

Supplementary File 3

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:

1) Table S1. Model Comparison on <i>H. sapiens</i> Training-testing dataset.....	(2)
2) Table S2. Model Comparison on <i>H. sapiens</i> Blind Evaluation dataset	(3)
3) Table S3. Comparison of methods using <i>H. sapiens</i> benchmarking dataset.....	(4)
4) Table S4. Comparison of methods using <i>M. musculus</i> benchmarking dataset.....	(4)
5) Table S5. Comparison of methods using <i>C. elegans</i> benchmarking dataset.....	(4)
6) Table abbreviations for Table S1 to Table S6.....	(5)
7) Table S6. System compatibility evaluation of ChemEXIN on random genes from <i>H. sapiens</i> and <i>M. musculus</i>	(6)
8) Table S7. Details on the Genome Annotation/Assembly versions, and number of sites for organisms under study.....	(7)

Table S1. Model Comparison on *H. sapiens* Training-testing dataset.

Model	CLASS CDS (0)				CLASS EXON-START (1)				CLASS EXON-END (2)				Average Sensitivity	Average Specificity	Average F1-Score	Average Precision	Average Accuracy
	TP ^a	FP ^b	TN ^c	FN ^d	TP	FP	TN	FN	TP	FP	TN	FN					
DTC ^e	25,164	19,956	103,354	17,069	40,660	19,689	84,106	21,088	40,629	19,445	84,536	20,933	0.638	0.82	0.637	0.636	0.762
GBC ^f	16,528	12,480	110,830	25,705	48,020	20,680	83,115	13,728	47,641	20,194	83,787	13,921	0.648	0.835	0.646	0.657	0.785
K-NNC ^g	23,122	22,609	100,701	19,111	34,614	31,695	72,100	27,314	29,428	24,075	79,906	32,134	0.529	0.76	0.526	0.526	0.684
LRC ^h	18,502	14,977	108,333	23,731	47,449	18,138	85,657	14,299	47,162	19,315	84,666	14,400	0.658	0.839	0.657	0.662	0.789
RFC ⁱ	23,883	7,965	115,345	18,350	48,962	19,057	84,738	12,786	48,195	17,481	86,500	13,367	0.714	0.861	0.719	0.735	0.821
SVMC ^j	30,545	16,775	106,535	11,688	49,295	9,406	94,389	12,453	49,414	10,108	93,873	12,148	0.775	0.892	0.772	0.772	0.854
3DCNN ^k	30,388	16,107	107,271	11,777	49,785	9,760	94,413	11,585	50,205	9,298	94,237	11,803	0.781	0.895	0.778	0.778	0.858

Table S2. Model Comparison on *H. sapiens* Blind Evaluation dataset.

Model	CLASS CDS (0)				CLASS EXON-START (1)				CLASS EXON-END (2)				Average Sensitivity	Average Specificity	Average F1-Score	Average Precision	Average Accuracy
	Tp ^a	Fp ^b	TN ^c	FN ^d	TP	FP	TN	FN	TP	FP	TN	FN					
DTC ^e	6,820	6,468	33,531	13,180	13,187	10,192	29,807	6,813	13,225	10107	29,893	6,774	0.554	0.777	0.543	0.548	0.703
GBC ^f	7,579	3,946	36,053	12,421	15,576	8,695	31,304	4,424	15,480	8,723	31,277	4,519	0.644	0.822	0.628	0.646	0.763
K-NNC ^g	7,661	7,282	32,717	12,339	11,184	13,859	26,140	8,816	9,638	10,375	29,625	10,361	0.475	0.737	0.472	0.480	0.650
LRC ^h	8,696	4,868	35,131	11,304	15,368	7,337	32,662	4,632	15,377	8,353	31,647	4,622	0.657	0.829	0.647	0.655	0.772
RFC ⁱ	5,570	2,509	37,490	14,430	15,839	10,466	29,533	4,161	15,687	9,928	30,072	4,312	0.618	0.809	0.590	0.635	0.746
SVM ^j	13,798	5,388	34,611	6,202	15,975	4,034	35,965	4,025	16,111	4,693	35,307	3,888	0.765	0.882	0.764	0.764	0.843
3DCNN ^k	14,001	5,222	34,777	5,999	16,225	4,202	35,797	3,775	16,191	4,158	35,842	3,808	0.774	0.887	0.773	0.773	0.849

Table S3. Comparison of methods using *H. sapiens* benchmarking dataset.

