LncCat: An ORF Attention Model to Identify LncRNA Based on Ensemble Learning Algorithm of CatBoost by Sequence information

Hongqi Feng[1], Shaocong Wang[1], Xuemei Hu[3], Sen Yang[1, 2, *], Chenyang Zhu[1]

[1]School of Computer Science and Artificial Intelligence Aliyun School of Big Data School of Software, Changzhou University, Changzhou 213164, China;

[2]Changzhou No.2 People's Hospital, the Affiliated Hospital of Nanjing Medical University, Changzhou 213164, China
[3]Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China;
*Correspondence: ys@cczu.edu.cn (Sen. Yang);

Supplementary File

Table S1 Detailed description of all features used in experiment

Type	No.	variable	Features	dimension	Description
codon-related	1	codon_num	stop codon count	1	The number of stop codons in transcript.
	2	codon_frequency	stop codon frequency	1	The number of stop codons divided by transcript length.
	3	fickett	Fickett TESTCODE Score	1	Position value of X (A, T, C, G) is calculated as: X_1 = Number of Xs in position 0, 3, 6, … X_2 = Number of Xs in position 1, 4, 7, … X_3 = Number of Xs in position 2, 5, 8, … $X_{pos} = \frac{MAX(X_1, X_2, X_3)}{MIN(X_1, X_2, X_3)+1}$ A_{pos}, T_{pos}, C_{pos}, G_{pos} and the percentage composition of each base are calculated. These eight values are converted into probabilities (p) of coding using a lookup table in the original article. Each probability is multiplied by a weight (w) for the respective base. Finally, the Fickett score is calculated as: $\text{Fickett Score} = \sum_{i=1}^{8} p_i w_i$.
GC	4	gc_content	GC content	3	GC content in the 1st, 2nd and 3rd position of codons.
	5	gc_frame_list	GC1_frame_score GC2_frame_score GC3_frame_score	3	The variance of GC-content in the 1st, 2nd and 3rd among three reading frames.

transcript-related	6	CTD	CTD	30	The nucleotide composition (C) describes the percent composition of each nucleotide (A, T, C, G) in a transcript sequence. The nucleotide distribution (D) describes five relative positions along the transcript sequence of each nucleotide, with the 0 (first one), 25%, 50%, 75% and 100% (last one). The nucleotide transition (T) describes the percent frequency with the conversion of four nucleotides between adjacent positions.
	7	distance_list	Euclidean-distance to lncRNA Euclidean-distance to PCT Logarithm-distance to lncRNA Logarithm-distance to PCT Euclidean-distance ratio Logarithm-distance ratio	6	$\text{EucDist.LNC}=\sqrt{\sum(\text{freq.seq}(i)\text{-freq.lnc}(i))^2}$, $\text{EucDist.PCT}=\sqrt{\sum(\text{freq.seq}(i)\text{-freq.pct}(i))^2}$, $\text{logDist.LNC}=\frac{1}{n}\sum\frac{\text{freq.seq}(i)}{\text{freq.lnc}(i)},i=1,2,\ldots,4^k$, $\text{logDist.PCT}=\frac{1}{n}\sum\frac{\text{freq.seq}(i)}{\text{freq.pct}(i)},i=1,2,\ldots,4^k$, $\text{EucDist.Ratio}=\frac{\text{EucDist.LNC}}{\text{EucDist.PCT}}$, $\text{logDist.Ratio}=\frac{\text{logDist.LNC}}{\text{logDist.PCT}}$, where freq.seq are the k-adjoining base(s) frequencies of an unevaluated sequence; freq.lnc are the average frequencies of lncRNAs' k-adjoining base(s); freq.pct are the average frequencies of PCTs' k-adjoining base(s); *i* denotes the different types of k-adjoining base(s), and n is the total number of the k-adjoining base(s) in one sequence.
	8	hexmaer	Hexamer score	1	For a given DNA sequence, the probability of the sequence under the coding and noncoding models is calculated. $F_c(h_i)$ and $F_{nc}(h_i)$ are calculated from coding and non-coding training sets respectively, refer to in-frame hexamer frequency (*i*=1, 2, ...,4096). For a given hexamer sequences $S=H_1, H_2, \ldots, H_m$, the hexamer score can be calculated as: $\text{Hexamer Score} = \frac{1}{m}\sum_{i=1}^{m}\log(\frac{F_c(h_i)}{F_{nc}(h_i)})$.
	9	3_mer	3-mer	64	The occurrences of k-length contiguous subsequences.
	10	2_mer	2-mer	16	
structure-related	11	mw	Molecular weight (Mw)	1	Molecular weight of the peptide.
	12	pi	Isoelectric point (pI)	1	Theoretical isoelectric point of the peptide.
	13	gravy	Gravy	1	Average hydrophilicity of the peptide.
	14	ii	Instability index	1	Average stability of the peptide.
ORF-related	15	ORF_length	The longest ORF length	1	Longest ORF length.
	16	ORF_cover	ORF coverage	1	The ratio of longest ORF and transcript lengths.
ORF-attention1	17	ORF1_length	ORF length	1	Length of ORF1.
	18	ORF1_fickett	Fickett TESTCODE	1	Fickett score of ORF1.

