Biophysical Journal
Article

Biophysical Society

Toward a Universal Structural and Energetic Model for Prokaryotic Promoters

Akhilesh Mishra,[1,2] Priyanka Siwach,[1,3] Pallavi Misra,[1] Bhyravabhotla Jayaram,[1,2,4,*] Manju Bansal,[5] Wilma K. Olson,[6] Kelly M. Thayer,[7] and David L. Beveridge[8]

[1]Supercomputing Facility for Bioinformatics & Computational Biology; [2]Kusuma School of Biological Sciences, Indian Institute of Technology, Delhi, India; [3]Department of Biotechnology, Chaudhary Devi Lal University, Sirsa, Haryana, India; [4]Department of Chemistry, Indian Institute of Technology, Delhi, India; [5]Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka, India; [6]Department of Chemistry & Chemical Biology and BioMaPS Institute for Quantitative Biology, Rutgers, Piscataway, New Jersey; [7]Department of Chemistry, Vassar College, Poughkeepsie, New York; and [8]Departments of Chemistry, Molecular Biology, and Biochemistry and Molecular Biophysics Program, Wesleyan University, Middletown, Connecticut

ABSTRACT With almost no consensus promoter sequence in prokaryotes, recruitment of RNA polymerase (RNAP) to precise transcriptional start sites (TSSs) has remained an unsolved puzzle. Uncovering the underlying mechanism is critical for understanding the principle of gene regulation. We attempted to search the hidden code in ~16,500 promoters of 12 prokaryotes representing two kingdoms in their structure and energetics. Twenty-eight fundamental parameters of DNA structure including backbone angles, basepair axis, and interbasepair and intrabasepair parameters were used, and information was extracted from x-ray crystallography data. Three parameters (solvation energy, hydrogen-bond energy, and stacking energy) were selected for creating energetics profiles using in-house programs. DNA of promoter regions was found to be inherently designed to undergo a change in every parameter undertaken for the study, in all prokaryotes. The change starts from some distance upstream of TSSs and continues past some distance from TSS, hence giving a signature state to promoter regions. These signature states might be the universal hidden codes recognized by RNAP. This observation was reiterated when randomly selected promoter sequences (with little sequence conservation) were subjected to structure generation; all developed into very similar three-dimensional structures quite distinct from those of conventional B-DNA and coding sequences. Fine structural details at important motifs (viz. −11, −35, and −75 positions relative to TSS) of promoters reveal novel to our knowledge and pointed insights for RNAP interaction at these locations; it could be correlated with how some particular structural changes at the −11 region may allow insertion of RNAP amino acids in interbasepair space as well as facilitate the flipping out of bases from the DNA duplex.

INTRODUCTION

An organism's complete set of genetic information is expressed in a highly regulated manner across time and space. Promoters are among the early players in regulation of gene expression. The promoter is the genomic sequence that acts as a platform for the assembly of RNA polymerase (RNAP) and other transcription factors and is located just upstream to coding sequence. Bacterial promoters consist of at least three RNAP recognition sequences: the −10 element, the −35 element, and the upstream promoter element. Sequence elements within or near to these regions contribute to regulation (1,2), but there is poor sequence conservation around these core elements (3,4). Recently discovered noncanonical transcripts in prokaryotes also have unconventional promoter location and architecture, as revealed by genome-wide transcriptional start site (TSS) mapping studies at single-nucleotide (nt) resolution (5). What guides the recruitment of transcriptional machinery so precisely to so many unconventional sites? Structural homology among different promoters, in which different sequences lead to similar structural variants, was considered as an alternative criterion quite early (3). Lately, DNA structural descriptors like DNA stability, stacking energy, A-philicity, propeller twist, and roll, among others, have been used to define/identify promoter regions to a certain extent in both prokaryotes and eukaryotes, and some structural properties were found to correlate well with gene expression (6–10). Though these studies make a significant contribution toward the understanding of promoter architecture,

Submitted June 15, 2018, and accepted for publication August 2, 2018.

*Correspondence: bjayaram@chemistry.iitd.ac.in

Akhilesh Mishra and Priyanka Siwach contributed equally to this work.

Editor: Tamar Schlick.

https://doi.org/10.1016/j.bpj.2018.08.002

things are far away from a universal model capable of explaining the underlying mechanism of transcription initiation at precise locations. RNAP is considered as the central component in transcription regulation, regulating by recognizing and binding to specific promoter sequences and facilitating unwinding of the DNA duplex near TSSs. With emerging reports on DNA structure regulating biological processes (11), a need arises to know whether promoter structure acts simply as a passive platform on which transcriptional machinery (RNAP and σ factors in bacteria) acts or whether it also regulates/directs/actively participates in transcription initiation.

The study was planned with two clear goals: to prepare complete structural and energetic profiles of TSSs and their adjoining regions in search of a universal model for prokaryotic promoter and to understand their implications on transcription initiation. The structure and dynamics of nucleic acids is guided by base sequence as well as by the sugar-phosphate backbone. Earlier attempts, mentioned above, have focused only on sequence and that too by taking only a few parameters; no attempts have been made toward a complete structural and energetic characterization of promoter regions. The last few decades have witnessed a revolutionary evolution in the analysis of nucleic acids' structure (12–20). We have previously reported that hydrogen-bond, stacking, and solvation energy show clear signatures of functional densities of DNA sequences (21–27).

For our study, we proceeded with nine backbone, eight interbasepair (inter-BP), six intrabasepair (intra-BP), five basepair axis (BP axis), and three energetic properties, adding to a total of 31 parameters, for exploring the genomic regions comprising primary TSSs of 12 microorganisms (belonging to both kingdoms, Archaea and Eubacteria, of prokaryotes). Numeric values of conformational parameters for the unique dinucleotides steps were obtained from crystal structures of B-DNA only (from the Nucleic Acid Database using Curves+ (17)), whereas in-house programs were used for energy parameters (27). Here, we report that these parameters provide unique structural and energetic signatures at TSSs. Our results offer to our knowledge fundamentally new insights into the active role of DNA structure and energetics at TSSs in transcription initiation and new pathways to explore transcriptional regulation in prokaryotes.

MATERIALS AND METHODS

Promoter and coding sequence data set preparation

A total of 16,519 primary TSS positions were selected from 12 organisms (Table 1). Sequences of 1001 nucleotides in length (spanning 500 nucleotides upstream and downstream of the TSS positioned at 0) for all selected TSS positions were extracted from the respective genome sequences. As a control data set, coding sequence (CDS) data for the respective organism were retrieved from the Ensembl Bacteria website. Out of 45,220 CDS sequences, only 6218 sequences had length greater than 1500 nt, from which we extracted 1001 central regions as a control data set for our analysis.

