PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2023, **25**, 7323

Molecular dynamics simulation-based trinucleotide and tetranucleotide level structural and energy characterization of the functional units of genomic DNA†

Dinesh Sharma,[a] Kopal Sharma,[a] Akhilesh Mishra,[a] Priyanka Siwach,[b] Aditya Mittal [a] and B. Jayaram *[ac]

Received 15th October 2022,
Accepted 9th February 2023

DOI: 10.1039/d2cp04820e

rsc.li/pccp

Genomes of most organisms on earth are written in a universal language of life, made up of four units – adenine (A), thymine (T), guanine (G), and cytosine (C), and understanding the way they are put together has been a great challenge to date. Multiple efforts have been made to annotate this wonderfully engineered string of DNA using different methods but they lack a universal character. In this article, we have investigated the structural and energetic profiles of both prokaryotes and eukaryotes by considering two essential genomic sites, *viz.*, the transcription start sites (TSS) and exon–intron boundaries. We have characterized these sites by mapping the structural and energy features of DNA obtained from molecular dynamics simulations, which considers all possible trinucleotide and tetranucleotide steps. For DNA, these physicochemical properties show distinct signatures at the TSS and intron–exon boundaries. Our results firmly convey the idea that DNA uses the same dialect for prokaryotes and eukaryotes and that it is worth going beyond sequence-level analyses to physicochemical space to determine the functional destiny of DNA sequences.

Introduction

With the advent of next-generation sequencing (NGS), considerable genomic data have been generated and annotation has been a challenge. Genome annotation, commonly known as DNA annotation, identifies and assigns a suitable function to the significant elements within the DNA, such as protein-coding mRNA genes (which may further be composed of introns and exons), non-coding RNA regions, promoters, enhancers, and silencers.[1–3] Molecular biology approaches provide us with reliable sequence annotation. Compared to these experimental methods, no other alternative methods are available to provide us with the detailed annotation of genes and the other regulatory elements[4,5] to the same level of reliability. However, experimental techniques are capital-intensive and time-consuming and thus create a massive gap between sequencing and annotation.[6] Recent computation-based methods are a promising avenue for bridging this gap.[3] Computational genome annotations involve cataloguing protein-coding and non-coding genes and additional functional elements involved in gene expression and regulation. These are emerging as the preferred choices for the fast characterization of newly assembled genomes.

Over the years, scientists have been constantly working to develop various computational algorithms for the annotation of different DNA elements. Among the various sites, promoters and exon–intron boundaries are a prime focus of the present research. Promoters are one of the genome's most essential components. These elements commence the transcription process by primarily binding to the RNA polymerase (RNAP). Together with the bound polymerase, these elements undergo several changes in their overall architecture.[7] The promoter's function, however, is not restricted to starting transcription. Also, these regions aid in appropriately identifying and confirming predicted genes in genome annotation. Sequence-based methods rely on capturing a consensus sequence around promoters and such approaches have been moderately successful.[8,9] Machine learning-based promoter predictors are developed through training over extensive sequence data. These tools offer good accuracy but are highly genome specific.[10–16] Despite the many insights gained over the years from such investigations, the creation of a promoter prediction tool capable of high performance across diverse organisms is still a long way off.[6]

Eukaryotes are complex organisms that have a membrane-bound nucleus and cell organelles. In these organisms, a gene

[a] *Supercomputing Facility for Bioinformatics & Computational Biology, Kusuma School of Biological Sciences, Indian Institute of Technology, Delhi, India*
[b] *Department of Biotechnology, Chaudhary Devi Lal University, Sirsa, Haryana, India*
[c] *Department of Chemistry, Indian Institute of Technology, Delhi, India. E-mail: bjayaram@chemistry.iitd.ac.in*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d2cp04820e

is composed of two essential elements, *viz.*, introns and exons. Introns are the noncoding regions of a gene (or mRNA) that are removed before the maturation of the mRNA through a process called splicing. The segments of the DNA (or mRNA) that finally become part of the mature mRNA and thus code for the proteins are exons.[7] Like promoter identification, detecting accurate intron–exon architecture within a gene is essential and has recently received significant attention in eukaryotic genome annotation.[17] The primary focus of categorization was on the splice sites (SSs) and junction sequences. Conventionally, it includes a G–T nucleotide pair at the 5′ end of the intron, and an A–G nucleotide pair at the 3′ end.[7,17] With the continuous characterization of SSs, many sequence alterations occur at these sites. Further, the identification of the exon–intron junctions becomes challenging due to the presence of alternative splicing. These challenges significantly reduce the reliance on SSs for detecting exon–intron boundaries.[18] Computational algorithms have been developed to recognize these junctions within the genes. Some tools use a score to predict the splice site, calculated against matrices created from vast sets of splice sites.[19–22] Other approaches, such as the Genscan[23] and Genomescan,[24] compare the splice site sequence signal (at the intron–exon boundaries) to known protein sequences in an integrated manner. Programming-based tools for *ab initio* gene prediction like Genomewise,[25] Augustus,[26] Fgenesh,[27] GeneParser,[28] GeneID[29] are built using advanced computational models like HMM or Dynamic model. These methods learn the characteristics of the huge training sequences and produce good results for a particular species or organism. For a sequence not in the training dataset, their accuracy decreases considerably.

Since there exists no universal model for the identification and characterization of these genomic elements, it is necessary to come up with a chemistry-based approach to annotation. In our previous research communications,[6,17,30–34] we have elucidated the importance of the structural and energy-based profiling of the DNA elements. Likewise, other groups have also focused on capturing the structural and energy signals of various regions by considering properties like free energy, A-philicity, curvature, bendability, inter-BP properties, and DNA duplex stability under stress.[80–84] Through these structural and energetic parameter analyses, it became clear that there are unique structure profiles and energy profiles for every component within the DNA. These profiles are universal for a particular element and thus can be utilized for their efficient recognition. Though the di-nucleotide-based parameters provide a specific structure and energy description of the genomic sites in consideration, they do not take into account the neighbouring effect of the flanking nucleotides. Thus, we miss out on the adjacency effects. We have utilized higher nucleotide steps in the present study, advancing our previous work.

Compared to our previous research,[6,17] which utilizes the X-ray-based dinucleotide data of B-DNAs to profile the TSSs in prokaryotes and exon–intron boundaries in humans, in this present work, we have taken into consideration the neighbouring effects by mapping structural and energetic parameters of all the unique tri and tetra-nucleotide steps.[6] The biophysical features utilized in this research have been computed through micro-second-long molecular dynamics (MD) simulations on all the possible tri- and tetranucleotide steps. Unlike the dinucleotide-based study, the total number of B-DNA structures currently available in the Nucleotide Database (NDB) does not describe the instances of all possible tri- and tetranucleotide steps. Atomistic molecular dynamics (MD) simulations used in this study are currently the only way to obtain robust and transferable parameters.[35] The profiling distinctly delineates the TSSs in both prokaryotes and eukaryotes and the exon–intron boundaries of all the protein-coding genes in humans. These increased nucleotide steps incorporate neighboring effects from the adjacent nucleotide and have provided us with new insights into the structure and energetics of DNA, different from the dinucleotide-based characteristics. The current study is concerned with the recognition and characterisation of the physicochemical signals at the TSS and exon–intron transitions. Our future aim is to develop a method for predicting these sites; hence, it is currently not viable to benchmark the current prediction algorithms. However, a thorough comparison of the three dimensions, *viz.*, sequence, energy and structure-based analysis made here, will strengthen the importance of using the physicochemical-based approaches as a ubiquitous model for annotation.

For this study, the influence of the neighbouring step has been accounted on the base pair and backbone structure/dynamics, and thus for more accurate calculations, these parameters (nine backbone and four BP-axes) have been mapped over all the possible trinucleotides (64), while the six intra-BP and three energy parameters were considered for all the unique trinucleotides (32). The inter-BP parameters have been computed over all the unique tetra-nucleotide steps (136) (represented in the parameter details file, ESI†).[78,79] The numerical values for the structural parameters are extracted from the μ-second long MD simulations, while in-house software is used to calculate the energy parameters. The sequence datasets comprise 16 519 and 197 356 primary TSS sites for promoter analysis in prokaryotes and eukaryotes, respectively. For a similar characterization of exon–intron in the human genome, we have used ∼0.33 million exon–intron boundaries for the exon start site and exon end site. Users can download the data from: **https://www.scfbio-iitd.res.in/Tri_Tetra/data.html**. The structural and energy-based descriptions of these DNA elements give us unique signatures for their characterization. To establish the universality of these biophysical signals, a fair comparison at the same sites has been made with a few widely used sequence-based genome annotation approaches (described in the Methods section). Our results firmly convey that the structural and energy signal patterns can undoubtedly be hidden signals within the DNA of both the prokaryotes and eukaryotes through which the genome conveys information about its functional features.

Experimental

Materials and methods

Parameters for characterizing genomic sequences. For the present study, we have considered 28 parameters. These include 25 structural and 3 energy variables. Among the structural

parameters are nine backbone parameters (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Chi, Phase and Amplitude), six intra-BP (Shear, Stretch, Stagger, Buckle, Propel and Opening), and four BP-axis (X-displacement, Y-displacement, Inclination and Tip), which have been mapped over the 64 possible trinucleotide steps. The values of the six inter-BP step structural parameters (Shift, Slide, Rise, Roll, Twist and Tilt) have been defined for all the 136 unique tetra-nucleotide steps. The values of these 25 parameters were obtained by averaging the last 500 ns from microsecond-long MD simulations following the methodology in ref. 35. Here, 13 oligomers, each with a length of 18 bp and GC terminals on each end were used. The reduced number of oligonucleotides relative to earlier training libraries makes it more feasible to produce multi-microsecond trajectories under different simulation environments.[35] These oligonucleotides containing all the instances of possible trinucleotides (64) and unique tetranucleotides (136 of a total of 256 tetranucleotides) were created using the leap program of AMBERTOOLS[36] and simulated using the pmemd.cuda[37] code from AMBER14.[36] Canonical duplexes for these oligonucleotides were generated using the Arnott B-DNA fiber parameters.[38] These were then solvated using an SPC/E[39] water model, maintaining a minimum distance of 10 Å from the edge of the defined box. The systems were neutralized by adding K^+Cl^- or Na^+Cl^- ions (150 mM), using the PARMBSC1[40] force field and Dang's parameters[41] to describe the DNA and the ions, respectively. Systems were then simulated in the *NPT* ensemble by using Particle–mesh Ewald corrections[42] and periodic boundary conditions for 1 μs. Bonds involving hydrogen were constrained using SHAKE.[43] The trajectory files are available at the BIGNASim[44] Server: **https://mmb.irbbarcelona.org/BIGNASim/**. The trajectories were processed using the cpptraj[45] module of AMBERTOOLS[36] and NAFlex.[46] CURVES+ and CANAL programs[47] were used to measure and analyze the helical parameters and backbone torsional angles as per the ABC (Ascona B-DNA Consortium) standards.[48–51] Hydrogen bond energy, stacking energy and solvation energy constitute the energy parameters included in our study, and their values were derived as per the methodology reported in our previous work.[52] File 1 (ESI†) contains the values of all 28 parameters for the unique tri and tetra-nucleotide steps that were computed above. All numerical conversions in this study have been made using this Table.

