A Machine-learning Augmented Physicochemical Approach to Eukaryotic Genome Annotation

Dinesh Sharma (E. No. 2019BLZ8642), Kusuma School of Biological Sciences, IIT Delhi

Abstract

Far from being a random string of nucleotides, DNA is a precisely organized polymer whose sequence and structure orchestrate development, metabolism, and the ability of organisms to adapt to their environment. Converting this raw molecular script into meaningful biological knowledge through genome annotation is a cornerstone of modern biology. Despite major advances in sequencing technologies over the past two decades, accurately annotating genomes—especially large and complex eukaryotic genomes—remains a significant challenge. Features such as repetitive elements, alternative splicing, distal regulatory interactions, and diverse chromatin environments make annotation particularly difficult, and traditional annotation approaches suffer from dependence on well-annotated reference genomes and limited applicability to non-model species.

This thesis presents a complementary approach to genome annotation that is grounded in the physicochemical properties of DNA and enhanced by machine learning (ML), providing a framework that is scalable, interpretable, and species independent. The central hypothesis is that functional genomic regions possess characteristic structural and energetic "signatures" determined by the intrinsic biophysical nature of DNA. To capture these signatures, multi-microsecond molecular dynamics simulations were performed across all possible trinucleotide and tetranucleotide steps, generating a high-resolution library of 28 parameters describing DNA base-pair axis, backbone flexibility, inter base-pair axis, intra-base pair axis, and energetics. Together, these descriptors create a comprehensive multidimensional representation of DNA that reveals information not apparent from sequence alone.

Applying this framework, the thesis first explores the physicochemical profiling of key genomic features, including transcription start sites (TSS) and exon-intron boundaries, across multiple prokaryotic and eukaryotic species. These regions demonstrate distinctive and conserved biophysical patterns, supporting their potential use as universal markers of genomic function. This finding lays the conceptual foundation for a structure-informed annotation strategy. Building upon these observations, a deep learning (DL) model, ChemEXIN, was developed to predict exon-intron junctions solely from physicochemical profiles, at the DNA level. By avoiding reliance on sequence alignment or organism-specific training data, ChemEXIN delivers accurate and transferable predictions across diverse eukaryotes. Its performance illustrates that DNA's biophysical behaviour alone can encode the signals necessary for identifying regulatory features, offering a novel approach for cases where conventional tools underperform.

To test the scalability of this paradigm, the framework was extended to profile approximately 4.5 million genomic loci across 11 diverse eukaryotic organisms. This large-scale effort lays the foundation for a futuristic biophysical genome workbench by systematically characterizing structural landscapes across diverse functional elements. The analysis reveals recurring trends and species-specific variations, highlighting the universality and interpretability of the physicochemical signals. Crucially, this foundational work demonstrates strong potential for annotating poorly characterized genomes—particularly those for which standard annotation pipelines prove inadequate.

Together, these contributions introduce and validate an ML–guided, biophysically grounded strategy for genome annotation. By shifting the lens from purely sequence-based models to those that integrate DNA's intrinsic properties, this thesis opens a new dimension in computational genomics—one that enhances annotation accuracy, supports comparative analysis, and bridges annotation gaps across the tree of life. The synthesis of MDS and DL paves the way for next-generation tools in genome annotation, enabling more comprehensive and interpretable insights into genome function and evolution.

References:

1) **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.
2) **Sharma, D.**, Aslam, D., Sharma, K., Mittal, A., & Jayaram, B. (2025). Exon–intron boundary detection made easy by physicochemical properties of DNA. Molecular Omics, 21(3), 226-239.
3) **Sharma, D.**, Aslam, D., Mittal, A., & Jayaram, B. (2025). Structure and dynamics dictate the functional destiny of genomic DNA across multiple organisms. International Journal of Biological Macromolecules, 147488.