Supplementary File 2

Exon-Intron Boundary Detection Made Easy by Physicochemical Properties of DNA

Dinesh Sharma[a], Danish Aslam[a], Kopal Sharma[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 (IIT Delhi), Hauz Khas, New Delhi, 110016, India.
b: Department of Chemistry, IIT Delhi, Hauz Khas, New Delhi, 110016, India.
*: bjayaram@chemistry.iitd.ac.in

Table of Contents:

Methodology S1: Principal Component Analysis (PCA) to elucidate the relative importance of parameters.

Principal Components Analysis (PCA) is a well-established technique for dimensionality reduction across various scientific fields. In this study, we applied PCA to our dataset comprising a combined seven features, each 50 units in length, including Backbone, BP-axis, Inter-BP axis, Intra-BP axis, Hydrogen bond energy, Stacking energy, and Solvation energy. This analysis aimed to elucidate the significance of these features and provide a framework for identifying the most influential variables. Below, we outline the steps involved in this analysis.

STEP 1: Data Preparation

Our analysis commenced with the normalisation of the 50-unit-long combined numerical profiles associated with each feature. This normalisation step was essential to ensure that each parameter contributed equally to the analysis by standardising their scales. Standardisation plays a crucial role in PCA, as it ensures that the variances of the initial variables are taken into account accurately.

STEP 2: PCA Execution

PCA was applied to the standardised data frame, involving the calculation of the covariance matrix, followed by the determination of its eigenvalues and eigenvectors. Eigenvalues represent the variance captured by each principal component, while eigenvectors specify the direction of these components.

STEP 3: Feature Importance Analysis

- We computed cos2 values for each feature with respect to each principal component. Cos2 values indicate the proportion of a feature's variance captured by a principal component. Features with high cos2 values on PC1 and PC2 are well-represented and hence deemed important. This analysis reveals how each feature contributes to the variance explained by the principal components.
- Using scree plots (Refer to Figure S1 and Figure S2), we visualized the variance explained by each principal component (PC). The first two PCs collectively accounted for over 80% of the variance. To delve deeper into the contribution of the seven parameters within these PCs, we plotted the individual contributions of each parameter in the first two PCs.
- Given that all parameters made substantial contributions to PC1 and PC2, we proceeded with incorporating all parameters into the downstream analysis.

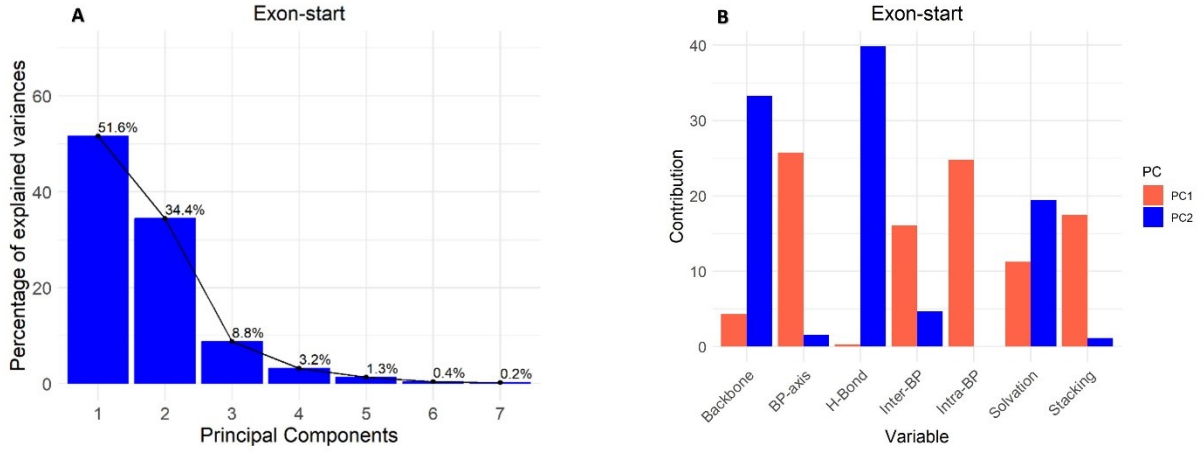

Figure S1: (A) Bar graph depicting percentage variance at exon-start explained by all the PCs. **(B)** Percentage contribution of all seven parameters at exon-start towards PC1 and PC2.

