Received: 7 May 2023 | Accepted: 19 July 2023

DOI: 10.1002/iub.2778

RESEARCH ARTICLE

IUBMB LIFE WILEY

LncRTPred: Predicting RNA–RNA mode of interaction mediated by lncRNA

Gourab Das | **Troyee Das** | **Sibun Parida** | **Zhumur Ghosh**

Division of Bioinformatics, Bose Institute, Kolkata, India

Correspondence
Zhumur Ghosh, Division of Bioinformatics, Bose Institute, P-1/12, C.I.T. Scheme-VII M, Kolkata 700054, India.
Email: zhumur@jcbose.ac.in

Present address
Sibun Parida, ITER, Shiksha O Anusandhan University, Bhubaneswar, Odisha 751030, India.

Funding information
Council of Scientific and Industrial Research, India, Grant/Award Number: 09/015(05170/2017-EMR-I; Department of Biotechnology, Government of India, Grant/Award Number: BT/PR40174/BTIS/137/45/2022; Indian Council of Medical Research, Government of India, Grant/Award Number: RBMH/FW/2020/10

Abstract

Long non-coding RNAs (lncRNAs) play a significant role in various biological processes. Hence, it is utmost important to elucidate their functions in order to understand the molecular mechanism of a complex biological system. This versatile RNA molecule has diverse modes of interaction, one of which constitutes lncRNA–mRNA interaction. Hence, identifying its target mRNA is essential to understand the function of an lncRNA explicitly. Existing lncRNA target prediction tools mainly adopt thermodynamics approach. Large execution time and inability to perform real-time prediction limit their usage. Further, lack of negative training dataset has been a hindrance in the path of developing machine learning (ML) based lncRNA target prediction tools. In this work, we have developed a ML-based lncRNA–mRNA target prediction model- 'LncRTPred'. Here we have addressed the existing problems by generating reliable negative dataset and creating robust ML models. We have identified the non-interacting lncRNA and mRNAs from the unlabelled dataset using BLAT. It is further filtered to get a reliable set of outliers. LncRTPred provides a cumulative_model_score as the final output against each query. In terms of prediction accuracy, LncRTPred outperforms other popular target prediction protocols like LncTar. Further, we have tested its performance against experimentally validated disease-specific lncRNA–mRNA interactions. Overall, performance of LncRTPred is heavily dependent on the size of the training dataset, which is highly reflected by the difference in its performance for human and mouse species. Its performance for human species shows better as compared to that for mouse when applied on an unknown data due to smaller size of the training dataset in case of mouse compared to that of human. Availability of increased number of lncRNA–mRNA interaction data for mouse will improve the performance of LncRTPred in future. Both webserver and standalone versions of LncRTPred are available. Web server link: http://bicresources.jcbose.ac.in/zhumur/lncrtpred/index.html. Github Link: https://github.com/zglabDIB/LncRTPred.

KEYWORDS

lncRNA–mRNA interaction, machine learning, real-time prediction, reliable negative data, target prediction

Abbreviations: lncRNA, long non-coding RNA; ML, machine learning; WAFNRLTG, weighted average fusion network representation learning for predicting LncRNA target genes.

1 | INTRODUCTION

Long non-coding RNAs (lncRNAs) are transcripts of >200 nucleotides in length.[1] It garnered lot of attention among researchers due to its diverse functionality[2,3] across various biological processes like regulation of gene expression,[4] cell cycle,[5] transcriptional and post-transcriptional processes,[6,7] X-chromosome inactivation[8] etc. They have also been reported to be involved in various diseases like cancer,[2] neurological disorders,[9] autoimmune diseases[10] and so on. Although being regarded as a master regulator for various physiological processes,[11] there remains a lot to explore regarding the working mechanism of these molecules. It has been well investigated that lncRNA accomplishes their function by interacting with various biological molecules like RNAs,[12] DNA[13] and proteins[14] which causes decay in mRNA, splicing, aberrations, alterations in protein stability and many more. Among these, RNA–RNA mode of interaction is very popular which has been documented in several reports as BACE1-antisense lncRNA(BACE1-AS) with its target BACE1 mRNA.[15] Similar to BACE1-AS, PTB antisense lncRNA (PTB-AS) modulates the expression of PTBP1 mRNA[16] by binding to the 3′ untranslated region (UTR) of PTBP1, which is an RNA-binding protein that promotes gliomagenesis.[17] Further, lncRNA named FGFR3 antisense transcript 1 (FGFR3-AS1) binds complementarily to its antisense FGFR3 gene, suggesting a potential regulatory effect of FGFR3-AS1 in the expression of the FGFR3 gene.[18] Hence, it is extremely important to identify this mode of interaction executed by lncRNAs. Several databases have been developed to archive the lncRNA–mRNA interactions.[19–21] The database compiled by Terai et al.[22] contains predicted lncRNA–mRNA interactions which are again limited to only one local base-pairing interaction for each lncRNA–RNA interaction. RAID contains lncRNA–mRNA interaction data obtained from literature base.[23] RISE[24] includes experimentally validated lncRNA–RNA interactions based on high-throughput sequencing methods.[25,26] Both of these suffer from the limitation of storing a very less number of interactions.

Till today, predicting lncRNA–mRNA interactions have been done mainly by adopting thermodynamics-based-free energy minimisation approach. One such algorithm is MechRNA[27] which mainly uses IntaRNA2,[28] based on seed constraints and interaction site accessibility for the lncRNA–mRNA interaction prediction. Antisense search approach (ASSA) uses sequence alignment and thermodynamic approach to calculate lncRNA target.[29] But, large execution time makes it difficult to use most of the thermodynamic energy minimisation-based target prediction tools.[30] LncTar[31] deploys the nearest-neighbour algorithm, considering free energy minimisation technique to detect the interactions. But only 10 lncRNA–mRNA interactions were taken as the experimentally validated dataset to test the prediction accuracy of LncTar. Fukunga et al. try to address these issues by developing the webserver—LncRRIsearch[30] for predicting human and mouse lncRNA targets. It uses RIblast,[32] a seed and extension method; pre-calculates and stores lncRNA–mRNA interaction score in database which is fetched when required for prediction. But, this limits its capability to make real-time predictions where the interactions are not pre-calculated. All these opened up the avenue to adopt machine learning (ML)-based approach to deal with such target prediction problems. Few ML-based tools are there which mainly predict lncRNA–miRNA[33,34] and lncRNA–protein interactions[14] but as far as the ML-based lncRNA–mRNA target prediction is concerned, the existence of such tool is quite rare.[35] Moreover, it is quite well-known that any supervised ML algorithms execute upon a properly labelled dataset. However, the lack of negative training dataset is the main hindrance on the path of developing ML-based lncRNA–mRNA prediction model. Weighted Average Fusion Network Representation Learning method-based model for predicting lncRNA Target Genes (WAFNRLTG) is one such model[35] which has been developed based on the assumption that highly similar lncRNAs tend to have similar interaction and lncRNAs indirectly regulate gene expressions via adjusting expressions of miRNAs. Such assumption may not be valid for all lncRNA–mRNA interactions. Further, random shuffling of unlabelled interactions, which is one of the widely followed techniques to generate negative training data,[33–35] is the basis of generating negative dataset in case of WAFNRLTG. But this procedure embeds noise within ML models as it incorporates erroneous learning.[36] This creates a limitation of this prediction model.

With all these challenges on board, we have come up with 'LncRTPred' which is a ML-based lncRNA–mRNA target prediction model. We have incorporated transcript sequence as features, constituting two aspects: generation of reliable negative training data and creation of ML models. The first part of our work involves the development of the training dataset from the experimentally validated lncRNA–mRNA experiments with a special approach to obtain the reliable negative datasets. Subsequently, the next part deals with the creation of the ML models and training them for precise lncRNA–mRNA target prediction. Finally, we have tested its performance against experimentally validated disease-specific lncRNA–mRNA interactions.