Crystal structures of B-DNA only

A total of 74 crystal structures of B-DNA, without any modification or association with protein or ligand molecule, were obtained from the Nucleic Acid Database (see Table S1).

Structural-parameter-value calculation

Twenty-eight parameters were selected: nine backbone (Alpha (α), Beta (β), Gamma (γ), Delta (δ), Epsilon (ε), Zeta (ζ), Chi (χ), Phase, and Amplitude), eight inter-BP (Shift, Slide, Rise, Tilt, Roll, Twist, H-Rise, and H-Twist), six intra-BP (Shear, Stretch, Stagger, Buckle, Propel, and Opening), and five BP axis (X Displacement, Y Displacement, Inclination, Tip, and Axis-Bend). The values for these parameters, for the crystal structures obtained above, were calculated using Curves+ (17). After calculating values for all the parameter 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. The parameter values for the unique dinucleotide steps thus obtained are provided in Table S2.

Energy-parameter-value calculation

The values for three energy parameters (viz. hydrogen-bond energy, stacking energy, and solvation energy) for the unique 10-dinucleotide steps was done as reported in our previous work (27).

Obtaining the structural and energy profile of each sequence

The calculated dinucleotide values for each parameter were used for getting the structural profile of the 1001-nt-long promoter and CDS sequence by performing a moving average calculation on a sliding window of 25-BP covering 24-dinucleotide steps. The same exercise was performed independently on all the selected sequences of primary promoter sequence and CDS sequence (as control) for all the 31 parameters.

Profile plotting of sequences

The plotting was performed using MATLAB software.

Normalization of values

To bring all the parameters on the same scale, the values were made dimensionless using normalization. The values were normalized between 0 and 1 by subtracting the minimal value of the profile from each value and then dividing the value with range of the profile (i.e., max–min).

Making derived structural criteria to define a sequence

The normalized values showing similar behavior were combined together to form two structural vectors: vector1 from 14 parameters showing peak (Stretch, Opening, Rise, Roll, Twist, H-Rise, H-Twist, β, γ, ε, Phase, Amplitude, hydrogen bond, stacking energy) and vector2 from 17 parameters showing cleft at TSS (X Displacement, Y Displacement, Inclination, Tip, Ax-Bend, Shear, Stagger, Buckle, Propel, Shift, Slide, Tilt, α, δ, ζ, χ, solvation).

TABLE 1 A Brief Description of the Selected Microorganisms along with the TSS and CDS Data Used in Our Study

Kingdom	Phylum	Microorganism	Genome Size, %GC content	Characteristic Features	Number of Primary TSS (reference)	Number of CDSs
Arche-bacteria	Euryarchaeota	*Methanolobus psychrophilus*	3.07 Mb, 44.6%	cold adaptive, methanogenic	1463 (40)	355
		Thermococcus kodakarensis	2.08 Mb, 52%	fermentative heterotroph, grows at 85°C	1248 (41)	208
		Halofrex volcanii	3.93 Mb, 65.63%	halophile	1723 (42)	425
Eubacteria	Actinobacteria	*Mycobacterium tuberculosis* H37Rv	4.38 Mb, 65.5%	pathogen, Gm +ve and −ve	1440 (43)	626
		Streptomyces coelicolor A3	9.05 Mb, 71.98%	soil dweller, Gm +ve	2771 (44)	1201
	Proteobacteria	*Helicobacter pylori*	1.63 Mb, 38.9%	pathogen, Gm −ve	816 (45)	227
		Salmonella enterica serovar Typhimurium	5.067 Mb, 52.09%	pathogen, Gm −ve	1871 (46)	624
		Escherichia coli	5.17 Mb, 50.6%	harmless gut microbe, Gm −ve	1222 (47)	577
		Pseudomonas aeruginosa PA14	6.58 Mb, 66.2%	pathogen, Ubiquitous, Gm −ve	2118 (48)	853
	Firmicutes	*Bacillus amyloliquefaciens*	3.95 Mb, 46.4%	soil dweller, Gm +ve	1062 (49)	393
	Chlamydiae	*Chlamydia pneumonia* CWL029	1.22 Mb, 40.6%	pathogenic, airborne, Gm −ve	357 (50)	198
	Cyanobacteria	*Synechocystis* sp. PCC6803	3.57 Mb, 47.7%	autotroph and heterotroph, Gm −ve	430 (51)	531
Total					16,519	6218

Gm, Gram stain; +ve, positive; −ve, negative.

Generating structures of promoter DNA

Twelve sequences (−75 to +25) were extracted with respect to randomly selected TSSs, one from each organism, and were subjected to structure generation using the X3DNA software package (28). Fine structures of five-nt-long motifs (from the −11, −35, and −70 regions) of *Bacillus amyloliquefaciens* were also generated. We first generated the generic B-DNA structure of selected sequences using the fiber tool of the 3DNA package and then analyzed the structures with the help of the find_pair and analyze tool. This command generated two parameter values files, the BP step-parameter file and the BP helical-parameter file. In the first step, we modified the BP step-parameter-value file using our predicted value and generated a modified Protein Data Bank (PDB) structure using the rebuild tool. Then, we again analyzed the modified PDB structure using the find_pair and analyze tool. This time, we modified the BP step helical parameter file of the BP steps using our predicted values and rebuilt the second-step-modified PDB structure. In this way, we are able to modify values for 18 DNA structural parameters including inter-BP, intra-BP, and BP axis parameters, i.e., all except the backbone angles and sugar-puckering variables. Because all parameters are correlated, it is assumed that these 18 structural parameters are sufficient to generate the structure of DNA sequence.

RESULTS AND DISCUSSIONS

All structural and energy parameters give signature profiles at TSSs

Primary promoter sequences were obtained by extracting 500 nucleotides both upstream and downstream from the given TSS from the complete genome sequence of each organism; CDSs were obtained from the Ensembl Bacteria website, and only the central region (1000 nucleotides long) of each CDS was taken (Table 1).

The numeric profiles of 31 structural and energy parameters were obtained for the pooled primary promoters (16,519) and the CDSs (6218) (see Materials and Methods) and are shown in Fig. 1. These pooled profiles were obtained by lining up all the promoter sequences with TSS at the same position, and all CDSs were also superimposed. Next, all the sequences were converted to numeric sequences for different structural parameters, and the average over all numeric sequences for each position is plotted (Fig. 1).

