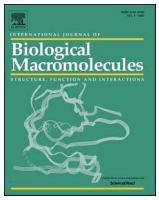
International Journal of Biological Macromolecules 329 (2025) 147488

Contents lists available at ScienceDirect

International Journal of Biological Macromolecules

journal homepage: www.elsevier.com/locate/ijbiomac



Structure and dynamics dictate the functional destiny of genomic DNA across multiple organisms

Dinesh Sharma [a], Danish Aslam [a], Aditya Mittal [a], B. Jayaram [a,b,*]

[a] Supercomputing Facility for Bioinformatics & Computational Biology (SCFBio) & Kusuma School of Biological Sciences, Indian Institute of Technology, Delhi, 110016, India
[b] Department of Chemistry, Indian Institute of Technology, Delhi, 110016, India

ARTICLE INFO

Keywords:
DNA
Genomic physical fingerprinting
Eukaryotic genome annotations
Sequence variability
Molecular dynamics simulations
Functional genomics

ABSTRACT

DNA is a dynamic molecule composed of numerous genic and regulatory elements that orchestrate cellular functions. Traditional methods often fail to provide accurate functional genome annotations because they do not effectively account for sequence variability within and across different organisms. To address this, we conducted an extensive genomic physical fingerprinting of ~4.6 million genomic elements. Utilizing multi-microsecond molecular dynamics simulation-derived parameters over higher nucleotide steps, we uncovered critical details about the physicochemical profiles of key DNA elements, including Coding Sequences, Promoters, Gene boundaries, Exon-Intron boundaries, Start Codons, and Stop Codons across the four diverse eukaryotic kingdoms. This exploration marks a significant advancement in functional genomics and opens new avenues for genome annotation, enhancing our understanding of eukaryotic genome complexity. We identified characteristic biophysical signals associated with key genomic sites across all 11 organisms studied. Our results demonstrate that closely related organisms exhibit similar patterns, underscoring the universality of our features. Furthermore, an investigation in human Enhancers and 5′/3’ UTR boundaries support this novel structural and energetic framework to annotate genomic elements with unprecedented accuracy. Lastly, sequence to structure and dynamics to function was an established axiom in proteomics. Our results demonstrate that this is true for nucleic acids as well.

1. Introduction

The completion of the Human Genome Project (HGP) in 2003 marked a pivotal moment in the field of genomics, igniting widespread interest and advancement in genome annotation [1]. Genome annotation, a fundamental task in molecular biology, involves a thorough identification of various genetic elements essential for understanding the functioning of organisms at the molecular level [2]. These elements encompass a diverse array, ranging from protein-coding mRNA genes to non-coding RNA genes, as well as regulatory elements like Promoters, Exons, Introns, Enhancers, Untranslated regions (UTRs), etc. Since its inception, genome annotation has evolved into a multifaceted endeavor driven by technological advancements and a growing appreciation of the complexities underlying genetic regulation [1,3,4].

Initially, sequence-based profiling approaches emerged as indispensable tools in genomic research. They offered unique insights into the functional elements embedded within DNA sequences. By analyzing sequence compositions and patterns, these methodologies provide valuable information about gene structure, genome elements, and evolutionary relationships [5–11]. However, despite their utility, sequence-based approaches are not without limitations. They often struggle to accurately predict complex molecular interactions and regulatory networks solely from inconsistent and promiscuous sequence information at these sites. In contrast, knowledge-based methodologies constitute another prominent avenue in genomic analysis, relying on advanced statistical and mathematical methods for modeling [12–15]. Though these methods form the backbone of ongoing genomic research, their effectiveness is frequently hampered by the scarcity of experimental data available for training. As a result, they tend to be genome-dependent and lack transparency in molecular interpretations.

* Corresponding author at: Supercomputing Facility for Bioinformatics & Computational Biology (SCFBio) & Kusuma School of Biological Sciences, Indian Institute of Technology, Delhi, 110016, India.
E-mail address: bjayaram@chemistry.iitd.ac.in (B. Jayaram).
URL: http://www.scfbio-iitd.res.in (B. Jayaram).

https://doi.org/10.1016/j.ijbiomac.2025.147488
Received 4 March 2025; Received in revised form 26 August 2025; Accepted 4 September 2025
Available online 5 September 2025

Although these studies represent significant strides in elucidating the properties of genomic elements and their annotations, developing a universal model that can accurately characterize these sites remains a formidable challenge. Molecular dynamics simulations (MDS) have long been employed to explore the conformational behavior, flexibility, and energetic properties of nucleic acids at atomic resolution, offering valuable insights into DNA's structural, dynamic and sequence-dependent behavior [16–30]. Building on these advances and taking a step further, we adopt a genomic physical fingerprinting approach grounded in the hypothesis that discrete functional units on genomic DNA may exhibit unique and distinguishable structural and energetic properties [25–36]. This hypothesis posits that these physicochemical properties, inherent within the atomic models of DNA, can potentially serve as discernible markers for differentiating genomic elements. Expanding on this, our previous research has focused extensively on investigating the structural and energy profiles of both the prokaryotic and eukaryotic genomes. Specifically, our studies have revealed distinct structural and energy signals associated with Coding Sequences (CDS), Promoters, Genes, and Exon-Intron (EI) junctions [25–27,29,30,34–36].

Yet, several key questions arise: Are there similar signals present at other genomic sites which are yet to be explored? Do these signals differ significantly across the various types of genomic elements? Can these patterns be generalized across different organisms, or are they specific to particular taxa or kingdoms? Are these physicochemical signals consistent throughout genomes of varying complexity, such as those of plants versus animals? Moreover, can evolutionary relationships among species be elucidated by examining biophysical patterns at key genomic elements, thereby revealing a physicochemical basis for functional conservation and divergence across lineages? Collectively, these questions underscore the need for a systematic investigation of biophysical signals across diverse organisms and genomic elements. Addressing this gap lies at the core of our study, which seeks to uncover underlying patterns and novel insights into genomic architecture.

To address these critical questions, we conducted a thorough analysis of the biophysical characteristics of key genomic elements across 11 organisms from the kingdoms: Animalia, Plantae, Fungi, and Protista. Our study focuses on 662,106 CDS, 192,714 Promoters, 375,238 Gene boundaries (Gene Start and Gene End), 2,858,091 EI boundaries (Exon Start and Exon End), 235,677 Start Codons, and 252,182 Stop Codons. These genomic sites play fundamental roles in gene regulation and expression, making them essential to our investigation. CDS encode protein sequences, while Promoters serve as transcription initiation sites. Gene and EI boundaries govern expression and facilitate mRNA splicing [37]. The Start Codon marks the initiation point for protein translation, and the Stop Codon signals its termination [37,38].

Akin to our recently published research, we have utilized the MDS-derived structural and energy parameters [27,36]. Through a similar profiling approach, we compared the biophysical characteristics at the sites across all the eukaryotic kingdoms. Later, these profiles underwent a site-specific differentiation to segregate the kingdoms and the organisms within them through visualization of energetics and structure. The in-depth analysis revealed distinct biophysical signatures for genomic elements, which are unique at all the studied genomic sites across organisms. Following this, we extended our analysis to include Enhancers and UTRs (5′ and 3′). Enhancers are regulatory DNA sequences that increase the transcription of associated genes by facilitating the binding of transcription factors [38]. UTRs at the 5′ and 3′ ends of mRNA transcripts play vital roles in post-transcriptional regulation, affecting mRNA stability, localization, and translation efficiency [38]. This exhaustive profiling of Enhancers and UTRs was confined to humans due to the lack of comprehensive data on these sites in other organisms.

Our findings demonstrate that the structural and energetic landscapes of DNA carry biologically meaningful patterns. The patterns are not only robust but also remarkably conserved across evolutionarily related species. These genomic physical fingerprints are distinct at functionally important elements such as CDS, Promoters, and EI junctions, yet remain consistent within phylogenetic lineages. This dual nature of conservation and specificity points to evolutionary pressures shaping these signatures, underscoring their role in preserving regulatory architecture across genomes.

By shifting attention from the linear arrangement of A, T, G, and C to the underlying physicochemical behavior of DNA, this study offers a complementary layer for interpreting genome function and organization. These three-dimensional (3D) properties capture regulatory features that sequence alone may overlook, especially in complex or poorly characterized genomes. In doing so, this work deepens our understanding of genome biology by revealing how structural and energetic properties shape functional organization. It also lays the conceptual and analytical foundation for a universal, interpretable genome annotation framework. Conclusively, this is a step towards building a scalable, physics-informed genome reader, one that can be effectively applied to both model and non-model organisms.

2. Material and methods

2.1. Sequence data collection

To proceed with the biophysical fingerprinting of the genomic elements, we began our analysis with the sequence data collection. We focused on obtaining comprehensive datasets for CDS, Promoters, Gene boundaries, EI boundaries, Start Codons, and Stop Codons across various eukaryotic organisms from the kingdoms: Animalia, Plantae, Fungi, and Protista. Detailed information on these organisms and the exact number of unique sites is listed in Table 1. We obtained the positions of all genome elements except Promoters from the annotation files of organisms downloaded from GENCODE [39], ENSEMBL [40,41], and NCBI [42]. Promoter sites were obtained from the Eukaryotic Promoter Database (EPDnew) [43] and the YeasTSS database [44].

The number of organisms analyzed varied across the four kingdoms due to the lack of a dedicated database for experimentally verified genomic sites. As a result, five animals, two plants, three fungi, and one protist were included in the final study.

For each organism, with the extracted genome site positions set to 0, we obtained a 501-nucleotide long sequence from the reference genome by traversing 250 nucleotides upstream and downstream of the site under consideration. Details on the annotation files (GTF/GFF3) and reference genomes (FASTA) for all organisms considered in this study are provided in Table S1 of Supplementary File 1.

Important to note here, CDS regions are used as internal controls throughout our analyses, given their well-established biophysical stability [25–27,36]. Their consistent and evolutionarily conserved structural and energetic properties allowed them to serve as reference baselines when visually and quantitatively comparing biophysical profiles across other elements (e.g., Promoters, EI junctions, etc.). This approach helped in interpreting relative changes across functional sites.

In conclusion, eight separate sequence datasets were created for each organism, corresponding to the studied genome elements. Additionally, in humans, the availability of data on experimentally verified non-coding genome elements, such as Enhancer sites from EnhancerAtlas 2.0 [45,46] and UTRs from the respective annotation file, allowed us to further support our hypothesis (details available in Table S2 of Supplementary File 1). The precise methodology for extracting sequences corresponding to each genomic element is available in Supplementary File 1.

2.2. Biophysical parameters

The structural and energetic profiles of the extracted DNA sequences were characterized using a comprehensive set of 28 physicochemical parameters, comprising 25 structural and 3 energy-related features [27,36]. The structural descriptors included four Base Pair (BP)-axis properties (X-displacement, Y-displacement, Inclination, and Tip), six

Table 1
A comprehensive summary of genomic sites analyzed across organisms in the four eukaryotic kingdoms.