Sequence datasets for the various studies. For the promoter analysis, 16 519 primary TSS sites[53–64] from 12 bacterial species belonging to 6 different phyla (Table 1) were retrieved to provide the sequence, structural, and energetic profiles around the TSS in prokaryotes. For a consensus sequence-based analysis, sequences of 101 nucleotides in length, covering −95 to +5 nucleotide positions, with the actual TSS at 0, were extracted from respective genome sequences. For the structural and energy-based characterization, sequences of 1001 nucleotides in length surrounding each selected TSS were retrieved from their respective genome sequences (from −500 to +500 nucleotides, with TSS positioned at 0). Similarly, to compare the profiles, Coding Sequences (CDS) with a length of >1500 nucleotides from the 12 microorganisms were considered to generate a control set of 6218 CDSs; a central section of 1001 nucleotides from these CDS sequences was considered for structural and energetic comparison with the TSS regions. Eukaryotic promoters were subjected to a similar investigation. For the characterization, 197 356 primary TSS sites[65] from all the chromosomes of 8 yeast species, covering both the Hemiascomycetes (budding yeast) and Schizosaccharomycetes (fission yeast), were retrieved from the YeasTSS web server[65] (Table 2 and Table S1, ESI†). Here, a 101-nucleotide-long sequence dataset, spanning −80 to +20 with TSS at 0, for the consensus study and a 1001 nucleotide long TSS, from −500 to +500 with TSS at 0 and CDS sequence dataset (comprising 16 715 CDS sequences, satisfying the >1500 nucleotide length condition) for the biophysical profiling were obtained following the exact methodology of prokaryotes.

To characterize the intron–exon boundary junctions, the human genome annotation file was retrieved from the GENCODE database. From this file, the start and end positions of 328 368 exons from all the protein-coding genes were considered. A series of exon–intron boundary sequence datasets were constructed using these genomic locations. For the sequence study, two datasets, each containing sequences of 51-nucleotide length, were created using the exon-start and exon-end, placed

Table 1 Prokaryotic species studied along with their TSS and CDS data

Phylum	Species	Genome size (Mb)	Number of primary TSS	Number of CDS
Euryarchaeota	*Methanolobus psychrophilus*	3.07	1463	355
	Thermococcus kodakarensis	2.08	1248	208
	Halofrex volcanii	3.93	1723	425
Actinobacteria	*Mycobacterium tuberculosis* H37Rv	4.38	1440	626
	Streptomyces coelicolor A3	9.05	2771	1201
Proteobacteria	*Helicobacter pylori*	1.63	227	227
	Salmonella enterica serovar Typhimurium	5.067	1871	624
	Escherichia coli	5.17	1222	577
	Pseudomonas aeruginosa PA14	6.58	2118	853
Firmicutes	*Bacillus amyloliquefaciens*	3.95	1062	393
Chlamydiae	*Chlamydia pneumonia* CWL029	1.22	357	198
Cyanobacteria	*Synechocystis* sp. PCC6803	3.57	430	531
Total			16 519	6218

Table 2 Yeast species investigated along with their TSS and CDS sites

Group	Species	Chromosomes	Number of TSS	Number of CDS
Budding yeast	*Candida albicans*	Chr1, Chr2, Chr3, Chr4, Chr5, Chr6, Chr7 and ChrR	26 545	2329
	Kluyveromyces lactis	ChrA, ChrB, ChrC, ChrD, ChrE and ChrF	34 655	1910
	Lachancea kluyveri	ChrA, ChrB, ChrC, ChrD, ChrE, ChrF, ChrG and ChrH	17 411	2086
	Naumovozyma castellii	Chr1, Chr2, Chr3, Chr4, Chr5, Chr6, Chr7, Chr8, Chr9 and Chr10	19 189	2144
	Saccharomyces cerevisiae	ChrI, ChrII, ChrIII, ChrIV, ChrV, ChrVI, ChrVII, ChrVIII, ChrIX, ChrX, ChrXI, ChrXII, ChrXIII, ChrXIV, ChrXV and ChrXVI	17 925	2223
	Saccharomyces paradoxus	ChrI, ChrII, ChrIII, ChrIV, ChrV, ChrVI, ChrVII, ChrVIII, ChrIX, ChrX, ChrXI, ChrXII, ChrXIII, ChrXIV, ChrXV and ChrXVI	29 690	2184
	Yarrowia lipolytica	ChrA, ChrB, ChrC, ChrD, ChrE and ChrF	27 793	2365
Fission yeast	*Schizosaccharomyces pombe*	ChrI, ChrII and ChrIII	24 148	1474
Total			197 356	16 715

at the 26th position, respectively. These datasets were organised based on the location of the genes on chromosomes (Table S2, ESI†) and were then used to identify a consensus sequence around the positions of concern. Different from the sequence datasets, two new datasets were created for the structural and energy-based characterization of the exon start and exon end regions, each containing 328 368 exon–intron boundary sequences. Dataset I had sequences of 401 nucleotides in length; these sequences were created by extracting 200 nucleotides upstream (representing the exon sequence) and downstream (representing the intron sequence) from the exon end position placed at 0. Similar to Dataset I, Dataset II was generated by considering the exon start positions; these sequences represent intron and exon sequences from −200 to −1 and +1 to +200, respectively. For the control dataset, we extracted sequences with similar lengths from the middle region of the exons that were longer than 1000 nucleotides (30 140 out of 328 368).

Sequence, structural and energy profiling. A consensus sequence around TSS for each prokaryotic and eukaryotic species was obtained for the 101 nucleotide long extracted regions using WebLogo3 software.[66] Likewise, to identify a consensus sequence around the exon start and exon end sites, we used the Jalview software[67] on a randomly selected gene of each chromosome from the sequence datasets comprising the 51 nucleotides. Based on the results and to get an exact consensus, we considered 2 trimer and 2 pentamer motifs *viz.* −1, 0, 1 and −2, −1, 0 for trimer motifs and −2, −1, 0, +1, +2 and −4, −3, −2, −1, 0 for pentamer motifs (0 being the exon start or end position, and motif position selected after trying various combinations). To get a comprehensive picture, we also did a position-specific percentage analysis of the undecamer sequence at the start and end sites.

All relevant sequence datasets were analyzed to obtain the structural and energy profiles corresponding to the TSS in prokaryotes and eukaryotes and intron–exon interfaces in the human genome. These datasets include the 1001 nucleotide long sequences containing TSS in prokaryotes and eukaryotes and the 401 nucleotide long sequences of Dataset I and Dataset II obtained from the exon-end and exon-start sites, respectively. A sliding window of one nucleotide was used to cover the whole sequence. The values of the trinucleotide and tetra-nucleotide parameters were utilized at each sliding window to convert the sequence into 28 numerical series. These numerical profiles of all sequences relating to a specific parameter were averaged for each position after applying a sliding window of 25 base pairs. A similar methodology was applied to the control sequences (CDS) in each category. The resulting 28 averaged profiles (corresponding to each parameter) were then used to plot and visualize the structural and energetic changes occurring around the TSS (in prokaryotes and eukaryotes) and exon–intron boundary sites.

Examining individual sequences for signal threshold. To explore the granularity of the signal trend in individual sequences, a threshold analysis was done. The threshold analysis validates the occurrence of patterns, as evident through the average plots corresponding to each parameter on individual sequences.

For both prokaryotes and eukaryotes, respective sequence datasets containing TSS were considered for each parameter. Based on the location of the pattern observed in the average plot of a particular parameter, positive (TSS) and negative (CDS) vectors were created from the individual sequences. For the positive vector, a window of 100 spanning −50 to +50 with TSS at 0 was considered. For the negative vector, a similar-sized window was taken after traversing 200 nucleotides downstream of the TSS.

In parallel, for a threshold analysis on the intron–exon boundaries, discrete sequences of the exon-start and exon-end datasets for each parameter were used. A 60-nucleotide long vector spanning −30 to +30 was extracted around the exon start and end positions to create the positive set for the respective group. To create a negative set corresponding to each position and sequence within both exon-start and exon-end datasets, we traversed 150 nucleotides from the positive position towards the exon region and extracted a similar length vector. For each pair of positive and negative sets, the area enclosed between them was compared with different threshold values for each parameter in the above-mentioned respective datasets. The idea is that those pairs in which the area confined within the vectors are small, *i.e.*, less than two standard deviations from the mean, and will have indistinguishable signals but the remaining pairs will have distinct profiles with significant magnitudes. As a result,

signals that fulfil the two standard deviation threshold criteria indicate that a TSS signal or exon–intron boundary signal is present in those specific sequences. The area under the curve calculations and the evaluation of thresholds are specified in the methodology in S1a and S1b and in Fig. S12–S15 (ESI†).

Normalization of parameter values. Using normalization, the values of all the parameters were made dimensionless.