2 | MATERIALS AND METHODS

The development of the LncRTPred has been executed in three steps which are as follows:

- Step 1: Building of training (positive) and test dataset using experimentally validated lncRNA–mRNA target datasets corresponding to human and mouse species.
- Step 2: Generation of reliable negative data from pre-processed features which adopts special approach to generate negative data.
- Step 3: Development of the ML-based model.

The entire workflow (mentioned above) is provided in Figure 1.

2.1 | Building a positive training dataset

Human: Experimental data depicting positive interaction between lncRNA and mRNA for human have been compiled from Supplementary information provided by Zhang et al.[37] The sequences are extracted from the Refseq database.[38] Data compilation has been done by screening those target genes having a single transcript, resulting in 1675 experimentally validated positive interactions spreading across 421 unique mRNAs and 134 lncRNAs with 310 transcripts.

Mouse: NPInter4[39] serves as the source for positive training dataset in case of mouse species. The corresponding sequences have been fetched from Ensembl Biomart.[40] The entire positive training dataset includes 88 lncRNAs and 399 mRNA transcripts executing 500 positive interactions.

The details about the input dataset for both the species are tabulated in Table 1. Interacting lncRNA and mRNA sequences constituting the training data for human and mouse species have been provided in Files S1 and S2, respectively.

2.2 | Generation of reliable negative data from pre-processed features

Reliable negative data have been generated in two steps.

Step 1 is the pre-processing stage, which determines the feature space of interacting RNAs in terms of feature selection, normalisation and principal component

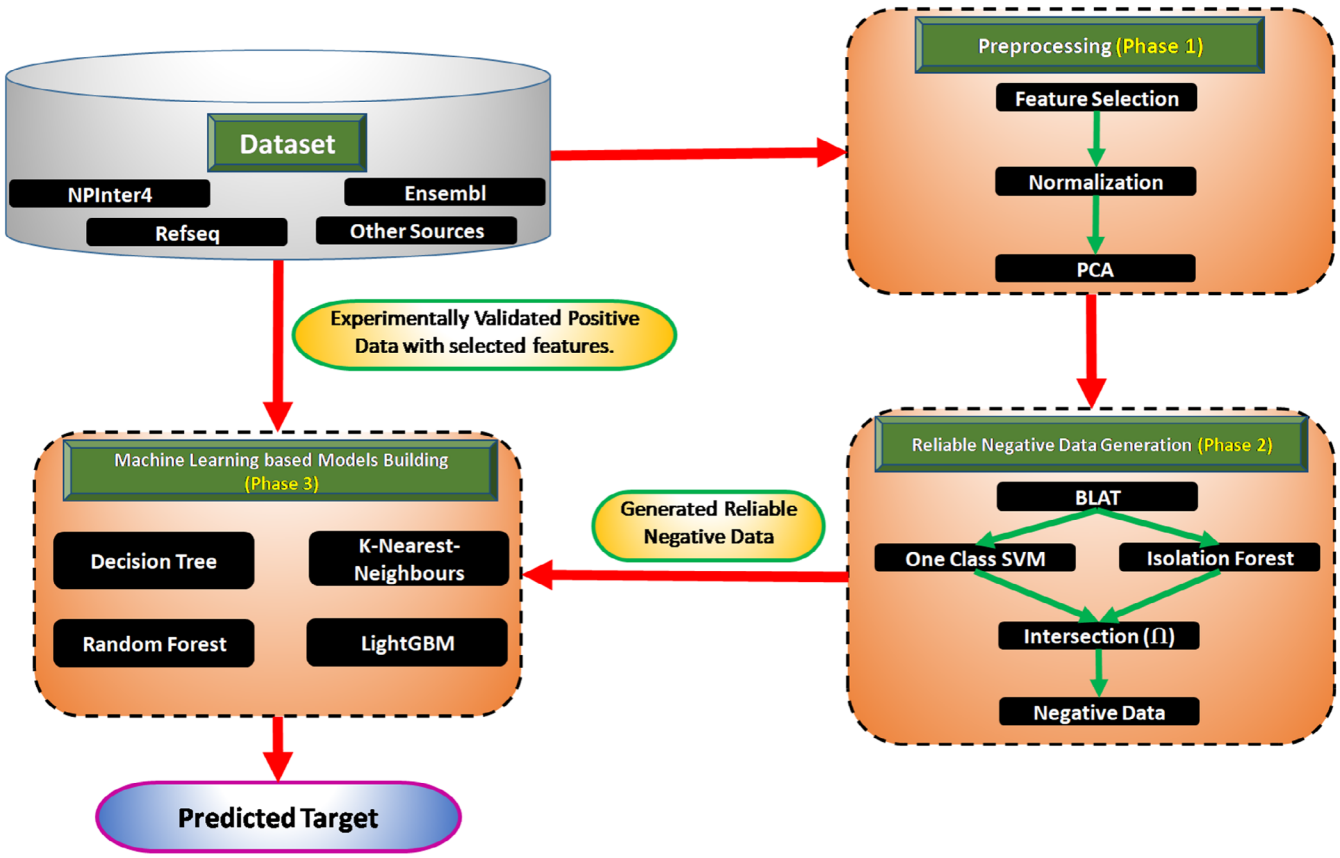

FIGURE 1 Flow Diagram of the entire work. The flow diagram depicts the various steps involved towards development of LncRTPred which includes building the training and test dataset along with creation of reliable negative data followed by development of the ML-based models.

TABLE 1 Input dataset used for developing LncRTPred.

	Human		Mouse	
Description	Sources	Data size (in #)	Sources	Data size (in #)
lncRNA transcripts	PMID: 29401217, NCBI RefSeq	310	NPInter 4, Ensembl biomart	88
mRNA transcripts		421		399
Possible interaction pairs		130,510		35,112
Experimentally validated lncRNA–mRNA pairs		1,675		500
Unlabelled pairs		128,835		34,612

analysis (PCA).[41,42] In step 2, reliable negative data are generated from the unlabelled content by incorporating The blast-like alignment tool (BLAT),[43] one-class support vector machine[44–46] and isolation forest.[47]

2.2.1 | Step 1

Pre-processing

Input datasets consist of raw long RNA sequences corresponding to both lncRNA and mRNA. Before incorporating various ML tools, it is required to convert those strings of data into numeric form by virtue of feature selection, normalisation and principal component analysis (PCA).

Feature selection

Raw sequences are provided as string of nucleotides which are not sufficient to be fed into ML models. Feature selection step generates bag of nucleotides in both simplex and complex forms. The set of features which have been considered are sequence length, GC percentage and counting the occurrences of single, double, triple and quadruple nucleotides which cumulatively produced 684 features for both lncRNA and mRNA are detailed in Table 2.

Normalisation

Unnormalised data produce range inflation for certain features resulting in slow convergence in various ML models. This has been addressed by normalisation technique given in Equation (1) corresponding to input data X with mean and std denoting the arithmetic average and standard deviation, respectively.

$$\text{Normalised}_X = \frac{X - \text{mean}(X)}{\text{std}(X)}. \quad (1)$$

We have incorporated the Scikit-learn tool[48] for normalisation using the above equations.

TABLE 2 Feature description of lncRNA and mRNA extracted from raw sequence data.