The abscissa shows the position relative to TSS, whereas the ordinate represents the numeric value of that parameter. As is clear from Fig. 1, all the parameters are capable of distinguishing the primary promoter sequences from CDSs. The promoter sequences show unique intrinsic value at the TSS and nearby regions, resulting in a sharp/broad peak/cleft at/near TSSs, and hence make a signature profile for that parameter (Fig. 1; for individual profiles of each organism, see Fig. S2).

As the sequence proceeds to the TSS, a gradual increase in the basal value (given by CDS and extreme upstream and downstream regions of TSS) is observed for 13 parameters (β, γ, ε, Phase, Amplitude, Rise, H-Rise, Roll, Twist, H-Twist, Stretch, Opening, and solvation energy) whereas 18 properties (α, δ, ζ, χ, Shift, Slide, Tilt, Shear, Stagger, Buckle, Propeller Twist, X Displacement (Xdis), Y

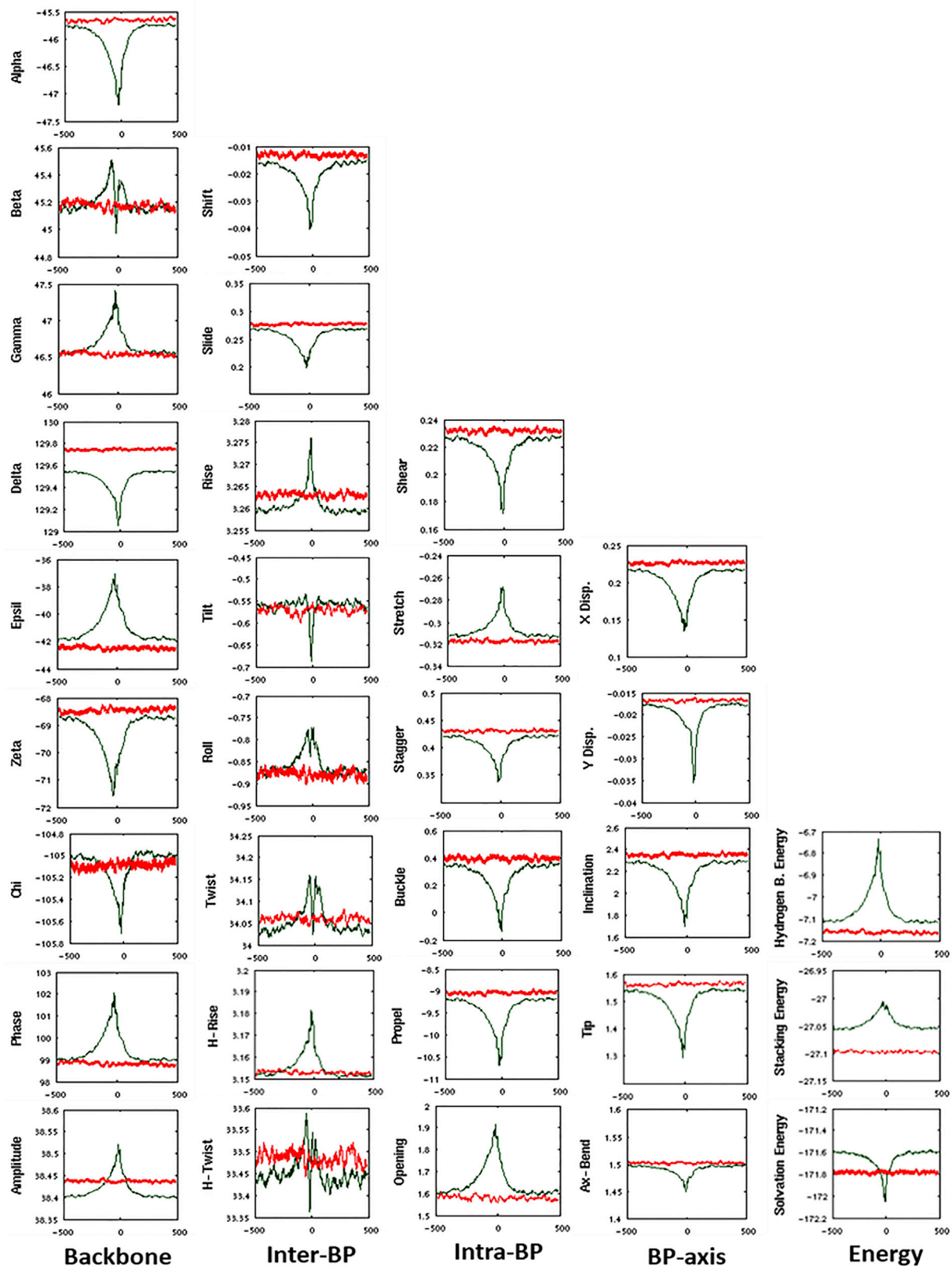

FIGURE 1 Structural and energy profiles of 1001-nt-long sequences having primary promoters (*green line*) and coding sequences (*red line*). Sequences having primary promoters (16,519) were lined up with TSSs at the same position ("0"), extending 500 nucleotides on both sides. Likewise, all CDSs were also superimposed. The ordinate represents the numeric value of that parameter, whereas the abscissa represents the nt position. To see this figure in color, go online.

Displacement (Ydis), Inclination, Tip, Ax-Bend, hydrogen bond energy, and stacking energy) show gradual decrease till TSS or its nearby upstream position and afterwards retrack back to basal values. Correlation exists among these structural properties, but ultimately each parameter contributes in its own way. The results obtained were analyzed to fulfill this need to know the impact of each parameter on the overall structure and shape of DNA at the TSS, as presented below.

Some properties adopt a very gradual change pattern spanning across a long distance (from the −250th ± 100 position through the TSS to +100th ± 50) in almost all the 12 prokaryotes. This category includes 24 properties—all torsion angles (α, β, γ, δ, ε, ζ, and χ) and sugar-puckering variables (Phase and Amplitude) of the sugar-phosphate backbone, all the five BP axis parameters (Xdis, Ydis, Inclination, Tip, and Ax-Bend; see Fig. S1), six inter-BP parameters (Shift, Slide, Roll, and H-Rise), four intra-BP properties (Shear, Stretch, Stagger, and Buckle), and two energy properties (hydrogen-bond and stacking energy (Fig. 1; Fig. S2). The second category belongs to those properties that give a very sharp signature profile, spanning across a small length of 30–35 nucleotides or less (−20 ± 5 to +10 ± 5); it includes seven parameters—four inter-BP (Rise, Tilt, Twist, H-Twist), two intra-BP (Propeller Twist and Opening), and one physicochemical property (solvation energy) (Fig. 1; Fig. S1). Either the required change in each of these properties at the TSS can be achieved by following their respective pattern or the change itself is needed across the respective distances.