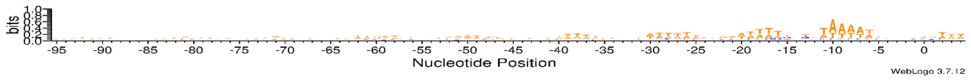

Fig. 1 Sequence consensus for *Helicobacter pylori*, a prokaryote (as obtained using WebLogo software), from position −95 to +5 with respect to TSS at 0. Some consensus is evident at position −10. However, on comparing the sequence consensus from all 12 organisms (ESI†), no specific pattern emerges.

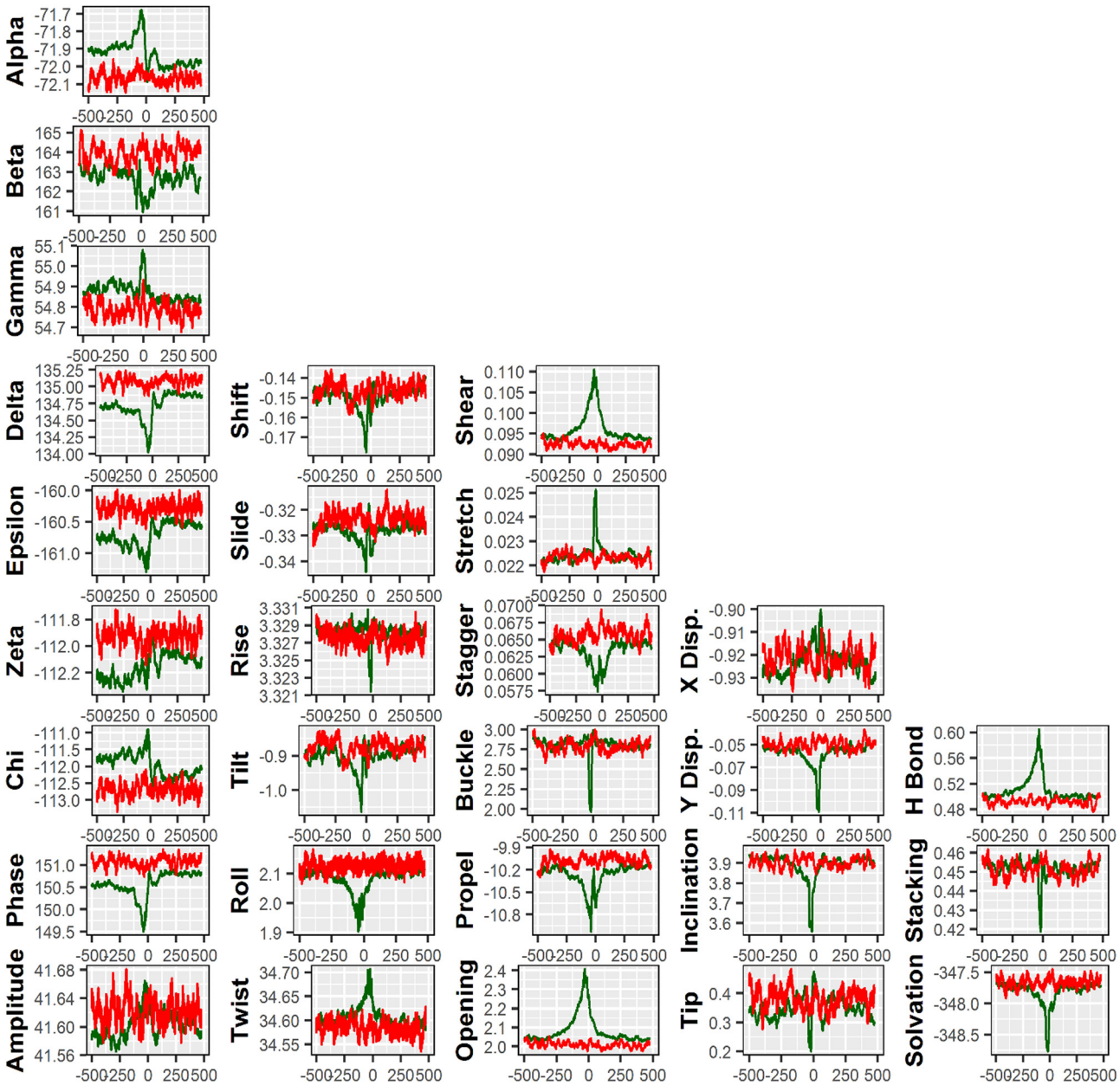

Fig. 2 Structure and energy profiles of *Helicobacter pylori*, a prokaryote. TSS sequences are shown in green, whereas red lines represent CDSs. The numeric value of the parameter is represented by the ordinate, while the nucleotide position relative to the TSS is shown by the abscissa. Parameters are represented in the parameter details file (ESI†). The correlation among the parameters (for the entire dataset) and the important features are present in the ESI.†

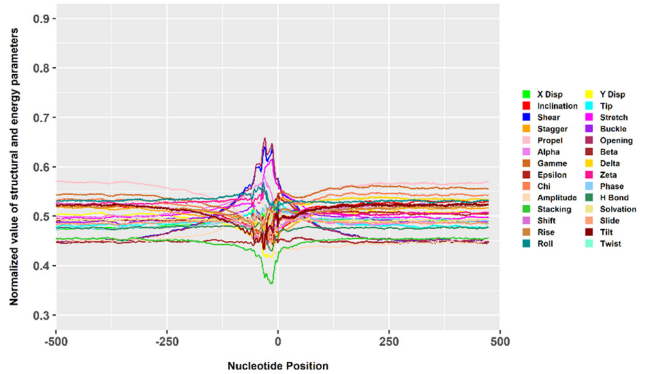

Fig. 3 Normalised and combined structure and energy graph for all sequences (16 519) containing TSS from 12 prokaryotic organisms.

These normalized values were used to generate a combined single plot of all structural and energy parameters.

Results and discussion

The profiling of the promoters and the intron–exon boundaries was carried out by inspecting three dimensions, *viz.*, sequence, structure, and energy. The results are detailed below in two separate sections corresponding to each element.

Promoters in both prokaryotes and eukaryotes present unique signals at the TSS

In prokaryotes, we considered the TSS containing promoter sequences from 12 different organisms belonging to archaebacteria and eubacteria. The consensus analysis was carried out around the TSS on the nucleotide sequences. For each organism, sequences were aligned from position −95 to +5, with the TSS at 0, and a consensus was derived using the WebLogo3 software. Fig. 1 shows a graphical representation of the consensus sequence for *Helicobacter pylori* (the results for the 11 other organisms are available in Fig. S1, ESI†). These graphs show that there is a consensus sequence specific to each organism, and no universal pattern is apparent. It is also well known that there is considerable promiscuity at the sequence level. Since it is evident from contemporary studies related to transcription that RNA polymerase's binding event and mechanism of action remain the same throughout the prokaryotic species, a universal signal within the DNA must exist for the precise binding of these enzymes to the promoter sites. To reveal this hidden signal, which is not evident from the consensus analysis of the combined sequences, we investigated the structural and energy profiles of the promoter sequences.

Our Lab has been investigating the energetics of DNA for the past 20 years, and we have demonstrated the importance of solvation, stacking, and hydrogen bond energy in identifying the distinct genomic functional units.[30–34,52,68,69] Apart from the energy parameters, in recent years, we have explored the complete structural profile of DNA in and around various important motifs by taking into consideration the individual parameters belonging to four major DNA structural categories.[6,17,30] For each energy and structural parameter, individual TSS and CDS sequences from respective organisms were converted to their numerical profiles; these were then averaged at each position and were used for plotting (Fig. 2 and 3). Fig. 2 shows that for all the parameters, a unique pattern is observed around the TSS (present at the 501th position and marked as 0 on the abscissa in the graphs) in comparison to the CDS and the extreme upstream and downstream regions of the TSS sequences for *Helicobacter pylori* (results for the remaining species are available in Fig. S2, ESI†). Fig. 3 is the combined graph for the structural and energy patterns obtained for all the sequences from all 12 organisms. These results support the scientific theories stating that the overall stability of DNA sequences, both evolutionary and thermodynamic, decreases near the promoter region.[30,70–74] The energy profiles in Fig. 2 and 3 show that the hydrogen bond energy is increasing, thereby reducing the melting temperature of DNA at the TSS site. The graph of stacking energy shows that the DNA at the TSS region becomes stiffer to provide a stable platform for RNA polymerase binding. The solvation energy is reduced as DNA makes a tertiary structure near the TSS region. A similar trend is observed for the structural parameters while proceeding toward the TSS region. Out of the total 25 structural parameters, an overall rise at or around the TSS is noticeable for the 10 parameters (Alpha, Gamma, Zeta, Amplitude, Twist, Shear, Opening, X-displacement, and Tip). The other 15 properties (Beta, Delta, Epsilon, Chi, Phase, Shift, Slide, Tilt, Roll, Stagger, Buckle, Propel, Y-displacement, and Inclination) show a decrease in their pattern at the TSS region. Different from our previous study related to di-nucleotides (Fig. S3, ESI†), here, 11 parameters show a different pattern of values upstream and downstream of the TSS, which may be a result of the neighbouring effect, which was not pronounced in the dinucleotide motif dataset; this may be helpful in better predicting the TSS sites.[6]

To strengthen the notion, we performed a similar sequence and biophysical profiling of the promoter sequences from eukaryotes. For the analysis, we considered 7 budding yeasts and 1 fission yeast. A methodology similar to the prokaryotic promoter characterization was followed. Considering the TSS position at 0, to obtain the consensus, we extracted 101 nucleotide length sequences from position −80 to +20. The positions here

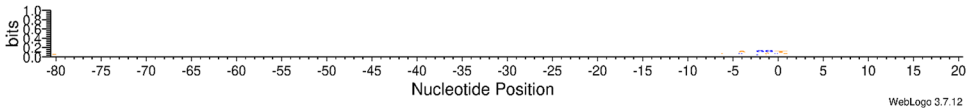

Fig. 4 Sequence consensus for *Kluyveromyces lactis*, a eukaryote (using WEbLogo software), from position −80 to +20 with respect to the TSS at 0. No specific consensus is evident around the TSS for any species (ESI†).