Feature descriptions	Number of features
Sequence length	1
GC percentage	1
Unique nucleotides (A, C, G, T) count	4
Two contiguous nucleotides (AA, AC, AG, AT, ..., etc.) count	16
Three contiguous nucleotides (AAA, AAC, AAG, AAT, ..., etc.) count	64
Four contiguous nucleotides (AAAA, AAAC, AAAG, AAAT, ..., etc.) count	256
lncRNA features count	1 + 1 + 4 + 16 + 64 + 256 = 342
mRNA features count	1 + 1 + 4 + 16 + 64 + 256 = 342
Total number of features	342 * 2 = 684

Principal component analysis

Principal component analysis[41,42] creates new set of orthonormal and uncorrelated features known as principal component by linear combinations of actual feature space to reduce data dimensionality after retaining maximum threshold variations. We have executed PCA technique upon original set of 684 features; plotted the explained variability of each principal component and finally employed the Elbow Method[49] (https://en.wikipedia.org/wiki/Elbow_method_(clustering)) to determine the number of principal component having retained maximum variations.

2.2.2 | Step 2

Non-availability of negative interaction data is a big bottleneck in the path of developing a supervised ML-based

algorithm in this case. In such cases, researchers normally perform random shuffling among the set of all molecules for generation of negative data which incorporate flaws in learning procedure and generate false prediction.[33,35,50] We have addressed this issue by generating reliable negative data from unlabelled pair of lncRNA and mRNA sequences by adopting three strategies: The BLAT, one-class support vector machine and isolation forest have been placed into execution.

The BLAT

BLAT[43] provides a faster version of pairwise sequence alignment tool that can be utilised for both RNA/DNA and protein/protein alignment. It provides output for both reverse complementary and also exactly matched sequences. Many unlabelled pairs of lncRNA and mRNA have been obtained. Sequence alignment governs the interactions among the biological molecules[51–53] which shows the roadmap for generating negative data from unlabelled set. We extracted those pair of data where the BLAT tool did not provide any output, which implies that there has been no similarity or complementarity between the lncRNA and mRNA sequences. Such sequence non-matching indicates an enhanced probability of non-interactive nature of the molecules. Overall, BLAT screens the unlabelled dataset to provide a set of reliable negative data. Outputs are subsequently fed into one-class support vector machine and isolation forest for further processing.

One-class support vector machine

Support vector machine (SVM)[44,45] is the maximum margin classifier that can be utilised to solve binary or multi-class classification problems. One-class SVM pertaining to the variation of SVM model[46] belongs to the group of unsupervised learning and behaves robustly in outlier detection. It is trained upon the positive dataset and validated against the screened data obtained from BLAT; depending upon feature set of fitted positive data and can detect outlier.

Isolation forest

Isolation forest[47] is a variation of random forest utilised for anomaly or outlier detection. It is trained with dataset containing one class and segregates the anomaly from the normal data. For each data point, isolation forest is averaged over the path length of individual isolation tree to determine anomalous data. Following the same strategy mentioned in the previous section, isolation forest is trained with the positive dataset and detects outlier from the filtered BLAT data.

Finally, the set of lncRNAs and mRNAs predicted as outliers by both one-class SVM and isolation forest are considered to be the reliable negative dataset (shown in Table 3).

TABLE 3 Data statistics corresponding to reliable negative data generation.

Description	Human	Mouse
# of lncRNA–mRNA interaction data	130,510	35,112
# of experimentally validated positive interactions	1,675	500
# of non-interacting pairs determined by BLAT	54,474	16,978
Predicted outliers by BLAT and one-class SVM	12,601	2083
Predicted outliers by BLAT and isolation forest	27,909	3,626
Commonly predicted outliers (reliable negative data) determined by BLAT, one-class SVM and isolation forest	9,685	1794

2.3 | ML-based model building

The positive and reliable negative dataset received from previous section have been split randomly into training and validation sets as provided in Table 4, in order to execute model fitting approach. The biggest constraint explicitly posed in this situation is the formation of imbalanced classification problem for both human and mouse species as reliable negative data outnumbered the positive ones. Multiple techniques have been incorporated in order to address this issue embedded across hyper-parameters of various ML models. In this paper, four ML models have been utilised for the prediction

TABLE 4 Training and validation data used for developing LncRTPred model.

Description	Human	Mouse
Size of dataset with label 1	1,675	500
Size of dataset with label 0	9,685	1794
Size of entire dataset	11,360	2,294
Training dataset		
Size of training dataset with label 1	1,300	400
Size of training dataset with label 0	9,200	1,690
Size of entire training dataset	10,500	2090
Validation dataset		
Size of validation dataset with label 1	375	100
Size of validation dataset with label 0	485	104
Size of entire validation dataset	860	204

Note: Dataset with label 1 corresponds to positive interactions. Dataset with label 0 corresponds to reliable negative data.

task; decision tree, K-nearest-neighbours (KNN), random forest and LightGBM. Among them we have incorporated the Scikit-learn tool to create first three models, whereas LightGBM model has its standalone implementation.

2.3.1 | Decision tree

Decision tree[54] embraces non-parametric, rule-based technique; proceeds by growing a tree, can be utilised to solve various prediction-based tasks.

2.3.2 | KNN

KNN[55] is also a non-parametric approach where any object is classified based on the instance voting of the nearest neighbours. It is very much important to determine the number of neighbours (K) in order to balance its performance. In our case, we considered number of neighbours to be 3 to classify a particular object.

2.3.3 | Random forest

Random forest[56] is an instance of ensemble learning computed by averaging individual constituent of large number of decision trees, that is, most voted output is the result. Constituent decision trees within random forest are trained upon random subset of features and data both. Such Ensemble approach generally gives better output compared to the individual model.

2.3.4 | LightGBM

LightGBM[57] is faster version of gradient boosting decision tree (GBDT). Rather than scanning through the whole dataset LightGBM utilises gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB) in leaf wise manner.

Further, LncRTPred cumulative_model_score is provided as the final output based on the prediction result generated by decision tree, KNN, random forest and LightGBM where the final output is considered to be positive if it is generated by at least three among the four models.

Various metrics are taken into consideration to determine the model performance corresponding to validation set viz., confusion matrix, accuracy, sensitivity, specificity, ROC-AUC, F1 score and Matthews correlation coefficient. Being a classification algorithm, target variable is discrete in terms of 1 as positive and 0 as negative interaction. Fundamental aspects associated with discrete-dependent variables are:

True positive
Both true or actual and predicted label are positive or 1.

True negative
Both true or actual and predicted label are negative or 0.

False positive
Here true or actual label is negative or 0 and predicted label is positive or 1.

False negative
Here true or actual label is positive or 1 and predicted label is negative or 0.

Confusion Matrix projects TP, TN, FP and FN across 2-by-2 matrix to analyse the strength and loopholes of the models at various target classes where primary diagonal of the matrix denotes the accurate prediction in terms of TN and TP and secondary diagonal denotes the wrong prediction in terms of FP and FN. Accuracy is the most basic performance metrics defined by Equation (2)

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (2)$$

But model performance cannot be justified by accuracy alone if there exists an imbalance distribution between two classes as it can achieve higher accuracy by correctly predicting the major classes resulting in biasness towards predicted results. Equations (3) and (4) define Sensitivity which is also known as TP rate or recall

TABLE 5 Unknown test dataset used for performance comparison between LncRTPred and LncTar.