Kingdom	Organism	Chromosomes	Total number of genomic sites in consideration across all chromosomes							
			CDS	Promoters	Gene start	Gene end	Exon start	Exon end	Start codon	Stop codon
Animalia	*Homo sapiens*	Chr 1–22, X, Y	243,333	29,598	20,004	19,950	358,289	358,103	31,373	48,244
	Mus musculus	Chr 1–19, X, Y	168,376	25,111	21,616	21,583	287,935	288,174	27,733	34,058
	Rattus norvegicus	Chr 1–20, X, Y	41,186	12,601	23,118	23,111	181,046	180,977	35,768	32,331
	Drosophila melanogaster	Chr 2 L, 3 L, 2R, 3R, 4, X, Y	54,994	16,972	13,729	13,734	69,288	69,008	16,332	16,107
	Caenorhabditis elegans	Chr I-V, X	8975	7120	19,965	19,968	131,931	131,804	24,323	21,240
Plantae	*Arabidopsis thaliana*	Chr 1–5	35,614	22,703	27,411	27,416	158,898	158,867	31,804	32,356
	Zea mays	Chr 1–10	56,183	17,081	39,033	39,033	212,442	212,143	45,489	45,020
Fungi	*Saccharomyces cerevisiae*	Chr I-XVI	12,416	5117	5976	5983	5877	5884	5973	6006
	Kluyveromyces lactis	Chr A-F	10,253	23,020	5076	5076	5016	5016	5080	5076
	Yarrowia lipolytica	Chr A-F	12,926	27,794	6446	6446	6203	6204	6500	6447
Protista	*Plasmodium falciparum*	Chr 1–14	17,850	5597	5282	5282	12,491	12,495	5302	5297
TOTAL SITES			662,106	192,714	187,656	187,582	1,429,416	1,428,675	235,677	252,182

Intra-BP (referred to as Intra from here onwards) organization features (Shear, Stretch, Stagger, Buckle, Propeller Twist, and Opening), and nine Backbone torsion angles (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Chi, Phase, and Amplitude), each mapped across all 64 possible trinucleotide combinations [27,36]. Additionally, six Inter-BP (referred to as Inter from here onwards) step parameters (Shift, Slide, Rise, Roll, Twist, and Tilt) were defined for the 136 unique tetranucleotide steps [27,36]. Given the lack of experimentally resolved B-DNA structures in the Nucleic Acid Database (NDB) that encompass all required tri- and tetranucleotide contexts, atomistic MDS were employed as the only feasible approach to derive accurate and consistent values for these parameters [27,36].

The structural parameter values were derived from MDS, wherein the final 500 ns of 1-μs-long trajectories were averaged to capture equilibrium properties [47]. Simulations were performed on a set of 13 ds-DNA oligomers, each 18 BPs in length and flanked by GC clamps [47]. These sequences were systematically designed to encompass all 64 trinucleotide and 136 unique tetranucleotide combinations. The canonical B-DNA models were built using the LEAP module of the AMBERTOOLS [48] and shaped according to Arnott B-DNA fiber parameters [49]. The systems were solvated using the SPC/E water model, ensuring a 10 Å minimum distance from the edge of the defined box [47]. Each system was then neutralized and ionized with K^+Cl^- or Na^+Cl^- at a physiological concentration of 150 mM, using the PARMBSC1 [50] force field and Dang's parameters [51] to describe the DNA and the ions, respectively. The simulations were run using the CUDA-enabled pmemd [52] engine from AMBER14 [48], employing standard practices like periodic boundary conditions, particle mesh Ewald (PME) [53] for electrostatics, and SHAKE [54] constraints on bonds involving hydrogens [47].

All simulations were carried out under constant pressure and temperature (NPT ensemble) for 1 μs per system [47]. The generated trajectories, available through the BIGNASim [55] database, were analyzed using cpptraj [56] and NAFlex [57]. DNA geometries and helical parameters were obtained using the CURVES+ and CANAL software [58]. All measurements were made in accordance with the Ascona B-DNA Consortium (ABC) guidelines to ensure consistency and comparability with prior structural studies [47]. Further details on obtaining the structural parameters can be found in ref. [27, 47] and in Supplementary File 1.

To compute the three energy-related parameters: H-bond, Stacking, and Solvation energies, across all 64 possible trinucleotide combinations, we adopted a previously established framework [29–31,34]. Specifically, we leveraged the comprehensive MDS dataset generated by the ABC group, which includes all tetranucleotide sequences [22–24,59–61]. Energy values for each trinucleotide were derived by averaging their occurrences across all tetranucleotides in which they appear, thus ensuring context-embedded estimation of energetic features. The detailed computational protocol for deriving these energy parameters has been described previously in references [30, 31]. The complete set of computed values for all 28 structural and energy parameters across unique tri- and tetra-nucleotide steps is provided in Supplementary File 2 (reference table).

2.3. Sequence to numerical profiles

We converted the DNA sequences into numerical profiles following the computed parameter values in our reference table. This conversion employed an overlapping mapping approach where tri- and tetranucleotides were assigned their corresponding structural and energy values from the start to the end of a sequence. This method generated 28 numerical profiles that reflect the complete biophysical attributes throughout the sequences.

Following the creation of numerical profiles, we used a sliding, non-overlapping window of 25 nucleotides to minimize noise. Within this window, we averaged the numerical values. This method was applied throughout the entire length of the profiles, yielding the noise-attenuated tri- and tetra-nucleotide profiles. We then employed a normalization on these sequence profiles, where each value was subtracted from the maximum value and divided by the range (maximum−minimum value). This process yielded normalized profiles for all sequences, facilitating more accurate comparisons.

To get absolute structural depictions, we integrated the individual structural feature profiles falling under the main class (BP-axis, Inter, Intra, or Backbone). These four combined structural profiles were then averaged across sequences and visualized along with the three individual averaged energy features at the genomic elements.

2.4. Visualization of numerical profiles and correlation analyses

Two sets of comparisons were performed using these numerical profiles (genomic physical fingerprints). In the first set, we compared the signals at different genomic sites by aggregating profiles for these sites within all organisms of the same kingdom. This allowed us to identify and analyze patterns specific to each kingdom. In the second set of comparisons, the combined profile for a particular genomic site across a specific organism was compared with its kingdom profile. This broader analysis enabled us to detect universal patterns and differences in the biophysical properties of these sites across all the explored eukaryotic organisms within a kingdom.

To elucidate the relationships between these biophysical parameters, we conducted correlation analyses across the kingdoms and the individual organisms. For this analysis, a 200-length segment was extracted from positions −100 to +100 relative to the genomic site.

2.5. Comparison of the genomic physical fingerprints

Following a visual inspection of the complete numerical profiles, a specific segment exhibiting the signature pattern of each genomic element was extracted from all corresponding genomic element sequences. The segment length for each parameter across genomic sites and kingdoms is specified in Tables S3 and S4 of Supplementary File 1. This extracted segment was then averaged within all sequences of the genomic element in consideration for each parameter. The averaged values of the three energy parameters (H-bond, Stacking, Solvation) and four structural parameters (BP-axis, Inter, Intra, and Backbone) were compared separately. A significance test was performed using one-way analysis of variance (ANOVA) to determine whether the energy and structural characteristics of the seven genomic elements differed significantly. This was followed by Tukey-Kramer's post-hoc test to account for sequence number variability, along with false discovery rate (FDR) correction to adjust *p*-value for multiple comparisons. Separate statistical reports were generated for the structural and energy parameters.

Following the establishment of significance among the genomic elements, a comparison was made to assess their separation based on biophysical profiles. A single data point for each parameter was derived by averaging the previously obtained segment-averaged values across all sequences of a particular genomic element. This representation characterized each genomic element in terms of the seven parameters, averaged across sequences and within the segment. Subsequently, principal component analysis (PCA) was performed, and the biophysical profiles of all the genomic elements were inspected across kingdoms and within organisms of the same kingdom. This analysis utilized the first three principal components (PCs), which together accounted for over 86 % of the variance in each case.

3. Results

3.1. The biophysical profiles of genomic elements are uniquely defined

Figs. 1–4 illustrate the trends of seven biophysical parameters across eight genomic sites within each of the four kingdoms. These site-specific profiles were generated by averaging the numerical profiles of sequences corresponding to each site across all organisms within the same kingdom. Figs. S1-S10 of Supplementary File 1 present the organism-specific genomic site profiles. A comparison of the individual trends

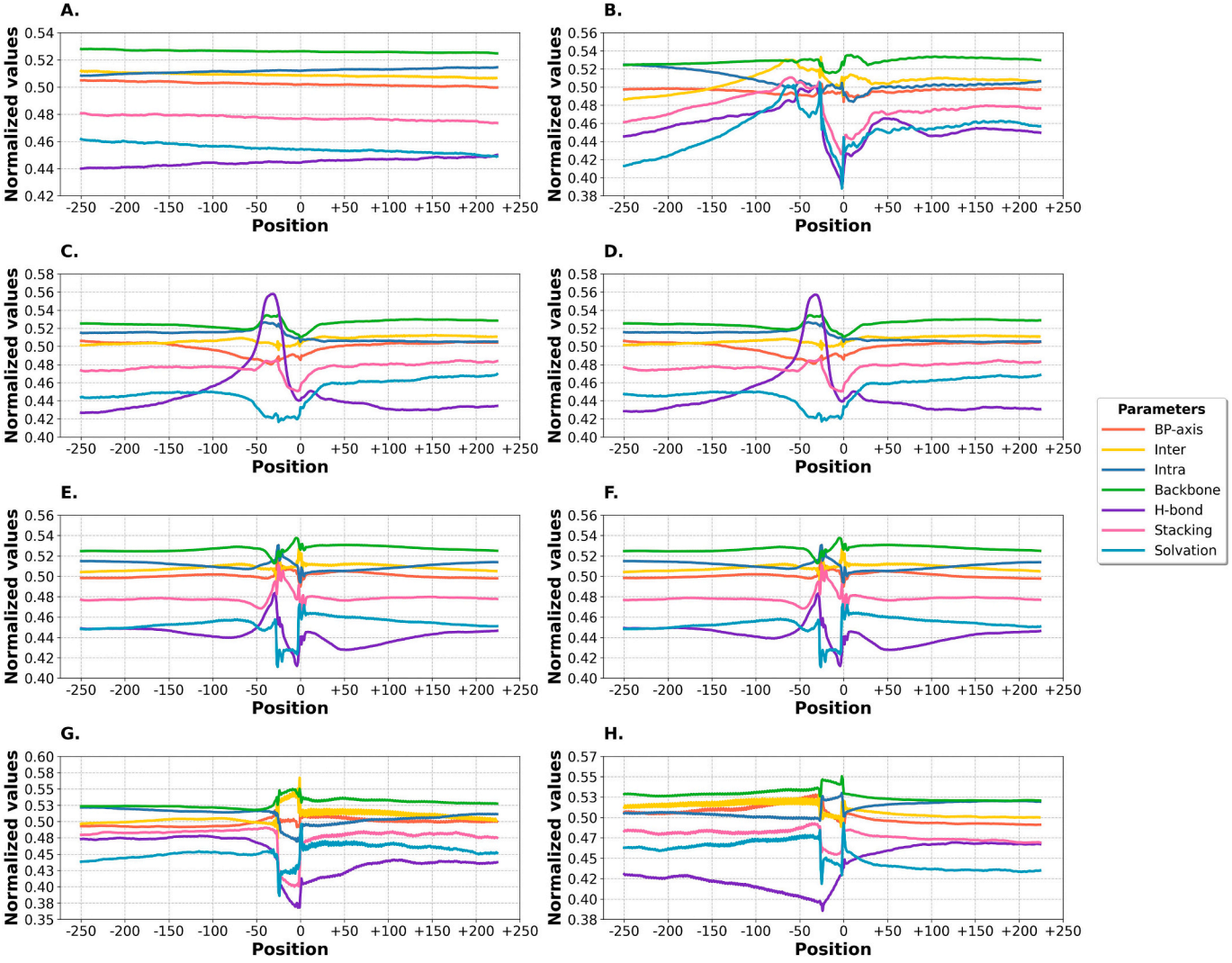


Fig. 1. Biophysical profiles of key genomic elements across the kingdom Animalia. The plots display normalized values of seven biophysical parameters (BP-axis, Inter, Intra, Backbone, H-bond, Stacking, Solvation) across eight genomic sites, with positions relative to the genomic site set to 0. Panels (A.) to (H.) correspond to the following genomic sites: (A.) CDS, (B.) Promoters, (C.) Gene Start, (D.) Gene End, (E.) Exon Start, (F.) Exon End, (G.) Start Codon, and (H.) Stop Codon. The x-axis represents the position relative to the site, and the y-axis represents the normalized biophysical value. The profiles reveal distinct structural and energetic patterns specific to each genomic element, indicating their critical roles within the kingdom.

for all parameters across all sites within an organism, relative to their kingdom profile, is provided in Figs. S11-S13 of Supplementary File 1.