are based on the promoter architecture in yeast[75] and thus differ from what we considered for the prokaryotes. Using the Weblogo3 software, we obtained the consensus for *Kluyveromyces lactis* (Fig. 4, consensus for the rest of the species is in Fig. S4, ESI†). These figures show little or no consensus at the promoter. Just like the prokaryotes, some consensus does occur at the organism level and fades out when all the sequences are considered. The consensus sequence-based approach is thus organism-specific and fails to establish a universal signal for promoter profiling. Structural and energy-based characterization of the eukaryotic promoter sequences reveals distinguishing patterns around the TSS in all species (Fig. 5, 6 and Fig. S5, ESI†). Fig. 5 depicts the structure and energy profile of *Kluyveromyces lactis*, while Fig. 6 shows the trend specific to each parameter, mapped over the combined TSS sequences and compared to the CDS sequences from all the organisms. For each parameter, a specific pattern is evident at the TSS. A sharp increasing or decreasing signal at the TSS following a low intense signal around −300 to −400 can be seen for each feature. It can be hypothesized that these characteristic signals are due to the presence of important motifs prior to the actual promoter; however, a thorough investigation is essential to validate the finding. Comparing the signals from prokaryotes with those of eukaryotes suggests that the trend for each parameter is not the same. Here, 11 structural parameters

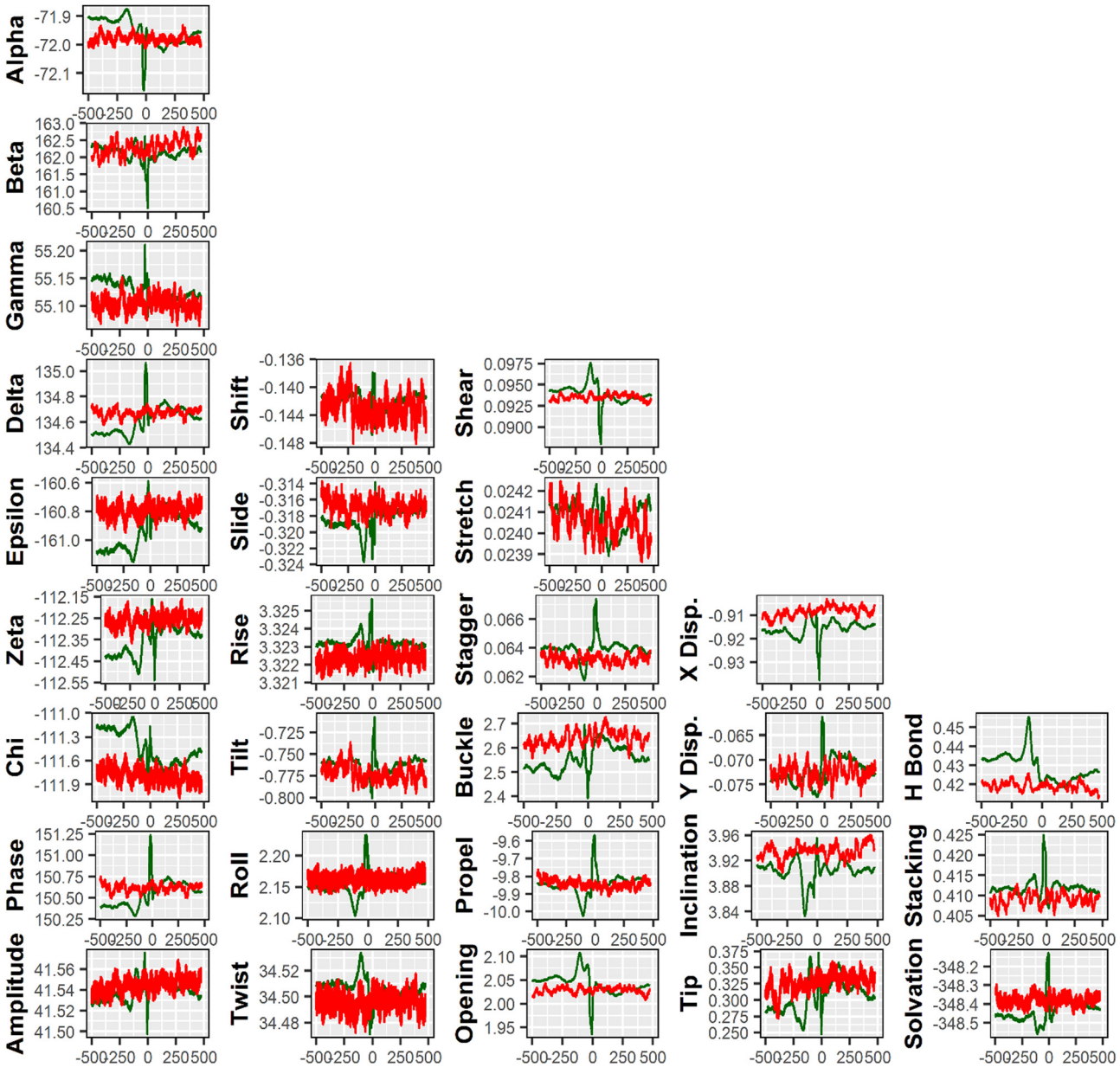

Fig. 5 Structure and energy profile of *Kluyveromyces lactis, a eukaryote*. TSS sequences are shown in green, whereas red lines represent CDSs. The numeric value of the parameter is represented by the ordinate, while the nucleotide position relative to TSS is shown by the abscissa. Parameters are represented in the parameter details file (ESI†). The correlation among the parameters (for the entire dataset) and the important features are present in the ESI.†

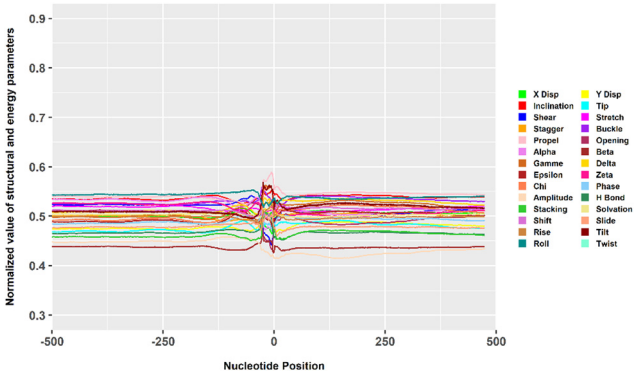

Fig. 6 Normalised and combined structure and energy graph for all sequences (197 356) containing TSS from the 8 eukaryotic species.

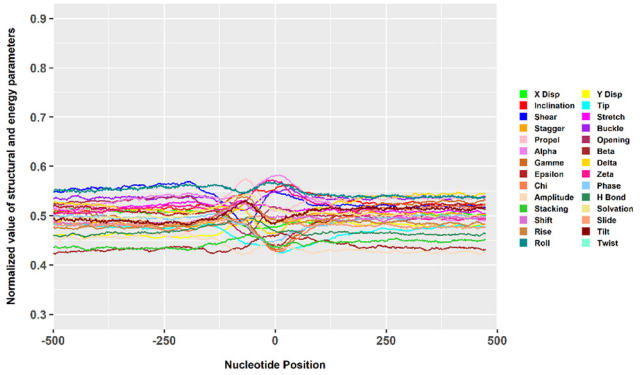

Fig. 7 Normalised and combined structure and energy graph for all sequences (5561) containing TSS from *Caenorhabditis elegans*.

(Alpha, Beta, Epsilon, Zeta, Amplitude, Slide, Twist, Shear, Stretch, Opening, and X-displacement) show decreasing behaviour at the TSS. An increasing nature at the TSS is evident for the rest of the 14 parameters (Gamma, Delta, Chi, Phase, Shift, Rise, Tilt, Roll, Stagger, Buckle, Propel, Y-displacement, Inclination, and Tip). This different behaviour of parameters can be attributed to the complex nature of eukaryotic promoters, which in the case of prokaryotes, have minimal motifs and regulatory elements; further investigation is necessary to comment on this contrasting behaviour of parameters. With the idea that a signal is said to be indispensable for usage only if it is captured at the basal level, we characterized the sequences on the chromosome level. Fig. S6 (ESI†) indicates that the structural and energy signals are also evident at the chromosome level in all the yeast species. Since yeasts are simple unicellular eukaryotic organisms that form a connecting link between the prokaryotes and the higher eukaryotes, it is necessary to characterize a multicellular higher eukaryote to see the pertinence of our approach. *Caenorhabditis elegans* is a common eukaryotic multicellular experimental model in biology.[76] For the structural and energy characterization of this multicellular eukaryote, we obtained the potential TSS sites for all the chromosomes of *Caenorhabditis elegans*[77] (TSS site information is in Table S7, ESI†). The CDS sites were also retrieved from the respective genome annotation files. Following the biophysical-based prokaryotic and yeast promoter characterization methodology, these sites were used for obtaining the 1001 nucleotide length sequences. The structural and energy profiling at the TSS of *Caenorhabditis elegans* for the combined sequences and chromosome sequences are presented in Fig. 7 and Fig. S7, S8 (ESI†). The signals are very smooth and the parameter plots reveal that the structural and energy profiles change at the TSS. The weak intensities of the signal can be attributed to the complexity of the higher eukaryotic promoter sites or the lower TSS sites known and considered for the study. However, the presence of a pattern corresponding to each parameter at the TSS sites indicates the robustness of the characterization approach. Based on the various results from the promoter characterization in both prokaryotes and eukaryotes, the structural and energy-based characterization approach outperformed the consensus sequence-based approaches, which tend to be organism-specific and non-universal. The promoter characterization for both the prokaryotes and the eukaryotes delineates the fact that several sequence variations occur at the TSS sites, and these cannot be followed for their efficient recognition. Despite sequence alterations at these sites, the physicochemical profiles for these regions are conserved and can be exploited for comprehensive annotations. Since the figures presented so far are the average plots (for all sequences throughout all organisms in prokaryotes and eukaryotes, all sequences at the organism level in prokaryotes and eukaryotes, and all sequences at the chromosome level in the case of eukaryotes only), it is crucial to know the comprehensiveness and sensitivity of this study on individual sequences. On applying the threshold methodology as explained in the Methods section, it was observed that for the specified position, where the trend was evident in the averaged plots, $>98\%$ of sequences in both prokaryotes and eukaryotes followed a similar trend. These results are detailed in Tables S3 and S4 and Fig. S12 and S13 (ESI†), with the graphs depicting the area enclosed for each pair of TSS and CDS vectors from prokaryotes and eukaryotes (a distribution plot of the area calculated for each pair of TSS vector and CDS vector is infeasible given the large volume of data).