	Human		Mouse	
Description	**Sources**	**Data size (in #)**	**Sources**	**Data size (in #)**
Unknown test interaction data	NPInter 4, Ensembl biomart	44	NPInter 4, Ensembl biomart	90
lncRNAs for unknown test dataset		12		25
mRNAs for unknown test dataset		37		90

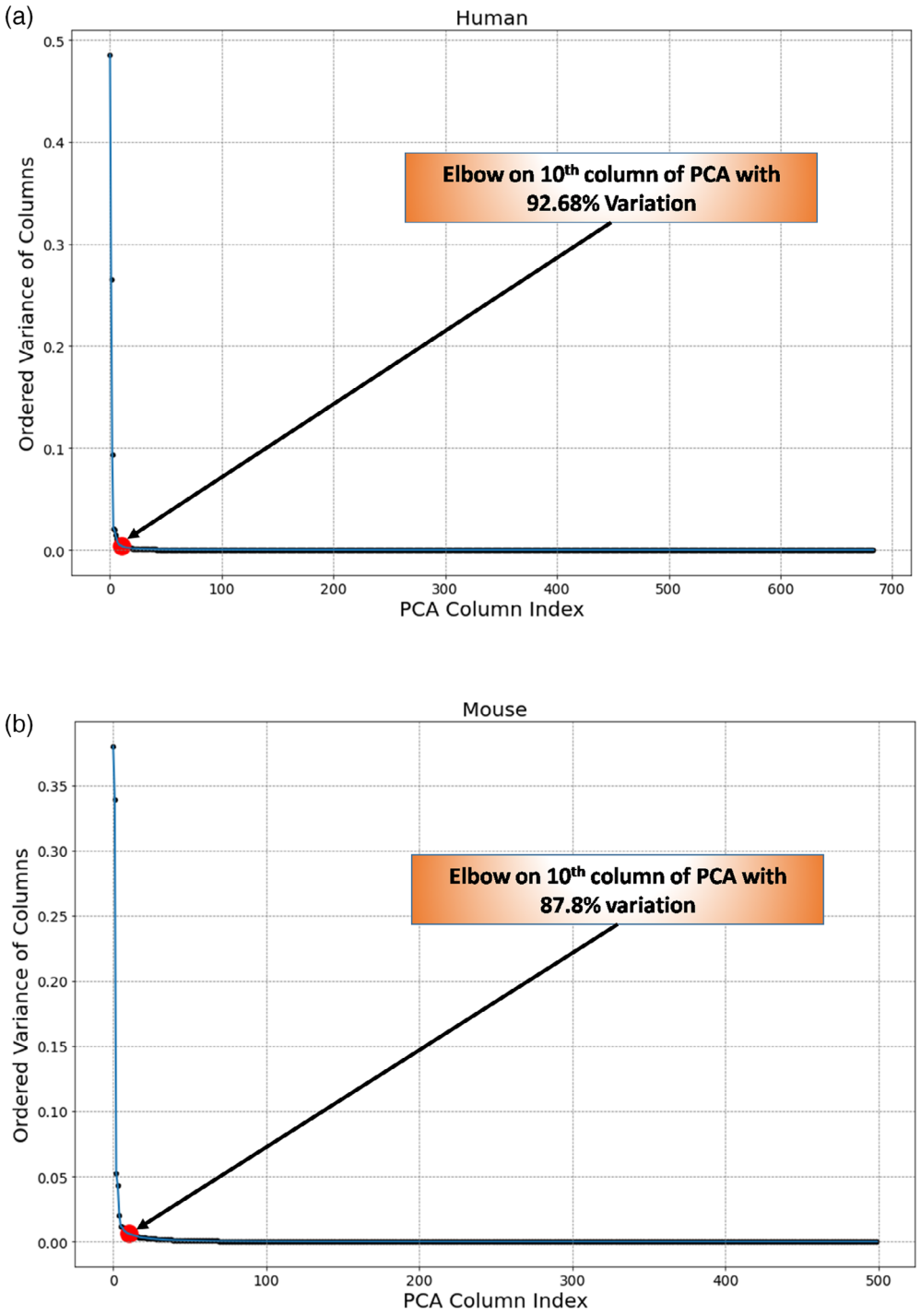

FIGURE 2 Elbow detection curve for (a) human species and (b) mouse species. It depicts explained variability among the features, governed by Principal Components which leads towards selecting optimum set of features.

and specificity or TN rate respectively. Basically, these two metrics segregate the performance of both positive and negative classes individually.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{3}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}. \tag{4}$$

Equation (5) defines FP rate for wrongly predicted data

$$\text{False positive rate} = \frac{\text{FP}}{\text{TN} + \text{FP}}. \tag{5}$$

We have also incorporated area under receiver operating characteristic Curve (ROC-AUC) performance metric corresponding which plots the TP rate (sensitivity) at

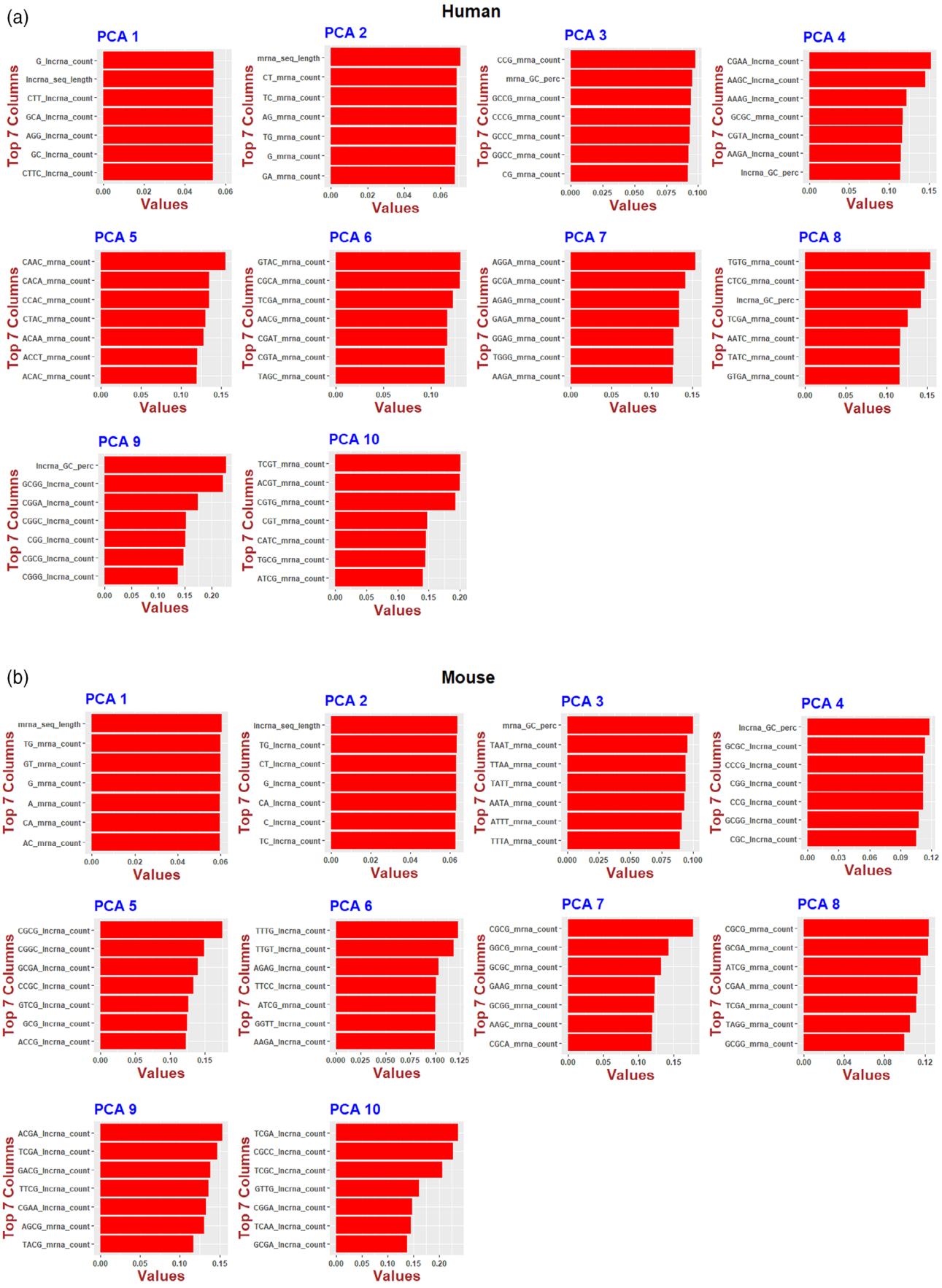

FIGURE 3 Feature importance of (a) human and (b) mouse across different principal components. Bar chart depicts the feature importance of top seven features across selected principal components for human and (B) mouse.