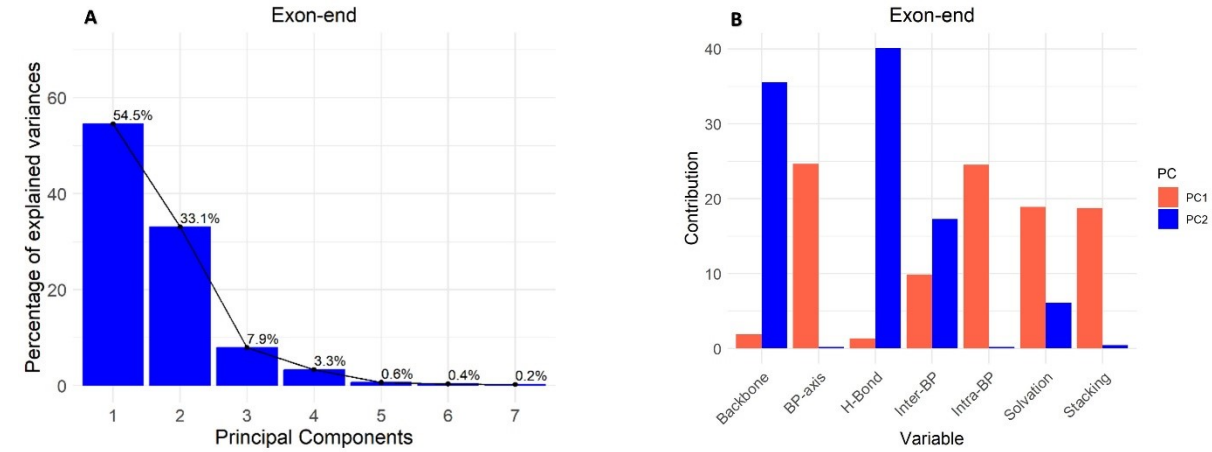

Figure S2: (A) Bar graph depicting percentage variance at exon-end explained by all the PCs. **(B)** Percentage contribution of all seven parameters at exon-end towards PC1 and PC2.

Methodology S2: 3D-CNN architecture

Comparative analyses across both the Training-testing (Supplementary File 3, Table S1) and Blind Evaluation (Supplementary File 3, Table S2) sets demonstrated similar performance between the 3D-CNN model and the SVM-based classifier. However, due to the three-dimensional nature of our dataset, encompassing biophysical features, sequence length, and nucleotide positions (tri-/tetra-) corresponding to each sequence, we adopted the 3D-CNN as the final model for the ChemEXIN prediction pipeline. Additionally, the receiver operating characteristic (ROC) curves for the Blind Evaluation set (Figure 4) reinforced the superiority of the 3D-CNN model. The area under the curve (AUC) values for all three

classes demonstrated a marginally improved performance in the 3D-CNN model (0.78, 0.85, 0.85) compared to the SVM classifier (0.78, 0.85, 0.84), indicating its enhanced predictive accuracy.

Architecture Description:

In the development phase of the 3D-CNN model, *H. sapiens* datasets were utilised. The Training-testing set encompassed three distinct classes (CDS, exon-start, exon-end), facilitating a multi-classification approach. A rigorous optimisation was undertaken to fine-tune critical hyperparameters for model performance. The learning rate governing the magnitude of parameter updates during optimisation was set to 0.001. This established a balance between convergence efficiency and model stability. Subsequently, the batch size, which determines the number of samples processed per iteration, was carefully chosen as 128, aiming to utilise computational resources while maintaining robust learning dynamics efficiently.

During the training cycle, potential overfitting was avoided by structuring it across 40 epochs. The multiple traversals ensured the model's generalisation capacity while monitoring the signs for overfitting. Architectural adjustments to optimise feature extraction capabilities were achieved by the customised number of filters and the size of convolutional kernels. The model employed a single Conv3D layer with 32 filters and a kernel size of (1, 1, 1), followed by a MaxPooling3D layer with a pool size of (1, 1, 1). Furthermore, the design of the dense layers, encompassing the number of neurons, and the incorporation of dropout regularisation (with a dropout rate of 0.5) were calibrated to stabilise the model's generalisation prowess. The ReLU activation function was employed throughout the model architecture to introduce non-linearity and enhance feature representation. Finally, the Adam optimiser, renowned for its adaptive learning rate properties, was strategically leveraged to facilitate effective model parameter updates throughout.

The resultant customised 3D-CNN model demonstrated fair predictive performance across the human datasets without compromising the information from the three dimensions. Identical models for *M. musculus* and *C. elegans* were trained and integrated into ChemEXIN.

