IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 13, 2021, accepted November 26, 2021, date of publication November 30, 2021,
date of current version December 27, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3131846

LNCRI: Long Non-Coding RNA Identifier in Multiple Species

SALEH MUSLEH[1], MOHAMMAD TARIQUL ISLAM[2], AND TANVIR ALAM[1]
[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[2]Department of Computer Science, Southern Connecticut State University, Connecticut, New Haven, CT 06515, USA

Corresponding author: Tanvir Alam (talam@hbku.edu.qa)

This work was supported by the College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar.

ABSTRACT The pervasive nature of long non-coding RNA (lncRNA) transcription in the mammalian genomes has changed our protein-centric view of genomes. But the identification of lncRNAs is an important task to discover their functional role in species. The rapid development of next-generation sequencing technology leveraged the opportunity to discover many lncRNA transcripts. However, the cost and time-consuming nature of transcriptomics verification techniques barred the research community from focusing on lncRNA identification. To overcome these challenges we developed LNCRI (Long Non-Coding RNA Identifier), a novel machine learning (ML)-based tool for the identification of lncRNA transcripts. We leveraged weighted k-mer, pseudo nucleotide composition, hexamer usage bias, Fickett score, information of open reading frame, UTR regions, and HMMER score as a feature set to develop LNCRI. LNCRI outperformed other existing models in the task of distinguishing lncRNA transcripts from protein-coding mRNA transcripts with high accuracy in human and mouse. LNCRI also outperformed the existing tools for cross-species prediction on chimpanzee, monkey, gorilla, orangutan, cow, pig, frog and zebrafish. We applied the SHAP algorithm to demonstrate the importance of most dominating features that were leveraged in the model. We believe our tool will support the research community to identify the lncRNA transcripts in a highly accurate manner. The benchmark datasets and source code are available in GitHub: http://github.com/smusleh/LNCRI.

INDEX TERMS Long non-coding RNA, lncRNA, mRNA, machine learning, sequence analysis.

I. INTRODUCTION

About About 2% of human genomic regions are involved in encoding proteins, and the rest are non-coding regions which do not finally produce proteins [1]. The intricating transcriptional landscape in humans has opened a new paradigm of pervasive transcription process which led to the discovery of novel non-coding RNAs and their role in cellular and functional processes. Long non-coding RNAs (lncRNAs), which are defined as a type of ncRNA having more than 200 nucleotides in length, have recently been shown to be evident in linking mutations in their sequence and their role in the dysregulation for many diseases [2]. H19 was one of the earliest discovered lncRNAs having very similar characteristics to protein coding genes considering the polyadenylation, splicing, localization, and the transcription mechanism in support of RNA polymerase II [3]. XIST was among the early examples of lncRNAs which are known to be involved in the silencing of X-chromosome [4]. Since these pioneering discoveries of H19 and XIST, many lncRNAs have been discovered in human and other mammalian species as well as in plants [5]. LncRNAs, predominantly being considered junk regions for decades [6], are now recognized as the most cryptic but functionally most crucial player in biomedical research. Many other lncRNAs are known to play a significant role in a multitude of human diseases, and readers are referred to the articles [2], [7]–[9] to have a deeper understanding of their role in multiple cellular processes and diseases.

The pervasive transcriptomics nature of the lncRNA in human [10] and mouse [11] has already been established by the FANTOM consortium in their catalogue of lncRNAs. The FANTOM Consortium proposed over 23K lncRNA genes with highly accurate 5' end [12]. Recent version (version 37 and version 26 for mouse, respectively) of GENCODE [13] release provides a list of 18K and

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

13K lncRNA genes from human and mouse, respectively. MiTranscriptome has collected over 58K lncRNA genes in humans [14], however, it is not well established if all of them have functional evidence. NONCODE version 6.0 prepared a collection of 96K and 87K lncRNA genes from human and mouse, respectively [15]. Based on the above-mentioned landmark projects on lncRNA, we can observe that the discovery of lncRNAs is continuing. Advances in next-generation sequencing (NGS) techniques have provided the scientific community to discover a large number of lncRNA transcripts and their cellular roles [12]. Although thousands of lncRNAs transcripts have already been discovered across species, there is a huge number of lncRNAs that are yet to be identified and annotated in multiple species and this is a challenging task. Firstly, lncRNAs may have similar biogenesis characteristics to those of messenger RNA (mRNA) as both are mainly transcribed the RNA Polymerase II [16]. Secondly, lncRNAs can even undergo the similar transcriptional and post-transcriptional processes as mRNA transcripts [17]. Thirdly, mRNA and lncRNA transcripts have similarities in terms of transcript length as well as splicing structure [18], [19], which make the lncRNA identification process complicated. Fourthly, the low-level expression of lncRNAs also hinder their discovery in multiple cells or tissues [16].

To overcome the challenges of the currently available experimental technologies, many computational methods, mainly machine learning (ML)-based techniques, have been leveraged to distinguish lncRNA transcripts from mRNA transcripts. The common formulation of this ML based approach is to put the transcripts from lncRNA and mRNA under a classification framework. There exist multiple ML-based methods for recognizing lncRNA transcripts from mRNA transcripts based on ML-based techniques. One of the pioneering works in this domain was CONC (''coding or non-coding''), where the authors applied support vector machine (SVM) based ML-model to distinguish mRNA transcripts which are responsible for producing proteins, from non-coding RNAs (ncRNA) transcripts [20]. The authors used mammalian and non-mammalian eukaryotic ncRNA transcripts from RNAdb [21], NONCODE [5] database and mRNA transcripts from SwissProt [22] database to train ML models. Lia *et al.* developed PLEK (''predictor of lncRNAs and mRNAs based on an improved k-mer scheme'') to recognize lncRNAs transcripts from mRNA transcripts by introducing an improved k-mer scheme [23]. The authors proposed k-mers (k=1 to 5) with sliding window sizes up to five with a step size of one to encode the whole transcripts. Then the proposed k-mer based features were fed into a SVM-based model to classify lncRNA transcripts from mRNA transcripts for human, mouse and other vertebrates. Sun *et al.* developed the CNCI (''Coding-Non-Coding Index'') tool to distinguish protein-coding transcripts from lncRNA transcripts based on the intrinsic composition of sequences [24]. The authors proposed a novel encoding approach of the sequence based on adjoining neighboring triplets representing all possible two consecutive triplets (64*64 possible combinations) from the sequence. Out of six reading frames, the authors selected only one most-like CDS (MLCDS) and used the length and score of MLCDS as a feature vector. Using the proposed features, SVM-based ML model was built to classify lncRNA transcripts and protein-coding transcripts. CNIT (''Coding-Non-Coding Identifying Tool''), an updated version of CNCI, was developed for the same purpose with higher accuracy and much faster speed [25]. Han *et al.* developed LncFinder, an lncRNA recognition system which recommend 19 different features summarizing the sequence composition, structural information, and physicochemical properties of the nucleotides to find lncRNA from human, mouse, chicken, zebrafish, and wheat [26].

Recently, deep learning (DL)-based models have been proposed for the identification of lncRNA transcripts. Baek *et al.* developed lncRNAnet [27], a DL-based model by combining convolutional neural network (CNN) and recursive neural network (RNN). RNN was used to detect the intrinsic nature of the input sequence. Variable-length sequences were handled using the Bucketing technique [28] along with k-mer embedding vectors, and the network was trained through the GloVe embedding [29] for the identification of lncRNA. Yang *et al.* developed LncADeep for the identification of partial and full-length lncRNA transcripts by incorporating hand-curated features like Fickett score, hexamer score, coding sequence (CDS) length etc. and then feeding into a deep belief network based model [30]. Tripath proposed DeepLNC, a deep neural network-based model that used k-mers (k=1 to 5) from input sequences as a feature set to classify lncRNA transcripts and mRNA transcripts [31]. Interested readers may check the reviews in [32], [33] which summarize different ML-based methods that have been proposed to identify lncRNAs.

