Supercomputing Facility for Bioinformatics and Computational Biology (SCFBio)
Indian Institute of Technology (IIT) Delhi, New Delhi, India

A Novel Physicochemical Property-Based Exon-Intron Boundary Prediction Method

Dinesh Sharma[1], Aditya Mittal[1] and B. Jayaram[1,2]

[1]*SCFBio, Kusuma School of Biological Sciences, IIT Delhi, New Delhi 110016, India*

[2]*Department of Chemistry, IIT Delhi, New Delhi 110016, India*

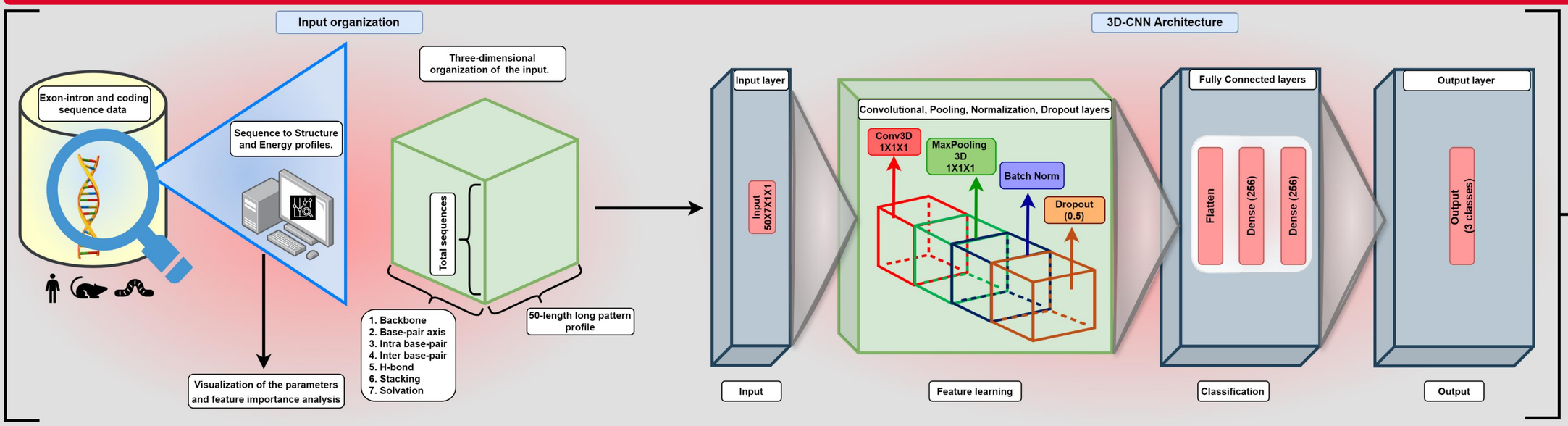
ABSTRACT

Eukaryotic genomes are composed of protein-coding genes that are organized into exons and introns. Exons are the coding sections of these genes, whereas introns are non-coding sequences interspersed between exons. Precisely identifying exon-intron boundaries is essential for understanding the intricate structure of eukaryotic genes and their influence on gene expression and disease. While computational methods using consensus sequences and sophisticated modeling have enhanced our insights on boundary annotations, they frequently encounter challenges due to complex genomic structures, sequence diversity, and alternative splicing events. Building on the hypothesis that the structure and energy properties of deoxyribonucleic acid (DNA) sequences reflect their functional roles, we utilized 28 Molecular dynamics (MD)-simulations derived biophysical properties to develop a deep-learning-based exon-intron boundary prediction tool, ChemEXIN. The tool has been meticulously trained on all exon-intron boundary junctions derived from protein-coding genes in humans, mice, and roundworms. Our model excels in predicting exon-intron boundaries, surpassing the performance of five widely recognized tools in this domain. Notably, the human model achieved an accuracy of 92.5%, with a sensitivity of 93.1% and a specificity of 91.9%. Impressively, our models maintain their superior predictive capabilities even when compared against mice, roundworms, and untrained protein-coding gene sequences. With possible consequences for understanding gene expression, regulation, and biomedical research, our study offers a breakthrough in exon-intron boundary annotations.

