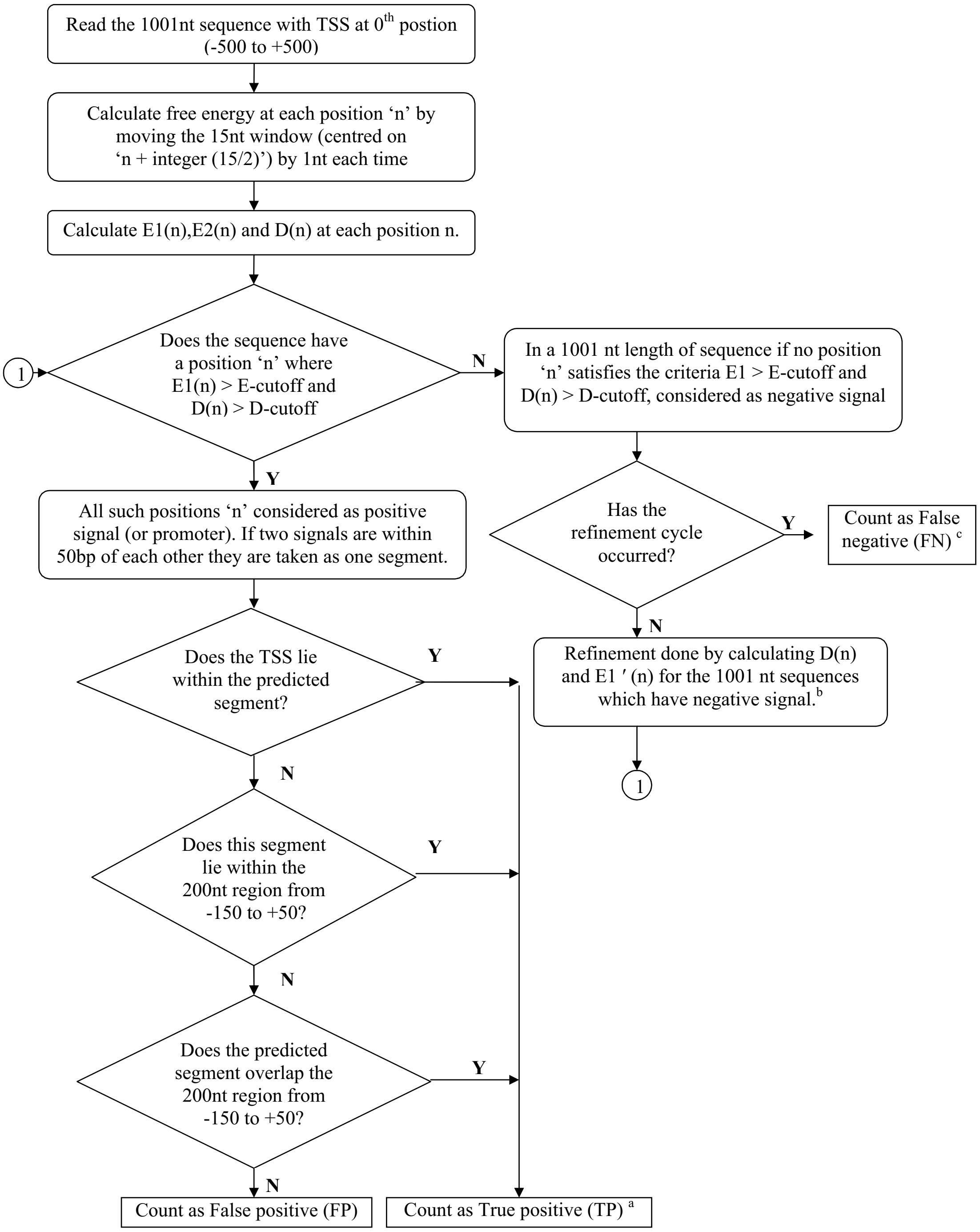
Read the 1001nt sequence with TSS at 0th postion (-500 to +500)
Calculate free energy at each position 'n' by moving the 15nt window (centred on 'n + integer (15/2)') by 1nt each time
Calculate E1(n),E2(n) and D(n) at each position n.
Does the sequence have a position 'n' where E1(n) > E-cutoff and D(n) > D-cutoff
1
N
In a 1001 nt length of sequence if no position 'n' satisfies the criteria E1 > E-cutoff and D(n) > D-cutoff, considered as negative signal
Y
All such positions 'n' considered as positive signal (or promoter). If two signals are within 50bp of each other they are taken as one segment.
Has the refinement cycle occurred?
Y
Count as False negative (FN) c
N
Refinement done by calculating D(n) and E1 ′ (n) for the 1001 nt sequences which have negative signal.b
1
Does the TSS lie within the predicted segment?
Y
N
Does this segment lie within the 200nt region from -150 to +50?
Y
N
Does the predicted segment overlap the 200nt region from -150 to +50?
Y
N
Count as False positive (FP)
Count as True positive (TP) a

Figure 1. A flowchart summarizing promoter prediction method. [a] If more than one region in the sequence satisfies the true positive (TP) criteria (shown in figure 2), then only the prediction which is nearest to the TSS is taken as TP. The other predictions are counted as false positives (FP). [b]The same procedure is followed to calculate TP and FP in case of the subsequent refinement cycle. [c]The sequences identified with a negative signal after the second refinement cycle are counted as false negative (FN).