In this study, we propose LNCRI (Long Non-Coding RNA Identifier), a novel ML-based pipeline to distinguish lncRNA transcripts from mRNA transcripts. To evaluate the prediction performance of LNCRI and compare it with other existing tools, we used benchmark datasets for multiple species. We found that LNCRI outperformed the existing state-of-the-art tools for lncRNA identification in human, mouse and eight other species. Our contribution in this work can be summarized as follows:

1) We have used the largest collection of lncRNA transcripts and mRNA transcripts from GENCODE and RefSeq for the identification of lncRNA transcripts in multiple species.
2) We proposed a novel combination of features representing weighted k-mer, pseudo nucleotide composition, hexamer usage bias, Fickett score, information of open reading frame, UTR regions, and HMMER score to distinguish lncRNA transcripts from protein-coding transcripts.
3) We proposed a CatBoost based model LNCRI, which achieved the best performance compared to other

existing methods considering multiple evaluation metrics for human and mouse.

4) LNCRI outperformed other existing models for eight other species in the cross-species transcript prediction task.

II. MATERIALS AND METHODS

A. DATA COLLECTION

We collected all the lncRNA transcripts from GENCODE release 37 for human and release 26 for mouse. GENCODE database, launched by Human Genome Research Institute (NHGRI) under a project named The ENCyclopedia Of DNA Elements (ENCODE), is one of the largest and most reliable sources for human and mouse functional elements [34]. The GENCODE database mainly incorporates four types of functional elements: (a) protein-coding genes, (b) pseudo genes, (c) long non-coding RNA genes, and (d) small non-coding RNA genes [35]. The genes/transcripts in GENCODE were annotated based on computational approach supported by manual annotation and experimental validation. For lncRNA annotation, GENCODE does not a apply strict 200bp length threshold, albeit very few annotated lncRNAs fall below this threshold [35]. For protein-coding transcripts, we collected all the transcripts from RefSeq [36]. RefSeq database, established by the National Center for Biotechnology in the United States, provides a comprehensive collection of well annotated, non-redundant set of sequences for mRNA transcripts and other non-coding RNA transcripts. RefSeq covers sequence from multiple species including human and mouse. From RefSeq, we considered only the transcript that had clear potential for clear protein-coding (i.e., having an NM RefSeqID) capability. Interested readers are suggested to check [37] to compare the annotation pipeline of GENCODE and RefSeq. All the transcript sequences were collected from the genome assembly version GRCh37 and GRCm10 for human and mouse, respectively. The collection contains 97,482 lncRNA transcripts and 104,760 protein-coding transcripts from human. The collection also contains 18,833 lncRNA transcripts and 37,907 protein-coding transcripts from mouse.

B. DATA SET PREPROCESSING

From the collected dataset, transcript sequences having characters other than ''A'', ''C'', ''G'', or ''T'' (represents U in the corresponding RNA) were discarded. Then, we removed the duplicate sequences. To avoid redundancy from the collected dataset, we applied CD-HIT [38]. As suggested in [39], sequences having more than 80% similarity, based on CD-HIT, were dropped to avoid any bias for the ML model. Moreover, sequences shorter than 200 nucleotides (nt) or longer than 3000 nt were removed from our analysis as prescribed in lncRNAnet [27]. To make both lncRNA and mRNA transcript datasets balanced, we down sampled the mRNA transcript dataset to be in bar with the lncRNA transcript dataset. After all the pre-processing steps, we had 43,839 and 43,383 transcripts from human lncRNA and mRNA, respectively. We also found 3,295 and 2,828 transcripts from mouse lncRNA and mRNA, respectively. Hereafter we will refer this collection of dataset from human and mouse as permissive dataset. To avoid any bias and to check if the CD-HIT cut-off has any significant effect on the performance of machine learning model, we generated another dataset for both human and mouse based on 60% CD-HIT cut-off. This cut-off generated 41,817 and 36,433 transcripts from human lncRNA and mRNA, respectively as well as 3,292 and 2,381 transcripts from mouse lncRNA and mRNA, respectively. Hereafter we will refer this dataset as stringent dataset.

C. K-MER RELATED FEATURES

For each sequence we counted the frequencies of mono-, di-, tri-consecutive nucleotides in the whole transcript body. Then we normalize the k-mer count by the sequence length and calibrated by the possible combination of k-mer. This generated 84 features (4 from mono-, 16 from di- and 64 from tri-consecutive nucleotides), for the development of ML models based on the following equation.

$$WeightedKmer_i = \frac{C_i}{L} * \frac{1}{4^{3-k}}; \quad k = 1, 2, 3 \tag{1}$$

where, C_i represents the count of k-mer in the transcript and L represents the transcript length. We also checked the observed/expected ratio of nucleotide combinations in lncRNAs and mRNAs as suggested in [40]. Supplementary File 01 provides the details of observed/expected ratio of all mono-, di- and tri-nucleotides.

D. PSEUDO K-TUPLE NUCLEOTIDE COMPOSITION (PseKNC)

The pseudo k-tuple nucleotide composition (PseKNC) reflects the physicochemical properties and sequence-order effects of nucleotides in DNAs [41], [42]. The sequence-order information is preserved through the physiochemical properties of the constituent oligonucleotides. The dimension of this feature vector is of (4k+λ) where k represents k-mer (having a positive integer value), and λ represents the highest counted rank of the correlation along a DNA sequence. In our case: k=3 and λ=10 were used to generate a 74-dimension feature vector.

E. ORF RELATED INFORMATION

Open reading frame (ORF) is a well-known property of mRNA transcripts. ORF information has been used in multiple previous studies [27], [30] to distinguish lncRNA transcripts from mRNA transcripts. We considered the length of the longest ORF from three forward frames, starting with the start codon (''ATG'') and ending with any of the stop codons (''TAG'', ''TAA'', or ''TGA''). We also considered the longest ORF coverage defined as the ratio of the longest ORF length to the whole transcript length. This provided a total of two features for the development of ML models.

F. HEXAMER USAGE BIAS

Hexamer has shown to be effective in the discrimination of protein-coding sequence from the non-coding sequence they can capture potential adjacent amino acids based on codons [43]. Inspired by this phenomena, we calculated the hexamer score representing the log-likelihood ratio of the presence of hexamer in coding to non-coding sequence. For the longest ORF, we calculated the average hexamer score of all the hexamers as suggested in [30] in the following way.

$$AverageHexamerScore = \frac{1}{n}\sum_{i=1}^{n}\log\frac{FC(h_i)}{FNC(h_i)} \tag{2}$$

where n is the total number of hexamers in a sequence. FC(h_i) and FNC(h_i), i=1, 2, …, 4096 represent the frequency of hexamer h_i in all the training coding and non-coding sequences, respectively. This provided one feature for the development of ML models.

G. FICKETT SCORE

In 1982, Fickett [44] demonstrated that coding regions may have asymmetric codon bias and nucleotide content that could be considered to distinguish non-coding regions from protein-coding regions. We calculated the nucleotide (A, C, G, or T) composition that is favored by the first, second and the third position of codon in a transcript.

$$A_{fickett} = \frac{Max(A_1, A_2, A_3)}{Min(A_1, A_2, A_3) + 1} \tag{3}$$

where A_1, A_2, A_3 represent the number of A in the first, second and the third position of codon in a transcript. We calculated the Fickett score for C, G, and T as well. This provided four features for the development of ML models.