The biophysical profiles of CDS across Animalia, Plantae, Fungi, and Protista exhibit a common stability trend, with minor fluctuations observed across the kingdoms. In Animalia and Plantae, the physicochemical profiles show notable consistency. This may reflect a conserved genomic organization essential for accurate and efficient protein translation [26,27,36]. In contrast, Fungi and Protista display more significant variability in their CDS profiles, which may be attributed to their low sequence number considered in this study. This may also indicate a broader range of structural conformations and a more diverse set of protein-coding strategies.

Overall, the stability observed in Figs. 1–4 and Figs. S1–S10 Supplementary File 1 suggests that energy and structural transitions within this region remain consistent. This consistency helps preserve the integrity required for efficient protein synthesis. Moreover, to investigate whether the signals observed in CDS regions (as extracted from annotation files) are consistent with those from internal exonic regions, we analyzed 501 BP segments centered on the midpoint of protein-coding exons longer than 1000 BP in humans. These regions lie well within EI boundaries and are unlikely to be influenced by junction-associated features. As shown in Fig. S33 (Supplementary File 1), the resulting biophysical profiles are smooth and consistent. This supports the notion that internal coding regions exhibit stable structural and energetic properties, providing a baseline contrast to the more pronounced signals observed at other regions discussed in subsequent sections.

The Promoter regions across the four kingdoms exhibit distinct biophysical trends, particularly around the transcription start sites (TSS). H-bond has a similar pattern among the energy parameters across all the kingdoms. A characteristic signal of varying intensity, comprising a sudden rise followed by a dip, is notable. Stacking and Solvation tend to have a similar trend across all kingdoms. However, this trend is not well defined and occurs in different regions of the profiles. Orchestrating on a similar note, all the structural parameters display unceremonious changes within the profiles across all the kingdoms. The abrupt biophysical changes observed within Promoters across all kingdoms can be attributed to their diverse architectural features [62,63].

Promoters are classified into three classes: core, proximal, and distal. The composition of core promoter elements, including TATA boxes, Initiator (Inr) elements, DNA replication-related elements (DRE), TFIIB recognition elements (BRE), Pause Buttons (PB), etc., vary significantly

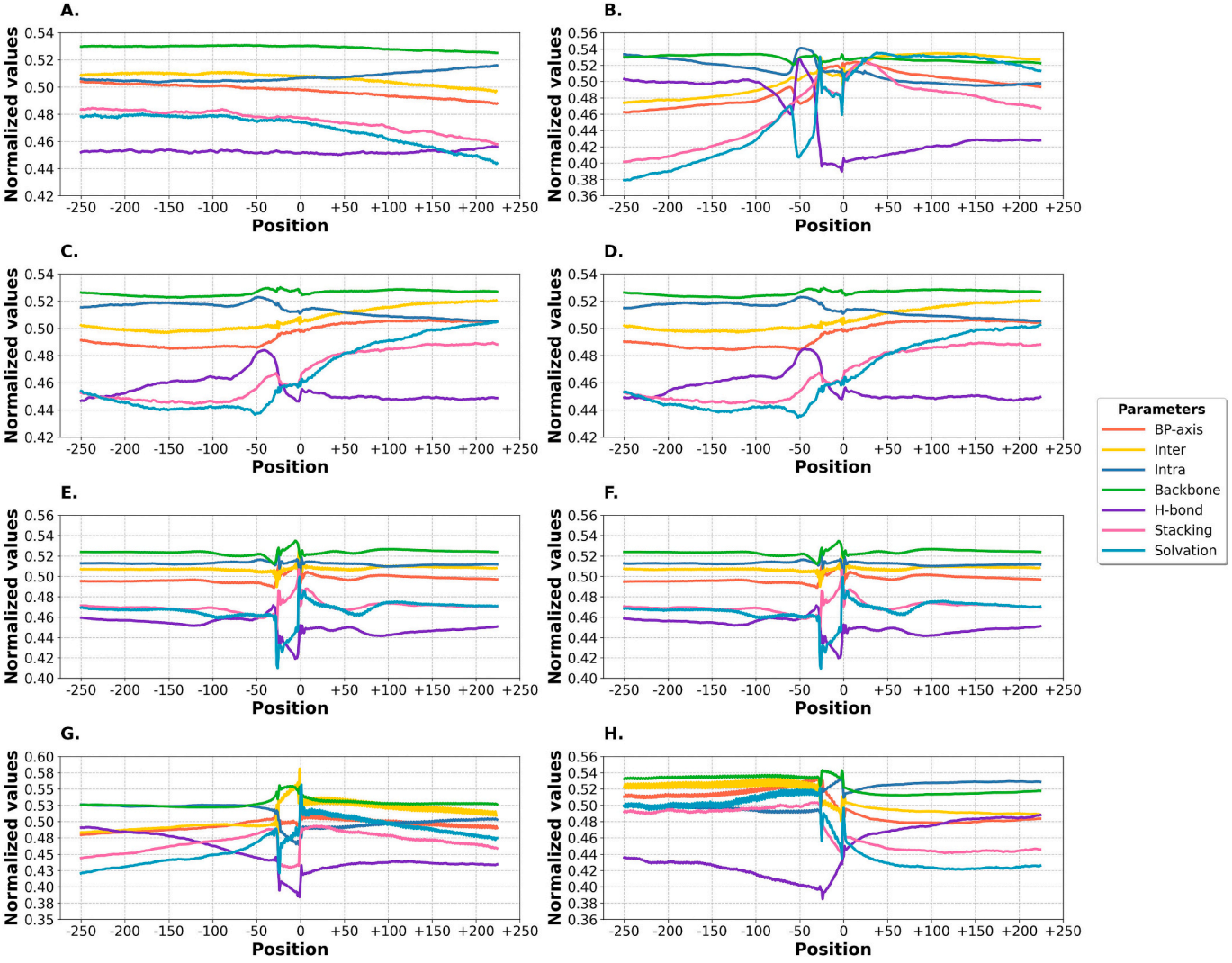


Fig. 2. Biophysical profiles of key genomic elements across the kingdom Plantae. The plots display normalized values of seven biophysical parameters (BP-axis, Inter, Intra, Backbone, H-bond, Stacking, Solvation) across eight genomic sites, with positions relative to the genomic site set to 0. Panels (A.) to (H.) correspond to the following genomic sites: (A.) CDS, (B.) Promoters, (C.) Gene Start, (D.) Gene End, (E.) Exon Start, (F.) Exon End, (G.) Start Codon, and (H.) Stop Codon. The x-axis represents the position relative to the site, and the y-axis represents the normalized biophysical value. The profiles reveal distinct structural and energetic patterns specific to each genomic element, indicating their critical roles within the kingdom.

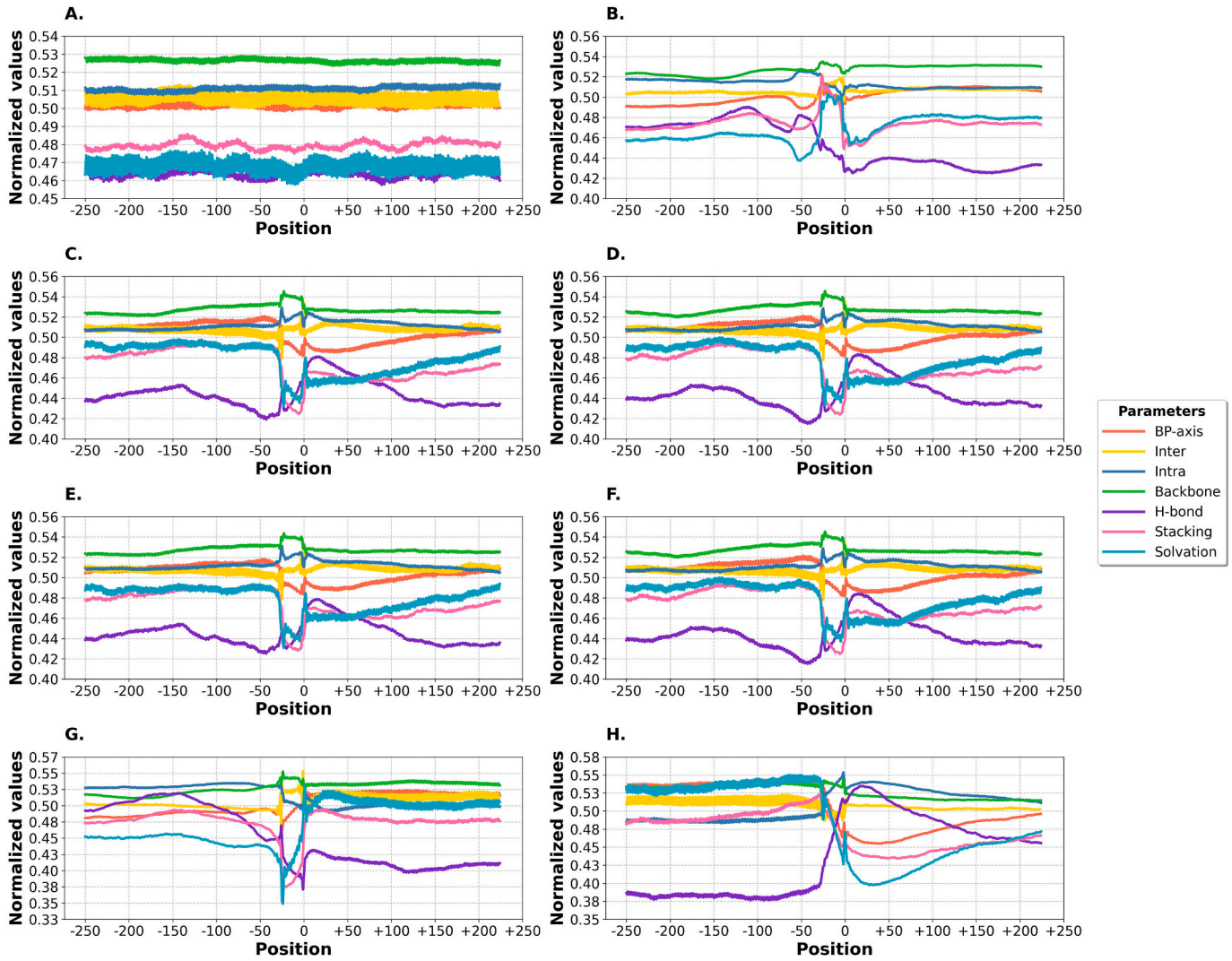


Fig. 3. Biophysical profiles of key genomic elements across the kingdom Fungi. The plots display normalized values of seven biophysical parameters (BP-axis, Inter, Intra, Backbone, H-bond, Stacking, Solvation) across eight genomic sites, with positions relative to the genomic site set to 0. Panels (A.) to (H.) correspond to the following genomic sites: (A.) CDS, (B.) Promoters, (C.) Gene Start, (D.) Gene End, (E.) Exon Start, (F.) Exon End, (G.) Start Codon, and (H.) Stop Codon. The x-axis represents the position relative to the site, and the y-axis represents the normalized biophysical value. The profiles reveal distinct structural and energetic patterns specific to each genomic element, indicating their critical roles within the kingdom.

among different organisms. In *Plasmodium falciparum* (Protista), TATA and BREd elements are predominant, whereas Inr elements are less frequent, and PBs are entirely absent [63]. In contrast, *Saccharomyces cerevisiae* (Fungi) shows a higher prevalence of Inr and BREd elements, along with a more balanced distribution of other core components. A similar pattern is observed in *Arabidopsis thaliana* and *Zea mays* (Plantae), though with slight variations [63]. Among animals, core promoter elements exhibit a high degree of similarity in *Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*, as demonstrated by their identical Promoter profiles in Fig. S1-S3 (Supplementary File 1). However, *Drosophila melanogaster* stands out with a notably higher frequency of DREs, which is absent in these mammals. *Caenorhabditis elegans* has the highest percentage of Inr. Additionally, PBs are more prevalent in Animalia than in other kingdoms [63]. These differences in core promoter composition, coupled with our consideration of all three promoter classes, underscore the distinct physicochemical profiles observed across kingdoms and the organisms.