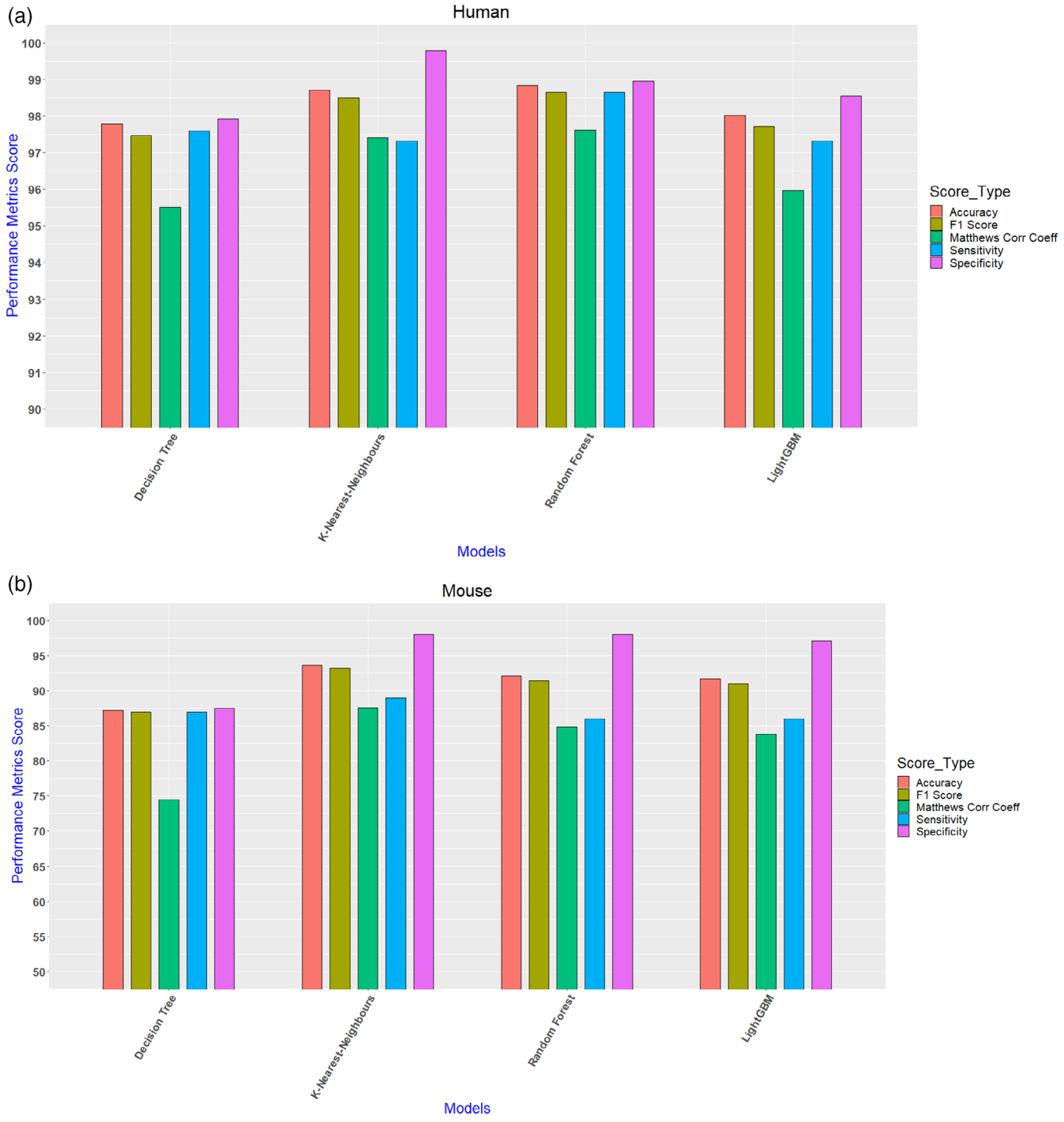

FIGURE 4 Performance comparison of the various ML Models used to develop LncRTPred for (a) human and (b) mouse. Bar chart showing the performance comparison among various LncRTPred models—decision tree, K-nearest neighbours, random forest and LightGBM on the validation Data in terms of accuracy, F1 score, Matthews correlation coefficient, sensitivity and specificity metrics corresponding to (a) human species and (b) mouse species.

Y-axis and FP rate at *X*-axis at different classification threshold where area under curve (AUC) depicts the performance of ML model by measuring the entire area between (0, 0) and (1, 1), so more the area covered under the curve signifies better performance. F1 Score is a versatile performance metric which corrects the drawbacks associated with accuracy. It is defined as harmonic mean of precision and recall, where Equations (6) and (7) defined the precision and F1 score respectively.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{6}$$

$$\text{F1 Score} = 2 * \frac{1}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}. \quad (7)$$

We have also validated our models using Mathews correlation coefficient (MCC) which finds the correlation coefficient between two binary variables, true class and the predicted class defined in Equation (8).

$$\text{MCC} = \frac{(\text{TP} * \text{TN}) - (\text{FP} * \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}. \quad (8)$$

The correlation coefficient in MCC ranges from +1 to −1, where +1 denotes perfect prediction, 0 denotes random prediction, whereas −1 denotes completely incorrect or wrong prediction. In this paper, we have projected the MCC score by multiplying with 100, such that they can be compared with other performance metrics.

We have compared the performance of LncRTPred with LncTar[31] which is the most popular lncRNA–mRNA interaction prediction tools based on free energy minimisation approach. For LncTar, the normalised free energy (ndG) cut-off of −0.1 (as per the recommendation of the LncTar authors) has been used to generate the energetically stable lncRNA–mRNA pairs. The prediction accuracy has been calculated by the percentage of TP predicted by LncRTPred and LncTar individually. It has been computed upon an unknown test dataset from NPInter4[39] and Ensembl Biomart,[40] provided in Table 5. It consists of 44 validated positive interactions involving 12 lncRNAs and 37 mRNAs for human and 25 lncRNAs and 90 mRNAs involved in 90 validated positive interactions corresponding to mouse species. Interacting lncRNA and mRNA sequences constituting the test data for human and mouse species have been provided in Files S3 and S4, respectively.

Non-availability of codebase containing pre-trained saved model (github link [https://github.com/HGDYZW/WAFNRLTG]) limited us to include WAFNRLTG tool[35] in the comparison.

Required software

Initially, Python packages have been incorporated to execute various analytics purpose in forms of preprocessing, negative data generation and model creation. Anaconda software belonging to python 3.8 containing bulk of the library incorporated for the analysis including numpy (min version: 1.19.1; max version: 1.20.3.), pandas (min version: 1.1.3; max version: 1.5.3) and Scikit-learn (min version: 0.23.1; max version: 0.24.2). Besides, LightGBM (version 2.3.1) library has been separately downloaded for building the corresponding model. Data visualisation has been done through Matplotlib from python and ggplot2 in R.