H. UTR REGIONS

For mRNA transcripts, untranslated regions (UTR) may contain some specific pattern compared to lncRNA transcripts [30]. To capture such characteristics, we first identified the longest ORF. Then we considered the upstream region of the start codon and the downstream region of the stop codon as 5' UTR and 3' UTR, respectively as suggested in [30]. Then we calculated the ratio of UTR length to the transcript length as coverage of UTR. In this way, we generated two features. We also computed the CG content of the 5' UTR and 3' UTR, which provided two more features for the development of ML models.

I. CONSERVATION SCORE

As lncRNAs genes are less conserved than the protein-coding genes, conservation profile would be a good distinguishing feature for separating mRNA transcripts from lncRNA transcripts. We aligned each transcript against Pfam [45] version 34 using HMMER [46]. Based on the alignment we extracted eleven features for the development of ML models. We considered bit score of overall sequence alignment and the matched domain, e-value for overall sequence alignment and the matched domain, length of query and target sequence, mean posterior probability reflecting how reliable the alignment was, etc. Additionally, we used the HMM alignment ration (ratio of the length of the aligned region to the input sequence) for the development of ML models.

J. DEVELOPMENT OF CLASSIFICATION MODELS

After the Data collection step, we used the selected features to build different classifiers to distinguish lncRANs from protein coding ones. We used Decision Tree (DT), Support Vector Machine (SVM), Artificial Neural Network (ANN), Random Forest (RF), eXtreme Gradient Boosting (XGB), CatBoost algorithms to classify these two types of transcripts. The mRNA and lncRNA transcripts were considered as the positive and the negative set, respectively for the development of ML models. We used Python Scikit-learn GridSearchCV for hyperparameter tuning with five-fold cross validation. The details of parameter optimization are provided in Supplementary File 02. We applied a five-fold cross validation technique to evaluate the performance of the model. We used 80% data as the training set and the remaining 20% data as the test set. We used the following performance evaluation metrics for the models:

$$Accuracy(A_{CC}) = \frac{(tp + tn)}{(tp + tn + fp + fn)} \tag{4}$$

$$Sensitivity(S_n) = \frac{tp}{tp + fn} \tag{5}$$

$$Specificity(S_p) = \frac{tn}{tn + fp} \tag{6}$$

$$MCC = \frac{(tp * tn - fp * fn)}{\sqrt{(tp+fp) * (fp+fn) * (tn+fp) * (tn+fn)}} \tag{7}$$

where tp, fp, tn and fn represent number of true positive, false positive, true negative and false negative samples, respectively predicted by the model.

III. RESULTS

A. FICKETT SCORE AND HEXAMER SCORE PATTERNS AT THE lncRNA AND mRNA TRANSCRIPTS

Fickett score represents the combined effect of nucleotide (nt) composition and their codon usage bias [43]. It also reflects the degree at which a nt is favored in codon positions. Figure 1 shows the distribution of Fickett scores for nts in human. The Fickett score was relatively high for C in lncRNA transcripts compared to mRNA transcripts (lncRNA:mRNA = 0.054 ± 0.017: 0.023 ± 0.016). On the other hand, Fickett score for T in was relatively high in mRNA compared to lncRNA (lncRNA:mRNA = 0.016 ± 0.011: 0.039 ± 0.032). This summarily represents higher codon usage bias of C and T in lncRNA and mRNA, respectively.

Figure 2 shows that the distribution of hexamer scores was relatively higher in mRNA transcripts compared to lncRNA transcripts representing the specific hexamer patterns that are prevalent in mRNA transcripts compared to lncRNA

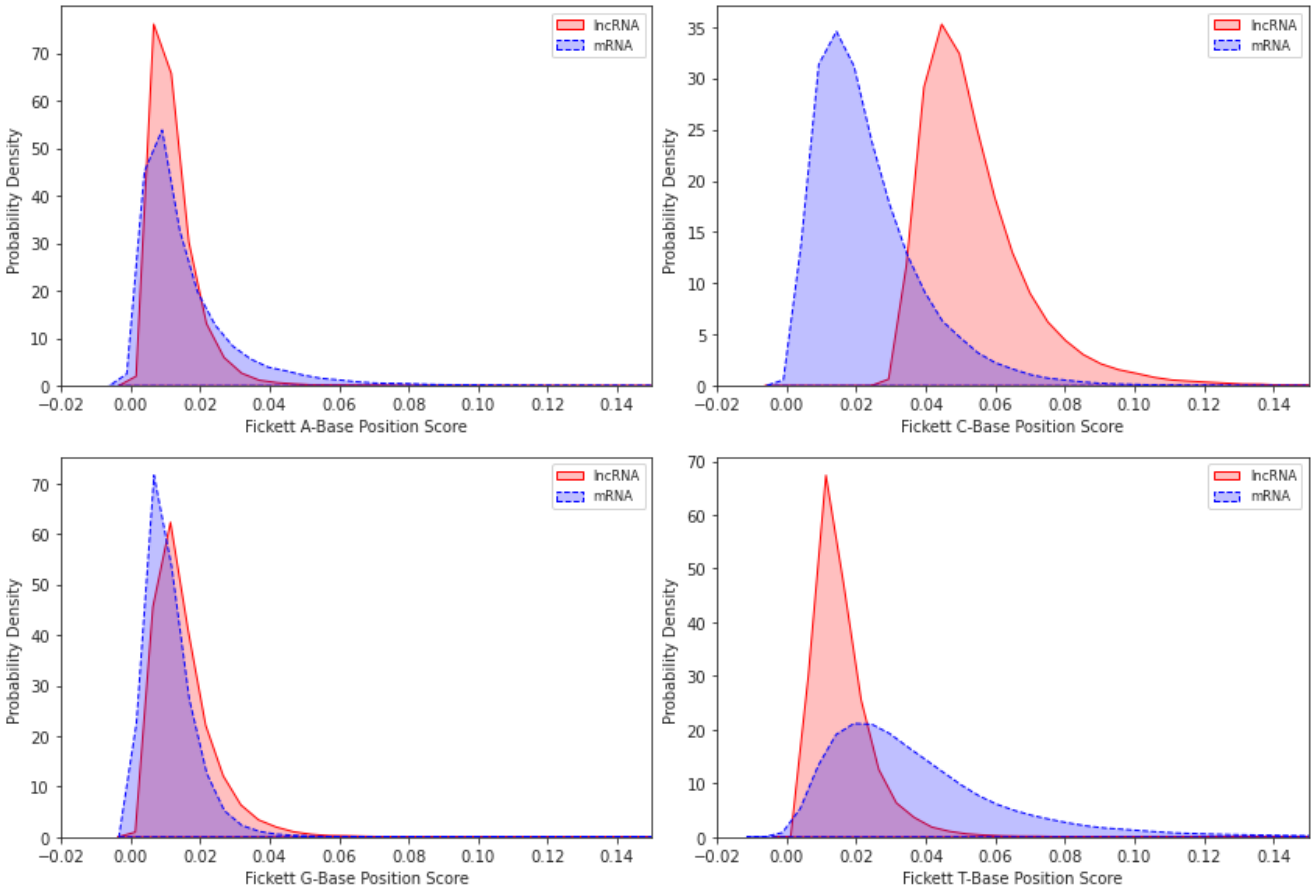


FIGURE 1. Distribution of Fickett score in human lncRNA transcripts and mRNA transcripts.

transcripts. It perfectly aligns with previous studies showing that protein-coding genes contain specific hexamer patterns, which are rare in the rest of the genome [47], [48].