Structural and energy profiles at intron–exon junctions

To widen the scope of our physicochemical characterization and to highlight its universality at all levels of the genome, a study similar to the promoters, incorporating the sequence, structure and energy characterization, was carried out for the more complex elements. Intron–exon profiling was done for the protein-coding genes in humans following the promoter characterization approach.

Using the Jalview software and the annotation dataset acquired from GENCODE, we conducted a consensus analysis of 328 365 human exon start/end sites. To achieve a particular sequence consensus, we characterized the exon start and exon end regions of a randomly selected gene from each chromosome. The results of the consensus analysis are presented in Fig. S8 (ESI†). From the results at the exon start and exon end, it is evident that there is a consensus, most likely in the form of a trimer or pentamer unit. To reveal these consensuses, we

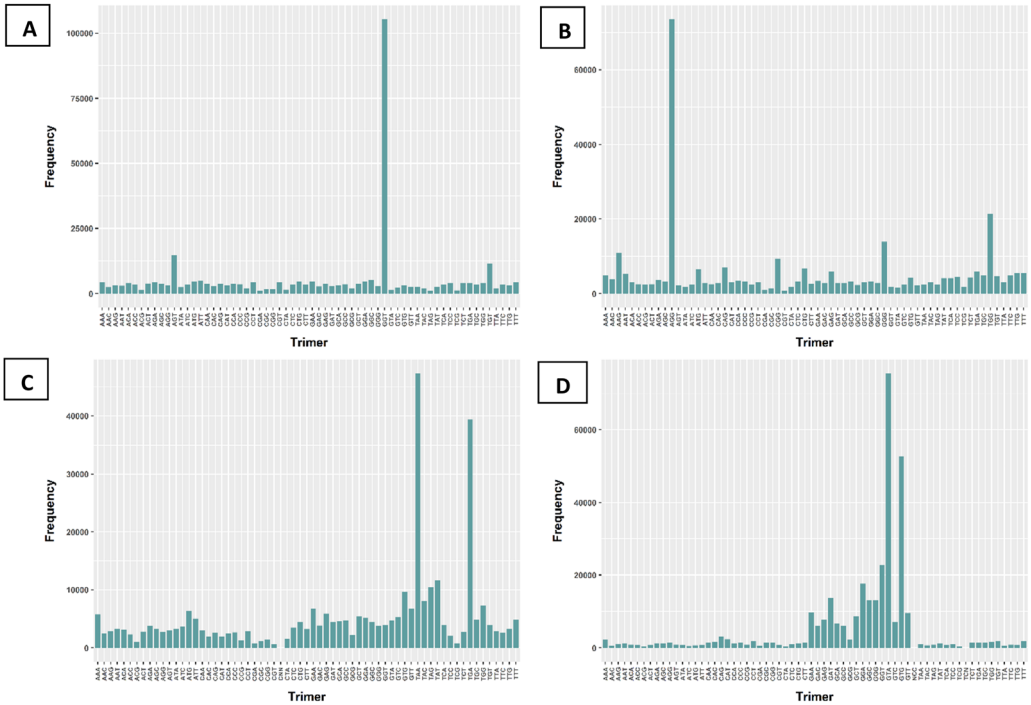

Fig. 8 Trimer frequency analysis (for all 328 364 sequences) at the exon start and exon end. (A and B) Represent trimer frequency at the exon start for position coordinates −1, 0, 1 and −2, −1, 0 respectively. (C and D) Are the trimer frequency at the exon end with position coordinates −1, 0, 1 and −2, −1, 0 respectively. The ordinate represents the frequency of a particular trimer, whereas the abscissa represents all the possible trimers (64).

carried out trimer and pentamer motif frequency analyses for the 51-nucleotide-long sequence dataset by placing the exon start/end site at the 26th position (considered as the 0th position for reference).

For the trimer motif analysis, we considered −1, 0, +1 and −2, −1, 0 positions and for the pentamer −2, −1, 0, 1, 2 and −4, −3, −2, −1, 0 positions were selected. From the frequency analysis of trinucleotide motifs (Fig. 8), the occurrences of TAA and TGA were higher at the −1, 0, and +1 positions than any other trimer, while GTA and GTG were higher at the −2, −1, and 0 positions of the exon-end sites, confirming the conventional wisdom of the occurrence of these nucleotides at the said sites.[7–17] These motifs undoubtedly have the potential to aid in the prediction of exon-end sites. However, taking into consideration the total size (328 364) of the data, the frequency of occurrence is not high, and further, there is no consistent motif to rely on. Contrary to this, trinucleotide motif-based exon-start site analysis at position −1, 0, 1 clearly showed a high occurrence of GGT trimer, while at position −2, −1, 0, there was a higher occurrence of the AGG trimer motif; the frequency, however, is still low considering the size of the data. A similar frequency analysis was performed at both sites for two pentanucleotide motifs with positions −2, −1, 0, 1, 2 and −4, −3, −2, −1, 0. The results show (Fig. 9) that the motif AGGTA (−4, −3, −2, −1, 0) has a higher frequency, and is better for predicting the exon end sites. At the exon start site, there was again a higher occurrence of the AGGTA pentanucleotide for the motif −2, −1, 0, 1, 2, facilitating the prediction of the concerned sites and confirming the conventional belief of the occurrence of the A–G nucleotide at the 3′ end of the intron.[7–17] To gain deeper insight into the nucleotide sequence pattern frequency at the exon–intron boundaries, an undecamer position motif extending from −5 to +5, with the exon start/end site positioned at 0, was considered (Fig. 10). The probability of the occurrence of A, T, G and C at each position was obtained. We found that G, G, and A were present with more than 50% probability at positions −4, 0 and 1, respectively, at the exon start. At the exon end, at only the −2 position, we had a greater than 50% occurrence of G. From the sequence-based analysis, various trinucleotide and pentanucleotide motifs occur frequently at the exon-start and exon-end regions. However, there was no specific consensus at these sites, and the occurrence is not uniform at the exon-start and exon-end sites. Since we could not find specific motif sequences with an absolute high consensus, we extended our study to the structural and energetic profiling of the exon–intron boundary elements by considering two separate datasets for exon start and exon end positions, respectively.

Using the trinucleotide and tetranucleotide parameter table, the 328 368-nucleotide sequences in both the exon-start position dataset and the exon-end position dataset, just like the promoter sequences, were converted to 25 structural and 3 energy numerical strings. For each parameter, these numerical profiles were averaged over all sequences and plotted. From the structure and energy description provided in Fig. 11 and 12 for exon start and exon end, respectively, the hydrogen bond energy showed a sudden increase in its value followed by a crest. It can be inferred from this trend that the structure at the boundary position initially became unstable and then restores

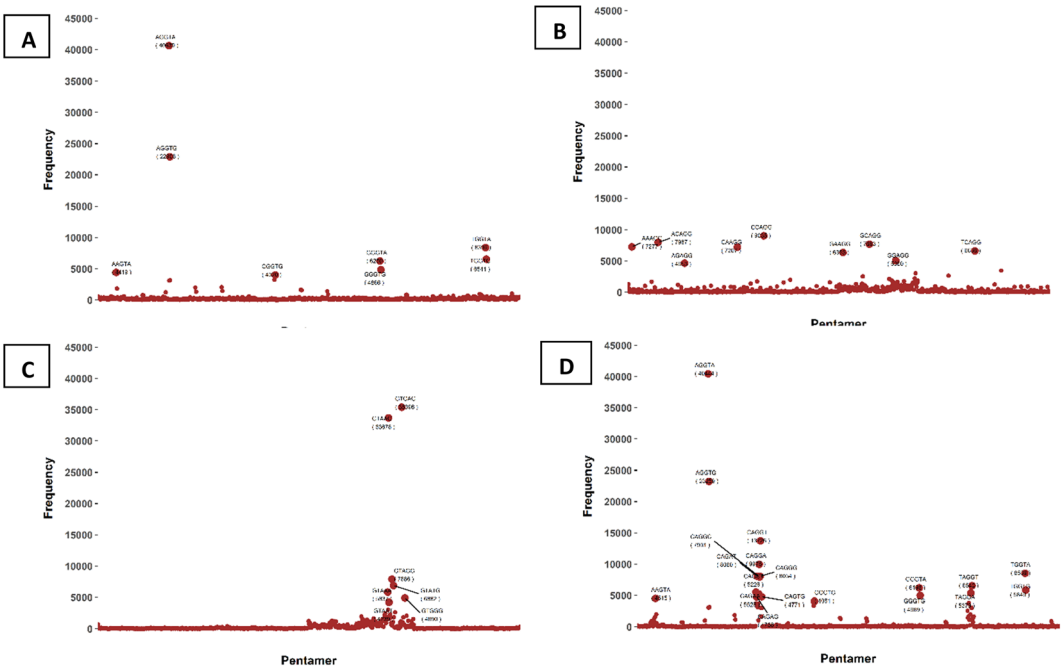

Fig. 9 Pentamer frequency analysis (for all 328 364 sequences) at the exon start and exon end. (A and B) Represent pentamer frequency at the exon start for position coordinates −2, −1, 0, 1, 2, and −4, −3, −2, −1, 0, respectively. (C and D) Represent pentamer frequency at the exon end for position coordinates −2, −1, 0, 1, 2, and −4, −3, −2, −1, 0 respectively. The ordinate represents the frequency of a particular pentamer, whereas the abscissa represents all the possible pentamers (1024).

its balance with the passage of the start/end sites. The stacking energy showed a peak at the boundary junction, thus decreasing the stiffness of DNA and making it more flexible. A sharp drop in solvation energy implies the formation of a secondary structure at the exon–intron boundaries. Contrary to these observations in the energy parameters, the middle portion of the exons (in red) has a constant energy profile throughout their lengths. These energy changes at the exon–intron junctions support the notion that the boundary elements are involved in the formation of a secondary structure in the RNA space, which facilitates the splicing event, and the DNA analyses here anticipate these events.