The B-DNA backbone is realized in two major conformer substates: BI and BII, with interconversion guided by coupled changes in two dihedral angles ε, ζ. The BI substate is characterized by lower value of ε and higher values of ζ (with $\varepsilon - \zeta < 0$), whereas the reverse (with $\varepsilon - \zeta > 0$) is true for the BII substate (29). Though the values of torsion angles ε and ζ observed in our study do not coincide with that of the canonical B-DNA, their dynamics, as the sequence proceeds toward the TSS, correlates with transition from the BI to BII substate (ε increases, whereas ζ decreases), and at TSS, ε attains the maximal while ζ has the minimal value (i.e., backbone appears to be in BII conformer) in all prokaryotes (Fig. 1; Fig. S2). BII is the less common substate of B-DNA, as has been observed in crystal structures and molecular dynamics simulations (30). Another way to define the backbone transitions is to look at α, γ angles, which are found to associate with canonical and noncanonical backbone states, with α decreasing while γ increases during transition from the canonical to noncanonical state (31). The similar negative coupling between α, γ angles was observed as the sequences proceed to TSS (with α decreasing and γ increasing) in all the selected prokaryotes, indicating a trend from canonical state to a noncanonical state, though the angles values were far from the standard values given for canonical/noncanonical (Fig. 1).

Basepairs of promoters show an increasing tendency to align on top of each other as the sequences move toward the TSS by gradually decreasing Shift and Slide values. However, the increased angular distance between basepairs toward the minor groove side (i.e., Roll) does not allow the basepairs to be in parallel. Roll dynamics exhibits some peculiar trends: although undergoing a gradual increase, it shows a sudden decrease near the −35th ± 10 position, followed by a sudden rise and then a slow decrease till past the TSS (Fig. 1). A similar trend was also observed for Twist and H-Twist, except that the sudden decrease followed by sudden increase was observed near the −10th ± 5 position. Rise increases while Tilt decreases across almost same span (−20 ± 5 to +10 ± 5) (Fig. 1). The inter-BP parameters, obtained from atomic molecular dynamics simulations, have been used earlier in promoter prediction algorithm (8).

Among the various intra-BP parameters, a gradual decrease is observed for Shear, Buckle, and Stagger, resulting in centrally aligned bases on the intersection of the x and y axes. The basepairs show a gradual increase in stretch (from ∼−250th ± 100 bp), with a peak near −10th ± 5 followed by gradual decrease. Propeller Twist shows a sharp decrease (making the basepairs more parallel to the y axis), whereas Opening shows a sharp increase at around the −10th ± 5 position. Propeller Twist has also been reported earlier as a differentiating property between promoters and nonpromoters (6,7).

The BP axis of promoter regions was observed to have lower values of translational (Xdis, Ydis) and rotational movements (Tip, Inclination) as well as of Ax-Bend compared to adjoining regions. A decrease in Xdis and Ydis moves the basepairs toward the center along the x axis and y axis. Likewise, decrease in rotational movements (Inclination and Tip) would orient the basepairs to adopt a perpendicular orientation to the axis. It can be said that the helix becomes narrow and rigid and bases more perpendicular to the axis as the sequences proceed to TSSs. Less bendability of promoter regions around TSSs has also been reported earlier (9,10). Less bendability of promoter DNA disfavors formation of a nucleoid in prokaryotes and nucleosomes in eukaryotes, making these regions more accessible to transcription machinery.

Among the three physicochemical properties, hydrogen bond energy and stacking energy exhibit a gradual decrease when the sequence moves toward the TSS till around −10th ± 5 position, afterwards showing a gradual increase till past TSS. A sharp increase in solvation energy was observed at around −10th ± 5 position of the promoter sequence. Lesser stability of the promoter region has also been reported earlier (10,26).

At the individual prokaryote level, it is observed that prokaryotes differ greatly in the mean genomic value and signal strength at TSSs for a given parameter, but the nature of

change is almost similar (Fig. S2). Further, the difference in the mean genomic value and the signal strength at TSS for a given parameter is not found to correlate with genome size, phylogeny, and %GC content.

Combining all parameters for obtaining a single criterion

These parameters were made statistically unitless so as to evaluate them on a single scale (see Materials and Methods). When these 31 normalized (dimensionless) parameters of all the 12 organisms (31 × 12) are plotted together on this new structural scale, a clear peak and cleft is observed at the TSS or its adjoining upstream region (Fig. 2).

The next step was to join together all the parameters so as to make a single structural criterion to define local DNA structure. As discussed in the previous section, some properties show a gradual increase, whereas others show a gradual decrease till the TSS. Two structural vectors were made by joining together all the parameters with similar behavior: vector1 from parameters showing peak and vector2 by combining parameters showing cleft (see Materials and Methods). Initially, we also thought of combining all the parameters by flipping up the sign of negative to positive to form a single derived vector. However, it seemed that both types of changes somehow compensate for effect of each other on the DNA's structural and energetic profile, as subtracting vector2 from vector1 will nearly lead to values of CDSs. So, it appeared more appropriate to us to present the graph as two vectors essentially carrying information on compensatory sets of parameters. When values of these two vectors were plotted for the promoters and CDSs, a three-line graph was obtained for all organisms: a single line for CDSs and two lines for promoter sequences (Fig. 3). One surprising and striking observation was that

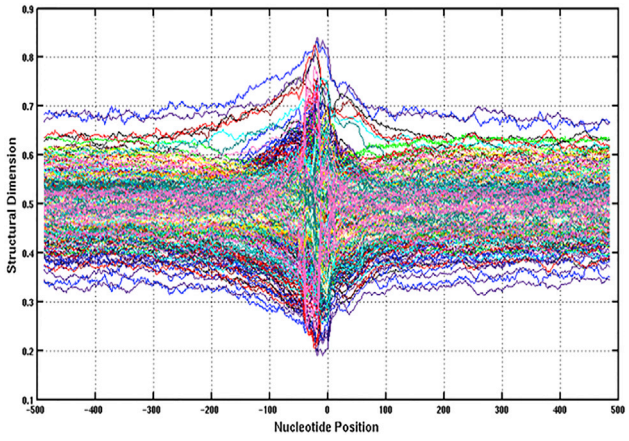

FIGURE 2 The normalized values of 31 parameters (of all the 12 organisms) versus nt position with respect to the TSS. Each organism was given a single color for all the 31 parameters. The plot represents 372 lines (31 × 12); a clear peak and cleft are observed at the TSS or its adjoining upstream region. To see this figure in color, go online.

despite their diversity, all organisms come to lie on the same position on this new structural scale (Fig. 3). The two vectors together give a uniform value of 0.5 for the CDSs of all organisms. For the promoter sequences, at the TSS, vector1 gives a peak of magnitude ranging from 0.57 to 0.63, whereas vector2 yields a cleft of magnitude 0.3–0.37 for all the organisms. The above observation strongly indicates that DNA speaks a universal language. The strength of the physical signals of DNA language at promoters has also been previously observed by combining experiments and simulations studies (32).