B. SEQUENCE PATTERNS AT THE UTR REGIONS OF lncRNA AND mRNA TRANSCRIPTS

3' UTR regions generally have longer length than the 5' UTR region [49] and the same pattern is observed for both lncRNA transcripts and mRNA transcripts (Figure 3a and 3b). For both 5' and 3' UTR, the length (UTR ratio) was relatively high for mRNA transcripts compared to the lncRNA transcripts. The GC content in the genic region of lncRNA [50] and in the promoter region of lncRNA models [33], [48] are not as enriched as protein coding genes. But the UTR regions of lncRNA transcripts are more GC enriched than the mRNA transcripts (Figure 3a and 3b). UTR with high GC content tend to enhance the gene regulation ability [49], which is one of the known functions of lncRNAs [51].

C. ORF AND HOMOLOGY OF lncRNA AND mRNA TRANSCRIPTS

Long putative ORF is highly unlikely to be present in any random sequence including noncoding sequences and ORF

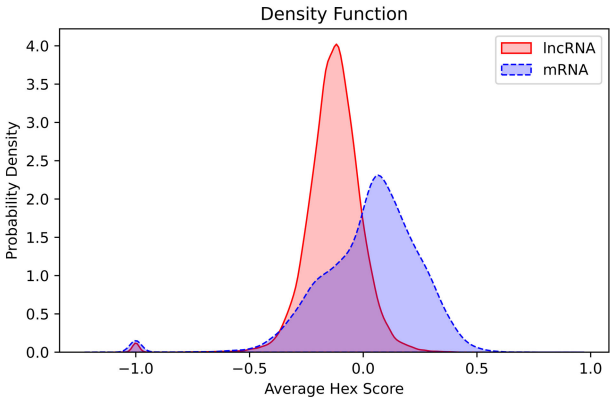
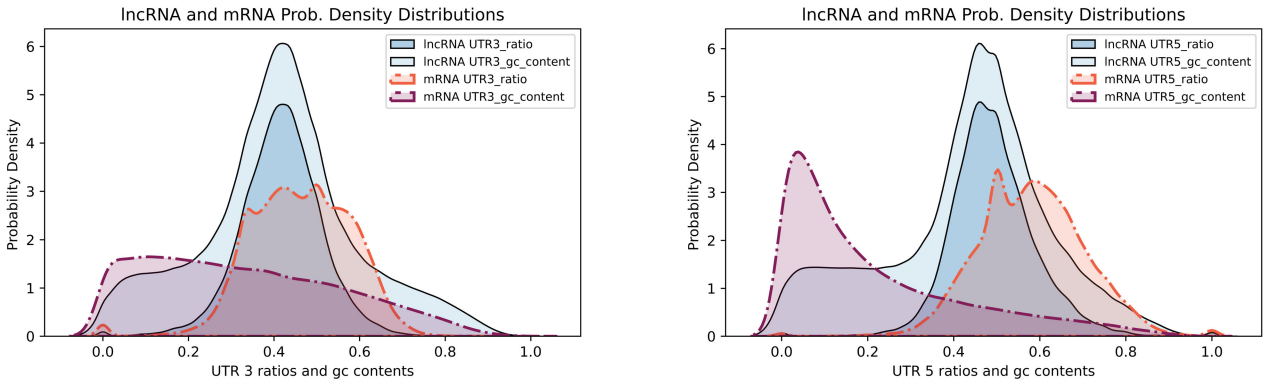


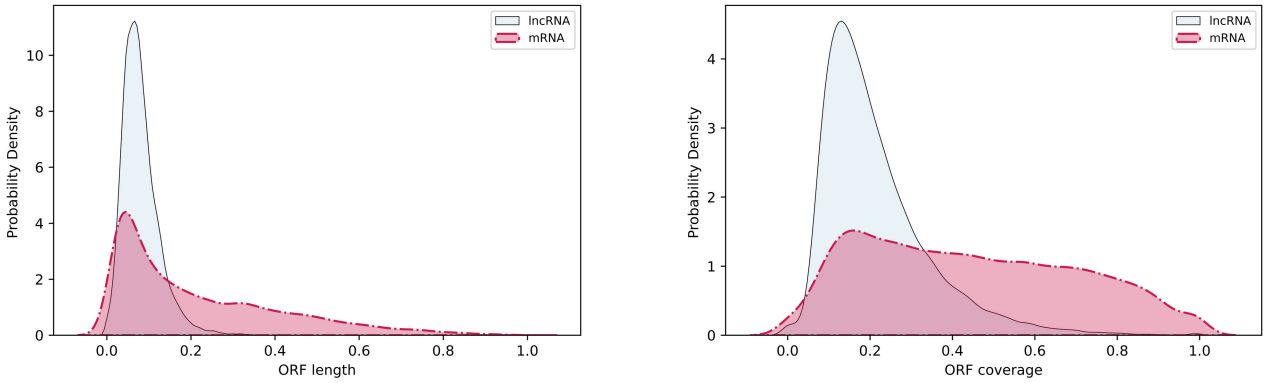
FIGURE 2. Distribution of Hexamer score in human lncRNA transcripts and mRNA transcripts.

over 100 length codons is usually considered to be a highly likely protein-coding sequence. Therefore, we observed higher ORF length and coverage in the protein coding transcripts Figures 4a and 4b, though some lncRNAs may have long ORF length. And this perfectly aligns with the known ORF length distribution in literature [50].



(a) (b)

FIGURE 3. Length and GC content distribution of UTR regions in transcripts. Figure 3a shows the ratios and GC contents at 3′ UTR, while Figure 3b shows ratios and GC contents at 5′ UTR.



(a) (b)

FIGURE 4. Distribution of ORF length and coverage in human lncRNA transcripts and mRNA transcripts. Figure 4a shows ORF Length, while Figure 4b shows ORF Coverage.

As protein coding genes are more conserved than lncRNA genes, searching homology using HMMER against Pfam database would provide higher similarity for mRNA genes than lncRNA genes and this pattern is clearly observed in Figure 5a and 5b. Supplementary File 03 provides the distribution of hexamer usage bias, Fickett score, open reading frame, UTR regions, HMMER score in mouse.

D. PERFORMANCE OF LNCRI IN HUMAN AND MOUSE DATASETS

Table 1 highlights the performance of ML models on different types of features that were used in the model. Based on ablation study, we can observe that the CatBoost based model performed the best among all the models we evaluated. The performance of the XGBoost and CatBoost models were very close but CatBoost based model slightly outperformed the XGBoost model (Table 1).

Among the type of features, weighted k-mer, PseKNC and UTR-based features had the distinguishing capability level of ∼85%, ∼85%, ∼80% Acc, respectively in both human and mouse (Figure 6). Fickett Score (Acc of ∼79%:∼73% in human:mouse) and Hexamer Score (Acc of ∼76%:∼73% in human:mouse) based features showed better performance in human compared to mouse (Figure 6). Interestingly we observed more distinguishing capability of ORF (Acc of ∼77%:∼84% in human:mouse) and HMMER based feature (Acc of ∼88%:∼97% in human:mouse) in mouse compared to the human (Figure 6). Combining all the features the CatBoost based model achieved the best performance with 93% Sn and 95% Sp for human and 97% Sn and 99% Sp for mouse (Figure 6, Table 1).