			Score		
	19	ORF1_gc	GC content	3	GC content of ORF1.
	20	ORF1_gc_frame_list	GC_frame_score	3	GC frame score of ORF1.
	21	ORF1_CTD	CTD	30	CTD of ORF1.
	22	ORF1_3_mer	3-mer	64	3_mer of ORF1.
	23	ORF1_hexamer	Hexamer score	1	Hexamer score of ORF1.
	24	ORF1_mw	Molecular weight (Mw)	1	Mw of ORF1.
	25	ORF1_gravy	Gravy	1	Gravy of ORF1.
	26	ORF1_ss	Secondary Structure	3	Calculate fraction of helix, turn and sheet. Amino acids in sheet: E, M, A, L. Amino acids in Turn: N, P, G, S. Amino acids in helix: V, I, Y, F, W, L.
	27	ORF1_EIIP		5	Nucleotides can be replaced with the following EIIP values: {A: 0.1260; C: 0.1340; G: 0.0806, T: 0.1335}. Let $X_e[n]$ be the EIIP indicator sequence of $Seq[n]$. Using FFT on $X_e[n]$ can get the corresponding power spectrum $\{S_e[k]\}(k=0, 1, 2, \ldots, N-1)$: $X_e[k]=\sum_{n=0}^{N-1} X_e[k]\, e^{-j\frac{2\pi kn}{N}}, S_e[k]=\left\|X_e[k]\right\|^2$ We extract quantile statistics features by calculating the top 0%, top 25%, top 50%, top 75% top 100% of the sorted power spectrum.
ORF-attention2	28	ORF2_length	ORF length	1	Length of ORF2.
	29	ORF2_fickett	Fickett TESTCODE Score	1	Fickett score of ORF2.
	30	ORF2_gc	GC content	3	GC content of ORF2.
	31	ORF2_gc_frame_list	GC_frame_score	3	GC frame score of ORF2.
	32	ORF2_CTD	CTD	30	CTD of ORF2.
	33	ORF2_3_mer	3-mer	64	3_mer of ORF2.
	34	ORF2_hexamer	Hexamer score	1	Hexamer score of ORF2.
	35	ORF2_mw	Molecular weight (Mw)	1	Mw of ORF2.
	36	ORF2_gravy	Gravy	1	Gravy of ORF2.
	37	ORF2_ss	Secondary Structure	3	Secondary Structure of ORF2.
	38	ORF2_EIIP		5	EIIP values of ORF1.
ORF-attention3	39	ORF3_length	ORF length	1	Length of ORF3.
	40	ORF3_fickett	Fickett TESTCODE Score	1	Fickett Score of ORF3.
	41	ORF3_gc	GC content	3	GC content of ORF3.
	42	ORF3_gc_frame_list	GC_frame_score	3	GC frame score of ORF3.
	43	ORF3_CTD	CTD	30	CTD of ORF3.
	44	ORF3_3_mer	3-mer	64	3_mer of ORF3.
	45	ORF3_hexame	Hexamer	1	Hexamer score of ORF3.