Since energy changes are not entirely responsible for bringing out the changes occurring at various sites in the genome, it is their union with the structural features that bring about any functional changes within the DNA. The structural profile of the exon start/end sites also supports the notion of the formation of a secondary structure. From the plots, it is evident that for all the parameters, there is a signature pattern at the exon start and end sites (in green) in contrast to the middle region of exons (represented in red), which mostly have a constant nature. Compared to the gradual rise and fall, as seen for most of the parameters near TSS regions, a combination of sharp peaks and clefts is observed at the intron–exon boundaries. For the 13 parameters (Alpha, Beta, Gamma, Chi, Slide, Shear, Buckle, Opening, Y-displacement, Inclination, Twist, Stretch, and X-displacement), a peak followed by a crest is evident from the sequences containing exon–intron

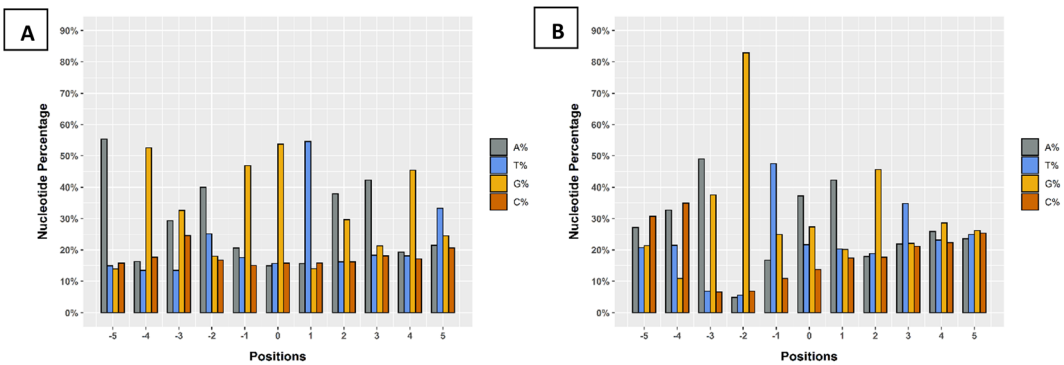

Fig. 10 Undecamer analysis (for all 328 364 sequences) at position −5 to +5 for (A) exon start and (B) exon end. The ordinate represents the frequency of A, T, G, or C, whereas the abscissa represents the nucleotide position relative to exon start/end site (at 0).

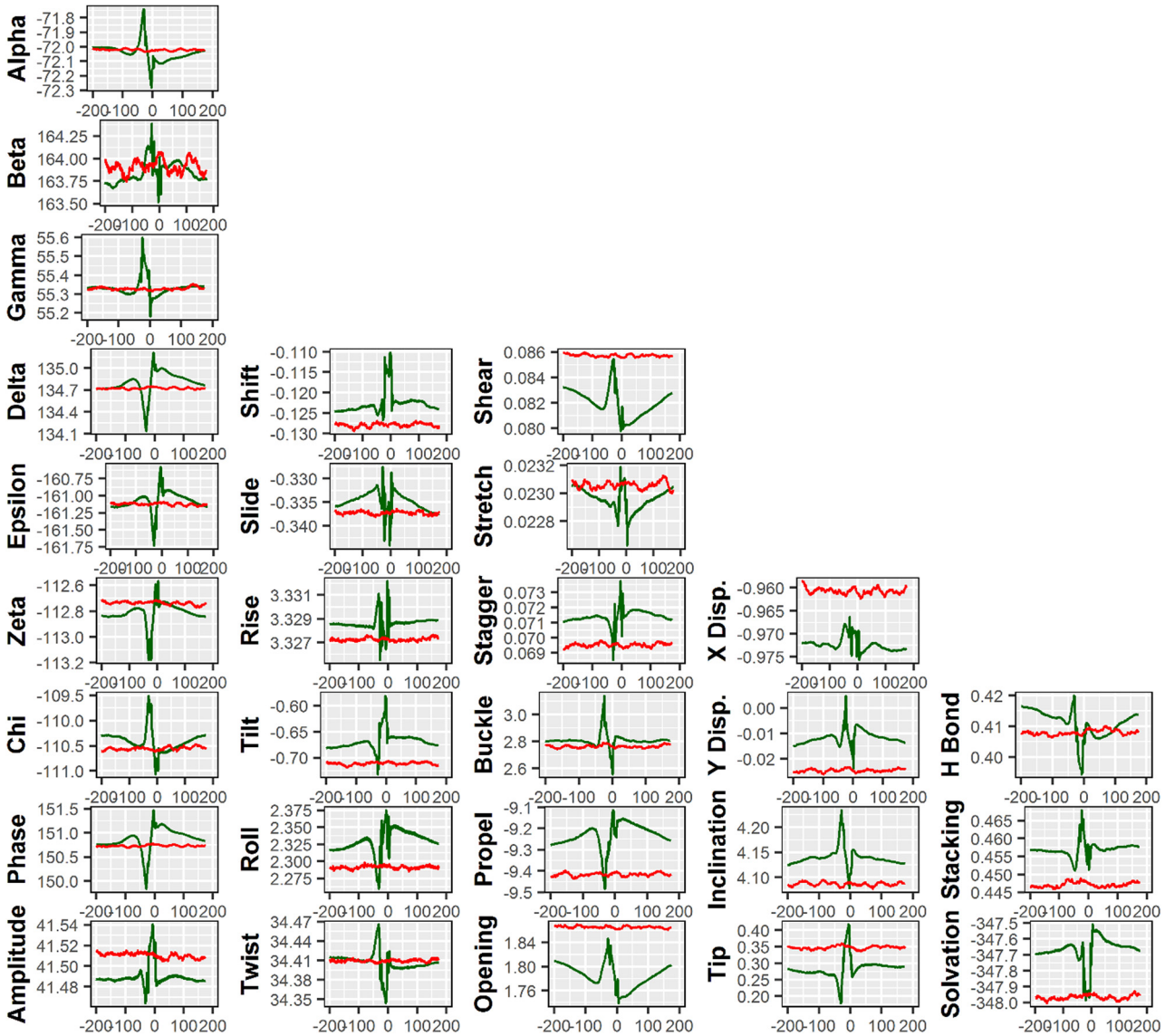

Fig. 11 Structure and energy profiles at exon start sites (for all 328 364 sequences). Sequences containing exon start sites are shown in green, whereas red lines represent CDSs. The numeric value of the parameter is represented by the ordinate, while the nucleotide position relative to the exon start is shown by the abscissa. Parameters are presented in the parameter details file (ESI†). The correlation among the parameters (for the entire dataset) and the important features are presented in the ESI.†

boundaries and a crest followed by a peak for the remaining 12 parameters (Delta, Epsilon, Zeta, Phase, Amplitude, Tilt, Stagger, Propel, Tip, Rise, Roll, and Shift). An absolute rise and fall were not observed for any of the parameters. The adjacency effect is evident for some of the parameters on comparing the exon start and exon end profiles with the dinucleotide-based profiles (Fig. S11, ESI†). These observations highlight the fact that the overall structural change is sudden at the boundary element of the exon and intron, and this change is not carried over larger distances. The observation made from all the physicochemical profiles emphasizes the fact that the intron–exon boundary elements are solely involved in the splicing event, and the structural and energy changes occurring in these regions facilitate the event. From all the graphs, it is apparent that the patterns of these parameters are highly similar at both the exon start and end sites. The threshold analysis results for the exon–intron and intron–exon junctions are presented in Tables S5, S6 and Fig. S13, S14 (ESI†). The analysis shows that for each parameter, the signal is evident for $>98\%$ of sequences and thus supports the notion that the signals are universal.

Conclusions

The molecular dynamics-derived trinucleotide and tetranucleotide-based structure and energy parameter characterization of TSS and exon–intron boundaries of both prokaryotes and eukaryotes point to the existence of a signature profile for these elements. Our