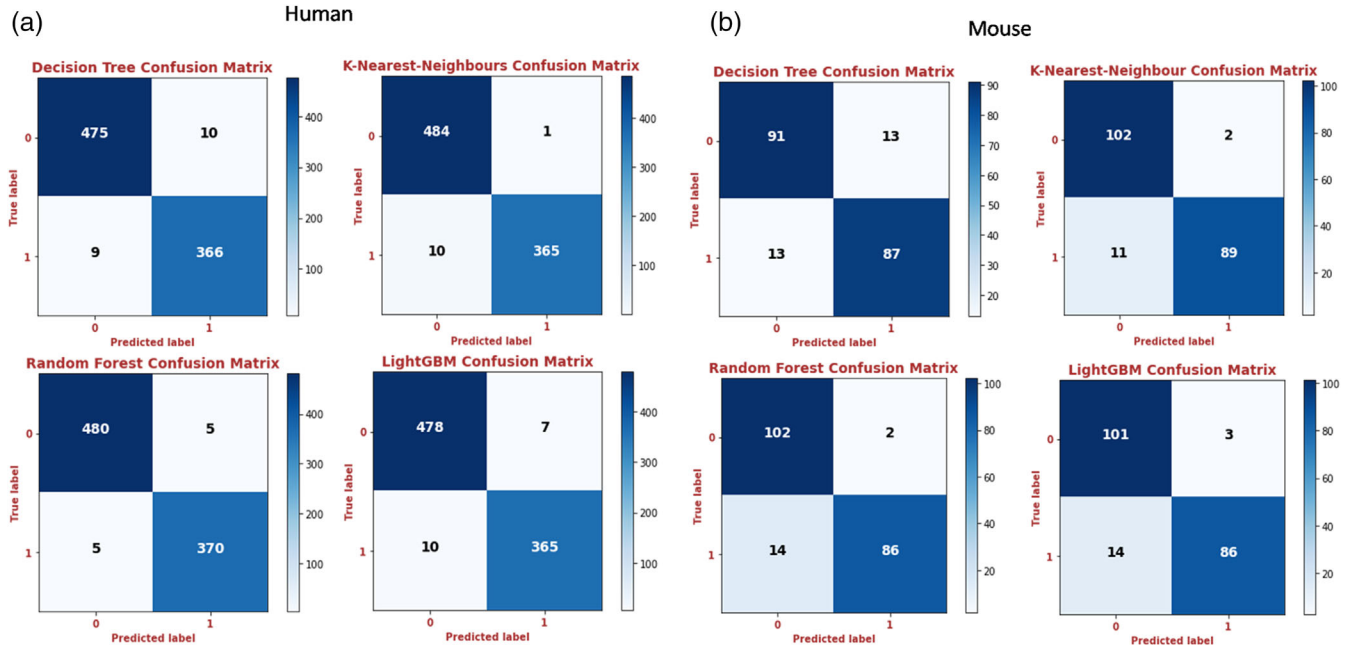

FIGURE 5 Confusion matrix showing performance statistics of LncRTPred corresponding to validation data for (a) human and (b) mouse. This matrix shows the TP and TN predictions representing the number of accurate predictions and FP and FN predictions representing the number of wrong predictions corresponding to validation data for human and mouse by the ML models (viz. decision tree, K-nearest neighbours, random forest and LightGBM) used to develop LncRTPred.

3 | RESULTS AND DISCUSSION

This part contains output from three sections: Feature selection, negative data generation and model building. The entire workflow is provided in Figure 1.

3.1 | Feature selection

As discussed above, we have generated 684 features for both human and mouse cumulatively from raw sequence data of lncRNAs and mRNAs. These features have been normalised and fed to the PCA tools in order to extract highly variable features by Elbow method. We achieved the elbow corresponding to the first 10 principal components retaining 92.68% and 87.8% variations for the features corresponding to human (shown in Figure 2a) and mouse species (shown in Figure 2b) respectively. Figure 3 shows the bar chart of all 10 principal components constituting the top seven features among 684 to have maximum impact on individual principal component, which justifies that both lncRNA and mRNA features play important roles in determining the principal components for both species.

3.2 | Reliable negative data generation

BLAT algorithm has been incorporated to identify non-interacting lncRNAs and mRNAs from the available unlabelled dataset. The resultant BLAT outputs are fed into both one-class SVM and isolation forest from which common set of outliers are considered to be the reliable negative dataset. The detailed statistics corresponding to the reliable negative data generation phase is tabulated in Table 3.

3.3 | Model building

The entire dataset has been split randomly into training and validation sets as provided in Table 4, in order to execute model fitting approach. Figure S1 shows the boxplots which help to identify the important features towards classifying the interacting and non-interacting set of lncRNAs and mRNAs. PCA_1, PCA_2, PCA_3, PCA_9 and PCA_1, PCA_2, PCA_3 and PCA_4 are able to capture the variations existing between the two classes for human and mouse species respectively.

As specified in Materials and Methods section, we have implemented four ML models to perform the prediction task: Decision tree, KNN, random forest and LightGBM. Due to the imbalance nature of the target class, ML model bias towards majorly existing class. This noise has been removed by tuning the hyperparameters associated with various models. Figure 4a,b depicts the comparative performance details of the various models corresponding to the validation data for human and mouse, respectively, which demonstrates that random forest and KNN outperform other models. Figure 5a,b shows the prediction statistics upon validation data in terms of confusion matrix for both human and mouse species, respectively. Figure 6a,b denotes the ROC-AUC curve to measure the ability of the classifier for both human and mouse species respectively. Figure 7a,b shows the horizontal bar plot depicting the feature importance generated by decision tree, random forest and LightGBM models, which corroborates closely with the boxplot interpretation corresponding to Figure S1a,b for both human and mouse species respectively.

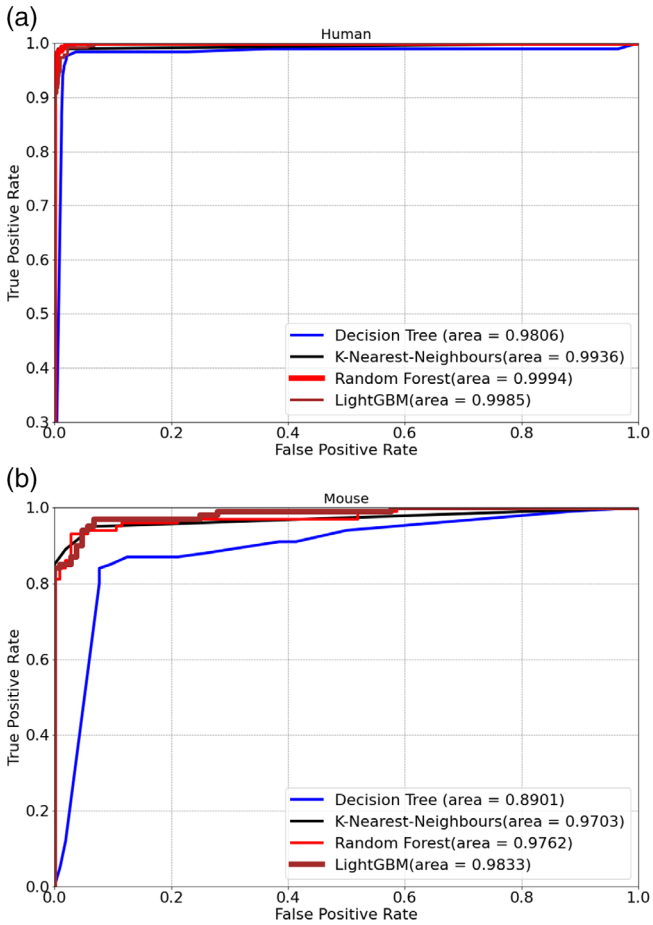

FIGURE 6 ROC-AUC Plot for (a) human and (b) mouse. ROC-AUC curve corresponding to the validation data showing the performance of the classification models used for developing LncRTPred, viz., decision tree, K-nearest neighbours, random forest and LightGBM by measuring their performance in terms of the entire two dimensional area under ROC curve from (0, 0) to (1, 1).