Different promoter sequences lead to similar structures

All the structural parameters act simultaneously to ultimately decide the DNA structure. The study was extended to generate structures of randomly selected promoter sequences, one from each organism. X3DNA software was used for generating structures using values of inter-BP, intra-BP, and BP axis for the unique di-nt steps generated during this study (Fig. 4).

For comparison, structure of one CDS randomly selected from *B. amyloliquefaciens* was also generated, and a canonical B-DNA structure was also taken. All the promoter sequences, despite poor sequence alignment (Fig S3), led to almost similar structures (in terms of showing some curvature), quite distinct from that of the CDS and canonical B-DNA (Fig. 4). To express this difference in some quantitative way, root mean-square deviation values of promoter backbone were calculated with reference to the CDS backbone using PyMOL software; all promoters exhibited root mean-square deviation values of above 12 Å with respect to the CDS. As is clear from Fig. 4, promoter regions adopt a slightly curved structure with variable groove dimensions throughout the length till the TSS; on the other hand, the CDS and generic B-DNA adopt a straight structure with nearly uniform groove dimensions. Each promoter, however, displayed its own style of structural distortions that might be unique to that organism or that particular gene, though much cannot be said at this stage. DNA curvature is a long-range secondary structural feature that can facilitate interaction between distant regions of DNA, leading to an indirect readout mechanism.

To have a closer look at the structural changes in promoter region, three-dimensional structures of five-nt-long motifs from important regions (−11, −35, and −70) of one promoter sequence (of *B. amyloliquefaciens*) were generated (Fig. 5). As is clear from Fig. 5, all the three regions show large deviations from the standard B-DNA in the arrangement of basepairs. At the −11 motif, a sharp increase in the vertical distance between basepairs at position −11 and −10 (4.4 and 4.6 Å in forward and reverse strand, respectively) is distinctly visible, the distance being significantly higher than 3.4 Å, the standard inter-BP

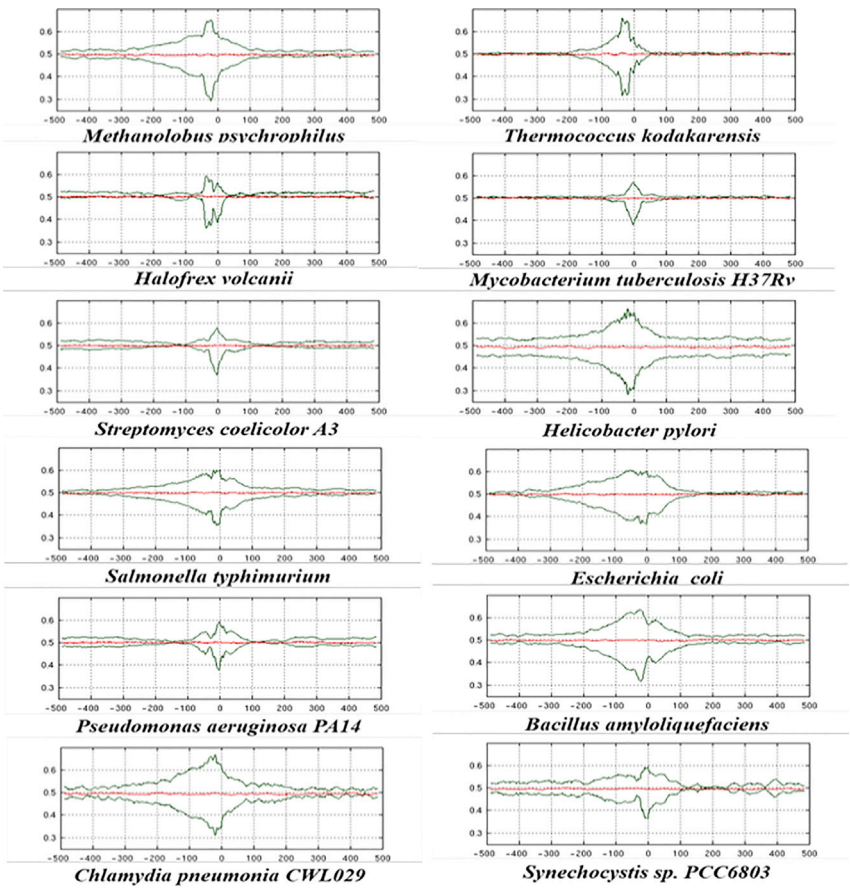

FIGURE 3 The derived structural vector profiles for the 12 organisms. Green lines represent sequences having the TSS at the "0" position, whereas the red line represents the CDSs. The green line showing a peak is vector1, and the green line showing a cleft is vector2; each is obtained by combining normalized values of parameters showing same behavior (see Materials and Methods). For CDS, both vectors give a single line graph (*red*). The ordinate represents the numeric value of the new structural scale, whereas the abscissa represents the nt position relative to TSS. To see this figure in color, go online.

distance in B-DNA structure. This supports the observation made for the sharp increase in Rise at the −11 region in Fig. 1. Another very interesting observation can be made from the Calladine model of this motif (Fig. 5 *d*). The BP at the −11th position shows high Stretch but low Twist, whereas consecutive basepairs on both sides exhibit high Twist. Similar behavior of Twist (low Twist position with high Twist on both sides) and Stretch for the −11 region was also recorded in Fig. 1.

The −35 motif displays remarkable deviations in the arrangement of basepairs (Fig. 5, *b* and *e*). The BP axis takes a slight bend at the −35th position. Basepairs at positions −34 and −33 show increased Stagger, whereas all basepairs show variable levels of Tilt, Roll, Shift, Slide, and Propeller Twist. It is difficult to interpret conclusively from these structures, but the −35 motif definitely seems to be a hot spot of different structural deviations. For the −70 region, an increase in angular distance from minor groove side (Roll) is visible, whereas a slight bending in axis is also observed (Fig. 5, *c* and *f*).