respect to TSS). The E-cutoff value was calculated from 611 *E. coli* promoter sequences of 101 nt length in which the TSS are more than 100 nt apart. In case of *B. subtilis*, it was calculated from 339 promoter sequences of same length. D-cutoff value corresponds to the difference between the average free energy for the 101 nt length known promoter

Table 1. Definition of thresholds of free energy values used to predict promoters in whole genome sequences

		E. coli	*B. subtilis*
Average free energy G calculated over whole genome sequence	Mean G	–20.096	–18.882
	Standard deviation (σ)	0.131	0.055
	G-cutoff (Mean+3σ)	–19.70	–18.72
Average free energy E calculated over promoter region of 101 nt length	Mean E	–18.637	–17.131
	Standard deviation (σ)	0.0001	0.0022
	E-cutoff (Mean+3σ)	–18.63	–17.13
	D-cutoff (E-cutoff – G-cutoff)	1.07	1.59
Cutoff values used earlier for promoter prediction (Kanhere and Bansal 2005 b)	E1	–18.15	
	D	1.90	

G specifies the average free energy over the entire genome. E is the average free energy over known promoter regions of 101 nt length (ranging from -80 to +20 with respect to TSS). All energy values are in kcal/mol and the standard deviation values are also indicated. E-cutoff and D-cutoff are the thresholds used to predict promoter regions in whole genome annotation.

sequences and the average free energy over the entire genome sequence (G). The average as well as the cutoff values are listed in table 1 (calculated after taking into consideration 3σ variation in these values). As a test case, these E-cutoff and D-cutoff values were used to check for promoter regions over the 251 *E. coli* promoter sequences and 283 *B. subtilis* promoter sequences, which are 500 nt apart. These cutoff values were then applied to entire genome sequence to get the whole genome annotation for promoter regions.

2.5 *Sensitivity and precision*

The sensitivity and precision for the predictions are calculated using the following formulae:

$$\text{Sensitivity} = \frac{\text{Number of true positive (TP)}}{\text{Number of true positive (TP) + Number of false negative (FN)}}$$

$$\text{Precision} = \frac{\text{Number of true positive (TP)}}{\text{Number of true positive (TP) + Number of false positive (FN)}}$$

2.6 *True positive, false positive and false negative definition for known promoter sequences*

A known promoter predicted correctly by our method is considered as a true positive. A promoter is considered to be predicted correctly if it meets at least one of the following three conditions, as illustrated in figure 2: (i) A transcription start site (TSS) lies within the predicted promoter region, (ii) Predicted promoter region lies within the 200 nt region spanning from -150 nt upstream of TSS to +50 nt downstream of TSS, (iii) Predicted promoter region overlaps with the 200 nt region mentioned above. (i.e. at least 20 nt of predicted promoter region overlap with the 200 nt region, if not the overhang region outside of 200 nt region is less than 20 nt in length).

All other predicted promoters are considered as false positives. For some of the known promoter sequences, our method did not locate any positive signal and these are considered as false negatives. The algorithm followed here is illustrated as a flowchart in figure 1.

2.7 *True positive and false positive definition for whole genome promoter prediction*

The algorithm followed to calculate true positives and false positives for the whole genome promoter prediction to identify promoters in the vicinity of known TSS is shown in figure 3. If a predicted promoter region lies within the upstream region of a known gene (–500 nt from the translation start site of a gene), then it is considered as true positive. On the other hand, if a predicted promoter region occurs within the coding region of gene, then it is considered as a false positive. The information for the 4474 annotated gene in the EcoCyc database is used in this analysis. Similar information is currently not available for the *B. subtilis* genome.

3. Results and discussion

3.1 *Promoter regions are less stable*

The sum of interaction between the constituent dinucleotides of a DNA contributes to its stability, which is in turn, a sequence dependent feature. Thus if one knows the relative contribution of each nearest neighbor interaction in the DNA, the overall stability for an oligonucleotide can be predicted from its sequence (Breslauer *et al* 1986). Based on this principle, and using the protocol outlined at the start of figure 1, the calculated average stability profile for *E. coli* and *B. subtilis* promoter sequences are shown in figure 4. Our earlier analysis with 227 known *E. coli* promoters, that are 500 nt apart and 89 *B. subtilis* promoters that are 500 nt apart (Kanhere and Bansal 2005a, b) indicated that promoters from both these bacteria, which have considerably different genome composition (A+T composition: *E. coli*