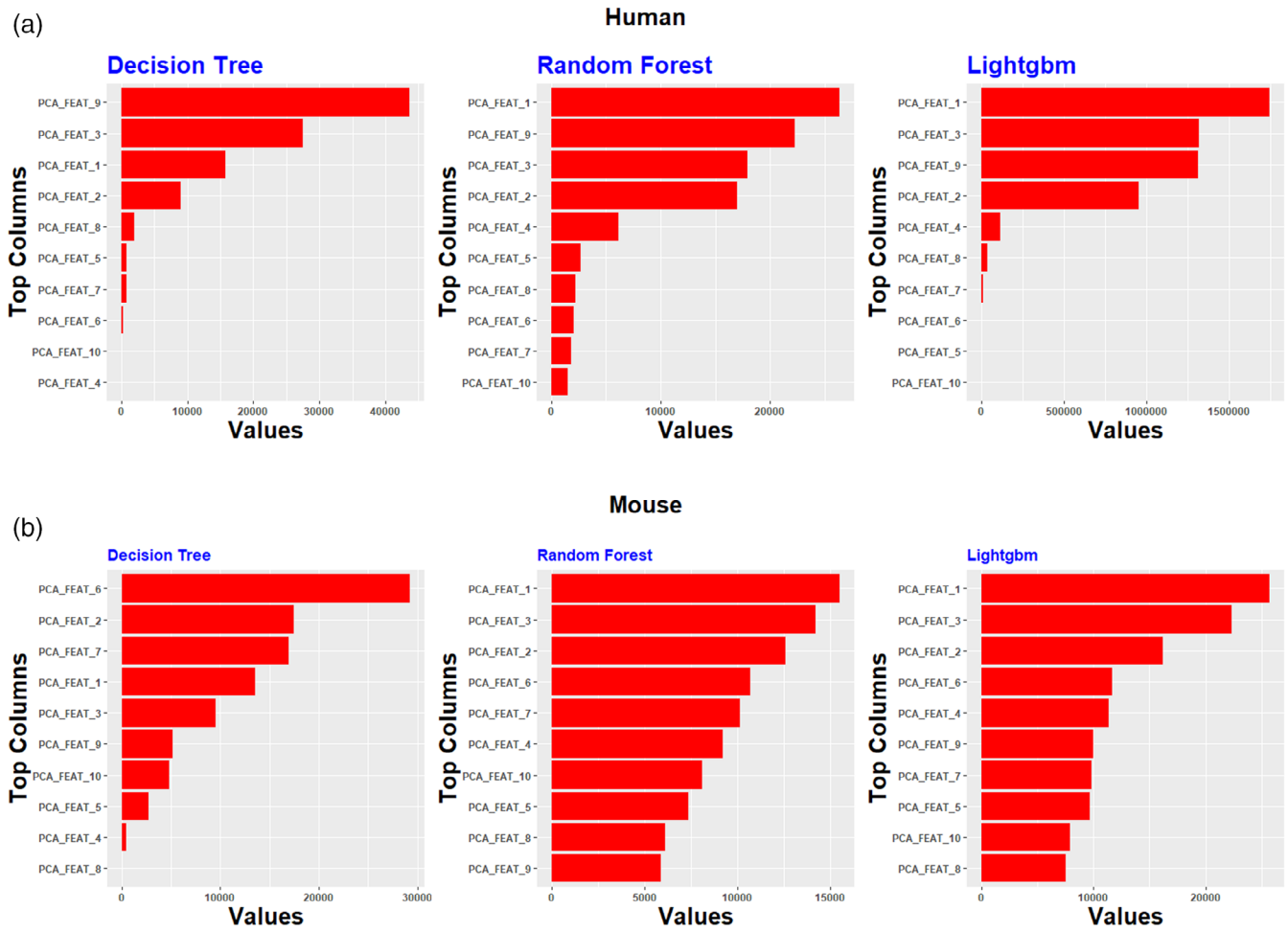

FIGURE 7 Feature importance plot generated by the classification models used by LncRTPred for (a) human and (b) mouse. Bar plots extracting the important features responsible for lncRNA–mRNA target identification task by the different models used for developing LncRTPred.

3.4 | Performance comparison with other prediction models

Efficacy of models corresponding to unknown positive test dataset (provided in Table 5) is measured by prediction accuracy as it contains only experimentally validated positive data. We have compared the prediction accuracy of LncRTPred with that of LncTar. Figure 8 clearly shows that LncRTPred outperforms LncTar.

3.5 | LncRTPred performance corresponding to disease-specific lncRNA–mRNA interactions

We have incorporated LncRTPred tool in the context of disease-specific lncRNA–mRNA target prediction. Here, we have considered four disease-specific RNA–RNA interactions, breast cancer, acute myeloid leukaemia (AML), cervical cancer and non-small cell lung cancer (NSCLC). Pan et al.[58] demonstrated that the interacting pair of NNT-AS1 and ZFP36 is upregulated in breast cancer tissue. Silencing NNT-AS1 lncRNA could induce cell apoptosis and prevent the progression of tumour by suppressing ZFP36. A positive expression correlation of SOX4 and HOXA-AS2 has been observed in AML patients. Qu et al.[59] reported SOX4 to be a downstream target of HOXA-AS2 as observed by the decrease in mRNA level of SOX4 upon HOXA-AS2 silencing in AML patients. HOXA-AS2 has been proposed to function as an oncogene by regulating the SOX4/PI3K/AKT pathway in AML. Wang et al.[60] described the impact of GAS5 gene stability upon Cervical Cancer. It has been found that GAS5-AS1 lncRNA is downregulated in cancer tissue. Basically, GAS5-AS1 interacted with GAS5 tumour suppressor gene to suppress tumour growth and metastasis in cervical cancer. Considering