	r	score		
46	ORF3_mw	Molecular weight (Mw)	1	Mw of ORF3.
47	ORF3_gravy	Gravy	1	Gravy of ORF3.
48	ORF3_ss	Secondary Structure	3	Secondary structure of ORF3.
49	ORF3_EIIP		5	EIIP values of ORF1.
Total dimensions			471	

Table S2 Performances of different methods on human data set

Methods	SEN	SPE	A CC	F1	MCC	PRE	AUC	AP
LncFinder	0.9624	0.9820	0.9722	0.9719	0.9446	0.9816	0.9934	0.9912
CPAT (re-trained)	0.9424	0.9684	0.9554	0.9548	0.9111	0.9676	0.9894	0.9876
CPAT (human model)	0.8944	**0.9856**	0.9400	0.9371	0.8837	**0.9842**	-	-
CPC2	0.9520	0.9708	0.9614	0.9610	0.9230	0.9702	0.9919	0.9889
PLEK (re-trained)	0.8768	0.9112	0.8940	0.8921	0.7885	0.9080	-	-
PLEK (default model)	**0.9972**	0.8848	0.9410	0.9441	0.8876	0.8964	0.9916	0.9936
CNCI	0.9732	0.9168	0.9450	0.9465	0.8914	0.9212	0.9422	0.9587
LncCat	0.9720	0.9748	**0.9734**	**0.9733**	**0.9468**	0.9747	**0.9961**	**0.9962**

Bold numbers indicate the highest value of the metrics.

Table S3 Performances of different methods on mouse data set

Method	SEN	SPE	ACC	F1	MCC	PRE	AUC	AP
LncFinder	0.9528	**0.9161**	**0.9344**	**0.9356**	**0.8695**	**0.9191**	**0.9710**	**0.9736**
CPAT (re-trained)	0.9245	0.8838	0.9042	0.9061	0.8090	0.8885	0.9559	0.9543
CPAT (mouse model)	0.8801	0.9061	0.8931	0.8917	0.7864	0.9036	-	-
CPC2	0.9289	0.7933	0.8611	0.8699	0.7290	0.8180	0.9273	0.9405
PLEK (re-trained)	0.8206	0.8150	0.8178	0.8183	0.6356	0.8160	0.8964	0.9067
PLEK (default model)	0.9000	0.7050	0.8025	0.8200	0.6168	0.7531	-	-
CNCI	**0.9644**	0.8622	0.9133	0.9175	0.8310	0.8750	0.9038	0.9245
LncCat	0.9428	0.9006	0.9217	0.9233	0.8441	0.9046	0.9699	0.9628

Bold numbers indicate the highest value of the metrics.

Table S4 Performances of different methods on wheat data set

Methods	SEN	SPE	ACC	F1	MCC	PRE	AUC	AP
LncFinder	**0.9495**	0.8860	0.9178	0.9203	0.8372	0.8928	0.9595	0.9677
CPC2	0.6600	**0.9145**	0.7873	0.7562	0.5941	0.8853	0.9158	0.9263

CPAT (re-trained)	0.8835	0.8120	0.8478	0.8530	0.6973	0.8245	0.9154	0.9121
CPAT (human model)	0.8605	0.8190	0.8398	0.8430	0.6801	0.8262	-	-
PLEK (re-trained)	0.8435	0.8660	0.8548	0.8531	0.7097	0.8629	0.9240	0.9297
PLEK (default model)	0.6680	0.7590	0.7135	0.6998	0.4288	0.7349	-	-
CNCI	0.8755	0.8580	0.8668	0.8679	0.7336	0.8604	0.8551	0.8637
LncCat	0.9435	0.9140	**0.9288**	**0.9298**	**0.8579**	**0.9165**	**0.9759**	**0.9706**

Bold numbers indicate the highest value of the metrics.