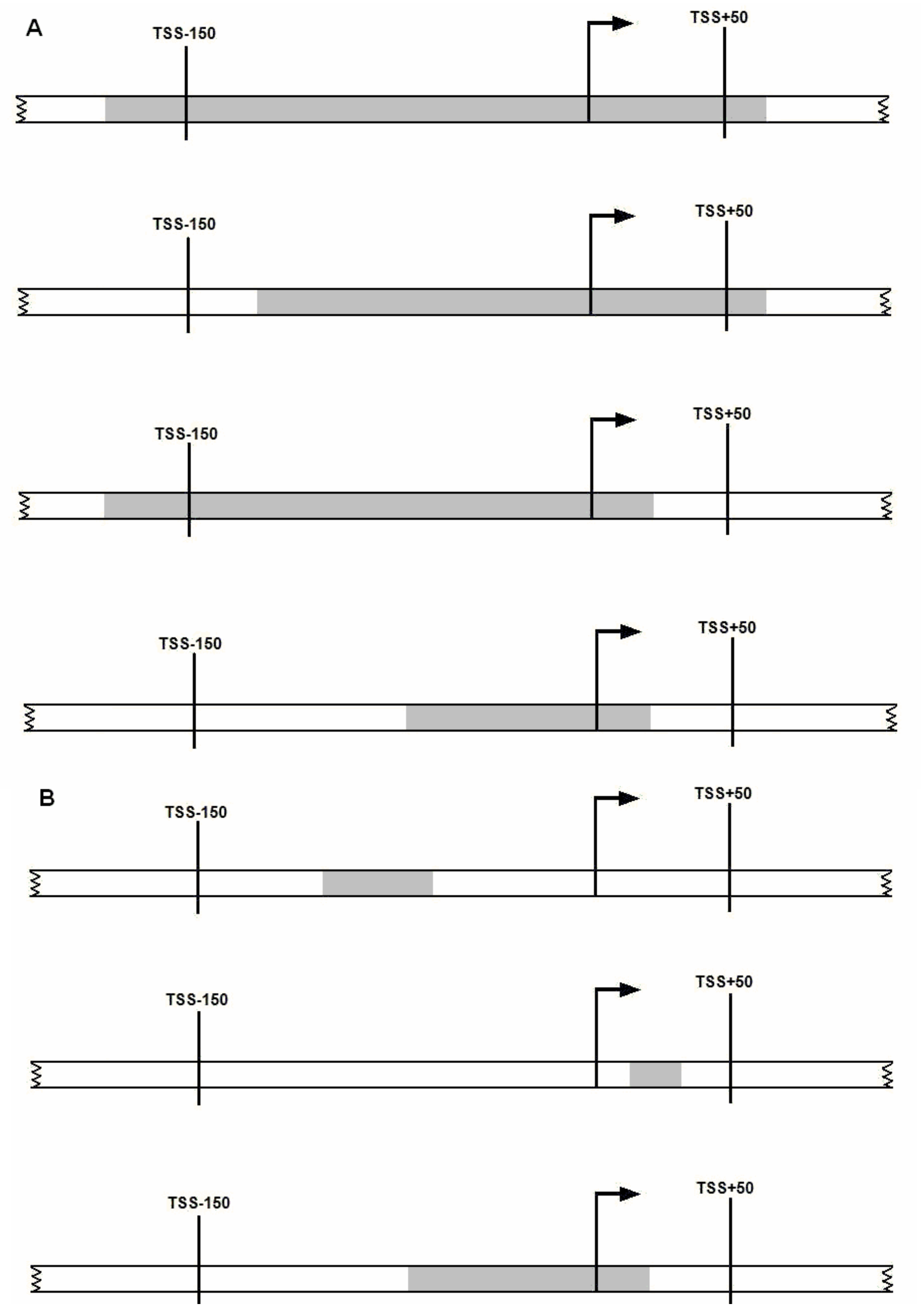

Figure 2(A,B). For caption, *see* p. 000.

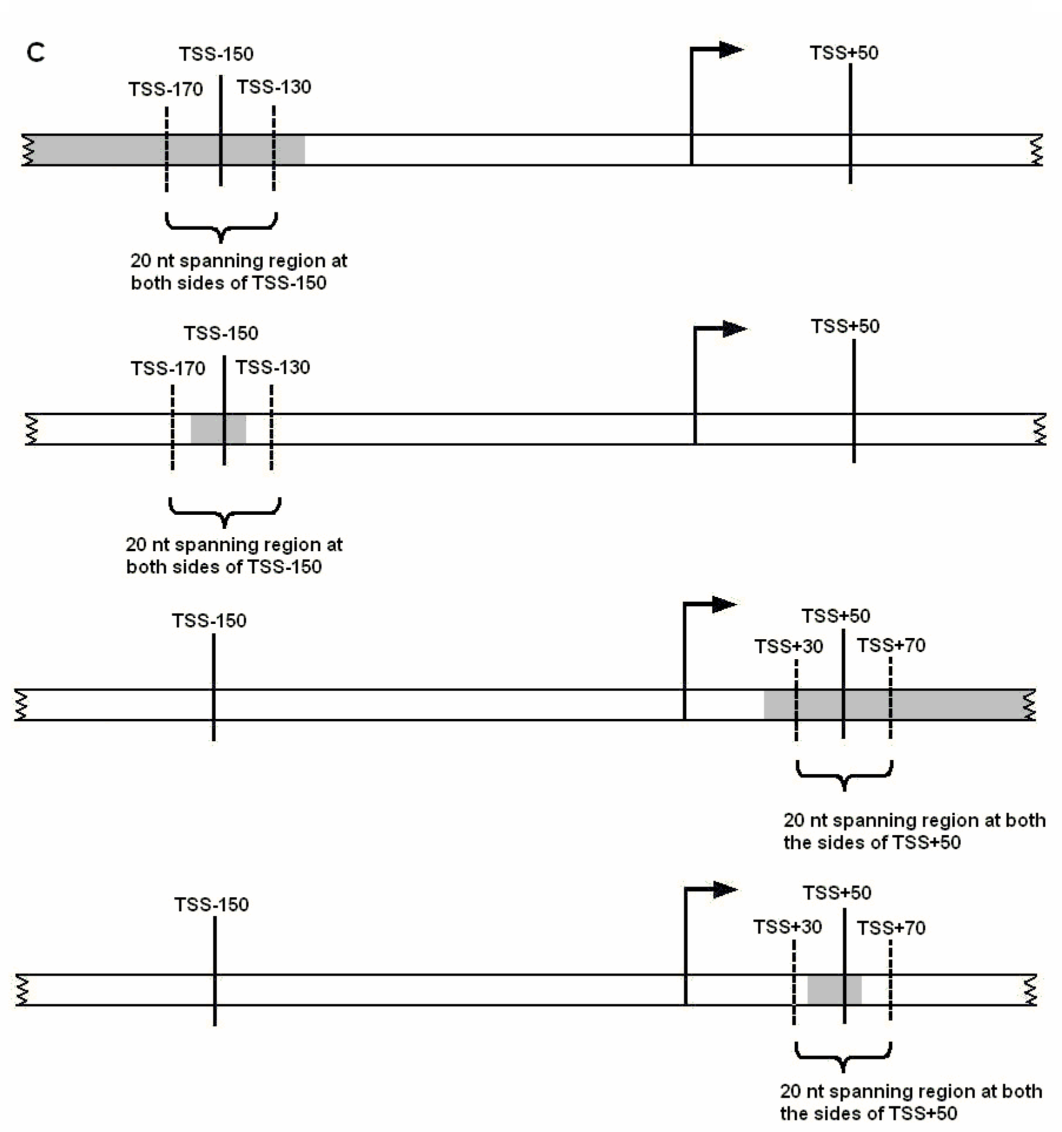
C
TSS-150
TSS-170
TSS-130
TSS+50
20 nt spanning region at both sides of TSS-150
TSS-150
TSS-170
TSS-130
TSS+50
20 nt spanning region at both sides of TSS-150
TSS-150
TSS+50
TSS+30
TSS+70
20 nt spanning region at both the sides of TSS+50
TSS-150
TSS+50
TSS+30
TSS+70
20 nt spanning region at both the sides of TSS+50