Implications for transcription initiation

In the light of results discussed above, it seems that the TSS and adjoining regions offer topographical signatures, which act as strong nucleating factors for inviting RNAP and transcription factors. The topographical landscape of DNA molecular shapes has been considered to provide an efficient means of indirect readout of DNA (shape recognition) (33). Efforts to correlate the changes observed in our study to the existing information about RNAP-promoter interactions at atomic level lead us to develop some fundamentally new insights, to our knowledge, about how these observed structural and energetic changes become instrumental in facilitating interaction with RNAP, which are explained below.

Further, the promoter structure and energetics seem to guide the subsequent interaction with various RNAP subunits and transcription factors. For instance, promoter DNA backbone undergoes a transition from BI to BII, resulting in the placement of phosphate toward the minor groove side; this might facilitate interaction with different domains of the σ subunit of RNAP, e.g., α-carbon-terminal domains of the σ subunit interact with the upstream promoter element using helix-hairpin-helix motifs (34) by hydrogen bonds between its backbone nitrogen and DNA backbone phosphate groups (35). Also, the −70 region and −35 regions, important for interaction with various subunits of RNAP, show lots of structural deviations. What is the need for such visible and significant deviations in the

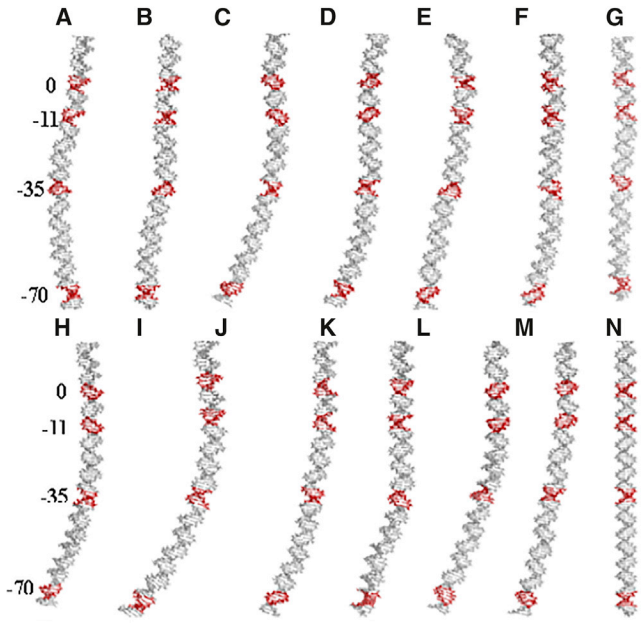

FIGURE 4 Three-dimensional structures of promoter regions (−75 to +25 with respect to TSS) belonging to different organisms: (*A*) *B. amyloliquefaciens*, (*B*) *C. pneumoniae*, (*C*) *E. coli*, (*D*) *H. volcanii*, (*E*) *H. pylori*, (*F*) *M. psychrophilus*, (*H*) *M. tuberculosis*, (*I*) *P. aeruginosa*, (*J*) *S. typhimurium*, (*K*) *S. coelicolor*, (*L*) *Synechocystis* sp.PCC6803, and (*M*) *T. kodakarensis*. For comparison, similar structures of the CDS region (*G*) and that of canonical B-DNA (*N*) are also given. To see this figure in color, go online.

basepair arrangements of these motifs? It demands a thorough investigation from many viewpoints. But at this stage, it seems that these changes may provide close access of various molecules of DNA to RNAP and other factors for required atomic interactions. Atomic details of the −35 element recognition by $\sigma 4$ of bacterial RNAP showed that helix-turn-helix motifs of $\sigma 4$ interact exclusively from the major groove side on both templates (36).

According to a recent report, promoter melting starts from within the −10 element (−12 to −7 nt position) by interaction with the σ_2 subunit of RNAP, resulting in a flipping out of the A_{-11} and T_{-7} bases of the nontemplate strand, which then get buried inside the pocket of the σ_2 subunit (37). Whether σ_2 actively disrupts the basepairs $-(A/T)_{-11}$ $(T/A)_{-7}$ by its aromatic amino acid shovels or passively captures transiently exposed bases remains to be established (37,38). Our study offers some novel insights, to our knowledge. At the −11 region, the vertical distance between two basepairs (Rise) displays a sharp increase (Fig. 1), particularly between the −11th and −10th position in one selected case (Fig. 5). This increased vertical distance between them might allow the aromatic amino acid shovels of σ_2 to enter in the inter-BP space. Such a significant increase in Rise is not observed in other regions of promoter. Further, Twist shows a typical behavior: a sharp increase somewhere around the −12th position followed by a sharp decrease and then another sharp increase (Fig. 1); the exact position of consecutive basepairs showing this pattern may vary from organism to organism, but it is observed in all selected organisms (Fig. S2; 17). For the selected promoter of *B. amyloliquefaciens*, it is −12,−11,−10 showing high, low, and high Twist, respectively. It seems possible that under such conditions, the middle-low Twist position comes under strong torsional stain due to adjoining high-Twist regions, and as a result, either it gets partially extruded out or amino acid shovels of σ_2 present in the inter-BP space find it easier to extrude this unstable position base. Further investigations are needed to confirm this hypothesis. The energetics profile shows the −10 region to be the most unstable, and it thus seems to facilitate promoter melting.

CONCLUSION

Prevalent thinking posits that RNAP is the key regulator of transcription initiation and after recognition and binding to the promoter DNA, it triggers a series of conformational changes in itself as well as in promoter DNA that are instrumental for transcription process initiation. However, the results obtained in the present study indicate that DNA exhibits changes in the overall structure at the TSS and nearby regions without any aid from RNAP and transcription factors. Some previous studies also report that DNA dynamically directs its own transcription (39). On the basis of the results obtained in our study, we conclude that DNA structure is a key regulator of transcription initiation; rather than acting as a passive platform on which RNAP acts to bring required changes, it assumes its structure and energetics on its own at the TSS and nearby regions so as to offer a conducive microenvironment to transcription machinery for precise recognition and atomic interactions needed for transcription initiation. Essentially, the message of the TSS is already built into the structure and energetics of DNA sequences. Further, we have used the values for unique dinucleotide steps in our study. Because conformational, energetics, and helical properties of a BP are strongly influenced by nearest neighbors (18), we expect even better manifestation of these signals if tetranucleotide and higher-order steps are considered instead of dinucleotides steps.