E. PERFORMANCE OF LNCRI AND OTHER EXISTING TOOLS ON PERMISSIVE DATASET

We compared the performance of LNCRI against state-of-the-art model for the classification of lncRNA transcripts from mRNA transcripts. We used the benchmark dataset for human and mouse to compare the performance of LNCRI against six other tools: CPC2, CNCI, PLEK, CNIT, CPAT and LncADeep. LNCRI outperformed all the tools for mRNA

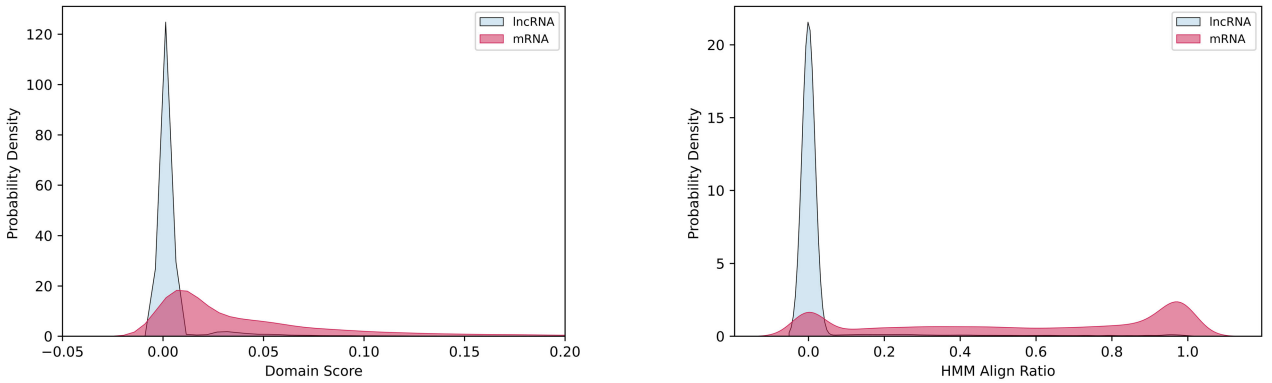


FIGURE 5. HMMER based feature distribution in human lncRNA transcripts and mRNA transcripts. Figure 5a shows domain score distribution, while Figure 5b shows align ratio distribution.

TABLE 1. Performance of ML models based on ablation study for Human and Mouse permissive dataset.

		Human						Mouse					
Feature (no of feature)	**Evaluation Metric**	**DT**	**SVM**	**ANN**	**RF**	**XGBoost**	**CatBoost**	**DT**	**SVM**	**ANN**	**RF**	**XGBoost**	**CatBoost**
K-mer (84)	S_n	0.68	0.82	0.85	0.61	0.82	0.82	0.67	0.82	0.75	0.73	0.84	0.84
	S_p	0.78	0.86	0.7	0.79	0.84	0.85	0.85	0.87	0.82	0.82	0.87	0.89
	A_{cc}	0.73	0.84	0.78	0.7	0.83	0.84	0.76	0.85	0.78	0.77	0.86	0.86
	MCC	0.47	0.67	0.56	0.41	0.66	0.68	0.52	0.69	0.57	0.55	0.71	0.72
PseKNC (74)	S_n	0.71	0.81	0.61	0.63	0.83	0.84	0.68	0.85	0.77	0.98	0.82	0.85
	S_p	0.75	0.85	0.70	0.78	0.85	0.86	0.81	0.88	0.85	0.08	0.90	0.88
	A_{cc}	0.73	0.83	0.65	0.70	0.84	0.85	0.75	0.86	0.81	0.53	0.86	0.86
	MCC	0.46	0.67	0.31	0.41	0.68	0.69	0.50	0.73	0.62	0.14	0.72	0.73
ORF (2)	S_n	0.62	0.52	0.52	0.62	0.68	0.69	0.83	0.83	0.80	0.76	0.84	0.83
	S_p	0.90	0.95	0.95	0.91	0.86	0.85	0.91	0.85	0.92	0.94	0.85	0.85
	A_{cc}	0.76	0.74	0.74	0.76	0.77	0.77	0.87	0.84	0.86	0.85	0.84	0.84
	MCC	0.54	0.52	0.52	0.55	0.55	0.55	0.74	0.69	0.72	0.71	0.69	0.69
Hexamer Score (1)	S_n	0.65	0.64	0.65	0.67	0.68	0.67	0.76	0.75	0.76	0.74	0.63	0.75
	S_p	0.86	0.87	0.86	0.85	0.85	0.85	0.71	0.70	0.72	0.73	0.68	0.70
	A_{cc}	0.76	0.76	0.75	0.76	0.76	0.76	0.74	0.73	0.74	0.74	0.66	0.73
	MCC	0.53	0.53	0.52	0.53	0.54	0.53	0.76	0.75	0.76	0.74	0.63	0.75
Fickett Score (4)	S_n	0.73	0.72	0.71	0.74	0.75	0.75	0.51	0.64	0.53	0.59	0.65	0.64
	S_p	0.84	0.85	0.85	0.83	0.84	0.83	0.95	0.82	0.94	0.84	0.83	0.82
	A_{cc}	0.78	0.79	0.78	0.78	0.79	0.79	0.73	0.73	0.73	0.72	0.74	0.73
	MCC	0.57	0.58	0.57	0.57	0.59	0.58	0.50	0.47	0.51	0.45	0.49	0.47
HMMER (11)	S_n	0.82	0.82	0.82	0.82	0.82	0.82	0.94	0.94	0.94	0.94	0.94	0.94
	S_p	0.93	0.93	0.93	0.93	0.95	0.95	1.00	1.00	1.00	1.00	1.00	1.00
	A_{cc}	0.88	0.88	0.88	0.88	0.88	0.88	0.97	0.97	0.97	0.97	0.97	0.97
	MCC	0.76	0.76	0.76	0.76	0.77	0.77	0.94	0.94	0.94	0.94	0.94	0.94
UTR Region (4)	S_n	0.69	0.73	0.71	0.75	0.75	0.75	0.73	0.76	0.71	0.64	0.76	0.76
	S_p	0.84	0.84	0.83	0.80	0.83	0.83	0.87	0.84	0.87	0.85	0.85	0.84
	A_{cc}	0.77	0.79	0.77	0.78	0.79	0.79	0.80	0.80	0.79	0.75	0.81	0.80
	MCC	0.54	0.58	0.55	0.55	0.58	0.59	0.61	0.60	0.59	0.51	0.62	0.60
All features (180)	S_n	0.89	0.83	0.89	0.85	0.92	0.93	0.97	0.94	0.95	0.94	0.977	0.977
	S_p	0.92	0.94	0.93	0.87	0.95	0.95	0.98	1	1	0.99	0.994	0.996
	A_{cc}	0.91	0.89	0.91	0.86	0.94	0.94	0.97	0.97	0.97	0.96	0.985	0.981
	MCC	0.81	0.77	0.82	0.72	0.87	0.88	0.94	0.94	0.94	0.93	0.971	0.962

transcript prediction (Table 2) for human and very close to lncADeep for lncRNA transcript prediction in human. For mouse, LNCRI outperformed all the tools we compared for mRNA and lncRNA transcripts.

F. PERFORMANCE OF LNCRI IN CROSS-SPECIES PREDICTION TASK

We also compared the performance of LNCRI against multiple species: Danio rerio (Zebrafish), Xenopus tropicalis (Frog), Bos taurus (Cow), Pan troglodytes (Chimpanzee), Sus scrofa (Pig), Macaca mulatta (Monkey), Gorilla gorilla (Gorilla), Pongo abelii (Orangutan). The benchmark datasets for multiple species was collected from [23]. For this purpose,

TABLE 2. Performance of LNCRI and other tools on permissive dataset.