Figure 2. Illustration of true positive criteria. In this figure gray shaded region represents the promoter region as predicted by our method. A promoter is considered to be predicted correctly (TP) if it meets one of three conditions. (i) A TSS lies within the predicted promoter region. (ii) Predicted promoter region lies within the 200nt region spanning from –150 nt upstream of TSS to +50 nt downstream of TSS. (iii) Predicted promoter region overlaps with the 200 nt region mentioned above (i.e. at least 20 nt of predicted promoter region overlap with the 200 nt region, if not the overhang region outside of 200 nt region is less than 20 nt long).

0.49 and *B. subtilis* 0.56), show similar features, though the *B. subtilis* plot showed a slightly broader and more symmetric low stability peak. In our current study with 283 *B. subtilis* promoter sequences, we find a very prominent peak around the transcription start position (figure 4B) similar to that seen for *E. coli* (figure 4A). A closer look at the promoter sequences from these two bacterial sequences also reveals that both have low stability peaks around -10 region, -50 region as well as around -35 region (figure 5). As shown in our earlier analysis there is a marked difference in stability of upstream and downstream regions with the average stability of upstream region being lower than that of downstream region. However, the 101 nt long promoter sequences, from the two bacteria, differ in their basal energy levels (by ~1.5 kcal/mol), which can be attributed to the differences in the A+T content (~7%) of the entire bacterial genomes and more so in the promoter regions (~9%). The frequency of each nucleotide occurring in the two genomes, as well as in various specified regions of these two bacterial sequences are shown in table 2.

A recent promoter prediction method based on DNA sequence and its structural response to superhelical stress