Focusing on the next four elements directly linked to genic regions, the energy and structural profiles of both Gene, and EI boundaries appear uniform across Animalia and Plantae. However, the strength of these signals varies between the two kingdoms. In Fungi, the biophysical profile intensities and patterns at these four sites match closely, with only slight variations in the flanking regions. This observation aligns with the fact that most genes in yeast (all Fungi species considered here) contain few Introns and exhibit reduced alternative splicing (AS) [64,65], likely contributing to a similar biophysical landscape at these boundary elements. In Protista, like Fungi, the biophysical profiles at the Gene and EI boundaries are nearly identical. This may be due to a low splicing rate of ≈5 % in their genome [66], contributing to the profile resemblance at these sites. However, these profiles exhibit contrasting behavior compared to other kingdoms. Notably, a rise is visible in the BP-axis and Solvation, while a dip, appearing in the Intra, becomes more profound in the H-bond. This counteractive association between H-bond and Solvation, assisted by an increase in Stacking across most kingdoms, may influence splicing events by temporarily destabilizing the EI boundaries [27]. However, this dynamic interplay of energy parameters at Gene boundaries extends across a broader region and may assist in their demarcation. The biological significance of these trends remains to be explored. Additionally, given that EI architecture is inherently sequence-dependent [37], it is critical to distinguish biophysical patterns that genuinely reflect this architecture from those that may arise due to compositional biases such as GC content or Exon structure. To

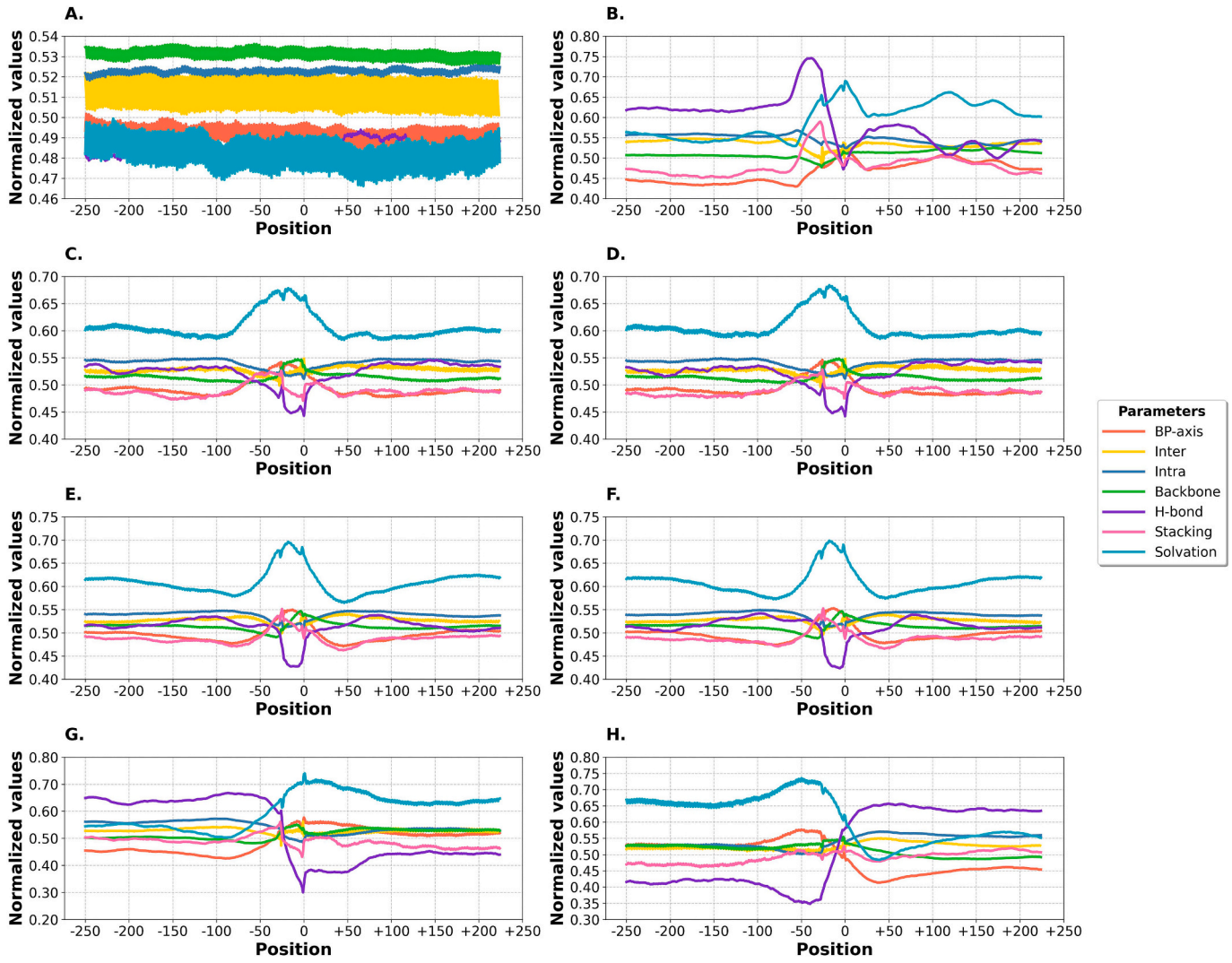


Fig. 4. Biophysical profiles of key genomic elements across the kingdom Protista. The plots display normalized values of seven biophysical parameters (BP-axis, Inter, Intra, Backbone, H-bond, Stacking, Solvation) across eight genomic sites, with positions relative to the genomic site set to 0. Panels (A.) to (H.) correspond to the following genomic sites: (A.) CDS, (B.) Promoters, (C.) Gene Start, (D.) Gene End, (E.) Exon Start, (F.) Exon End, (G.) Start Codon, and (H.) Stop Codon. The x-axis represents the position relative to the site, and the y-axis represents the normalized biophysical value. The profiles reveal distinct structural and energetic patterns specific to each genomic element, indicating their critical roles within the kingdom.

address this, we performed a control analysis using GC-matched shuffled sequences. These controls preserved both sequence length and nucleotide composition while disrupting the native sequence order. This approach (detailed in the Supplementary File 1) allowed us to assess whether the structural and energetic profiles observed at EI boundaries are truly architecture-specific. From the results (Fig. S34, S35 of Supplementary File 1), it is evident that parameters exhibited pronounced deviations from the shuffled background, specifically at the splice junction, across the four representative organisms. In many cases, *Z*-scores exceeded ±2 (Figs. S36-S43 of Supplementary File 1). These findings confirm that the observed biophysical signals are not random or composition-driven.

Continuing further, as we investigate the structural and energy landscape at the Start and Stop Codons, a consistent pattern within parameters emerges across all kingdoms. At the Start Codon, the energy profiles reveal a gradual decline in H-bond and Stacking energies. In contrast, Solvation energy in Animalia, Plantae, and Fungi exhibits a sudden drop followed by a rise. In Protista, Solvation shows a smooth upward transition beginning 50 nucleotides upstream of the Start Codon (marked as 0 on the x-axis in the plots) and remains elevated throughout the region.

The energy profiles at the Stop Codon display distinct and opposing trends to those of the Start Codon. The H-bond energy across all kingdoms shows a gradual increase at the Stop Codon, contrasting sharply with the behavior observed at the Start Codon. Stacking energy at the Stop Codon features a transition dip that is less pronounced than that at the Start Codon. The Solvation energy profile at the Stop Codon is marked by two characteristic shifts in the mid-region, with a gap length that coincides with the 25-nucleotide smoothing window used in our main analyses. Importantly, this spacing of approximately 20–25 nucleotides also appeared when profiles were smoothed with alternative window sizes of 20 and 30 nucleotides (Fig. S44 of Supplementary File 1), suggesting that the observed separation is unlikely to be an artefact of the smoothing procedure.

Structural transitions at the Start and Stop Codons show a noticeable drop in the BP-axis, with the decline more pronounced at the Start Codon than at the Stop Codon. The Intra and Backbone organization profiles exhibit shifts similar to those seen in the Solvation profiles, though with varying intensities. Specifically, these shifts are marked by an increase at the Start Codon and a decrease at the Stop Codon. Additionally, the Intra profiles demonstrate behavior opposite to the Inter and Backbone profiles across all kingdoms.

Remarkably, despite the various transitions observed in the biophysical profiles, the parameter landscapes on either side of the Start or Stop Codons are distinctly different across the kingdoms (Fig. 1–4 and Fig. S1-S10 of Supplementary File 1). This can be attributed to the fact that the Start and Stop Codons analyzed are associated with all protein-coding genes. Consequently, the region between the Start and Stop Codons, which represents the coding region in the expressed mRNA, exhibits lower H-bond energy, higher Solvation energy, and a stable Stacking energy state, indicating a more stable internal architecture.

To further elucidate the specific parametric changes at the eight genomic sites under study, we conducted a detailed correlation analysis on 200-length segments extracted around each site. The results of correlation analyses for the kingdoms and all the organisms are detailed in Figs. S14-S27 of Supplementary File 1. Considering an absolute correlation threshold of 0.70, multiple intercorrelations among the parameters occur at various sites with varying degrees across kingdoms and organisms. At the CDS, there is a high correlation between all the seven biophysical parameters across all the kingdoms. While these correlations are obvious in most organisms, they diminish in some cases. The correlations at other genomic sites are in accordance with the previous observations. Of the energy parameters analyzed, H-bond and Solvation show a consistent correlation at Stop Codons across all kingdoms. Overall, energy correlation patterns at most sites are similar between Animalia–Plantae and Fungi–Protista. For instance, Stacking and Solvation display similar correlation trends at both Exon Start and Exon End. Among the structural features, correlations between Inter-Intra occur at various common sites. However, the correlation at the Stop Codon is more pronounced. Correlation in BP-axis and Intra is evident throughout all studied organisms at Gene Start, Gene End, and Stop Codon. Backbone shows a regular correlation with Intra at Start Codon.

Investigating the cross-correlation between energy and structural parameters, we found that a change in one often leads to a transition in the other, and vice versa. Precisely, the H-bond and Backbone are aligned at the Exon Start, Exon End, Start Codon, and Stop Codon in Animalia-Plantae. A similar relation between the H-bond and BP-axis is seen at Gene Start, Gene End and Stop Codon throughout all kingdoms. This association is further extended to Exon Start and Exon End in Plantae-Fungi-Protista. An opposite correlation value trend at most of these sites for the same structural parameter, but with Solvation hints towards a complementary connection between H-bond and Solvation, as documented earlier. Having no prominent relation with other parameters, Stacking shows a notable relation with Intra at Stop Codon, except in Protista.