Method	TP ^a	FP ^b	TN ^c	FN ^d	Sensitivity	Specificity	F1-score	Precision	Accuracy
ChemEXIN	1,589	139	1,587	118	0.931	0.919	0.925	0.920	0.925
Spliceator	158	7,660	579	4,427	0.034	0.070	0.025	0.020	0.057
Fgenesh	357	560	321	546	0.395	0.364	0.392	0.389	0.380
geneid	456	1,599	230	1,810	0.201	0.126	0.211	0.222	0.168
Genscan	610	2,184	576	2,223	0.215	0.209	0.217	0.218	0.212
Augustus	352	326	1,972	1,900	0.156	0.858	0.240	0.519	0.511

Table S4. Comparison of methods using *M. musculus* benchmarking dataset.

Method	TP ^a	FP ^b	TN ^c	FN ^d	Sensitivity	Specificity	F1-score	Precision	Accuracy
ChemEXIN	1,601	411	1,584	391	0.804	0.794	0.800	0.796	0.799
Spliceator	166	6,005	20	4,541	0.035	0.003	0.031	0.027	0.017
Fgenesh	3	929	2	1,002	0.003	0.002	0.003	0.003	0.003
geneid	28	2,022	27	2,026	0.014	0.013	0.014	0.014	0.013
Genscan	17	2,747	26	2,787	0.006	0.009	0.006	0.006	0.008
Augustus	200	750	644	1,070	0.157	0.462	0.180	0.211	0.317

Table S5. Comparison of methods using *C. elegans* benchmarking dataset.

Method	TP ^a	FP ^b	TN ^c	FN ^d	Sensitivity	Specificity	F1-score	Precision	Accuracy
ChemEXIN	1,634	146	1,651	138	0.922	0.919	0.920	0.918	0.920
Spliceator	75	4,854	4	2,914	0.025	0.001	0.019	0.015	0.010
Fgenesh	111	1,290	8	1,435	0.072	0.006	0.075	0.079	0.042
geneid	89	2,029	18	2,031	0.042	0.009	0.042	0.042	0.026
Genscan	139	2,629	15	2,753	0.048	0.006	0.049	0.050	0.028
Augustus	642	1,720	390	2,188	0.227	0.185	0.247	0.272	0.209

In table abbreviations for Table S1 to Table S6:

^aTP: True Positive.

^bFP: False Positive.

^cTN: True Negative.

^dFN: False Negative.

^eDTC: Decision Tree Classifier.

^fGBC: Gradient Boosting Classifier.

^gK-NNC: K-Nearest Neighbours Classifier.

^hLRC: Logistic Regression Classifier.

ⁱRFC: Random Forest Classifier.

^jSVMC: Support Vector Machine Classifier.

^k3DCNN: Three-dimensional Convolutional Neural Network.

Table S6. System compatibility evaluation of ChemEXIN on random genes from *H. sapiens* and *M. musculus*.

Organism	Gene	Length (nt ^a)	Time (sec ^b) (Windows ^c)	Time (sec) (Linux ^d)	Time (sec) (MacBook ^e)
<i>H. sapiens</i>	DMD ^f	2,220,382	285.99	301.77	77.57
	BDNF ^g	188,307	30.19	31.46	12.89
	NEU1 ^h	10,881	10.18	7.41	7.14
<i>M. musculus</i>	RP1 ⁱ	409,685	49.5	58.46	15.95
	CDK6 ^j	189,524	28.06	29.75	9.59
	SCAF8 ^k	83,888	14.73	15.85	7.77

^ant: Nucleotides.

^bsec: Seconds.

^cWindows (Windows 10): RAM: 8gb; Processor: intel core i5 5th Gen (Without Graphic card).

^dLinux (Ubuntu 22.04): RAM: 8gb; Processor: intel core i3 11th Gen (Without Graphic card).

^eMacBook (macOS Sonoma, version 14): RAM: 16gb, Processor: M2 Pro (12 core CPUs and 19 core GPUs).

^fDMD: Dystrophin (muscular dystrophy, Duchenne and Becker types).

^gBDNF: Brain-derived neurotrophic factor.

^hNEU1: Neuraminidase-1.

ⁱRP1: Retinitis Pigmentosa-1.

^jCDK6: Cyclin-dependent kinase-6.

^kSCAF8: SR-related CTD associated factor-8.

Table S7. Details on the Genome Annotation/Assembly versions, and number of sites for organisms under study.

Organism	Genome Annotation Build (NCBI Accession)	Genome Assembly (strain/version)	Number of sites		Chromosomes
<i>Homo sapiens</i>	GRCh38, version 23 (Ensembl 81)	GRCh38.p3	CDS	30140	Chr 1-22, X, Y
			ES	328368	
			EE	328368	
<i>Mus musculus</i>	GRCm39, version M33 (Ensembl 110)	GRCm39	CDS	21762	Chr 1-19, X, Y
			ES	289749	
			EE	289749	
<i>Caenorhabditis elegans</i>	WBcel235 (GCF_000002985.6)	WBcel235	CDS	1761	Chr I-V, X
			ES	121545	
			EE	121545	