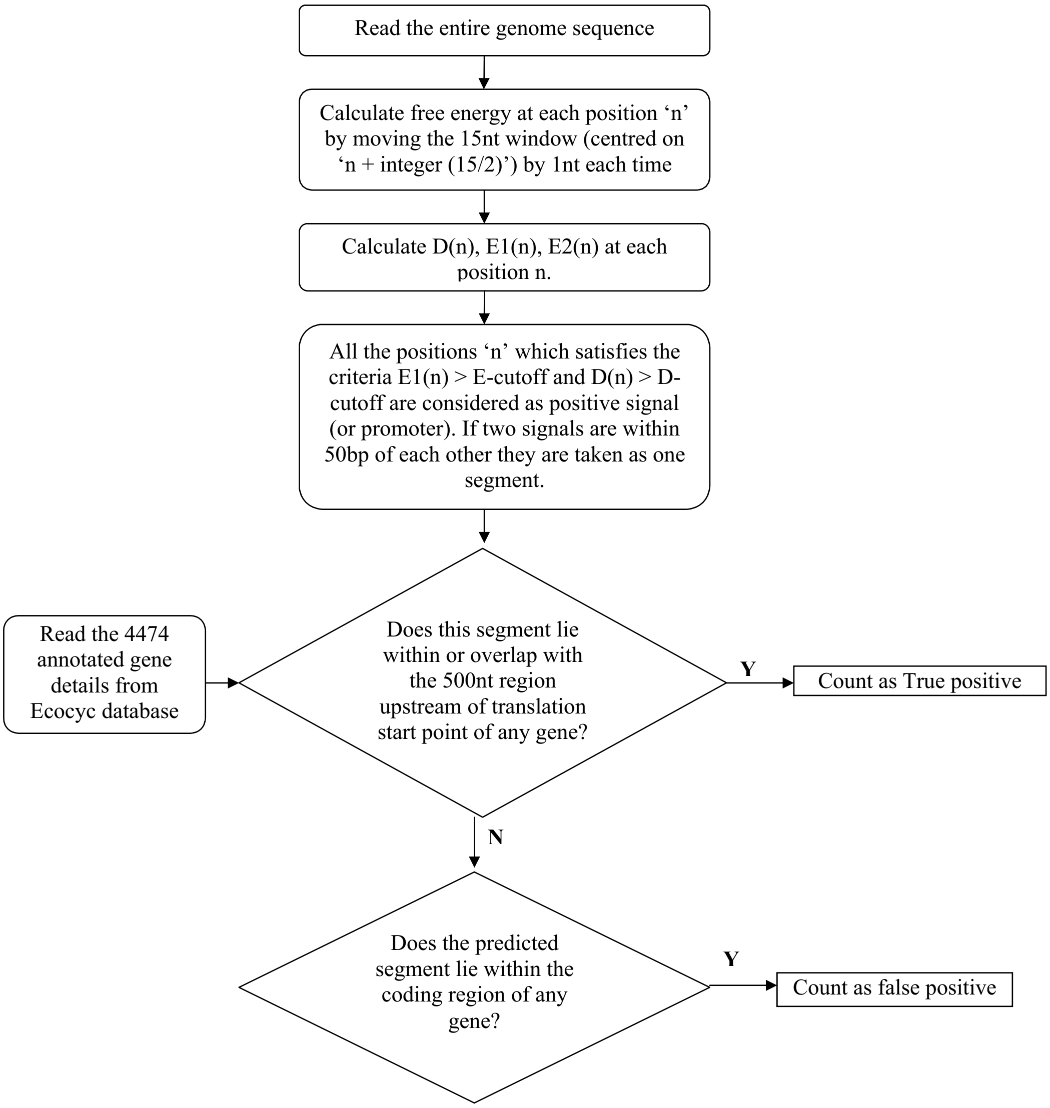
Read the entire genome sequence
Calculate free energy at each position 'n' by moving the 15nt window (centred on 'n + integer (15/2)') by 1nt each time
Calculate D(n), E1(n), E2(n) at each position n.
All the positions 'n' which satisfies the criteria E1(n) > E-cutoff and D(n) > D-cutoff are considered as positive signal (or promoter). If two signals are within 50bp of each other they are taken as one segment.
Read the 4474 annotated gene details from Ecocyc database
Does this segment lie within or overlap with the 500nt region upstream of translation start point of any gene?
Y
Count as True positive
N
Does the predicted segment lie within the coding region of any gene?
Y
Count as false positive

Figure 3. Flow chart representation for the promoter prediction method followed for analysis of whole genome sequences of prokaryotes.

(Wang and Benham 2006) had found that the regions between –174 and +57 are significantly destabilized and the maximum destabilization occurs at position –49 relative to the transcription start site. The strand separation (i.e. destabilization) occurs in this region because of high density of A+T nucleotides around the TSS region despite the fact that superhelicity couples all sites together (Botchan 1976; Kowalskai *et al* 1988). We also see in figure 4 that the low stability region spans about 150 nt (–100 to +50 about the TSS position) with low stability peaks occurring around –10 region, –50 region and a subsidiary peak around –35 region (shown in the close-up view in figure 5). However the -10 region is overall the least stable (average free energy -17.8 kcal/mol in case of *E. coli* and -15.7 kcal/mol in case of *B. subtilis*) as compared to the flanking regions.

3.2 *Promoter prediction analysis across known promoter sequences*

Promoter prediction has been carried over the known *E. coli* and *B. subtilis* promoter sequences of 1001 nt length. The methodology described above (*see* § 2.3) was used to predict promoters. Compared to our earlier method, the prediction accuracy is considerably enhanced in the present study. Detailed analysis over 1001 nt long known promoter