3.2. Statistical analyses reveal significant differences in biophysical profiles across genomic sites

Building on the extensive visualization results (Figs. 1–4, Figs. S1-S13 of Supplementary File 1), supported by correlation plots (Figs. S14-S27 of Supplementary File 1), our analysis reveals a distinct structure and energy landscape at each genomic site. These profiles show several intercorrelations among the parameters, suggesting a conserved internal architecture even with variations in DNA sequences within and across organisms. The intensity of these signals is largely consistent among organisms within the same kingdom and across organisms with similar genomic complexity across kingdoms. Although the patterns of these parameters are comparable, they show characteristic shifts, either narrow or broad, within organisms, as reflected in the kingdom profiles. For CDS, consistent with our previous studies [25–27,36], the structure and energy profiles showed no significant variation, with parameters remaining constant, and some profiles dominated by noise. The high correlation among parameters was also evident, but the lack of distinct patterns and the intended use of CDS as an internal control led us to exclude it from further analysis. We therefore focused on the remaining seven genomic sites and performed significance tests to better capture the differences in biophysical profiles across kingdoms and organisms. These findings highlight the potential of structural and energetic signals as informative features for genome annotation. To move beyond descriptive insights and evaluate their significance more rigorously to evaluate their predictive utility, we next performed statistical validation.

To statistically confirm that these structure and energy profiles are unique to each genomic site, we extracted site-specific segments corresponding to the parameters (Tables S3 and S4 of Supplementary File 1) from the kingdom profiles and performed a one-way ANOVA across kingdoms and organisms. The ANOVA results were further validated using Tukey-Kramer's post-hoc test, which accounted for the disparities arising from unbalanced datasets associated with genomic sites. We conducted separate energy- and structure-based one-way ANOVA tests to easily interpret and present the statistical reports. Additionally, FDR correction was applied to the resulting *p*-values to control for multiple testing and reduce the likelihood of false positives.

Figs. 5 and 6 present the mean and standard deviation of the three energy parameters: H-bond, Stacking, and Solvation and four structure parameters: BP-axis, Inter, Intra, and Backbone across seven genomic sites within the four kingdoms ((A.) Animalia, (B.) Plantae, (C.) Fungi, and (D.) Protista), respectively. The plots clearly illustrate the comparative biophysical profiles of these parameters at key genomic locations, with the error bars representing the standard deviations. The plots highlight the conserved nature of these profiles at specific genomic sites across different kingdoms and variability, especially in less structurally constrained organisms within Fungi and Protista. In both Animalia and Plantae, the mean values of energies remain relatively close at most genomic sites, with moderate variability (as indicated by the standard deviations). This consistency points towards a stable energy environment across these genomic sites.

The structural profiles, on the other hand, are conserved to a greater extent, which might reflect more conserved regulatory and structural mechanisms in these kingdoms. These findings suggest that while core biophysical characteristics are preserved, the degree of variability and flexibility in these parameters differs significantly across the lower kingdoms. This variation is likely shaped by evolutionary pressures and the need to adapt to diverse environments. Tables 2 and 3 are the one-way ANOVA tables corresponding to these figs. A comprehensive statistical summary for all kingdoms and organisms, along with the corresponding ANOVA tables, Tukey-Kramer's post-hoc analysis, and FDR-adjusted *p*-value results for each parameter within the energy and structure categories, is provided in Supplementary Files 3–6.

The Tukey-Kramer test across all kingdoms and organisms revealed significant differences in biophysical profiles across genomic sites, with only a few discrepancies. These results were further supported by FDR-adjusted *p*-values, which closely matched the unadjusted values. Among all comparisons, only two cases: a) Exon End vs. Gene End in *Plasmodium falciparum* and b) Exon Start vs. Promoters in *Caenorhabditis elegans* were not statistically significant after FDR correction. As noted earlier, the profiles for Gene Start and Gene End, as well as Exon Start and Exon End, remain consistently similar across all kingdoms and organisms. This pattern is further supported by the results of the post-hoc analysis (and FDR). Additionally, in Fungi, these four sites: Gene Start, Gene End, Exon Start, and Exon End exhibit no significant differences in their biophysical profiles, indicating identical structural and energy characteristics.

After the significance tests among genomic elements, we tried segregating the genomic sites across all kingdoms based on their seven biophysical profiles. A single data point corresponding to each parameter was calculated for each site by averaging the segment-averaged values across all sequences. This projected each genomic element in terms of the seven parameters. Since the parameters have significant multiple intercorrelations, we analyzed the genomic elements using PCA, and the first three PCs were considered across all cases. These components accounted for over 86 % of the variance (percentage variance and loadings tables are summarized in Table S5 and Table S6 of

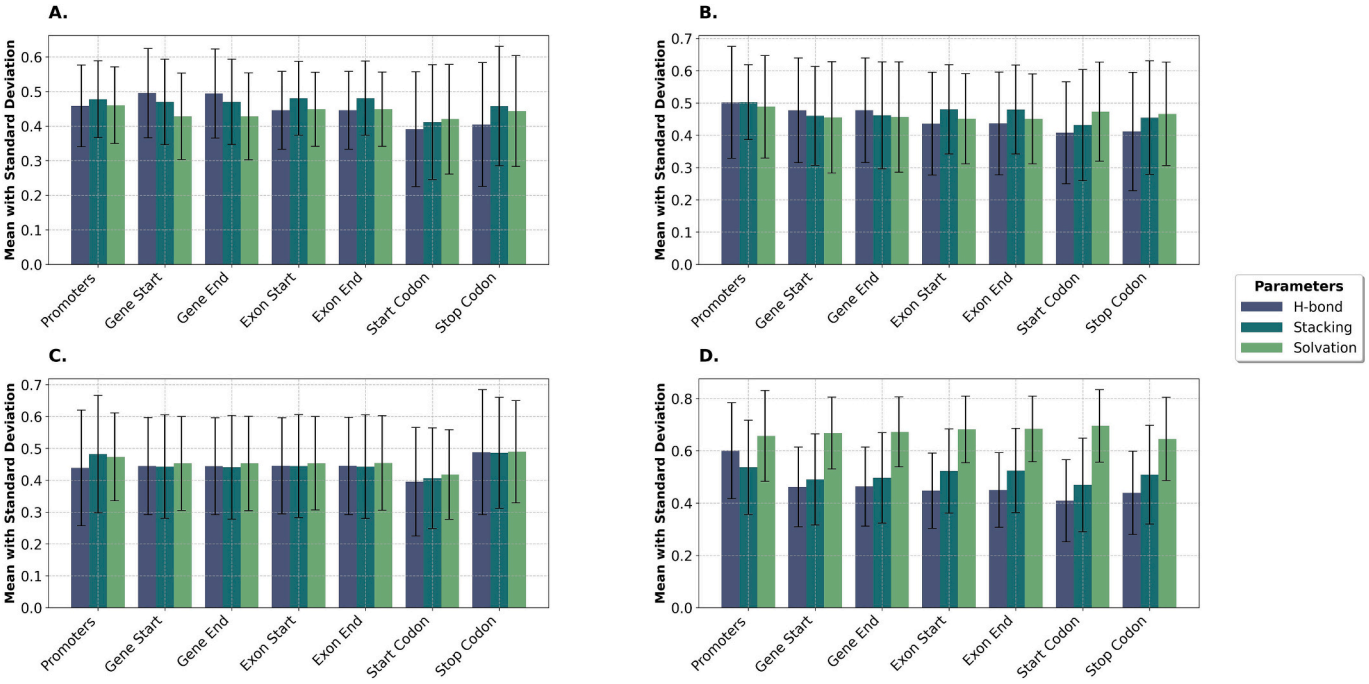


Fig. 5. A comparative analysis of energy parameters across genomic sites in four eukaryotic kingdoms. The bar plots represent the mean and standard deviation of the three energy parameters—H-bond, Stacking, and Solvation—across seven genomic sites: Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, and Stop Codon. Panels (A.) to (D.) correspond to the kingdoms: (A.) Animalia, (B.) Plantae, (C.) Fungi, and (D.) Protista. The data indicate the degree of conservation and variability in energy profiles across different genomic sites within each kingdom, highlighting both shared and kingdom-specific biophysical characteristics.

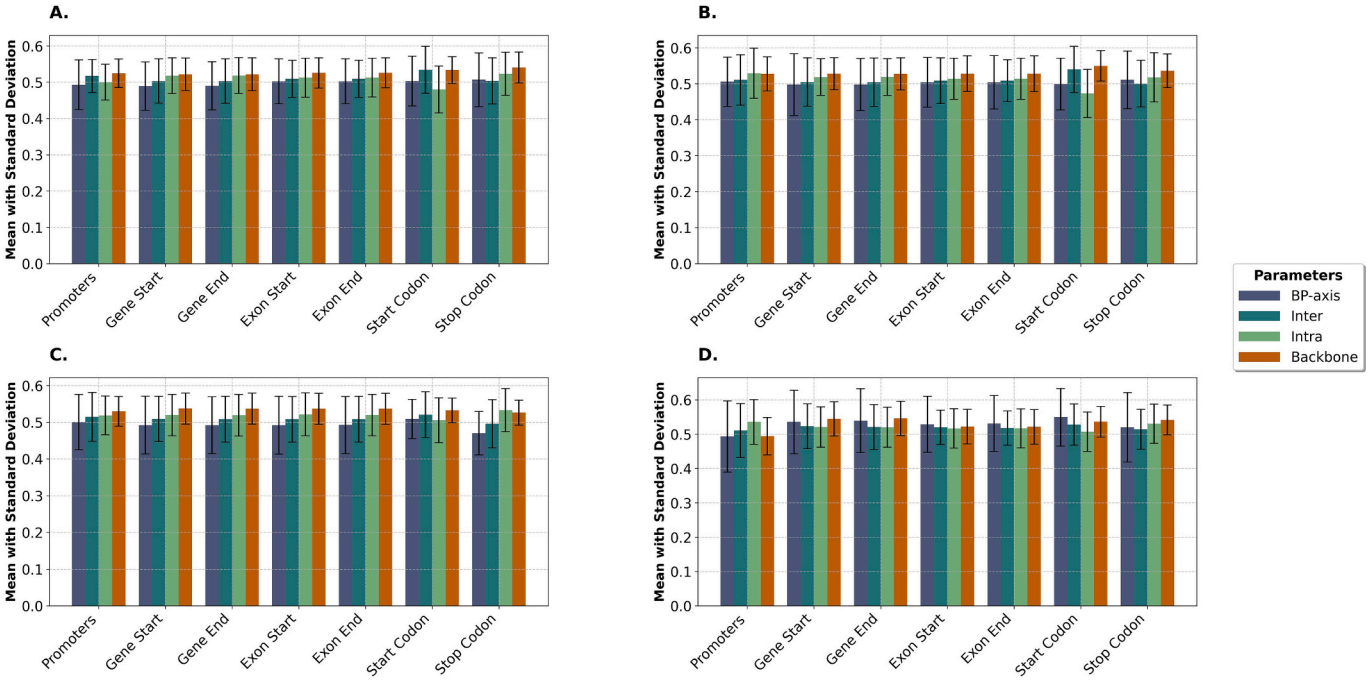


Fig. 6. A comparative analysis of structure parameters across genomic sites in four eukaryotic kingdoms. The bar plots represent the mean and standard deviation of the four structure parameters—BP-axis, Inter, Intra, and Backbone—across seven genomic sites: Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, and Stop Codon. Panels (A.) to (D.) correspond to the kingdoms: (A.) Animalia, (B.) Plantae, (C.) Fungi, and (D.) Protista. The data indicate the degree of conservation and variability in structure profiles across different genomic sites within each kingdom, highlighting both shared and kingdom-specific biophysical characteristics.