	Accuracy on Human		Accuracy on Mouse	
Tool	**mRNA**	**lncRNA**	**mRNA**	**lncRNA**
CPC2	60.00	95.36	74.00	95.00
CNCI	75.00	97.85	66.03	96.83
PLEK	57.29	95.25	62.62	92.21
CNIT	74.50	**98.50**	66.00	98.00
CPAT	79.00	96.86	78.00	96.00
LncADeep	92.90	97.90	84.00	96.00
LNCRI	**93.00**	95.58	**97.70**	**99.60**

we trained the model using Human dataset and fed the other species dataset into the trained model for inference purpose only. For evaluating cross-species prediction performance of LNCRI, we compared it against two other tools: CNCI and

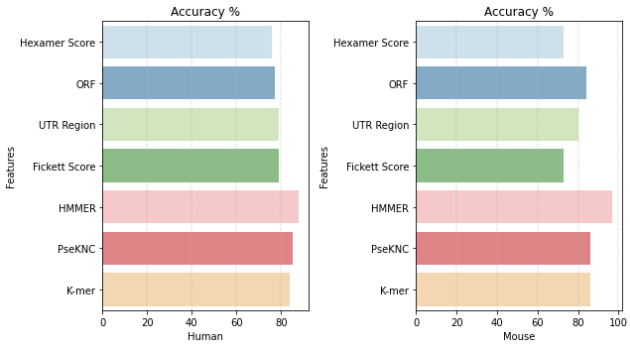


FIGURE 6. Results based on ablation study for CatBoost based model.

TABLE 3. Results on cross-species prediction task for LNCRI and other tools.

Species	Data	Transcript	CNCI	PLEK	LNCRI
Danio rerio (Zebrafish)	mRNA	14493	95.3	91.3	**100.00**
	lncRNA	419	89.3	90.9	**99.80**
Xenopus tropicalis (Frog)	mRNA	8874	92.9	94.5	**99.90**
	lncRNA	279	99.7	**100.00**	**100.00**
Bos taurus (Cow)	mRNA	13190	94.3	94.8	**99.96**
	lncRNA	182	**100.00**	99.5	99.46
Pan troglodytes (Chimpanzee)	mRNA	1906	90.2	87.1	**100.00**
	lncRNA	1166	100.0	**99.9**	99.91
Sus scrofa (Pig)	mRNA	3978	93.4	85.1	**99.95**
	lncRNA	241	95.9	98.3	**99.58**
Macaca mulatta (Monkey)	mRNA	5709	92.0	85.0	**99.98**
	lncRNA	359	99.7	**100.0**	99.16
Gorilla gorilla (Gorilla)	mRNA	33025	87.4	83.8	**99.98**
	lncRNA	367	99.7	99.7	**99.72**
Pongo abelii (Orangutan)	mRNA	3401	93.4	98.0	**99.97**
	lncRNA	392	99.8	**100.0**	98.97

PLEK. For almost all species LNCRI outperformed the tools that we tested (Table 3).

G. PERFORMANCE OF LNCRI AND OTHER EXISTING TOOLS ON STRINGENT DATASET

To avoid any bias and to check if the CD-HIT cut-off has any significant effect on the performance of LNCRI, we generated stringent datasets for both human and mouse based on 60% CD-HIT cut-off. The performance of LNCRI against other existing tools based on the stringent datasets is highlighted in Table 4. We can observe that LNCRI performed almost at the similar level in both permissive and stringent human datasets (Table 2 and Table 4). As like the human permissive dataset, LNCRI and CNIT performed the best for mRNA and lncRNA transcript prediction, respectively for the stringent dataset (Table 2 and Table 4). For mouse stringent dataset, the performance of LNCRI and the majority of other existing tools dropped slightly (Table 4) compared to their performance on permissive dataset (Table 2). But LNCRI performed the best for mouse mRNA transcript prediction task with 94.43% Acc. For lncRNA transcript prediction in mouse, performance of LNCRI (96.05% Acc) was very close to the highest performing tool CNCI (97.04% Acc) (Table 4).

IV. DISCUSSIONS

In this article we proposed LNCRI, a ML based model to identify lncRNAs in human, mouse and eight other species.

TABLE 4. Performance of LNCRI and other tools on stringent dataset.

	Accuracy on Human		Accuracy on Mouse	
Tool	**mRNA**	**lncRNA**	**mRNA**	**lncRNA**
CPC2	60.43	95.44	65.33	95.56
CNCI	75.11	97.72	63.31	**97.04**
PLEK	57.52	95.22	63.31	94.20
CNIT	77.77	**98.87**	64.24	96.59
CPAT	78.18	96.59	68.67	96.29
LncADeep	83.65	98.80	78.79	96.60
LNCRI	**92.21**	96.04	**94.43**	96.05

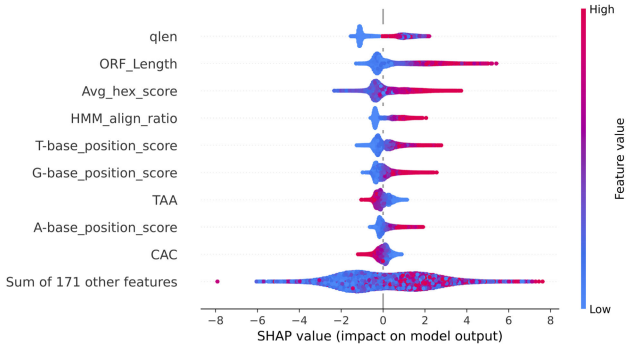


FIGURE 7. Summary plot highlighting the most influential features based on SHAP values for distinguishing lncRNA transcripts from mRNA in human.

The proposed model considered seven different types of features, namely: Weighted k-mer, PseKNC, ORF, Fickett Score, UTR information, Hexamer Score, and HMMER features to distinguish lncRNAs from mRNAs. LNCRI demonstrated better performance in the assigned task of classifying lncRNA transcripts from mRNA transcripts in both human and mouse compared to other existing tools (Table 2, Table 4). As LNCRI performed almost at the similar level for both the permissive and the stringent dataset, it is highly unlikely to have any bias in the proposed model. Moreover, LNCRI trained on human dataset achieved high Acc in the cross-species prediction task for almost all species we tested, indicating the effectiveness of LNCRI in poorly annotated species (Table 3).

To explain the proposed LNCRI model, we investigated the features that contributed in the CatBoost-based model most in distinguishing lncRNA transcripts from mRNA transcripts in human. We leveraged the SHapley Additive exPlanations (SHAP) algorithm [52] to identify the features that contributed most in this task. Figure 7 highlighted the top-ranked nine features based on these SHAP values as identified from the boosting model. The positive SHAP values for the influential features drive the model towards the mRNA class, whereas the negative SHAP values influence the model towards the lncRNA class. Among these dominant features, two features were from the weighted k-mer group: TAA and CAC. The higher values of TAA and CAC drive the model towards lncRNA prediction (Figure 7). The obs/exp ration of TAA and CAC were also higher in lncRNA transcripts compared to mRNA transcripts (Supplementary

File 01). Figure 7 also showed that the Fickett score for T-, G- and A-base positions were identified as influential features and the impact of T was more dominant than G and A. Hexamer score and ORF length were also influential features and they hold relatively lower values in lncRNA. Query sequence length (qlen) and conserved region alignment ration (HMM_align_ratio) were also suggested by the SHAP algorithm as dominating features indicating the importance of incorporating sequence length and alignment length information into the prediction model.

V. CONCLUSION

This article proposes LNCRI, a novel ML-based model to distinguish lncRNA transcripts from mRNA transcripts in human, mouse, and other species. LNCRI outperformed many of the existing state-of-the-art tools for lncRNA transcript identification in the considered species. Considering the low expression level and evolving annotations of lncRNA, its identification is a challenging task. To overcome the challenges, we have used the most extensive collection of lncRNA and mRNA transcripts to build a highly accurate ML-based model for the task of lncRNA identification. We believe LNCRI will provide more insights into lncRNAome by enabling the discovery of lncRNA transcripts with increased accuracy.

SUPPLEMENTARY FILES

1) Supplementary File 01: Obs/Exp ration of k-mers
2) Supplementary File 02: Details for parameter optimization for both human and mouse models.
3) Supplementary File 03: Graphs represent probability density function for Fickett, Average HEX, ORF, UTR, and HMM Features and SHAP values in Mouse.