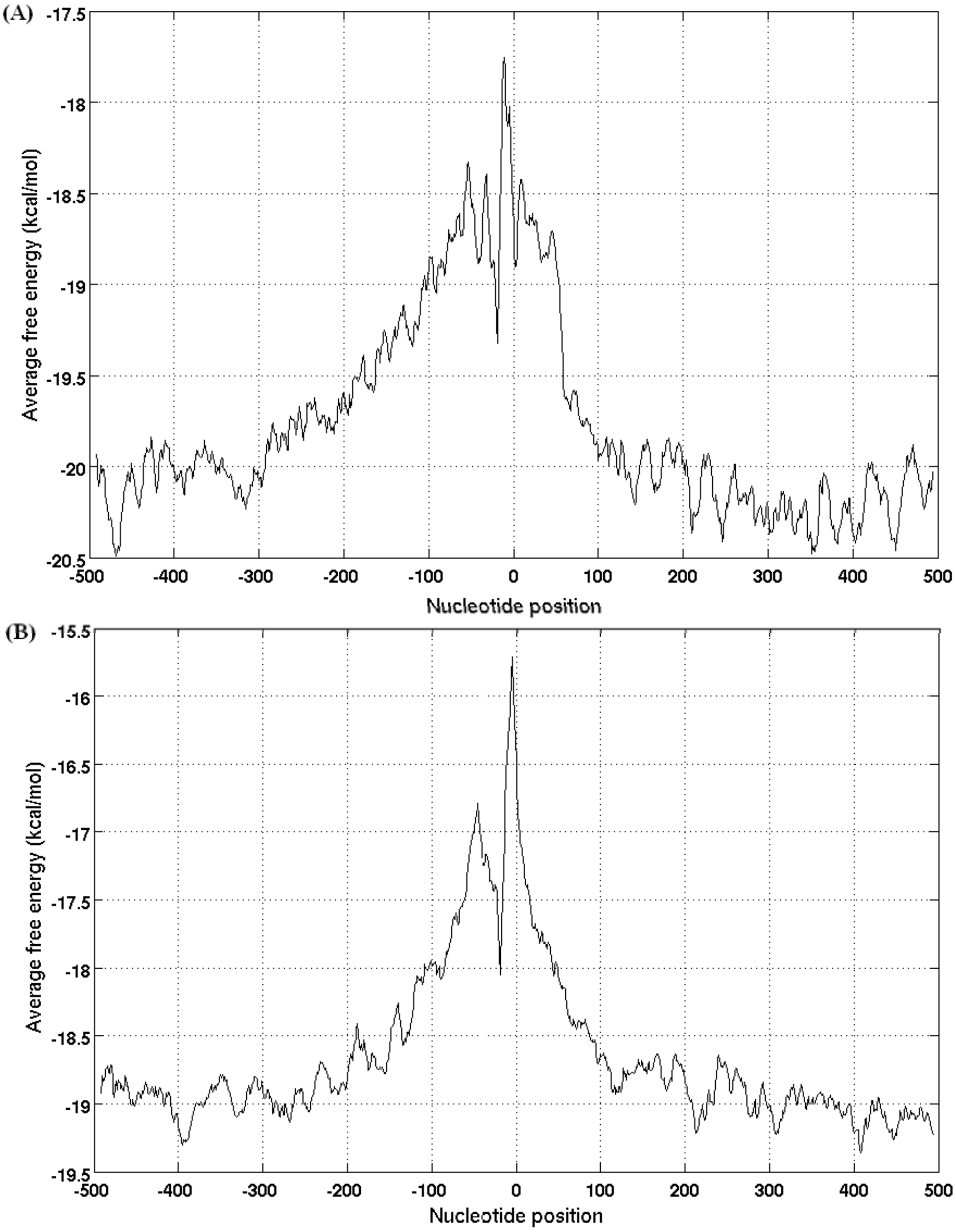

Figure 4. Average free energy profile around bacterial TSS. The figure shows the average free energy profiles of (**A**) *E. coli* (251 promoters) and (**B**) *B. subtilis* promoters (283 promoters) that are more than 500 nt apart. The free energy profile was plotted from 500 nt upstream to 500 nt downstream of transcription start site which corresponds to position 0. The nucleotide sequence position is shown on x-axis and average free energy value along y-axis. More negative values of free energy indicate greater stability, while the smaller values indicate lower stability.

sequences of *E. coli* is shown in table 3. The current method is 99% sensitive to a promoter signal and 52% accurate in case of *E. coli*. In case of *B. subtilis*, which has a different genome composition compared to *E. coli*, the present method is 94% sensitive to a promoter signal and gives a higher accuracy of 58%, using the cut-off values listed in table 1.

3.3 *Whole genome annotation of promoter regions*

Since we cannot define a false negative signal when promoter prediction is done over entire genome sequence, the present method was used to predict promoter regions without carrying out the subsequent refinement cycle, for