Supplementary File 1) and allowed us to explore the genomic fingerprints across kingdoms and within organisms of the same kingdom (Fig. 7 and Figs. S28-S31 of Supplementary File 1).

Fig. 7 clearly shows how the biophysical profiles of different genomic elements cluster across the four eukaryotic kingdoms. The first three PCs effectively separate the genomic elements based on their structural and energy parameters. In the plot, a distinct clustering of elements such as Promoters, Start Codon, and Stop Codon is observed, with notable segregation between kingdoms. Protista shows the most distinct separation, suggesting a unique biophysical landscape compared to the other

Table 2
One-way ANOVA summary for energy parameters across seven genomic sites.

Kingdom	Parameter	F-statistic	p-value	Significance
Animalia	H-bond	12,828.304	0	TRUE
	Stacking	8116.357	0	TRUE
	Solvation	2349.940	0	TRUE
Plantae	H-bond	3107.266	0	TRUE
	Stacking	1953.665	0	TRUE
	Solvation	697.293	0	TRUE
Fungi	H-bond	441.046	0	TRUE
	Stacking	625.822	0	TRUE
	Solvation	470.191	0	TRUE
Protista	H-bond	922.960	0	TRUE
	Stacking	115.359	2.918×10^{-145}	TRUE
	Solvation	90.842	6.716×10^{-114}	TRUE

Table 3
One-way ANOVA summary for structure parameters across seven genomic sites.

Kingdom	Parameter	F-statistic	p-value	Significance
Animalia	BP-axis	1681.819	0	TRUE
	Inter	5966.869	0	TRUE
	Intra	10,145.020	0	TRUE
	Backbone	4089.634	0	TRUE
Plantae	BP-axis	315.297	0	TRUE
	Inter	3549.309	0	TRUE
	Intra	6710.966	0	TRUE
	Backbone	2584.094	0	TRUE
Fungi	BP-axis	182.512	3.141×10^{-230}	TRUE
	Inter	95.978	2.716×10^{-120}	TRUE
	Intra	117.438	1.197×10^{-147}	TRUE
	Backbone	164.220	4.443×10^{-207}	TRUE
Protista	BP-axis	225.764	9.549×10^{-286}	TRUE
	Inter	53.199	1.013×10^{-65}	TRUE
	Intra	150.596	3.025×10^{-190}	TRUE
	Backbone	800.828	0	TRUE

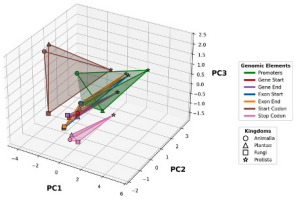


Fig. 7. PCA of biophysical profiles across eukaryotic genomic elements. The 3D plot shows the separation of seven genomic elements—Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, and Stop Codon—based on their structural and energy parameters across four kingdoms: Animalia (circles), Plantae (triangles), Fungi (squares), and Protista (stars). The axes represent the first three PCs (PC1, PC2, and PC3), which together account for 90.71 % of the variance. The clustering and distinct separation of genomic elements, particularly in Protista, indicate the unique biophysical landscapes that differentiate these elements within and across kingdoms. This analysis underscores the potential of biophysical signals in effectively distinguishing genomic elements in diverse eukaryotic organisms.

kingdoms. The clustering pattern also highlights that the Gene Start-Gene End and Exon Start-Exon End share close mapping based on their biophysical profiles across all kingdoms. This is again in agreement with the previous visualization and statistical reports. The results suggest that the biophysical properties of genomic elements serve to differentiate between them effectively. Additionally, they highlight evolutionary divergences across kingdoms. Higher eukaryotes tend to cluster more closely together, indicating the universality and conservation of these biophysical signals.

3.3. Expanding profiling horizons: Human enhancers and UTRs

Building on our findings, we extended the profiling analysis to two additional non-coding human genomic elements, Enhancers and UTRs (5′ and 3′). To carry out this characterization, we followed an approach similar to the Gene and EI boundaries by obtaining and profiling the Start and End positions of these elements from their respective sources (summarized in Table S2 of Supplementary File 1).

The parameter profiles of the extensively studied human genomic sites are present in Fig. S32 (Supplementary File 1). The detailed statistical report (ANOVA, Tukey-Kramer's post-hoc, and FDR adjustment) on the Enhancers and UTRs segment, with pattern lengths ranging from −75 to +50 in energy and − 100 to +50 in structure, is presented in Supplementary File 7. Figs. 8 and 9 display the mean and standard deviation of the energy and structural parameters at these sites. They also compare these values to those of the seven previously examined genomic sites.

The statistical tests, along with the multivariate line plot visualization of the seven biophysical properties at these 14 human genomic sites, as depicted in Fig. 10, visibly show the separation of the elements from one another. It is worth noting that the signals at 5' UTR Start/End and 3' UTR Start/End differ from each other. However, the fingerprints are identical within the same type of UTR (Start matches the End in both the UTRs). These characteristic physicochemical signals associated with the 5′ and 3' UTRs point towards their distinct roles in gene expression [38,67,68]. 3' UTRs are responsible for mRNA translation/degradation and localization [67], whereas 5' UTRs influence translation by undergoing secondary and tertiary structural changes [68]. Enhancer signals at both the Start and End are distinct. All the parameter profiles at the Enhancer Start have a prominent rise and drop, while these trends show an opposite trend (a prominent drop) at the Enhancer End. Though these signals can help demarcate the Enhancer boundaries, the precise biological impact of these patterns is yet to be investigated.

4. Discussion

Despite significant advances in genome annotation over the past two decades, existing methodologies remain limited by their reliance on species-specific training datasets, homology to annotated reference genomes, and high-quality transcriptomic data. These constraints are particularly challenging in the context of non-model organisms, where such resources are often sparse or unavailable, leading to incomplete or inaccurate annotations.

In this foundational work, we profiled the biophysical landscapes of millions of genomic loci across 11 diverse eukaryotic species, revealing statistically significant and biologically coherent patterns. Steady profiles observed within CDS are consistent with the physicochemical stability required during transcription. In contrast, the sharp biophysical transitions at EI junctions and Promoter sites likely facilitate accurate recognition by the splicing and transcriptional machinery.

Further, these trends show strong cross-species evolutionary conservation. Organisms within the same kingdom exhibit similar profiles, suggesting selective pressure to maintain essential regulatory architectures. At the same time, distinct site-specific signatures across distant species highlight the fundamental role of structural and energetic properties in defining functional genomic elements. Together, these findings position DNA's biophysical features as promising, yet underutilized, markers for advancing genome annotation.

While this study highlights the foundational potential of biophysical signatures for genome annotation, its relevance goes beyond exploratory profiling. We have already demonstrated the practical utility of this approach through our EI boundary prediction tool [36], which is both more accurate and computationally efficient than existing sequence-based methods. Conclusively, we aim to develop an integrated annotation workbench that expands to other functional elements using the distinct biophysical signals identified here. Without demanding

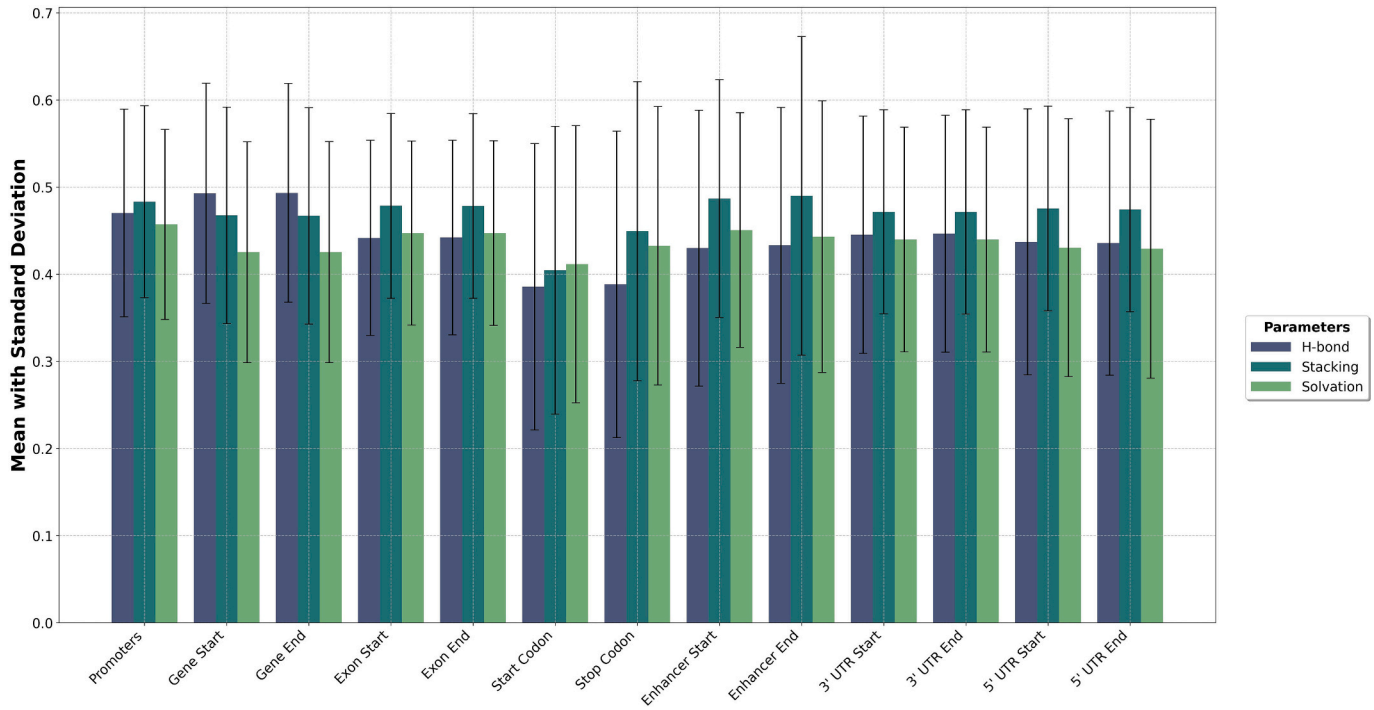


Fig. 8. Comparative analysis of energy parameters across genomic elements in humans. The bar plot shows the mean and standard deviation of the three energy parameters—H-bond, Stacking, and Solvation—across 14 genomic sites: Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, Stop Codon, Enhancer Start, Enhancer End, 3' UTR Start, 3' UTR End, 5' UTR Start, and 5' UTR End. The data highlight the energy landscape of these genomic elements, showcasing the variability and conservation of biophysical properties within human genomic sites.