REFERENCES

[1] E. Lander, L. Linton, B. Birren, C. Nusbaum, M. Zody, J. Baldwin, and K. Devon, "Initial sequencing and analysis of the human genome," *Nature*, vol. 409, no. 6822, pp. 860–921, 2001.

[2] H. Y. Chang, "Long noncoding RNAs and human disease," *Trends Cell Biol.*, vol. 21, no. 6, pp. 354–361, 2011.

[3] C. I. Brannan, E. C. Dees, R. S. Ingram, and S. M. Tilghman, "The product of the H19 gene may function as an RNA," *Mol. Cellular Biol.*, vol. 10, no. 1, pp. 28–36, 1990, doi: 10.1128/mcb.10.1.28-36.1990.

[4] M. F. Lyon, "Gene action in the X-chromosome of the mouse (Mus musculus L.)," *Nature*, vol. 190, no. 4773, pp. 372–373, Apr. 1961, doi: 10.1038/190372a0.

[5] C. Liu, "NONCODE: An integrated knowledge database of non-coding RNAs," *Nucleic Acids Res.*, vol. 33, pp. D112–D115, Dec. 2004.

[6] L. E. Orgel and F. H. C. Crick, "Selfish DNA: The ultimate parasite," *Nature*, vol. 284, no. 5757, pp. 604–607, Apr. 1980.

[7] X. Shi, M. Sun, H. Liu, Y. Yao, and Y. Song, "Long non-coding RNAs: A new frontier in the study of human diseases," *Cancer Lett.*, vol. 339, no. 2, pp. 159–166, Oct. 2013.

[8] Y. Huang, R. Regazzi, and W. Cho, *Emerging Roles of Long Noncoding RNAs in Neurological Diseases and Metabolic Disorders*. Lausanne, Switzerland: Frontiers Media, Jul. 2015.

[9] N. Lin and T. M. Rana, *Dysregulation of Long Non-coding RNAs in Human Disease*, A. M. Khalil and J. Coller, Eds. New York, NY, USA: Springer, 2013, pp. 115–136, doi: 10.1007/978-1-4614-8621-3_5.

[10] P. Carninci *et al.*, "The transcriptional landscape of the mammalian genome," *Science*, vol. 309, no. 5740, pp. 1559–1563, Sep. 2005.

[11] Y. Okazaki *et al.*, "Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs," *Nature*, vol. 420, no. 6915, pp. 563–573, Dec. 2002. [Online]. Available: https://www.nature.com/articles/nature01266

[12] C.-C. Hon *et al.*, "An atlas of human long non-coding RNAs with accurate 5' ends," *Nature*, vol. 543, no. 7644, pp. 199–204, Mar. 2017.

[13] T. Derrien, R. Johnson, G. Bussotti, A. Tanzer, S. Djebali, H. Tilgner, G. Guernec, D. Martin, A. Merkel, D. Knowles, J. Lagarde, L. Veeravali, X. Ruan, Y. Ruan, T. Lassmann, P. Carninci, J. Brown, L. Lipovich, J. González, and R. Guigó, "The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression," *Genome Res.*, vol. 22, pp. 1775–1789, Sep. 2012.

[14] M. K. Iyer, Y. S. Niknafs, R. Malik, U. Singhal, A. Sahu, Y. Hosono, T. R. Barrette, J. R. Prensner, J. R. Evans, S. Zhao, A. Poliakov, X. Cao, S. M. Dhanasekaran, Y.-M. Wu, D. R. Robinson, D. G. Beer, F. Y. Feng, H. K. Iyer, and A. M. Chinnaiyan, "The landscape of long noncoding RNAs in the human transcriptome," *Nature Genet.*, vol. 47, no. 3, pp. 199–208, Mar. 2015. [Online]. Available: https://www.nature.com/articles/ng.3192

[15] Y. Zhao, H. Li, S. Fang, Y. Kang, W. Wu, Y. Hao, Z. Li, D. Bu, N. Sun, M. Q. Zhang, and R. Chen, "NONCODE 2016: An informative and valuable data source of long non-coding RNAs," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D203–D208, Jan. 2016.

[16] L. Statello, C.-J. Guo, L.-L. Chen, and M. Huarte, "Gene regulation by long non-coding RNAs and its biological functions," *Nature Rev. Mol. Cell Biol.*, vol. 22, no. 2, pp. 96–118, Feb. 2021. [Online]. Available: https://www.nature.com/articles/s41580-020-00315-9

[17] C. P. Ponting, P. L. Oliver, and W. Reik, "Evolution and functions of long noncoding RNAs," *Cell*, vol. 136, no. 4, pp. 629–641, 2009.

[18] M. Guttman and J. L. Rinn, "Modular regulatory principles of large non-coding RNAs," *Nature*, vol. 482, no. 7385, pp. 339–346, 2012. [Online]. Available: https://www.nature.com/articles/nature10887

[19] I. Ulitsky and D. P. Bartel, "LincRNAs: Genomics, evolution, and mechanisms," *Cell*, vol. 154, no. 1, pp. 26–46, Jul. 2013.

[20] J. Liu, J. Gough, and B. Rost, "Distinguishing protein-coding from non-coding RNAs through support vector machines," *PLoS Genet.*, vol. 2, no. 4, p. e29, Apr. 2006. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1449884/

[21] K. C. Pang, S. Stephen, P. G. Engström, K. Tajul-Arifin, W. Chen, C. Wahlestedt, B. Lenhard, Y. Hayashizaki, and J. S. Mattick, "RNAdb—A comprehensive mammalian noncoding RNA database," *Nucleic Acids Res.*, vol. 33, pp. D125–D130, Dec. 2004.

[22] B. Boeckmann, A. Bairoch, R. Apweiler, M.-C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, and M. Schneider, "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003," *Nucleic Acids Res.*, vol. 31, no. 1, pp. 365–370, Jan. 2003.

[23] A. Li, J. Zhang, and Z. Zhou, "PLEK: A tool for predicting long non-coding RNAs and messenger RNAs based on an improved *k*-mer scheme," *BMC Bioinf.*, vol. 15, no. 1, p. 311, Sep. 2014, doi: 10.1186/1471-2105-15-311.

[24] L. Sun, H. Luo, D. Bu, G. Zhao, K. Yu, C. Zhang, Y. Liu, R. Chen, and Y. Zhao, "Utilizing sequence intrinsic composition to classify protein-coding and long non-coding transcripts," *Nucleic Acids Res.*, vol. 41, no. 17, p. e166, Sep. 2013.

[25] J.-C. Guo, S.-S. Fang, Y. Wu, J.-H. Zhang, Y. Chen, J. Liu, B. Wu, J.-R. Wu, E.-M. Li, L.-Y. Xu, L. Sun, and Y. Zhao, "CNIT: A fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W516–W522, Jul. 2019.

[26] S. Han, Y. Liang, Q. Ma, Y. Xu, Y. Zhang, W. Du, C. Wang, and Y. Li, "LncFinder: An integrated platform for long non-coding RNA identification utilizing sequence intrinsic composition, structural information and physicochemical property," *Briefings Bioinf.*, vol. 20, no. 6, pp. 2009–2027, Nov. 2019.

[27] J. Baek, B. Lee, S. Kwon, and S. Yoon, "LncRNAnet: Long non-coding RNA identification using deep learning," *Bioinformatics*, vol. 34, no. 22, pp. 3889–3897, 2018.