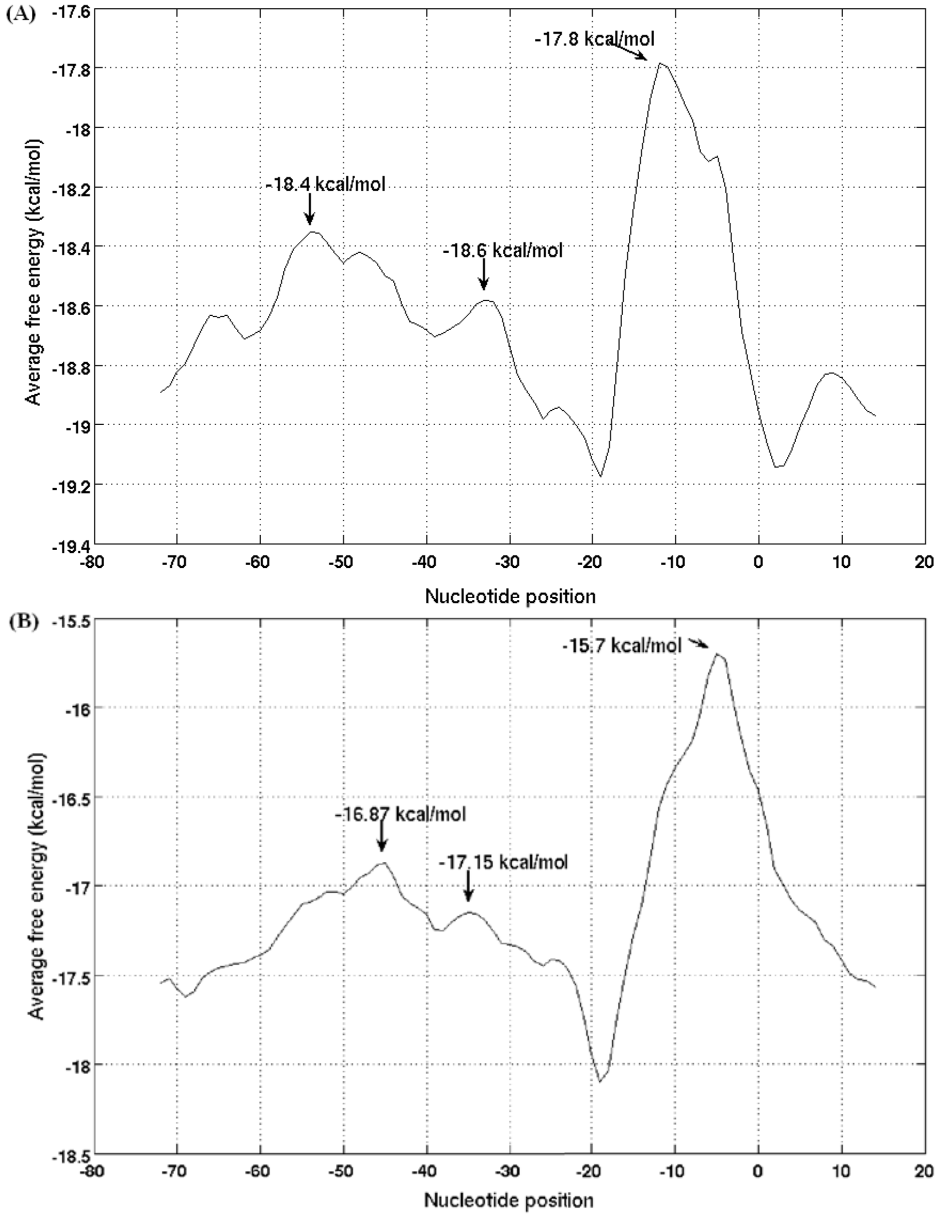

Figure 5. Free energy profile across the low stability regions near the bacterial TSS. A close-up view of average free energy profile for the low stability regions near the TSS, that are more than 100 nt apart. (**A**) *E. coli* (611 promoters) and (**B**) *B. subtilis* promoters (339 promoters). The free energy profile has been plotted from 80 nt upstream to 20 nt downstream of the TSS, which is positioned at 0. The arrow marks indicate the three lower stability peaks (near the –10, –35 and –50 regions) and the corresponding energy values are given above the arrow marks.

false negatives. In the current study we report the results on analysis of promoter regions over the entire *E. coli* genome.

The algorithm described above (*see* § 2.7 and figure 3) is used to calculate true positives and false positives for the whole genome promoter prediction, to identify promoters in the vicinity of known genes. The 4474 annotated gene information for *E. coli* from EcoCyc database is used in this analysis. The precision is found to be comparable to that obtained for known promoter sequences. Based on our definitions of true positive and false positive we achieved an accuracy level (i.e. precision) of 49% for the whole genome annotation of promoter regions in *E. coli* (table 4).

3.3.1 *Comparison with other structure based methods for whole genome annotation of promoter regions:* A recently proposed promoter prediction method based on

Table 2. The average frequency of individual nucleotides and A+T content of *E. coli* and *B. subtilis* across entire genome, in the promoter sequences as well as upstream and downstream region near the TSS are listed

		E. coli					*B. subtilis*				
		A	T	G	C	A+T	A	T	G	C	A+T
Whole genome		0.25	0.24	0.25	0.25	0.49	0.28	0.28	0.22	0.22	0.56
Upstream region	–80 to 0	0.28	0.30	0.20	0.21	0.58	0.33	0.34	0.17	0.16	0.67
	–500 to 0	0.26	0.26	0.24	0.23	0.52	0.30	0.28	0.21	0.20	0.59
Downstream region	0 to +20	0.30	0.26	0.23	0.21	0.56	0.38	0.26	0.23	0.13	0.64
	0 to +500	0.25	0.25	0.26	0.24	0.50	0.31	0.26	0.24	0.19	0.57
Entire length of promoter	–80 to +20	0.29	0.29	0.21	0.21	0.58	0.35	0.32	0.18	0.15	0.67
sequence	–500 to +500	0.25	0.26	0.25	0.24	0.51	0.31	0.27	0.23	0.19	0.58

The composition was calculated for 101 nt (–80 to +20 with respect to TSS) and 1001 nt (–500 to +500 with respect to TSS) length of promoter sequences; 611 promoter sequences from *E. coli* and 339 promoter sequences from *B.subtilis* were obtained when the TSS are 100 nt apart promoter sequences, while 251 promoter sequences from *E. coli* and 283 promoter sequences from *B. subtilis* were taken where the TSS are at least 500 nt apart promoter sequences.

Table 3. Analysis of promoter prediction results

	Earlier method *	Present method: first cycle	Second cycle refinement	Results from present method after second cycle of refinement for FN

Cutoff values (in kcal/mol)	E1 = –18.15
	D = 1.9
		E-cutoff = -18.64
		D-cutoff = 1.07
			E-cutoff = –18.64
			D-cutoff = 1.07	

Neighboring region criteria	50 nt (for E1 calculation) and 100 nt (for E2 calculation)	100 nt (for both E1 and E2 calculation)	50 nt (for E1 calculation) and 100 nt (for E2 calculation)	
No. of promoter sequences analysed	251	251	38[b]	251
TP	187	191	27	218
FP	201	156	43	199
FN	23	38 [b]	3	3
Sensitivity	0.8905	0.8341	-	0.9864
Precision	0.4820	0.5504	-	0.5228

The prediction has been done for 251 *E. coli* promoter sequences of 1001 nt length and at least 500 nt apart using two methods.
*In the earlier method D and E1 cutoff values (indicated in table 1) were chosen so that the precision is maximum. [b]The sequences in which no promoter signal is identified in the first cycle of iteration are considered for the second refinement cycle.