Table S1-4 Performances of different methods on chicken data set

Methods	SEN	SPE	ACC	F1	MCC	PRE	AUC	AP
LncFinder (re-trained)	0.9413	0.9294	0.9353	0.9357	0.8707	0.9302	0.9681	0.9674
CPAT (re-trained)	0.9191	0.8989	0.9091	0.9103	0.8183	0.9017	0.9595	0.9495
CPAT (human model)	0.8276	0.9058	0.8666	0.8617	0.7355	0.8986	-	-
CPC2	0.7694	0.9213	0.8453	0.8326	0.6987	0.9071	0.9324	0.9278
PLEK (re-trained)	0.9019	0.9081	0.9050	0.9047	0.8100	0.9075	0.9637	0.9651
PLEK (default model)	0.8906	0.7900	0.8403	0.8480	0.6841	0.8092	-	-
CNCI	0.9267	0.9112	0.9190	0.9195	0.8380	0.9125	0.9091	0.9177
LncCat	**0.9787**	**0.9643**	**0.9715**	**0.9717**	**0.9431**	**0.9648**	**0.9956**	**0.9948**

Bold numbers indicate the highest value of the metrics.

Table S6 Performances of different methods on zebrafish data set

Methods	SEN	SPE	ACC	F1	MCC	PRE	AUC	AP
LncFinder (re-trained)	0.8813	0.8975	0.8894	0.8885	0.7789	0.8958	0.9213	0.9236
CPAT (re-trained)	0.8400	0.8750	0.8575	0.8550	0.7154	0.8705	0.9243	0.9318
CPAT (zebrafish model)	0.7638	**0.9225**	0.8431	0.8296	0.6951	**0.9079**	-	-
CPC2	**0.9150**	0.7875	0.8513	0.8602	0.7083	0.8115	0.9158	0.9213
PLEK (re-trained)	0.7988	0.8275	0.8131	0.8104	0.6265	0.8224	0.8935	0.8931
PLEK (default model)	0.8400	0.6825	0.7613	0.7787	0.5291	0.7257	-	-
CNCI	0.8700	0.8747	0.8723	0.8722	0.7447	0.8744	0.8897	0.8951
LncCat	0.9000	0.8885	**0.8942**	**0.8950**	**0.7885**	0.8900	**0.9532**	**0.9484**

Bold numbers indicate the highest value of the metrics.