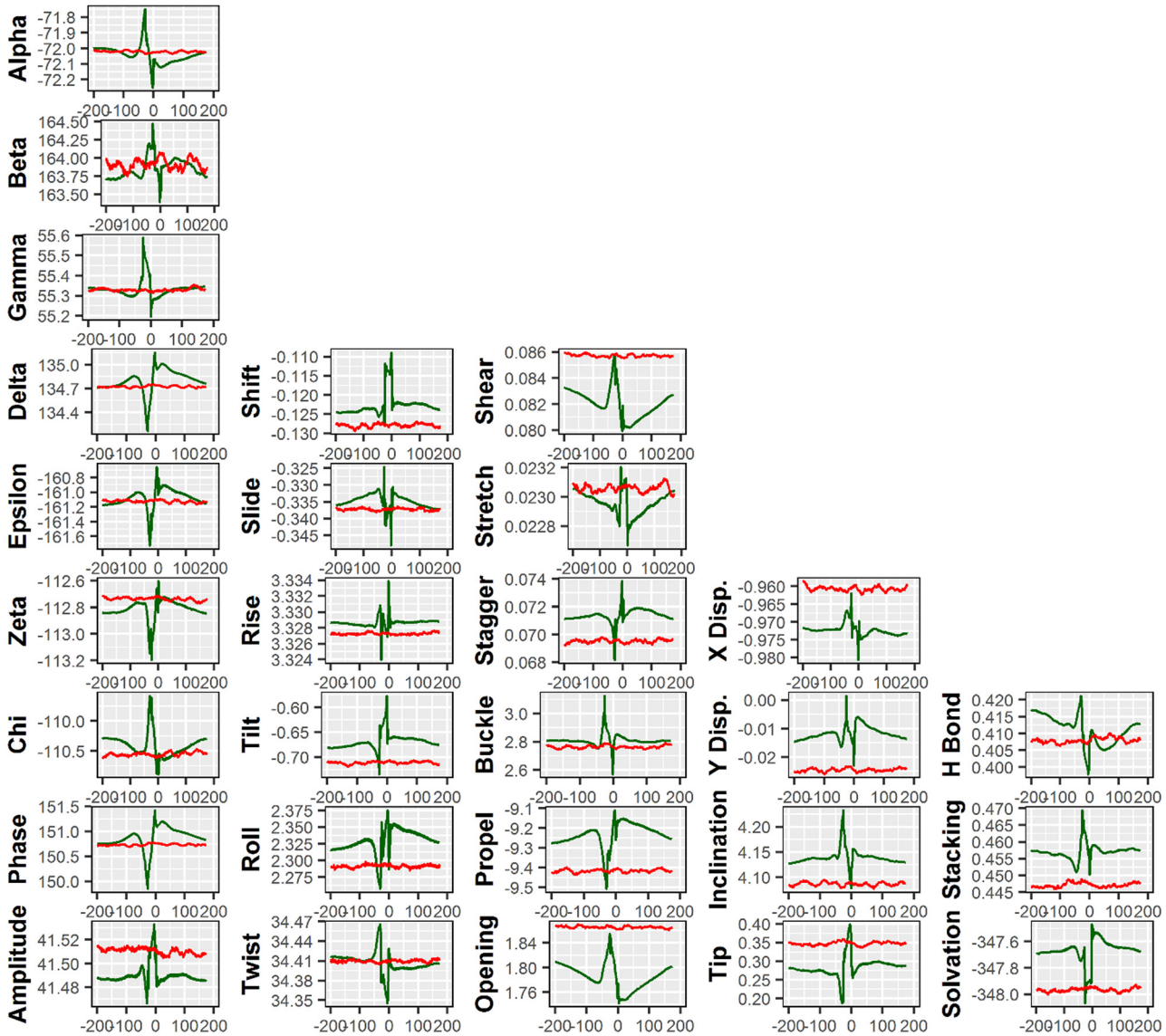

Fig. 12 Structure and energy profiles at exon end sites (for all 328 364 sequences). Sequences containing exon ends are shown in green, whereas red lines represent CDSs. The numeric value of the parameter is represented by the ordinate, while the nucleotide position relative to the exon end is shown by the abscissa. Parameters are represented in the parameter details file (ESI†). The correlation among the parameters (for the entire dataset) and the important features are presented in ESI.†

analysis reveals that while sequences may show variations at various sites within the DNA, the structure and energy profiles imparted by these sequences are always conserved. By building in the neighbourhood effects as compared to the dinucleotide, the tri- and tetranucleotide-based intrinsic signals investigated here appear to be more robust and unique. These intrinsic signals can assist in efficiently identifying and confirming the desired sites. Also, these signals could be utilized for designing improved genome annotation tools.

Author contributions

The project was designed by B Jayaram, Akhilesh Mishra, Priyanka Siwach, and Dinesh Sharma. Data collection and analysis were performed by Dinesh Sharma, Kopal Sharma, and Akhilesh Mishra. The results were analysed, and the text was written by Dinesh Sharma, Kopal Sharma, Akhilesh Mishra, and B Jayaram. Some of the ideas in the study came from Aditya Mittal.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

We gratefully acknowledge Dr Modesto Orozco and Dr Federica Battistini from the Molecular Modelling and Bioinformatics

group at Institute for research in biomedicine, Barcelona, for providing us with the structural MD simulation data utilized in the present study. B Jayaram is thankful to Profs D. L. Beveridge, Richard Lavery, and Krystyna Zakrzewska for many helpful discussions on the topic. Funding from the Department of Biotechnology, Government of India is gratefully acknowledged. Dinesh Sharma is a recipient of a senior research fellowship from the Council of Scientific & Industrial Research (CSIR), Government of India.

References

1 D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2001.
2 J. W. Fickett, The gene identification problem: an overview for developers, *Comput. Chem.*, 1996, **20**, 103–118.
3 G. D. Stormo, Gene-finding approaches for eukaryotes, *Genome Res.*, 2000, **10**, 394–397.
4 C. Mathé, M. F. Sagot, T. Schiex and P. Rouzé, Current methods of gene prediction, their strengths and weaknesses, *Nucleic Acids Res.*, 2002, **30**, 4103–4117.
5 J. E. Allen, M. Pertea and S. L. Salzberg, Computational gene prediction using multiple sources of evidence, *Genome Res.*, 2004, **14**, 142–148.
6 A. Mishra, S. Dhanda, P. Siwach, S. Aggarwal and B. Jayaram, A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics, *Bioinformatics*, 2020, **36**, 2375–2384.
7 J. Watson, T. Baker, S. Bell, A. Gann, M. Levine and R. Losick, *Molecular Biology of the Gene*, Pearson, 7th edn, 2013.
8 M. Dekhtyar, A. Morin and V. Sakanyan, Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes, *BMC Bioinf.*, 2008, **9**, 1–16.
9 P. É. Jacques, S. Rodrigue, L. Gaudreau, J. Goulet and R. Brzezinski, Detection of prokaryotic promoters from the genomic distribution of hexanucleotide pairs, *BMC Bioinf.*, 2006, **7**, 1–14.
10 A. Jong, H. Pietersma, M. Cordes, O. P. Kuipers and J. Kok, PePPER: a webserver for prediction of prokaryote promoter elements and regulons, *BMC Genomics*, 2012, **13**, 1–10.
11 S. D. A. Silva, S. Echeverrigaray and G. J. Gerhardt, BacPP: bacterial promoter prediction—a tool for accurate Sigma-factor specific assignment in enterobacteria, *J. Theor. Biol.*, 2011, **287**, 92–99.
12 H. Y. Lai, Z. Y. Zhang, Z. D. Su, W. Su, H. Ding and W. Chen, iProEP: a computational predictor for predicting promoter, *Molecular Therapy-Nucleic Acids*, 2019, **17**, 337–346.
13 I. A. Shahmuradov, R. Mohamad Razali, S. Bougouffa, A. Radovanovic and V. B. Bajic, bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli, *Bioinformatics*, 2017, **33**, 334–340.
14 V. S. A. Salamov and A. Solovyevand, Automatic annotation of microbial genomes and metagenomic sequences, in *Metagenomics and its applications in agriculture, biomedicine and environmental studies*, 2011, pp. 61–78.
15 R. K. Umarov and V. V. Solovyev, Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks, *PLoS One*, 2017, **12**, 0171410.
16 P. Umesh, J. K. Dubey, R. V. Karthika, B. S. Cherian, G. Gopalakrishnan and A. S. Nair, A novel sequence and context based method for promoter recognition, *Bioinformation*, 2014, **10**, 175.
17 A. Mishra, P. Siwach, P. Misra, S. Dhiman, A. K. Pandey and P. Srivastava, Intron exon boundary junctions in human genome have in-built unique structural and energetic signals, *Nucleic Acids Res.*, 2021, **49**, 2674–2683.
18 Y. Liu, M. Gonzàlez-Porta, S. Santos, A. Brazma, J. C. Marioni and R. Aebersold, Impact of alternative splicing on the human proteome, *Cell Rep.*, 2017, **20**, 1229–1241.
19 P. Senapathy, M. B. Shapiro and N. L. Harris, Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project, *Methods Enzymol.*, 1990, **16**, 252–278.
20 S. Brunak, J. Engelbrecht and S. Knudsen, Prediction of human mRNA donor and acceptor sites from the DNA sequence, *J. Mol. Biol.*, 1991, **220**, 49–65.
21 G. Yeo and C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, *J. Comput. Biol.*, 2004, **11**, 377–394.
22 K. Sahashi, A. Masuda, T. Matsuura, J. Shinmi, Z. Zhang and Y. Takeshima, In vitro and in silico analysis reveals an efficient algorithm to predict the splicing consequences of mutations at the 5′ splice sites, *Nucleic Acids Res.*, 2007, **35**, 5995–6003.
23 R. Ramakrishna and R. Srinivasan, Gene identification in bacterial and organellar genomes using GeneScan, *Comput. Chem.*, 1999, **23**, 165–174.
24 R. F. Yeh, L. P. Lim and C. B. Burge, Computational inference of homologous gene structures in the human genome, *Genome Res.*, 2001, **11**, 803–816.
25 E. Birney, M. Clamp and R. Durbin, GeneWise and genomewise, *Genome Res.*, 2004, **14**, 988–995.
26 M. Stanke and B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, *Nucleic Acids Res.*, 2005, **33**, 465–467.
27 A. A. Salamov and V. V. Solovyev, Ab initio gene finding in Drosophila genomic DNA, *Genome Res.*, 2000, **10**, 516–522.
28 E. E. Snyder and G. D. Stormo, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, *Nucleic Acids Res.*, 1993, **21**, 607–613.
29 R. Guigó, S. Knudsen, N. Drake and T. Smith, Prediction of gene structure, *J. Mol. Biol.*, 1992, **226**, 141–157.
30 A. Mishra, P. Siwach, P. Misra, B. Jayaram, M. Bansal and W. K. Olson, Toward a universal structural and energetic model for prokaryotic promoters, *Biophys. J.*, 2018, **115**, 1180–1189.
31 P. Singhal, B. Jayaram, S. B. Dixit and D. L. Beveridge, Prokaryotic gene finding based on physicochemical