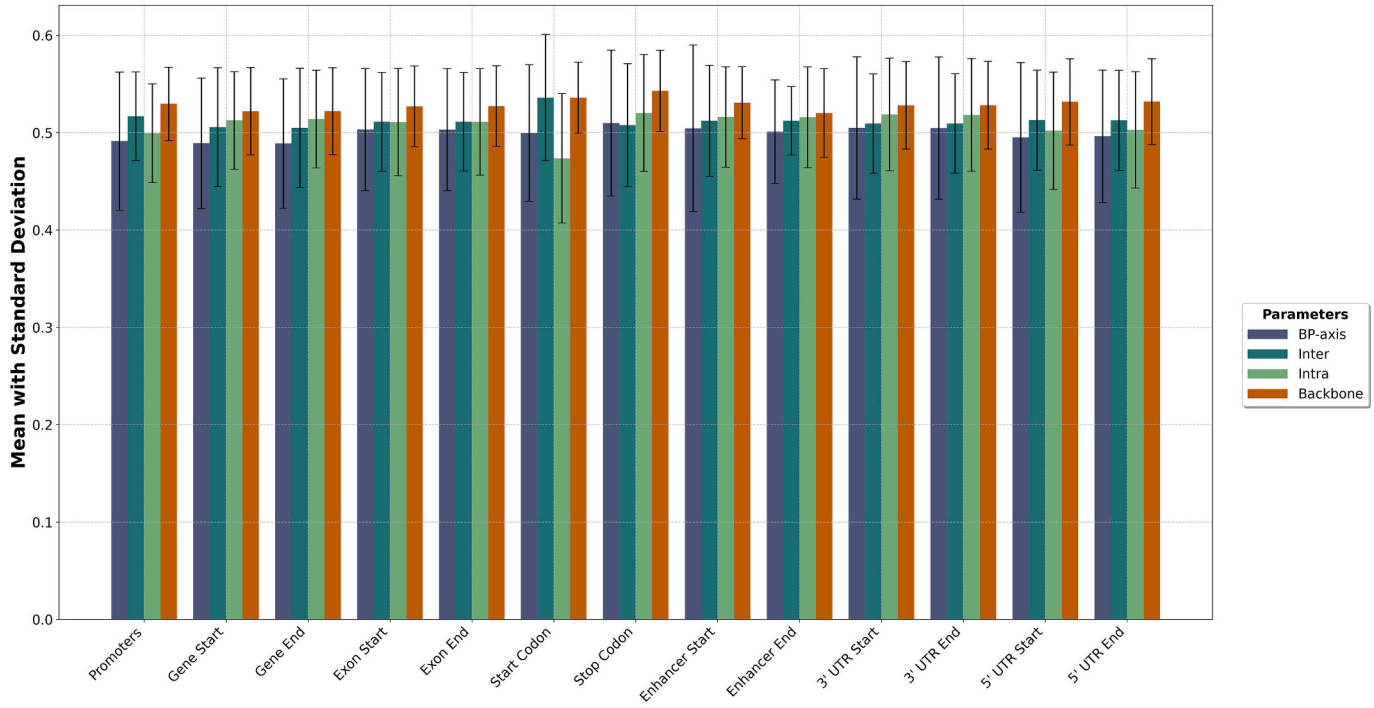


Fig. 9. Comparative analysis of structure parameters across genomic elements in humans. The bar plot shows the mean and standard deviation of the four structure parameters—BP-axis, Inter, Intra, and Backbone—across 14 genomic sites: Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, Stop Codon, Enhancer Start, Enhancer End, 3' UTR Start, 3' UTR End, 5' UTR Start, and 5' UTR End. The data highlight the structure landscape of these genomic elements, showcasing the variability and conservation of biophysical properties within human genomic sites.

excessive computational power, this accessible and scalable platform will offer a practical alternative to conventional methods, even for annotating non-model genomes with limited reference data.

Though this study provides compelling insights, a few limitations warrant mention. Due to the lack of uniformly curated datasets, elements like Enhancers and UTRs could only be analyzed in humans,

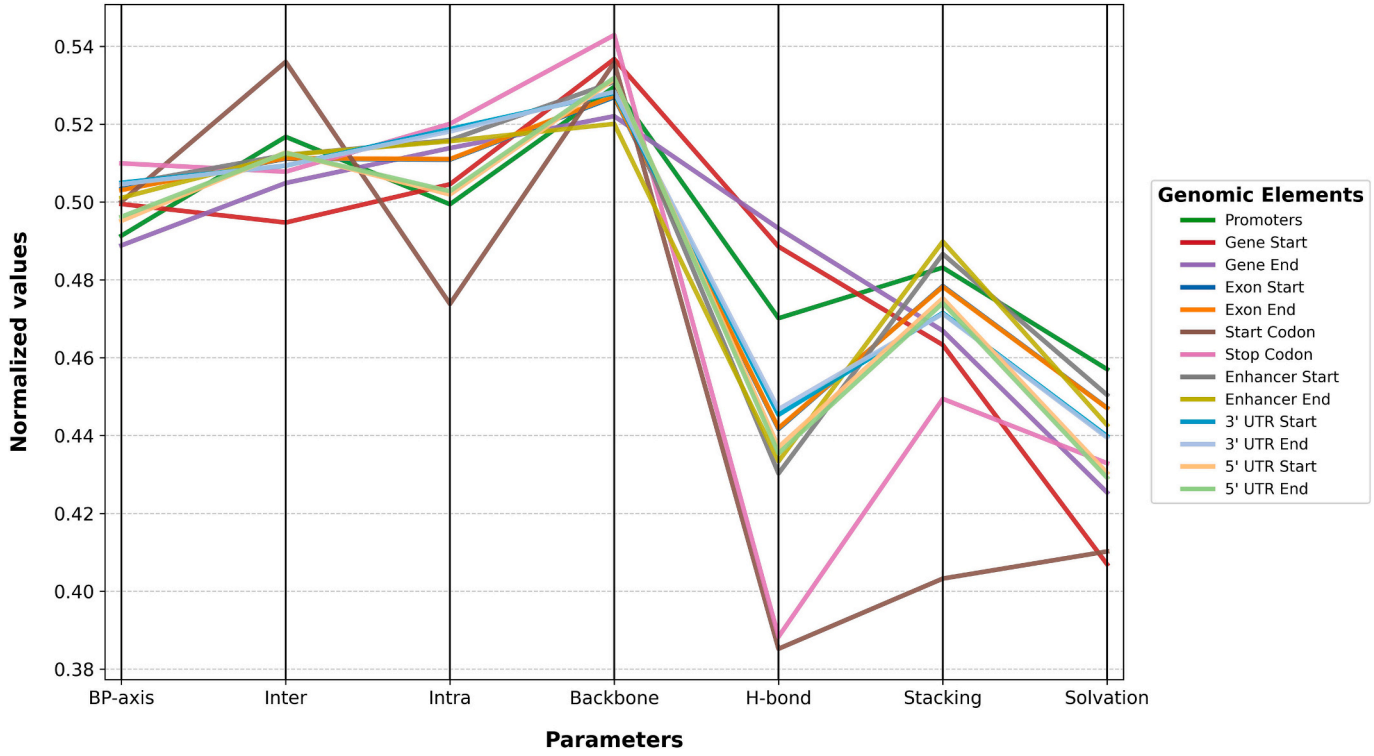


Fig. 10. Multivariate line plot of biophysical parameters across human genomic elements. The plot shows the normalized values of seven biophysical parameters (BP-axis, Inter, Intra, Backbone, H-bond, Stacking, Solvation) for 14 genomic sites: Promoters, Gene Start, Gene End, Exon Start, Exon End, Start Codon, Stop Codon, Enhancer Start, Enhancer End, 3' UTR Start, 3' UTR End, 5' UTR Start, and 5' UTR End. Each colored line represents a specific genomic element, illustrating the similarities and differences in their biophysical landscapes. This visualization highlights how distinct genomic elements in humans exhibit unique but sometimes overlapping biophysical profiles.

limiting cross-species comparisons. Similarly, analysis of Protista relied solely on *Plasmodium falciparum*, which may not reflect the full diversity of the clade and could contribute to outlier trends. Additionally, signal averaging across loci, though useful for capturing dominant patterns, may miss subtle but biologically relevant variations.

Irrespective of these constraints, the observed trends remain statistically robust and reproducible across millions of loci and multiple kingdoms. Rather than weakening the findings, these limitations highlight opportunities for higher-resolution studies. As more comprehensive, experimentally validated data becomes available, particularly for underrepresented elements and taxa, our framework is well-positioned to evolve into a more detailed and inclusive genome annotation platform.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijbiomac.2025.147488.

CRediT authorship contribution statement

Dinesh Sharma: Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Danish Aslam:** Writing – original draft, Methodology, Investigation, Data curation. **Aditya Mittal:** Writing – review & editing, Supervision, Conceptualization. **B. Jayaram:** Writing – review & editing, Supervision, Conceptualization.

Declaration of competing interest

The authors have nothing to declare.

Acknowledgements

The authors express their heartfelt thanks to Dr. Modesto Orozco and Dr. Federica Battistini from the Molecular Modeling and Bioinformatics group at the Institute of Research in Biomedicine, Barcelona, Spain, for providing the structural MD simulation data that was integral to this research. Additionally, we acknowledge the CoE support from the Department of Biotechnology (DBT), Government of India, to SCFBio, IIT Delhi. DS is grateful to the Council of Scientific & Industrial Research (CSIR), India, for providing him with the Senior Research Fellowship.

Data availability

The sequence datasets for all the genomic sites across all the studied organisms are available at http://www.scfbio-iitd.res.in/biophysical/Biophysical_Profiling_Datasets.tar

References

[1] L. Hood, L. Rowen, The human genome project: big science transforms biology and medicine, Genome Med. 5 (2013) 1–8.
[2] J.L. Reed, I. Famili, I. Thiele, B.O. Palsson, Towards multidimensional genome annotation, Nat. Rev. Genet. 7 (2) (2006) 130–141.
[3] G.F. Ejigu, J. Jung, Review on the computational genome annotation of sequences obtained by next-generation sequencing, Biology 9 (9) (2020) 295.
[4] Z. Chen, N.U. Ain, Q. Zhao, X. Zhang, From tradition to innovation: conventional and deep learning frameworks in genome annotation, Brief. Bioinform. 25 (3) (2024) bbae138.
[5] E. Birney, M. Clamp, R. Durbin, GeneWise and genomewise, Genome Res. 14 (5) (2004) 988–995.
[6] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints, Nucleic Acids Res. 33 (suppl_2) (2005) W465–W467.
[7] T. Brůna, A. Lomsadze, M. Borodovsky, GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins, NAR Genomics Bioinform. 2 (2) (2020) lqaa026.
[8] V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters, Genome Biol. 7 (2006) 1–12.
[9] E. Blanco, G. Parra, R. Guigó, Using geneid to identify genes, Curr. Protoc. Bioinform. 18 (1) (2007), 4–3.