[28] X.-Q. Liu, B.-X. Li, G.-R. Zeng, Q.-Y. Liu, and D.-M. Ai, "Prediction of long non-coding RNAs based on deep learning," *Genes*, vol. 10, no. 4, p. 273, Apr. 2019. [Online]. Available: https://www.mdpi.com/2073-4425/10/4/273

[29] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, Oct. 2014, pp. 1532–1543. [Online]. Available: https://aclanthology.org/D14-1162

[30] C. Yang, L. Yang, M. Zhou, H. Xie, C. Zhang, M. D. Wang, and H. Zhu, "LncADeep: Anab initiolncRNA identification and functional annotation tool based on deep learning," *Bioinformatics*, vol. 34, no. 22, pp. 3825–3834, Nov. 2018.

[31] R. Tripathi, S. Patel, V. Kumari, P. Chakraborty, and P. K. Varadwaj, "DeepLNC, a long non-coding RNA prediction tool using deep neural network," *Netw. Model. Anal. Health Informat. Bioinf.*, vol. 5, no. 1, p. 21, Jun. 2016, doi: 10.1007/s13721-016-0129-2.

[32] N. Amin, A. McGrath, and Y.-P.-P. Chen, "Evaluation of deep learning in non-coding RNA classification," *Nature Mach. Intell.*, vol. 1, no. 5, pp. 246–256, May 2019. [Online]. Available: https://www.nature.com/articles/s42256-019-0051-2

[33] T. Alam, H. R. H. Al-Absi, and S. Schmeier, "Deep learning in LncRNAome: Contribution, challenges, and perspectives," *Non-Coding RNA*, vol. 6, no. 4, p. 47, Nov. 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7711891/

[34] E. A. Feingold and L. Pachter, "The ENCODE (encyclopedia of dna elements) project," *Science*, vol. 306, no. 5696, pp. 636–640, 2004.

[35] A. Frankish, M. Diekhans, A.-M. Ferreira, R. Johnson, I. Jungreis, J. Loveland, J. M. Mudge, C. Sisu, J. Wright, J. Armstrong, and I. Barnes, "Gencode reference annotation for the human and mouse genomes," *Nucleic Acids Res.*, vol. 47, no. D1, pp. D766–D773, 2019.

[36] K. D. Pruitt, T. Tatusova, G. R. Brown, and D. R. Maglott, "NCBI reference sequences (RefSeq): Current status, new features and genome annotation policy," *Nucleic Acids Res.*, vol. 40, no. D1, pp. D130–D135, Jan. 2012. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3245008/

[37] A. Frankish, B. Uszczynska, G. R. Ritchie, J. M. Gonzalez, D. Pervouchine, R. Petryszak, J. M. Mudge, N. Fonseca, A. Brazma, R. Guigo, and J. Harrow, "Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction," *BMC Genomics*, vol. 16, no. S8, pp. 1–11, Dec. 2015.

[38] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, "CD-HIT Suite: A web server for clustering and comparing biological sequences," *Bioinformatics*, vol. 26, no. 5, pp. 680–682, 2010.

[39] Z.-D. Su, Y. Huang, Z.-Y. Zhang, Y.-W. Zhao, D. Wang, W. Chen, K.-C. Chou, and H. Lin, "ILoc-lncRNA: Predict the subcellular location of lncRNAs by incorporating octamer composition into general PseKNC," *Bioinformatics*, vol. 34, pp. 4196–4204, Jun. 2018.

[40] T. Alam, Y. A. Medvedeva, H. Jia, J. B. Brown, L. Lipovich, and V. B. Bajic, "Promoter analysis reveals globally differential regulation of human long non-coding RNA and protein-coding genes," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e109443. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.01094

[41] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "RepDNA: A Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, Apr. 2015.

[42] Z. Chen, P. Zhao, F. Li, T. T. Marquez-Lago, A. Leier, J. Revote, Y. Zhu, D. R. Powell, T. Akutsu, G. I. Webb, K.-C. Chou, A. I. Smith, R. J. Daly, J. Li, and J. Song, "ILearn: An integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data," *Briefings Bioinf.*, vol. 21, no. 3, pp. 1047–1057, May 2020.

[43] L. Wang, H. J. Park, S. Dasari, S. Wang, J.-P. Kocher, and W. Li, "CPAT: Coding-potential assessment tool using an alignment-free logistic regression model," *Nucleic Acids Res.*, vol. 41, no. 6, p. e74, Apr. 2013.

[44] J. W. Fickett, "Recognition of protein coding regions in DNA sequences," *Nucleic Acids Res.*, vol. 10, no. 17, pp. 5303–5318, 1982.

[45] R. D. Finn, P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry, A. L. Mitchell, S. C. Potter, M. Punta, M. Qureshi, A. Sangrador-Vegas, G. A. Salazar, J. Tate, and A. Bateman, "The Pfam protein families database: Towards a more sustainable future," *Nucleic Acids Res.*, vol. 44, no. D1, pp. D279–D285, Jan. 2016.

[46] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: Interactive sequence similarity searching," *Nucleic Acids Res.*, vol. 39, pp. W29–W37, Jul. 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3125773/

[47] T. Alam, M. Alazmi, R. Naser, F. Huser, A. A. Momin, V. Astro, S. Hong, K. W. Walkiewicz, C. G. Canlas, R. Huser, A. J. Ali, J. Merzaban, A. Adamo, M. Jaremko, M. Jaremko, V. B. Bajic, X. Gao, and S. T. Arold, "Proteome-level assessment of origin, prevalence and function of leucine-aspartic acid (LD) motifs," *Bioinformatics*, vol. 36, pp. 1121–1128, Oct. 2019.

[48] L. Roberts, N. Steele, C. Reeves, and G. King, "Training neural networks to identify coding regions in genomic DNA," in *Proc. 4th Int. Conf. Artif. Neural Netw.*, 1995, pp. 399–403.

[49] H. Liu, J. Yin, M. Xiao, C. Gao, A. S. Mason, Z. Zhao, Y. Liu, J. Li, and D. Fu, "Characterization and evolution of 5' and 3' untranslated regions in eukaryotes," *Gene*, vol. 507, no. 2, pp. 106–111, Oct. 2012.

[50] F. Niazi and S. Valadkhan, "Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs," *RNA*, vol. 18, no. 4, pp. 825–843, Apr. 2012.

[51] A. E. Kornienko, P. M. Guenzl, D. P. Barlow, and F. M. Pauler, "Gene regulation by the act of long non-coding RNA transcription," *BMC Biol.*, vol. 11, no. 1, p. 59, May 2013, doi: 10.1186/1741-7007-11-59.

[52] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017, *arXiv:1705.07874*.

SALEH MUSLEH is currently pursuing the Ph.D. degree with the College of Science and Engineering, Hamad Bin Khalifa University. He has a considerable experience in data science research for discovering and communicating the value data science brings to organizational decision making and optimizing opportunities. His research interests includes developing and applying machine learning algorithms to support bioinformatics research community to answer complex questions from big data. His current research focusing on long non-coding RNA (LncRNA) identification using novel machine learning-based pipeline purely based on sequence information.

MOHAMMAD TARIQUL ISLAM is currently an Assistant Professor with the Department of Computer Science, Southern Connecticut State University. He has published at notable peer-reviewed conferences and journals, such as *Computer Vision and Pattern Recognition*, International Conference on Image Processing, International Conference on Bioinformatics and Biomedicine, and *International Journal on Image and Video Processing*. His primary research interests include computer vision, deep learning, and applied bioinformatics. He was a recipient of several grants on the application of deep learning in computer vision and has supervised multiple graduate students in their pursuit of the master's degree.

TANVIR ALAM is currently an Assistant Professor with the College of Science and Engineering, Hamad Bin Khalifa University. Among his notable research works are on the transcription regulation of non-coding RNAs and their roles in different diseases. His research work also centered around the application of artificial intelligence (AI) on the diagnosis and prognosis of communicable and non-communicable diseases. He is a member of FANTOM Consortium. He also served as a reviewer for a number of international conferences and reputed journals.