characteristics of codons calculated from molecular dynamics simulations, *Biophys. J.*, 2008, **94**(11), 4173–4183.
32 S. Dutta, P. Singhal, P. Agrawal, R. Tomer, Kritee, E. Khurana and B. Jayaram, A physicochemical model for analyzing DNA sequences, *J. Chem. Inf. Model.*, 2006, **46**(1), 78–85.
33 G. Khandelwal, R. A. Lee, B. Jayaram and D. L. Beveridge, A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes, *Biophys. J.*, 2014, **106**(11), 2465–2473.
34 G. Khandelwal and J. Bhyravabhotla, A phenomenological model for predicting melting temperatures of DNA sequences, *PLoS One*, 2010, **5**(8), e12433.
35 P. D. Dans, A. Balaceanu, M. Pasi, A. S. Patelli, D. Petkevičiūtė and J. Walther, The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules, *Nucleic Acids Res.*, 2019, **47**, 11090–11102.
36 D. A. Case, V. Babin, J. Berryman, R. M. Betz, Q. Cai, D. S. Cerutti, T. E. Cheatham Iii, T. A. Darden, R. E. Duke, H. E. Gohlke and A. W. Goetz, *Amber 14*, 2014, pp. 1–826.
37 R. Salomon-Ferrer, A. W. Gotz, D. Poole, S. Le Grand and R. C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald, *J. Chem. Theory Comput.*, 2013, **9**, 3878–3888.
38 S. Arnott and D. W. L. Hukins, Optimised parameters for A-DNA and B-DNA, *Biochem. Biophys. Res. Commun.*, 1972, **47**, 1504–1509.
39 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, The missing term in effective pair potentials, *J. Phys. Chem.*, 1987, **91**, 6269–6271.
40 I. Ivani, P. D. Dans, A. Noy, A. Pérez, I. Faustino and A. Hospital, Parmbsc1: a refined force field for DNA simulations, *Nat. Methods*, 2016, **13**, 55–58.
41 L. X. Dang, Mechanism and thermodynamics of ion selectivity in aqueous solutions of 18-*crown*-6 ether: a molecular dynamics study, *J. Am. Chem. Soc.*, 1995, **117**, 6954–6960.
42 T. Darden, D. York and L. Pedersen, Particle mesh Ewald: an $N\cdot\log(N)$ method for Ewald sums in large systems, *J. Chem. Phys.*, 1993, **98**, 10089–10092.
43 J. P. Ryckaert, G. Ciccotti and H. J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes, *J. Comput. Phys.*, 1977, **23**, 327–341.
44 A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra and P. D. Dans, BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data, *Nucleic Acids Res.*, 2016, **44**, 272–278.
45 D. R. Roe and T. E. Cheatham III, PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data, *J. Chem. Theory Comput.*, 2013, **9**(7), 3084–3095.
46 A. Hospital, I. Faustino, R. Collepardo-Guevara, C. Gonzalez, J. L. Gelpí and M. Orozco, NAFlex: a web server for the study of nucleic acid flexibility, *Nucleic Acids Res.*, 2013, **41**, 47–55.
47 R. Lavery, M. J. H. P. D. Moakher, J. H. Maddocks, D. Petkeviciute and K. Zakrzewska, Conformational analysis of nucleic acids revisited: curves+, *Nucleic Acids Res.*, 2009, **37**, 5917–5929.
48 M. Pasi, J. H. Maddocks, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, P. D. Dans, B. Jayaram, F. Lankas, C. Laughton and J. Mitchell, μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA, *Nucleic Acids Res.*, 2014, **42**(19), 12272–12283.
49 D. L. Beveridge, G. Barreiro, K. S. Byun, D. A. Case, T. E. Cheatham III, S. B. Dixit, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks and R. Osman, Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d (CpG) steps, *Biophys. J.*, 2004, **87**(6), 3799–3813.
50 S. B. Dixit, D. L. Beveridge, D. A. Case, T. E. Cheatham 3rd, E. Giudice, F. Lankas, R. Lavery, J. H. Maddocks, R. Osman, H. Sklenar and K. M. Thayer, 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.*, 2005, **89**(6), 3721–3740.
51 R. Lavery, K. Zakrzewska, D. Beveridge, T. C. Bishop, D. A. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton and J. H. Maddocks, 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.*, 2010, **38**(1), 299–313.
52 A. Singh, A. Mishra, A. Khosravi, G. Khandelwal and B. Jayaram, Physico-chemical fingerprinting of RNA genes, *Nucleic Acids Res.*, 2017, **45**, 47.
53 J. Li, L. Qi, Y. Guo, L. Yue, Y. Li and W. Ge, *et al.*, Global mapping transcriptional start sites revealed both transcriptional and post-transcriptional regulation of cold adaptation in the methanogenic archaeon Methanolobus psychrophilus, *Sci. Rep.*, 2015, **5**(1), 1–9.
54 D. Jäger, K. U. Förstner, C. M. Sharma, T. J. Santangelo and J. N. Reeve, Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis, *BMC Genomics*, 2014, **15**, 1–15.
55 J. Babski, K. A. Haas, D. Näther-Schindler, F. Pfeiffer, K. U. Förstner and M. Hammelmann, *et al.*, Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq), *BMC Genomics*, 2016, **17**(1), 1–19.
56 T. Cortes, O. T. Schubert, G. Rose, K. B. Arnvig, I. Comas and R. Aebersold, Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in Mycobacterium tuberculosis, *Cell Rep.*, 2013, **5**, 1121–1131.
57 Y. Jeong, J. N. Kim, M. W. Kim, G. Bucca, S. Cho and Y. J. Yoon, The dynamic transcriptional and translational landscape of the model antibiotic producer Streptomyces coelicolor A3 (2), *Nat. Commun.*, 2016, **7**, 1–11.
58 C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiß and A. Sittka, The primary transcriptome of the major human pathogen Helicobacter pylori, *Nature*, 2010, **464**, 250–255.
59 C. Kröger, S. C. Dillon, A. D. Cameron, K. Papenfort, S. K. Sivasankaran and K. Hokamp, The transcriptional landscape and small RNAs of Salmonella enterica serovar

Typhimurium, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, 1277–1286.
60 R. Hershberg, G. Bejerano, A. Santos-Zavaleta and H. Margalit, PromEC: an updated database of Escherichia coli mRNA promoters with experimentally identified transcriptional start sites, *Nucleic Acids Res.*, 2001, **29**, 277.
61 O. Wurtzel, D. R. Yoder-Himes, K. Han, A. A. Dandekar, S. Edelheit and E. P. Greenberg, The single-nucleotide resolution transcriptome of Pseudomonas aeruginosa grown in body temperature, *PLoS Pathog.*, 2012, **8**(9), e1002945.
62 Y. Liao, L. Huang, B. Wang, F. Zhou and L. Pan, The global transcriptional landscape of Bacillus amyloliquefaciens XH7 and high-throughput screening of strong promoters based on RNA-seq data, *Gene*, 2015, **571**, 252–262.
63 M. Albrecht, C. M. Sharma, M. T. Dittrich, T. Müller, R. Reinhardt and J. Vogel, The transcriptional landscape of Chlamydia pneumoniae, *Genome Biol.*, 2011, **12**, 1–15.
64 M. Kopf, S. Klähn, I. Scholz, J. K. Matthiessen, W. R. Hess and B. Voß, Comparative analysis of the primary transcriptome of Synechocystis sp. PCC 6803, *DNA Res.*, 2014, **21**, 527–539.
65 J. McMillan, Z. Lu, J. S. Rodriguez, T. H. Ahn and Z. Lin, YeasTSS: an integrative web database of yeast transcription start sites, *Database*, 2019, **2019**, baz048, DOI: **10.1093/database/baz048**.
66 G. E. Crooks, G. Hon, J. M. Chandonia and S. E. Brenner, WebLogo: a sequence logo generator, *Genome Res.*, 2004, **14**, 1188–1190.
67 A. Waterhouse, J. Procter, D. A. Martin and G. J. Barton, Jalview: visualization and analysis of molecular sequences, alignments, and structures, *BMC Bioinf.*, 2005, **6**, 1.
68 G. Khandelwal and B. Jayaram, DNA–water interactions distinguish messenger RNA genes from transfer RNA genes, *J. Am. Chem. Soc.*, 2012, **134**, 8814–8816.
69 G. Khandelwal, J. Gupta and B. Jayaram, DNA-energetics-based analyses suggest additional genes in prokaryotes, *J. Biosci.*, 2012, **37**, 433–444.
70 A. Kumar and M. Bansal, Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression, *DNA Res.*, 2017, **24**(1), 25–35.
71 V. Brázda, R. C. Laister, E. B. Jagelská and C. Arrowsmith, Cruciform structures are a common DNA feature important for regulating biological processes, *BMC Mol. Biol.*, 2011, **12**(1), 1–6.
72 K. Yanagi, G. G. Privé and R. E. Dickerson, Analysis of local helix geometry in three B-DNA decamers and eight dodecamers, *J. Mol. Biol.*, 1991, **217**(1), 201–214.
73 M. A. El Hassan and C. R. Calladine, The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme, *J. Mol. Biol.*, 1995, **251**(5), 648–664.
74 W. K. Olson, A. A. Gorin, X. J. Lu, L. M. Hock and V. B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, *Proc. Natl. Acad. Sci. U. S. A.*, 1998, **95**(19), 11163–11168.
75 H. Tang, Y. Wu, J. Deng, N. Chen, Z. Zheng, Y. Wei, X. Luo and J. D. Keasling, Promoter architecture and promoter engineering in Saccharomyces cerevisiae, *Metabolites*, 2020, **10**(8), 320.
76 D. L. Riddle, T. Blumenthal, B. J. Meyer and J. R. Priess, *C. Elegans II*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor (NY), 2nd edn, 1997.
77 T. L. Saito, S. I. Hashimoto, S. G. Gu, J. J. Morton, M. Stadler, T. Blumenthal, A. Fire and S. Morishita, The transcription start site landscape of C. elegans, *Genome Res.*, 2013, **23**(8), 1348–1361.
78 X. J. Lu and W. K. Olson, 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures, *Nucleic Acids Res.*, 2003, **31**(17), 5108–5121.
79 X. J. Lu and W. K. Olson, 3DNA: a versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures, *Nat. Protoc.*, 2008, **3**(7), 1213–1227.
80 T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé and Y. Van de Peer, Generic eukaryotic core promoter prediction using structural features of DNA, *Genome Res.*, 2008, **18**(2), 310–323.
81 K. Florquin, Y. Saeys, S. Degroeve, P. Rouze and Y. Van de Peer, Large-scale structural analysis of the core promoter in mammalian and plant genomes, *Nucleic Acids Res.*, 2005, **33**(13), 4255–4264.
82 J. R. Goñi, A. Pérez, D. Torrents and M. Orozco, Determining promoter location based on DNA structure first-principles calculations, *Genome Biol.*, 2007, **8**(12), 1.
83 V. Rangannan and M. Bansal, High-quality annotation of promoter regions for 913 bacterial genomes, *Bioinformatics*, 2010, **26**(24), 3043–3050.
84 H. Wang and C. J. Benham, Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress, *BMC Bioinf.*, 2006, **7**(1), 1–5.