[10] A. de Jong, H. Pietersma, M. Cordes, O.P. Kuipers, J. Kok, PEPPER: a webserver for prediction of prokaryote promoter elements and regulons, BMC Genomics 13 (2012) 1–10.
[11] I.A. Shahmuradov, R. Mohamad Razali, S. Bougouffa, A. Radovanovic, V.B. Bajic, bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli, Bioinformatics 33 (3) (2017) 334–340.
[12] E.E. Snyder, G.D. Stormo, Identification of coding regions in genomic DNA sequences: an application of dynamic programming and neural networks, Nucleic Acids Res. 21 (3) (1993) 607–613.
[13] S.L. Salzberg, A.L. Delcher, S. Kasif, O. White, Microbial gene identification using interpolated Markov models, Nucleic Acids Res. 26 (2) (1998) 544–548.
[14] R.K. Umarov, V.V. Solovyev, Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks, PLoS One 12 (2) (2017) e0171410.
[15] N. Scalzitti, A. Kress, R. Orhand, T. Weber, L. Moulinier, A. Jeannin-Girardon, P. Collet, O. Poch, J.D. Thompson, Spliceator: multi-species splice site prediction using convolutional neural networks, BMC bioinformatics 22 (2021) 1–26.
[16] K. Yanagi, G.G. Privé, R.E. Dickerson, Analysis of local helix geometry in three B-DNA decamers and eight dodecamers, J. Mol. Biol. 217 (1) (1991) 201–214.
[17] M.A. El Hassan, C.R. Calladine, The assessment of the geometry of dinucleotide steps in double-helical DNA; a new local calculation scheme, J. Mol. Biol. 251 (5) (1995) 648–664.
[18] W.K. Olson, A.A. Gorin, X.J. Lu, L.M. Hock, V.B. Zhurkin, DNA sequence-dependent deformability deduced from protein–DNA crystal complexes, Proc. Natl. Acad. Sci. 95 (19) (1998) 11163–11168.
[19] D.L. Beveridge, G. Barreiro, K.S. Byun, D.A. Case, T.E. Cheatham, S.B. Dixit, E. Giudice, F. Lankas, R. Lavery, et al., Molecular dynamics simulations of the 136 unique tetranucleotide sequences of DNA oligonucleotides. I. Research design and results on d (CpG) steps, Biophys. J. 87 (6) (2004) 3799–3813.
[20] S.B. Dixit, D.L. Beveridge, D.A. Case, T.E. Cheatham, E. Giudice, F. Lankas, R. Lavery, J.H. Maddocks, R. Osman, H. Sklenar, et al., 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 (6) (2005) 3721–3740.
[21] R. Lavery, M.J.H.P.D. Moakher, J.H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: Curves+, Nucleic Acids Res. 37 (17) (2009) 5917–5929.
[22] R. Lavery, K. Zakrzewska, D. Beveridge, T.C. Bishop, D.A. Case, T. Cheatham III, S. Dixit, B. Jayaram, F. Lankas, C. Laughton, et al., 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 (1) (2010) 299–313.
[23] D.L. Beveridge, T.E. Cheatham, M. Mezei, The ABCs of molecular dynamics simulations on B-DNA, circa 2012, J. Biosci. 37 (2012) 379–397.
[24] 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, et al., μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA, Nucleic Acids Res. 42 (19) (2014) 12272–12283.
[25] A. Mishra, P. Siwach, P. Misra, B. Jayaram, M. Bansal, W.K. Olson, K.M. Thayer, D. L. Beveridge, Toward a universal structural and energetic model for prokaryotic promoters, Biophys. J. 115 (7) (2018) 1180–1189.
[26] A. Mishra, P. Siwach, P. Misra, S. Dhiman, A.K. Pandey, P. Srivastava, B. Jayaram, Intron exon boundary junctions in human genome have in-built unique structural and energetic signals, Nucleic Acids Res. 49 (5) (2021) 2674–2683.
[27] D. Sharma, K. Sharma, A. Mishra, P. Siwach, A. Mittal, B. Jayaram, Molecular dynamics simulation-based trinucleotide and tetranucleotide level structural and energy characterization of the functional units of genomic DNA, Phys. Chem. Chem. Phys. 25 (10) (2023) 7323–7337.
[28] J. Li, R. Rohs, Deep DNAshape webserver: prediction and real-time visualization of DNA shape considering extended k-mers, Nucleic Acids Res. 52 (W1) (2024) W7–W12.
[29] S. Dutta, P. Singhal, P. Agrawal, R. Tomer, Khurana Kritee, B. Jayaram, A physicochemical model for analyzing DNA sequences, J. Chem. Inf. Model. 46 (1) (2006) 78–85.
[30] P. Singhal, B. Jayaram, S.B. Dixit, D.L. Beveridge, Prokaryotic gene finding based on physicochemical characteristics of codons calculated from molecular dynamics simulations, Biophys. J. 94 (11) (2008) 4173–4183.
[31] G. Khandelwal, B. Jayaram, DNA–water interactions distinguish messenger RNA genes from transfer RNA genes, J. Am. Chem. Soc. 134 (21) (2012) 8814–8816.
[32] G. Khandelwal, J. Gupta, B. Jayaram, DNA-energetics-based analyses suggest additional genes in prokaryotes, J. Biosci. 37 (2012) 433–444.
[33] G. Khandelwal, R.A. Lee, B. Jayaram, D.L. Beveridge, A statistical thermodynamic model for investigating the stability of DNA sequences from oligonucleotides to genomes, Biophys. J. 106 (11) (2014) 2465–2473.
[34] A. Singh, A. Mishra, A. Khosravi, G. Khandelwal, B. Jayaram, Physico-chemical fingerprinting of RNA genes, Nucleic Acids Res. 45 (7) (2017) e47.
[35] A. Mishra, S. Dhanda, P. Siwach, S. Aggarwal, B. Jayaram, A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics, Bioinformatics 36 (8) (2020) 2375–2384.
[36] D. Sharma, D. Aslam, K. Sharma, A. Mittal, B. Jayaram, Exon–intron boundary detection made easy by physicochemical properties of DNA, Mol. Omics 2025 (21) (2025) 226–239.
[37] J. Watson, T. Baker, S. Bell, A. Gann, M. Levine, R. Losick, Molecular Biology of the Gene, 7th edition, Cold Spring Harbor Laboratory Press; CSHL, NY, 2013. ISBN-13: 978–0–321-76243-6.
[38] L.W. Barrett, S. Fletcher, S.D. Wilton, Regulation of eukaryotic gene expression by the untranslated gene regions and other non-coding elements, Cell. Mol. Life Sci. 69 (2012) 3613–3634.
[39] A. Frankish, M. Diekhans, I. Jungreis, J. Lagarde, J.E. Loveland, J.M. Mudge, C. Sisu, J.C. Wright, J. Armstrong, I. Barnes, et al., GENCODE 2021, Nucleic Acids Res. 49 (D1) (2021) D916–D923.
[40] P.W. Harrison, M.R. Amode, O. Austine-Orimoloye, A.G. Azov, M. Barba, I. Barnes, A. Becker, R. Bennett, A. Berry, J. Bhai, et al., Ensembl 2024, Nucleic Acids Res. 52 (D1) (2024) D891–D899.
[41] D.M. Bolser, D.M. Staines, E. Perry, P.J. Kersey, Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomic Data, in: Plant Genomics Databases: Methods and Protocols, 2017, pp. 1–31.
[42] E.W. Sayers, E.E. Bolton, J.R. Brister, K. Canese, J. Chan, D.C. Comeau, R. Connor, K. Funk, C. Kelley, S. Kim, et al., Database resources of the national center for biotechnology information, Nucleic Acids Res. 50 (D1) (2022) D20–D26.
[43] R. Dreos, G. Ambrosini, R. Cavin Périer, P. Bucher, EPD and EPDnew, high-quality promoter resources in the next-generation sequencing era, Nucleic Acids Res. 41 (D1) (2013) D157–D164.
[44] J. McMillan, Z. Lu, J.S. Rodriguez, T.H. Ahn, Z. Lin, YeasTSS: an integrative web database of yeast transcription start sites, Database 2019 (2019) baz048.
[45] T. Gao, J. Qian, Enhanceratlas 2.0: an updated resource with enhancer annotation in 586 tissue/cell types across nine species, Nucleic Acids Res. 48 (D1) (2020) D58–D64.
[46] M.F. Melgar, F.S. Collins, P. Sethupathy, Discovery of active enhancers through bidirectional expression of short transcripts, Genome Biol. 12 (2011) 1–11.
[47] P.D. Dans, A. Balaceanu, M. Pasi, A.S. Patelli, D. Petkevičiūtė, J. Walther, M. Orozco, The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules, Nucleic Acids Res. 47 (21) (2019) 11090–11102.
[48] D.A. Case, V. Babin, J.T. Berryman, R.M. Betz, Q. Cai, D.S. Cerutti, T. E. Cheatham III, T.A. Darden, R.E. Duke, H. Gohlke, A.W. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T.S. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K.M. Merz, F. Paesani, D.R. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C.L. Simmerling, W. Smith, J. Swails, R.C. Walker, J. Wang, R.M. Wolf, X. Wuand, P.A. Kollman, AMBER 14, University of California, San Francisco, 2014.
[49] S. Arnott, D.W.L. Hukins, Optimised parameters for a-DNA and B-DNA, Biochem. Biophys. Res. Commun. 47 (6) (1972) 1504–1509.
[50] I. Ivani, P.D. Dans, A. Noy, A. Pérez, I. Faustino, A. Hospital, M. Orozco, Parmbsc1: a refined force field for DNA simulations, Nat. Methods 13 (1) (2016) 55–58.
[51] 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. 117 (26) (1995) 6954–6960.
[52] R. Salomon-Ferrer, A.W. Gotz, D. Poole, S. Le Grand, R.C. Walker, Routine microsecond molecular dynamics simulations with AMBER on GPUs. 2. Explicit solvent particle mesh Ewald, J. Chem. Theory Comput. 9 (9) (2013) 3878–3888.
[53] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: an N log (N) method for Ewald sums in large systems, J. Chem. Phys. 98 (1993) 10089.
[54] J.P. Ryckaert, G. Ciccotti, H.J. Berendsen, Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes, J. Comput. Phys. 23 (3) (1977) 327–341.
[55] A. Hospital, P. Andrio, C. Cugnasco, L. Codo, Y. Becerra, P.D. Dans, J.L. Gelpi, BIGNASim: a NoSQL database structure and analysis portal for nucleic acids simulation data, Nucleic Acids Res. 44 (D1) (2016) D272–D278.
[56] D.R. Roe, T.E. Cheatham III, PTRAJ and CPPTRAJ: software for processing and analysis of molecular dynamics trajectory data, J. Chem. Theory Comput. 9 (7) (2013) 3084–3095.
[57] A. Hospital, I. Faustino, R. Collepardo-Guevara, C. Gonzalez, J.L. Gelpí, M. Orozco, NAFlex: a web server for the study of nucleic acid flexibility, Nucleic Acids Res. 41 (W1) (2013) W47–W55.
[58] R. Lavery, M.J.H.P.D. Moakher, J.H. Maddocks, D. Petkeviciute, K. Zakrzewska, Conformational analysis of nucleic acids revisited: curves+, Nucleic Acids Res. 37 (17) (2009) 5917–5929.
[59] J.S. McCaskill, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, Biopolymers: Orig. Res. Biomol. 29 (6–7) (1990) 1105–1119.
[60] M. Peyrard, T. Dauxois, H. Hoyet, C.R. Willis, Biomolecular dynamics of DNA: statistical mechanics and dynamical models, Phys. D: Nonlinear Phenom. 68 (1) (1993) 104–115.
[61] J. SantaLucia Jr., A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics, Proc. Natl. Acad. Sci. 95 (4) (1998) 1460–1465.
[62] Y.Y. Yamamoto, H. Ichida, T. Abe, Y. Suzuki, S. Sugano, J. Obokata, Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis, Nucleic Acids Res. 35 (18) (2007) 6219–6226.
[63] A. Vanaja, V.R. Yella, Delineation of the DNA structural features of eukaryotic core promoter classes, ACS Omega 7 (7) (2022) 5657–5669.
[64] K.B. Hooks, D. Delneri, S. Griffiths-Jones, Intron evolution in Saccharomycetaceae, Genome Biol. Evol. 6 (9) (2014) 2543–2556.
[65] S. Muzafar, R.D. Sharma, N. Chauhan, R. Prasad, Intron distribution and emerging role of alternative splicing in fungi, FEMS Microbiol. Lett. 368 (19) (2021) fnab135.

[66] L.M. Yeoh, C.D. Goodman, V. Mollard, E. McHugh, V.V. Lee, A. Sturm, A. Cozijnsen, G.I. McFadden, S.A. Ralph, Alternative splicing is required for stage differentiation in malaria parasites, Genome Biol. 20 (2019) 1–13.
[67] C. Mayr, Regulation by 3′-untranslated regions, Annu. Rev. Genet. 51 (1) (2017) 171–194.
[68] K. Leppek, R. Das, M. Barna, Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them, Nat. Rev. Mol. Cell Biol. 19 (3) (2018) 158–174.