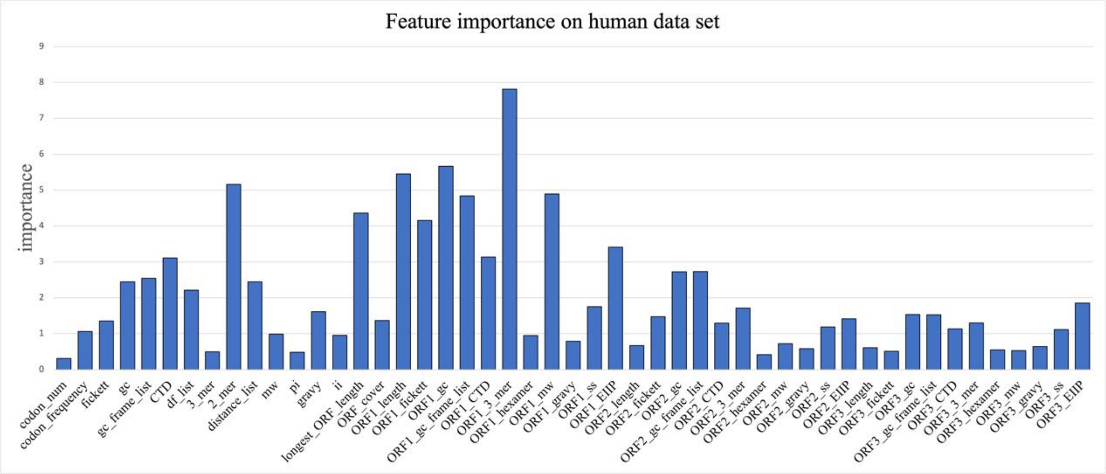

Figure S1 Feature importance on human data set

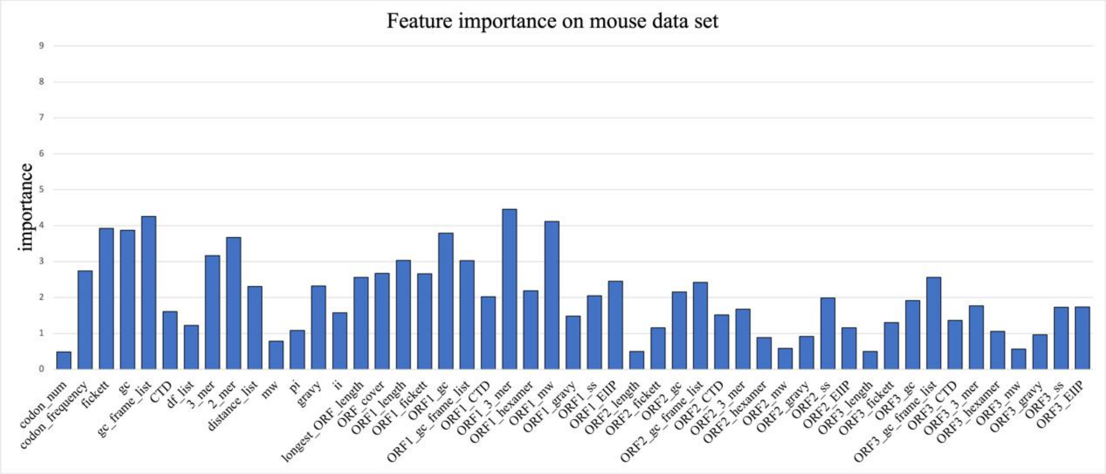

Figure S2 Feature importance on mouse data set

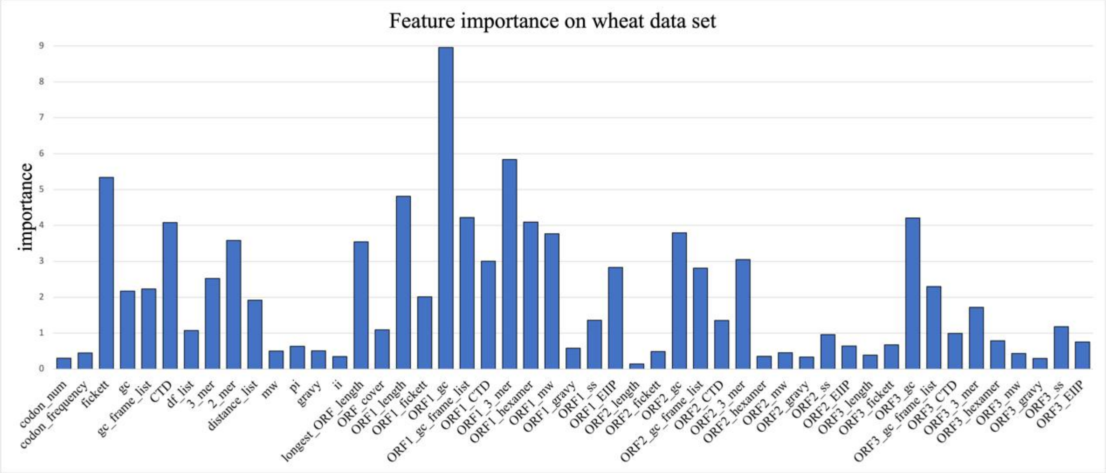