Table 4. Results of whole genome annotation of promoters in *E. coli*

Cutoff values (in kcal/mol)	Earlier method
	E1 = –18.15
	D = 1.9
		Present method
		E-cutoff = –18.64
		D-cutoff = 1.07

Total no. of annotated genes	4474	4474
Total No. of promoters predicted	5556	4636
TP	2451	2023
FP	2602	2098
Precision	0.485	0.491

stress–induced duplex destabilization of DNA sequence (or SIDD) (Wang and Benham 2006) attained a reliability of only 37%, when this property alone was used as a distinctive structural attribute to identify promoter sequences in the E coli genome. The authors evaluated their predicted promoter regions against the 927 documented TSSs from Regulon database (Salgado *et al* 2004). By our method, a reliability of 70% has been observed for *E. coli* genome annotation of promoter regions, when it was cross verified against the 960 TSSs of *E. coli* from EcoCyc database. Reliability has increased by 8% when compared to the predictions by our earlier method (Kanhere and Bansal 2005b) for promoter prediction, across the 960 TSSs listed in EcoCyc database.

4. Conclusion

Differences in DNA stability between neighboring regions can be used to annotate entire genome sequences for promoter regions. This enhanced version of our earlier algorithm for promoter prediction gives higher precision; also the method performs better in predicting promoter regions in *B. subtilis*, which has quite different nucleotide composition as compared to *E. coli*. Reliable promoter prediction is obtained when the differences between the free energy of the upstream region of known *E. coli* promoter sequences and the entire *E. coli* genome are used as thresholds to search for promoters in the whole *E. coli* genome sequence. The results are better as compared to other methods used for annotation of promoter regions of prokaryotic genomes. However the method needs to be improved further to eliminate the occurrence of predicted promoter regions within the coding regions (false positives here). This can be achieved by combining the stability criteria with a weight matrix analysis over the nucleotide sequences in the upstream regions. In general the method can be used for annotation of promoter regions in any prokaryotic genome where only limited experimental data is available on promoters and transcription start sites.

Acknowledgements

We wish to thank Yuko Makita, Human Genome Center, The Institute of Medical Science, University of Tokyo, Japan for her timely help in sending *B. subtilis* transcription start site details. This work was supported by Department of Biotechnology, New Delhi.

References

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

Botchan P 1976 An Electron Microscopic Comparison of Transcription on Linear and Superhelical DNA; *J. Mol. Biol.* **105** 161–176

Breslauer K J, Frank R, Blocker H and Marky L A 1986, Predicting DNA duplex stability from the base sequence; *Proc. Natl. Acad. Sci. USA* **83** 3746–3750

Bucher P 1990 Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences; *J. Mol. Biol.* **212** 563–578

Fickett J W and Hatzigeorgiou A G 1997, Eukaryotic promoter recognition; *Genome Res.* **7** 861–878

Harley C B and Reynolds R P 1987 Analysis of *E. coli* promoter sequences; *Nucleic Acids Res.* **15** 2343–2361

Hutchinson G B 1996 The prediction of vertebrate promoter regions using differential hexamer frequency analysis; *Comput. Appl. Biosci.* **12** 391–398

Kanhere A and Bansal M 2003 Identification of additional 'punctuation marks' in genomic DNA; *Proc. FAOBMB Bangalore* **139** 7–11

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

Kanhere A and Bansal M 2005b A novel method for prokaryotic promoter prediction based on DNA stability; *BMC Bioinformatics* **6** 1471–2105

Keseler I M , Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M and Karp PD 2005 EcoCyc: A comprehensive database resource for *Escherichia coli*; *Nucleic Acids Res.* **33** D334–D377

Kowalski D, Natale D and Eddy M 1988 Stable DNA unwinding, not "breathing," accounts for single-strand-specific nuclease hypersensitivity of specific A+T-rich sequences; *Proc. Natl. Acad. Sci. USA* **85** 9464–9468

Makita Y, Nakao M, Ogasawara N and Nakai K 2004 DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics; *Nucleic Acids Res.* **32** D75–D77

Margalit H, Shapiro B A, Nussinov R, Owens J and Jernigan RL 1988 Helix stability in prokaryotic promoter regions; *Biochemistry* **27** 5179–5188

Ohler U and Niemann H 2001 Identification and analysis of eukaryotic promoters: recent computational approaches; *Trends Genet.* **17** 56–60

Pedersen A G, Baldi P, Chauvin Y and Brunak S 1999 The biology of eukaryotic promoter prediction – a review; *Comput. Chem.* **23** 191–207

Prestridge D S 1995 Predicting Pol II promoter sequences using transcriptional factor binding sites; *J. Mol. Biol.* **249** 923–932

Reese M G 2001 Application of time-delay neural network to promoter annotation in the drosophila melanogaster genome; *Comput. Chem.* **26** 51–56

Salgado H, Gama-Castro S, Martinez-Antonio A, Diaz-Peredo E, Sanchez-Solano F, Peralta-Gil M, Garcia-Alonso D, Jimenez-Jacinto V *et al* 2004 RegulonDB (version 4.0), Transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12; *Nucleic Acids Res.* **32** D303–D306

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

Staden R 1984 Computer methods to locate signals in nucleic acid sequences; *Nucleic Acids Res.* **12** 789–800

Vollenweider H J, Fiandt M and Szybalski W 1979 A relationship between DNA helix stability and recognition sites for RNA polymerase; *Science* **205** 508–511

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** 1471–2105

Werner T 1999 Models for prediction and recognition of eukaryotic promoters; *Mammal. Genome* **10** 168–175

*e*Publication: 15 June 2007