Methodology S3: Overview of benchmarking output files.

In the course of this study, a crucial aspect involved a comprehensive benchmarking of the novel approach, ChemEXIN, and several commonly used gene structure organisation prediction tools, including Spliceator, Fgenesh, geneid, Genscan, and Augustus. The goal was to fairly compare these tools based on various performance matrices, such as prediction accuracy, F1 score, specificity, sensitivity, and precision. The main text reports the overall findings, while this file provides a precise overview of the output files generated by each of the selected tools.

1. **Spliceator:** The output encompassed details such as the splicing site type (either Donor or Acceptor), the precise position of the site, reliability scores equal to or exceeding 98 as set by default, and the corresponding sequence stretch.

2. **Fgenesh:** The initial section of the output file contains metadata comprising the sequence length, the count of predicted genes, and the number of predicted exons. Subsequently, attributes such as gene number, strand orientation (forward or reverse), feature name, start and end positions, confidence score, open reading frame (ORF), and feature-length are provided. This is followed by the predicted protein sequences. The feature types listed include CDSi (initial coding sequence), CDSf (full coding sequence), CDSl (last coding sequence), PolA (polyadenylation site), and TSS (transcription start site).

3. **geneid:** The output file provides predictions for each sequence from the input file, beginning with the sequence length, optimal gene structure (number of genes), gene confidence score, and gene strand in the metadata section. The primary attributes consist of exon type (internal, first, terminal, single), start and end positions, reliability score of each exon, and the gene strand (forward or reverse), accompanied by a unique identifier (gene identifier) for each prediction.

4. **Genscan:** The output file commences with the cumulative length of all sequences considered together, followed by the percentage of C-G content within the sequences. Subsequently, details regarding the parametric matrix specific to the organism are provided. The predictions of genes/exons follow this section. Key attributes include gene number/exon number for reference, the type of predicted signal (such as initial exon, internal exon, terminal exon, single-exon gene, promoter, poly-A-signal, etc.), the strand orientation (forward or reverse), start and end positions of the signal, signal length, reading frame, acceptor splice site score, donor splice site score, coding region score, probability, and signal score.

5. **Augustus:** The output file starts with metadata followed by predictions for all sequences provided in the input file. The metadata section includes details on the version of Augustus utilised during implementation, citation information, the organism-specific parametric configuration file, and the presence of hints files if provided. The predictions section provides detailed annotations such as gene, transcript, start_codon, stop_codon, 3'-UTR, 5'-UTR, transcription start site (tss), transcription termination site (tts), intron, and exon type (initial, internal, terminal, CDS) if predicted. Additionally, it includes coding sequence, protein sequence, and evidence to support the predictions.

Methodology S4: MDS methodology for obtaining structural parameters (detailed version available in Ref. "58" and "61" of main text)

The structural parameters consist of four BP-axis attributes (X-displacement, Y-displacement, Inclination, Tip), six Intra-BP attributes (Shear, Stretch, Stagger, Buckle, Propel, Opening), and nine

Backbone attributes (Alpha, Beta, Gamma, Delta, Epsilon, Zeta, Chi, Phase, Amplitude), derived for 64 tri-nucleotide steps. A total of 136 distinct tetra-nucleotide steps were used to measure six Inter-BP attributes, namely Shift, Slide, Rise, Roll, Twist, and Tilt. Values of the above structural variables were derived by calculating the mean of the last 500 ns in microsecond-long MDS. 18-bp long, 13 oligomers, having GC at both ends were used. The leap program from AMBERTOOLS was used to build oligo-nucleotides comprising all tri-nucleotide instances and unique tetra-nucleotides. Oligos were simulated in AMBER14 using pmemd.cuda and later processed to form canonical duplexes based on Arnott B-DNA fiber parameters. Solvation was accomplished using the SPC/E water model, with a distance of 10 Å from the box boundary. To neutralize the systems, 150 mM (K+/Na+) Cl- ions were added, with the PARMBSC1 force field used for the DNA and Dang's parameters for ions. The systems underwent simulation in the NPT ensemble for a duration of 1 μ-second, employing Particlemesh Ewald corrections and periodic boundary conditions. The H-Bonds were restricted using the SHAKE algorithm. The final trajectory files (available at https://mmb.irbbarcelona.org/BIGNASim/) were processed using AMBERTOOLS (cpptraj) and NAFlex. The backbone torsional angles and helical parameters were assessed through CURVES+ and CANAL programs following standards set by the ABC (Ascona B-DNA Consortium).