BACKGROUND

- Eukaryotic genes consist of protein-coding exons and non-coding introns. During gene expression, introns are removed, and exons are spliced to form mature mRNA. Proper identification of exon-intron boundaries is critical, as splicing errors can lead to genetic disorders.
- Traditional methods, such as sequence-based approaches and computational tools, often face challenges due to species variability and alternative splicing. Recent RNA-based tools and chromatin studies have improved boundary detection but still have limitations.
- ChemEXIN is a novel DNA biophysical property-based tool that identifies exon-intron boundaries by analyzing DNA's structural and energetic properties. It uses molecular dynamics features like backbone conformation, base pair organization, and energy interactions to detect precise exon-intron boundaries.
- Trained on data from *Homo (H.) sapiens*, *Mus (M.) musculus*, and *Caenorhabditis (C.) elegans*, ChemEXIN outperforms traditional methods across protein-coding and non-coding genes. Its robust, versatile performance enhances genome annotation through a biophysical perspective.

METHODS



RESULTS

Figure 1

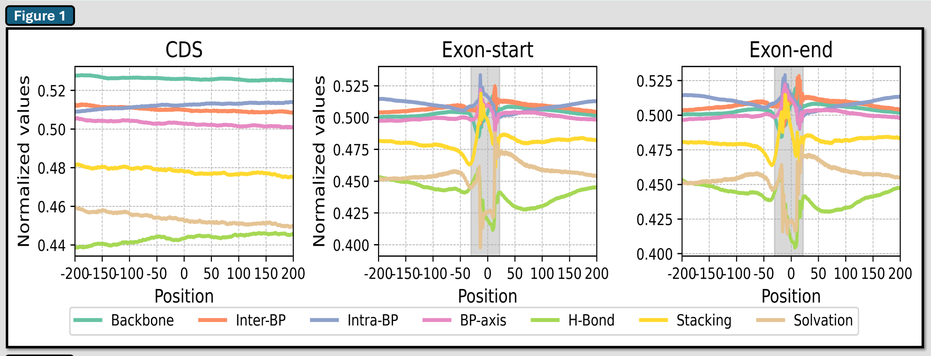


Figure 2

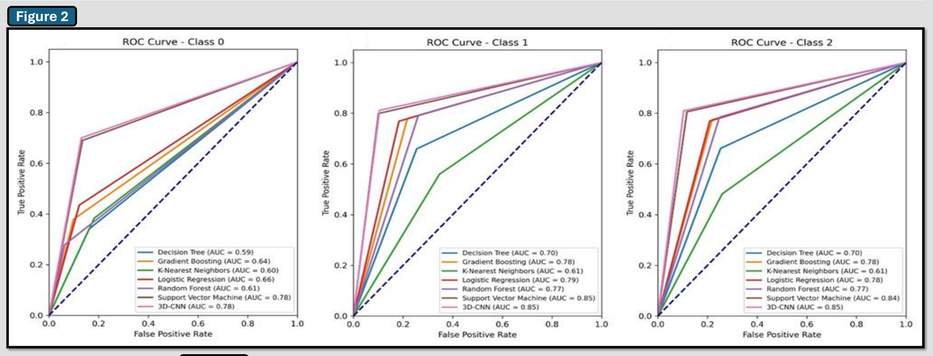


Figure 3

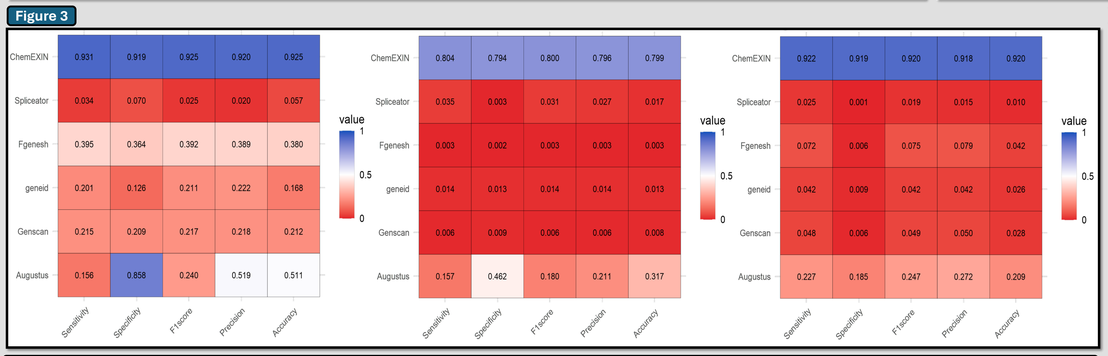


Figure 4

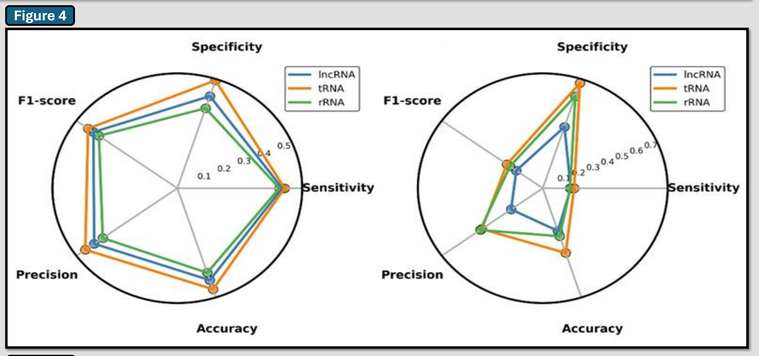


Figure and Table Captions

Figure 1: Profiles of the seven normalized structural and energetic parameters at the exon-intron junctions (Exon-start and Exon-end). Each line represents a major structural or energetic parameter. The four structural parameters were obtained by combining the individual parameters within that category. These parameters show the actual structural and energy change at the two boundaries.

Figure 2: AUROC depicting AUC scores for all three classes. Class 0: CDS, Class 1: Exon-start, and Class 2: Exon-end across all classifiers employed over the Blind-Evaluation set.

Figure 3: Heatmaps depicting the performance of ChemEXIN against other tools across all three organisms. Left: *H. sapiens* Middle: *M. musculus* Right: *C. elegans.*

Figure 4: Performance evaluation of ChemEXIN against Augustus on non-protein coding gene datasets. Left: ChemEXIN, Right: Augustus.