Data availability

We have considered 16,519 primary promoter sequences and 6218 CDSs from 12 organisms (Table 1). The user can download the complete set of organism-specific promoter sequence and CDS used in this analysis from our website (http://www.scfbio-iitd.res.in/software/data_TSS.jsp). The rest of the data are available in the Supporting Material.

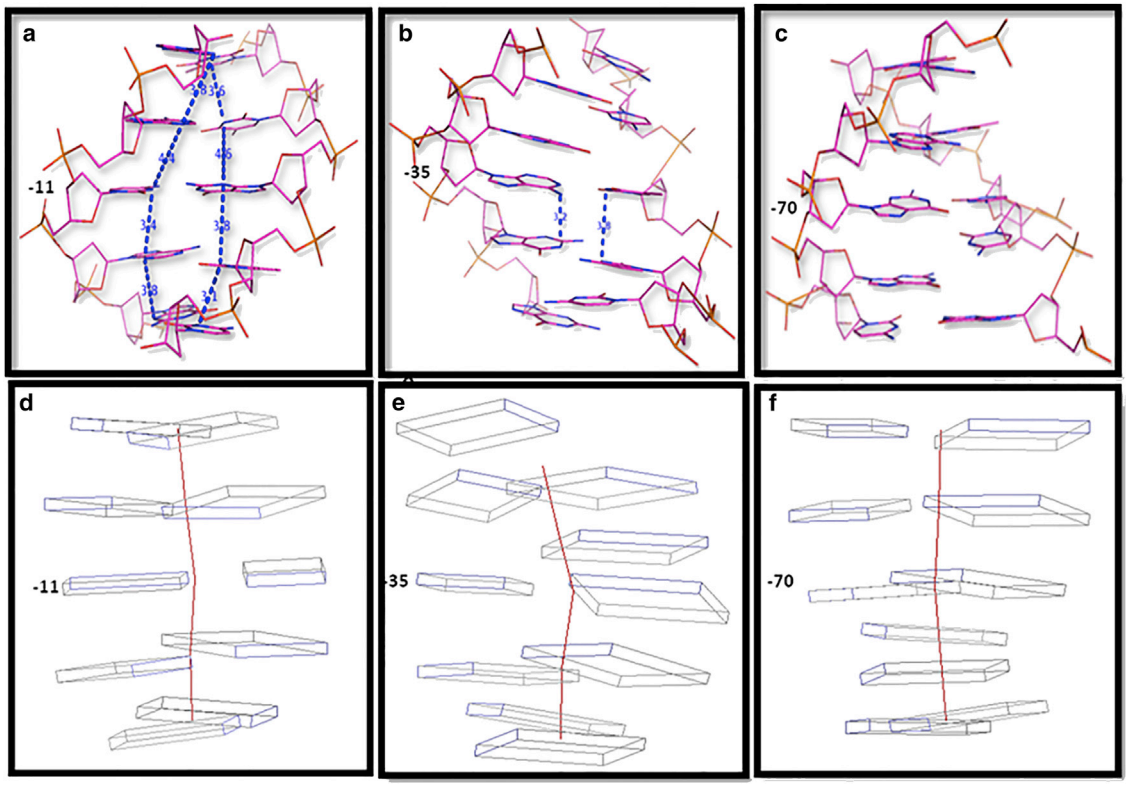

FIGURE 5 Three-dimensional structures of five-nt-long motifs of the −11, −35, and −70 regions from one randomly selected promoter sequence of *B. amyloliquefaciens*. (*a–c*) represent the line model structures of −11, −35, and −70, whereas (*d–f*) represent their respective Calladine and Drew model structures. To see this figure in color, go online.

SUPPORTING MATERIAL

Three figures and two tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30924-X.

AUTHOR CONTRIBUTIONS

B.J., P.S., and A.M. designed the project. P.S., A.M., and P.M. collected the data. P.S., A.M., and B.J. analyzed the results and wrote the manuscript. M.B., W.K.O., K.M.T., and D.L.B. contributed some of the ideas and software used in the study and critically read the manuscript.

ACKNOWLEDGMENTS

The authors thank Professors Richard Lavery and Krystyna Zakrzewska for their helpful suggestions. P.S. extends thanks to Chaudhary Devi Lal University for granting sabbatical leave to her.

Support from the Department of Biotechnology, Government of India to the Supercomputing Facility for Bioinformatics and Computational Biology, Indian Institute of Technology, Delhi, is gratefully acknowledged. A.M. is a recipient of senior research fellowship from University Grants Commission, Goverment of India.

REFERENCES

1. Pribnow, D. 1975. Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci. USA.* 72:784–788.
2. Harley, C. B., and R. P. Reynolds. 1987. Analysis of *E. coli* promoter sequences. *Nucleic Acids Res.* 15:2343–2361.
3. Lisser, S., and H. Margalit. 1994. Determination of common structural features in Escherichia coli promoters by computer analysis. *Eur. J. Biochem.* 223:823–830.
4. Levo, M., and E. Segal. 2014. In pursuit of design principles of regulatory sequences. *Nat. Rev. Genet.* 15:453–468.
5. Wade, J. T., and D. C. Grainger. 2014. Pervasive transcription: illuminating the dark matter of bacterial transcriptomes. *Nat. Rev. Microbiol.* 12:647–653.
6. Abeel, T., Y. Saeys, …, Y. Van de Peer. 2008. Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.* 18:310–323.
7. Florquin, K., Y. Saeys, …, Y. Van de Peer. 2005. Large-scale structural analysis of the core promoter in mammalian and plant genomes. *Nucleic Acids Res.* 33:4255–4264.
8. Goñi, J. R., A. Pérez, …, M. Orozco. 2007. Determining promoter location based on DNA structure first-principles calculations. *Genome Biol.* 8:R263.
9. Meysman, P., J. Collado-Vides, …, K. Laukens. 2014. Structural properties of prokaryotic promoter regions correlate with functional features. *PLoS One.* 9:e88717.
10. Kumar, A., and M. Bansal. 2017. Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression. *DNA Res.* 24:25–35.
11. Brázda, V., R. C. Laister, …, C. Arrowsmith. 2011. Cruciform structures are a common DNA feature important for regulating biological processes. *BMC Mol. Biol.* 12:33.
12. Yanagi, K., G. G. Privé, and R. E. Dickerson. 1991. Analysis of local helix geometry in three B-DNA decamers and eight dodecamers. *J. Mol. Biol.* 217:201–214.