Figure S3 Feature importance on wheat data set

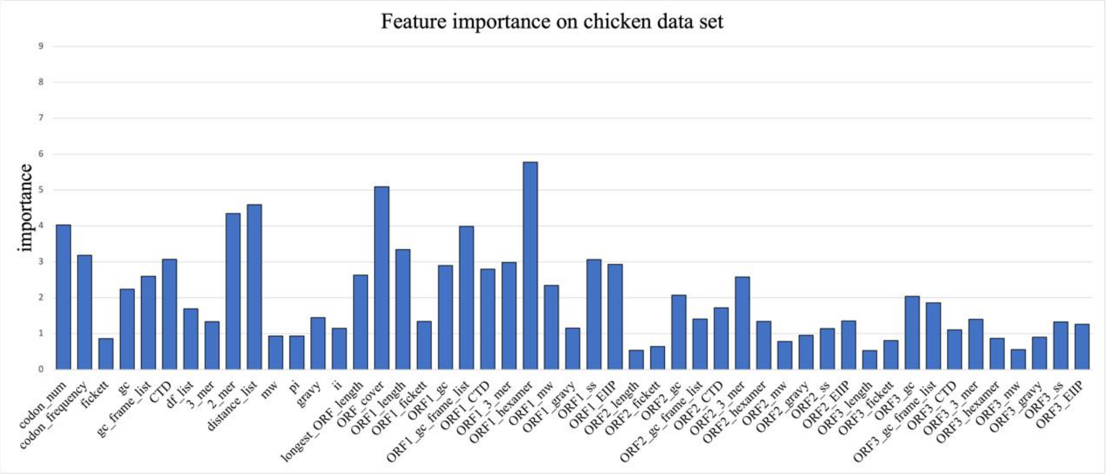

Figure S4 Feature importance on chicken data set

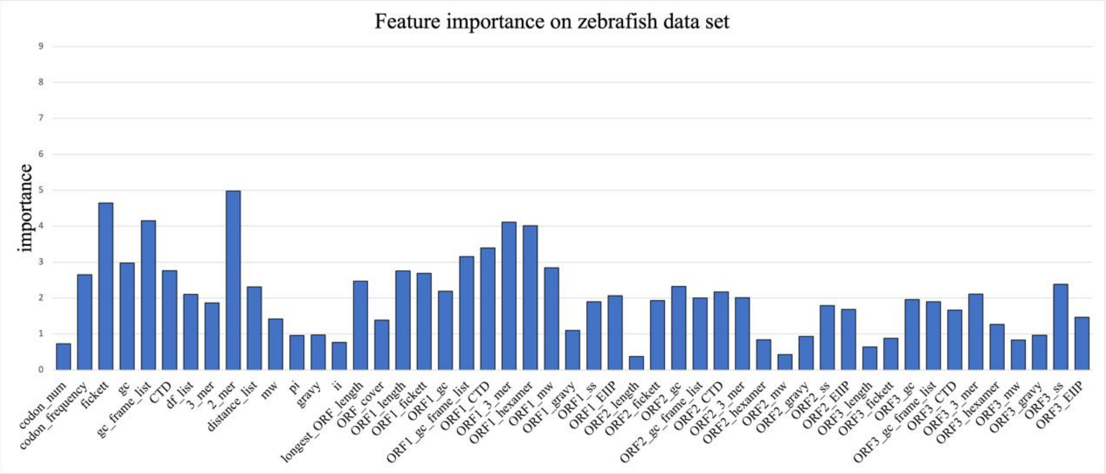

Figure S5 Feature importance on zebrafish data set