Table 1: Speed evaluation of ChemEXIN on random genes in humans and mice.

Table 1

Organism	Gene	Length (nt)	Predicted Sites	Average time (sec)
H. sapiens	DMD	2,220,382	47	221.77
	BDNF	188,307	28	24.84
	NEU1	10,881	8	8.24
M. musculus	RP1	409,685	26	41.30
	CDK	189,524	9	22.46
	SCAF8	83,888	15	12.78

CONCLUSIONS

- Leveraging a blend of biophysical parameters and 3D-CNN, ChemEXIN notably outperformed existing DNA sequence-based gene organization prediction tools.
- The performance of ChemEXIN highlights the robustness and potential of biophysical parameters for broader applications in genome annotations, even in contexts where traditional tools fail.

REFERENCES

- Jayaram, B., Sharma, D., Aslam, D., Sharma, K., & Mittal, A. (2024). Exon-Intron Boundary Detection Made Easy by Physicochemical Properties of DNA.
- Sharma, D., Sharma, K., Mishra, A., Siwach, P., Mittal, A., & Jayaram, B. (2023). Molecular dynamics simulation-based trinucleotide and tetranucleotide level structural and energy characterization of the functional units of genomic DNA. Physical Chemistry Chemical Physics, 25(10), 7323-7337.
- Mishra, A., Siwach, P., Misra, P., Dhiman, S., Pandey, A. K., Srivastava, P., & Jayaram, B. (2021). Intron exon boundary junctions in human genome have in-built unique structural and energetic signals. Nucleic Acids Research, 49(5), 2674-2683.
- Mishra, A., Dhanda, S., Siwach, P., Aggarwal, S., & Jayaram, B. (2020). A novel method SEProm for prokaryotic promoter prediction based on DNA structure and energetics. Bioinformatics, 36(8), 2375-2384.
- Mishra, A., Siwach, P., Misra, P., Jayaram, B., Bansal, M., Olson, W. K., ... & Beveridge, D. L. (2018). Toward a universal structural and energetic model for prokaryotic promoters. Biophysical Journal, 115(7), 1180-1189.

ACKNOWLEDGEMENTS

- Prof. Modesto Orozco, IRB Barcelona, Spain.
- IIT-Delhi, Hauz Khas, New Delhi, India.
- Department of Biotechnology (DBT), Govt. of India, India.
- Department of Science and Technology (DST), Govt. of India, India.
- Council of Scientific & Industrial Research (CSIR), Govt. of India, India.

Presented at: Biological Data Science 2024

Cold Spring Harbor Laboratory