13. el Hassan, M. A., and C. R. Calladine. 1995. The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme. *J. Mol. Biol.* 251:648–664.
14. Olson, W. K., A. A. Gorin, …, V. B. Zhurkin. 1998. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes. *Proc. Natl. Acad. Sci. USA.* 95:11163–11168.
15. Beveridge, D. L., G. Barreiro, …, M. A. Young. 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.
16. Dixit, S. B., D. L. Beveridge, …, P. Varnai. 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.
17. Lavery, R., M. Moakher, …, K. Zakrzewska. 2009. Conformational analysis of nucleic acids revisited: curves+. *Nucleic Acids Res.* 37:5917–5929.
18. Lavery, R., K. Zakrzewska, …, J. Sponer. 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.
19. Beveridge, D. L., T. E. Cheatham, III, and M. Mezei. 2012. The ABCs of molecular dynamics simulations on B-DNA, circa 2012. *J. Biosci.* 37:379–397.
20. Pasi, M., J. H. Maddocks, …, R. Lavery. 2014. μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic Acids Res.* 42:12272–12283.
21. Dutta, S., P. Singhal, …, B. Jayaram. 2006. A physicochemical model for analyzing DNA sequences. *J. Chem. Inf. Model.* 46:78–85.
22. Singhal, P., B. Jayaram, …, D. L. Beveridge. 2008. Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations. *Biophys. J.* 94:4173–4183.
23. Khandelwal, G., and J. Bhyravabhotla. 2010. A phenomenological model for predicting melting temperatures of DNA sequences. *PLoS One.* 5:e12433.
24. Khandelwal, G., and B. Jayaram. 2012. DNA-water interactions distinguish messenger RNA genes from transfer RNA genes. *J. Am. Chem. Soc.* 134:8814–8816.
25. Khandelwal, G., J. Gupta, and B. Jayaram. 2012. DNA-energetics-based analyses suggest additional genes in prokaryotes. *J. Biosci.* 37:433–444.
26. Khandelwal, G., R. A. Lee, …, D. L. Beveridge. 2014. A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes. *Biophys. J.* 106:2465–2473.
27. Singh, A., A. Mishra, …, B. Jayaram. 2017. Physico-chemical fingerprinting of RNA genes. *Nucleic Acids Res.* 45:e47.
28. Lu, X. J., and W. K. Olson. 2003. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* 31:5108–5121.
29. Temiz, N. A., D. E. Donohue, …, J. R. Collins. 2012. The role of methylation in the intrinsic dynamics of B- and Z-DNA. *PLoS One.* 7:e35558.
30. Heddi, B., N. Foloppe, …, B. Hartmann. 2006. Quantification of DNA BI/BII backbone states in solution. Implications for DNA overall structure and recognition. *J. Am. Chem. Soc.* 128:9170–9177.
31. Várnai, P., D. Djuranovic, …, B. Hartmann. 2002. α/γ transitions in the B-DNA backbone. *Nucleic Acids Res.* 30:5398–5406.
32. Durán, E., S. Djebali, …, M. Orozco. 2013. Unravelling the hidden DNA structural/physical code provides novel insights on promoter location. *Nucleic Acids Res.* 41:7220–7230.
33. Rohs, R., S. M. West, …, B. Honig. 2009. The role of DNA shape in protein-DNA recognition. *Nature.* 461:1248–1253.
34. Ross, W., A. Ernst, and R. L. Gourse. 2001. Fine structure of E. coli RNA polymerase-promoter interactions: α subunit binding to the UP element minor groove. *Genes Dev.* 15:491–506.
35. Doherty, A. J., L. C. Serpell, and C. P. Ponting. 1996. The helix-hairpin-helix DNA-binding motif: a structural basis for non-sequence-specific recognition of DNA. *Nucleic Acids Res.* 24:2488–2497.
36. Campbell, E. A., O. Muzzin, …, S. A. Darst. 2002. Structure of the bacterial RNA polymerase promoter specificity sigma subunit. *Mol. Cell.* 9:527–539.
37. Feklistov, A., and S. A. Darst. 2011. Structural basis for promoter-10 element recognition by the bacterial RNA polymerase σ subunit. *Cell.* 147:1257–1269.
38. Feklistov, A., B. Bae, …, S. A. Darst. 2017. RNA polymerase motions during promoter melting. *Science.* 356:863–866.
39. Choi, C. H., G. Kalosakas, …, A. Usheva. 2004. DNA dynamically directs its own transcription initiation. *Nucleic Acids Res.* 32:1584–1590.
40. Li, J., L. Qi, …, X. Dong. 2015. Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon Methanolobus psychrophilus. *Sci. Rep.* 5:9209.
41. Jäger, D., K. U. Förstner, …, J. N. Reeve. 2014. Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics.* 15:684.
42. Babski, J., K. A. Haas, …, J. Soppa. 2016. Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics.* 17:629.
43. Cortes, T., O. T. Schubert, …, D. B. Young. 2013. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis. *Cell Reports.* 5:1121–1131.
44. Jeong, Y., J. N. Kim, …, B. K. Cho. 2016. The dynamic transcriptional and translational landscape of the model antibiotic producer *Streptomyces coelicolor* A3(2). *Nat. Commun.* 7:11605.
45. Sharma, C. M., S. Hoffmann, …, J. Vogel. 2010. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature.* 464:250–255.
46. Kröger, C., S. C. Dillon, …, J. C. Hinton. 2012. The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl. Acad. Sci. USA.* 109:E1277–E1286.
47. Hershberg, R., G. Bejerano, …, H. Margalit. 2001. PromEC: an updated database of *Escherichia coli* mRNA promoters with experimentally identified transcriptional start sites. *Nucleic Acids Res.* 29:277.
48. Wurtzel, O., D. R. Yoder-Himes, …, S. Lory. 2012. The single-nucleotide resolution transcriptome of *Pseudomonas aeruginosa* grown in body temperature. *PLoS Pathog.* 8:e1002945.
49. Liao, Y., L. Huang, …, L. Pan. 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.
50. Albrecht, M., C. M. Sharma, …, T. Rudel. 2011. The transcriptional landscape of Chlamydia pneumoniae. *Genome Biol.* 12:R98.
51. Kopf, M., S. Klähn, …, B. Voß. 2014. Comparative analysis of the primary transcriptome of *Synechocystis* sp. PCC 6803. *DNA Res.* 21:527–539.