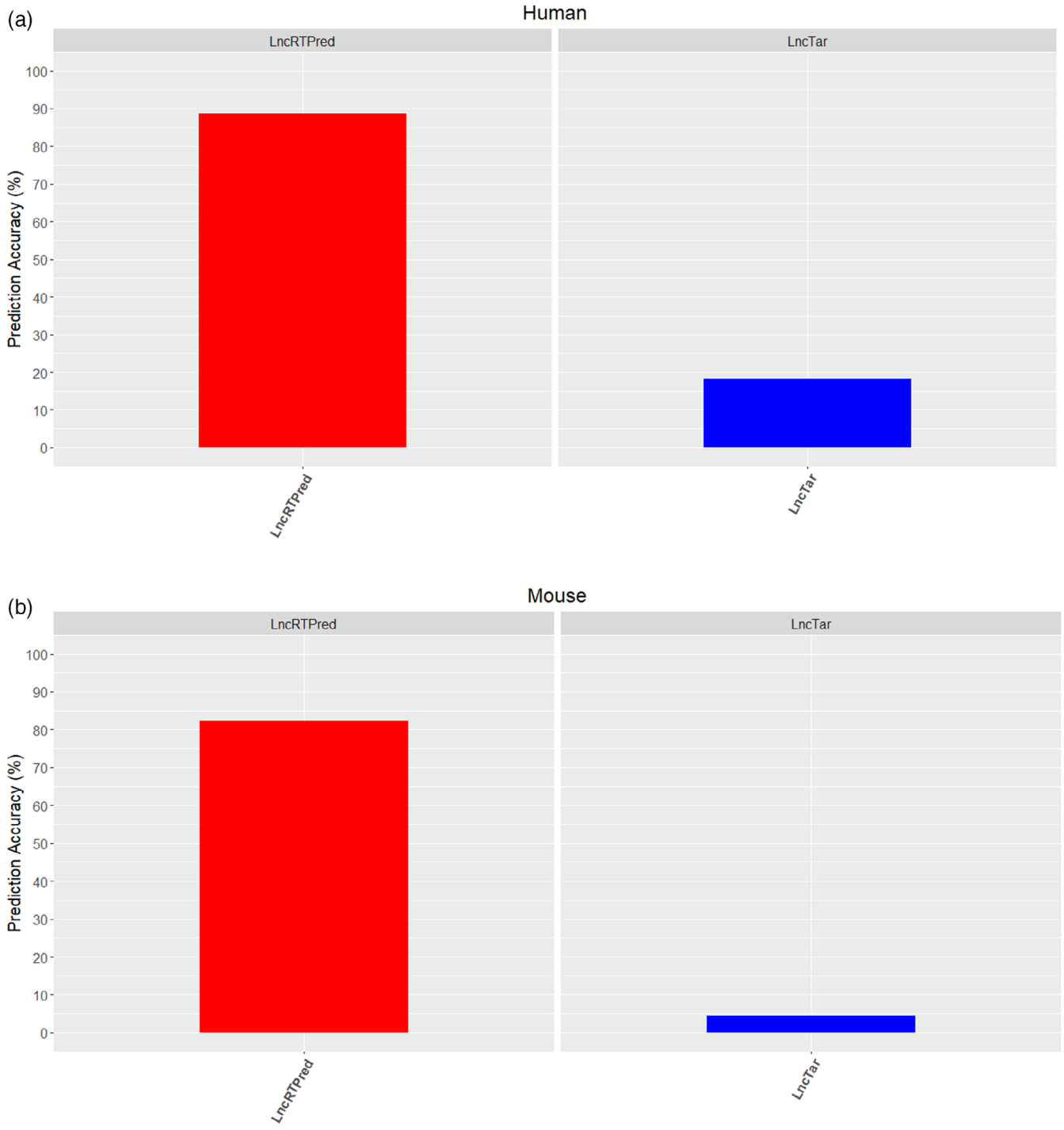

FIGURE 8 Performance comparison of LncRTPred and LncTar on unknown test data for (a) human and (b) mouse. Bar plot depicting the performance measure between LncRTPred cumulative_model_score and LncTar output (using the normalised free energy (ndG) cut-off of −0.1) corresponding to the validation data.

NSCLC, Liu et al.[61] demonstrated that HOTAIR lncRNA is significantly overexpressed in NSCLC tissue, it promotes tumour growth and metastasis by downregulating HOXA5 mRNA infer their interaction in NSCLC.

LncRTPred tool predicted the interaction among all these reported lncRNA–mRNA pairs to be positive (which corroborates with the experimental results) and the percentage of positive interaction predicted by four ML models named decision tree, KNN, random forest and LightGBM is shown in Figure 9. File S5 contains the details about the individual model and cumulative_model_score for each pair.

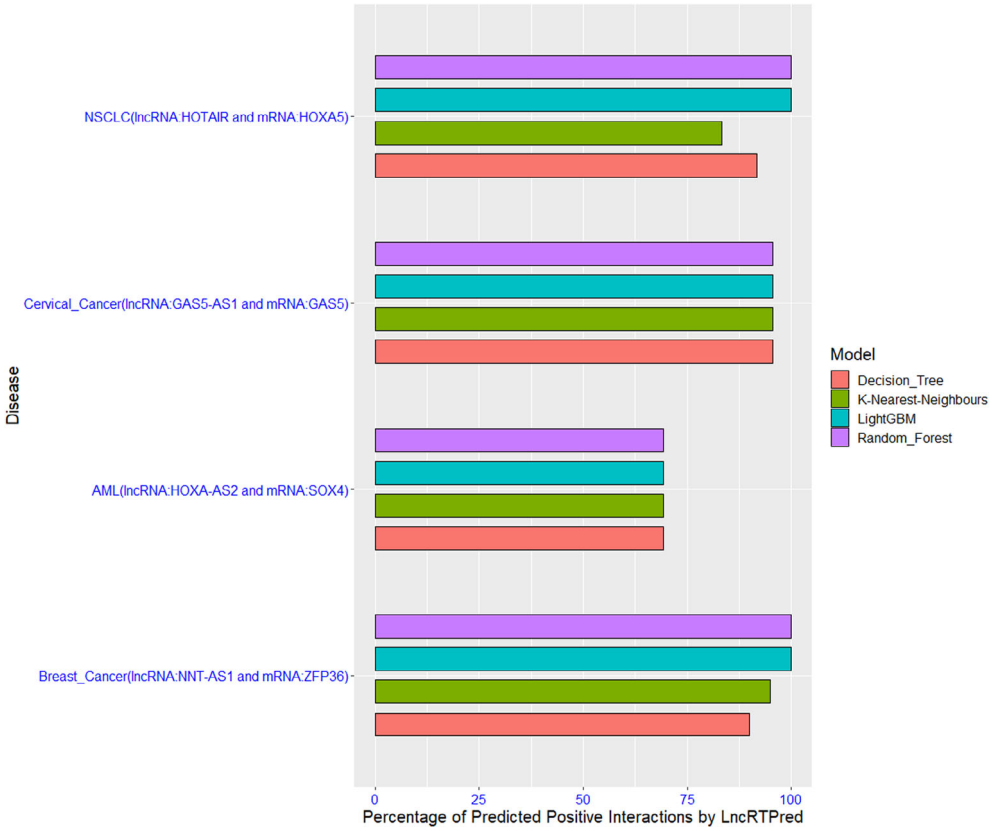

FIGURE 9 Performance testing of LncRTPred by predicting disease-specific lncRNA–mRNA target interactions. Horizontal Bar plot depicting the percentage of predicted positive interaction associated with multiple transcripts of well-known lncRNA–mRNA corresponding to AML, breast cancer, cervical cancer and NSCLC generated by LncRTPred.

4 | CONCLUSION

Being one of the most abundant classes of non-coding RNA molecules within the system, the regulatory role of lncRNA is widespread and their interacting partners are multiple. Extensive studies on lncRNAs over the years have unearthed that the intricate association of lncRNAs with their binding mRNA partners are responsible for bringing regulatory changes in both the normal and diseased conditions. The number of detected lncRNAs is enormous though the knowledge of their versatile functions is yet to be elucidated. In this regard, predicting their target mRNA partners could shed light on the RNA–RNA mode of interaction executed by the lncRNAs.

In this paper, we have developed an ML-based lncRNA-target prediction protocol named 'LncRTPred'. The creation of such prediction model identifies the determining factor for lncRNA–mRNA interaction. But the biggest obstruction in incorporating supervised ML models for such target prediction is the non-availability of reliable negative data. Here, while developing LncRTPred, we have adopted a novel approach to identify the non-interacting lncRNA and mRNAs from the unlabelled dataset using BLAT analysis. This has been further filtered by different computational techniques like one-class SVM and isolation forest to get the reliable set of outliers.

Overall, performance of LncRTPred as other ML models is heavily dependent on the size of the training dataset. From the results section, it is quite evident that its performance for human species shows better as compared to that for mouse when applied on an unknown data. This is due to the smaller size of the training dataset in case of mouse compared to that of human. In future, availability of significant number of lncRNA–mRNA interaction data corresponding to other species including mouse will allow us to make our model robust and versatile towards precise lncRNA-target prediction in case of mouse. This will also open up provisions for predicting lncRNA–mRNA target interaction for additional species. Further, it is essential to track its performance by in-depth wet lab validations subsequently.

ACKNOWLEDGEMENTS

This work is funded by the Bioinformatics Centre project at Bose Institute, Kolkata sanctioned by the Department of Biotechnology, Government of India, vide the sanction no. BT/PR40174/BTIS/137/45/2022 as well as the Council of Scientific and Industrial Research (CSIR) (sanction no. 09/015(05170/2017-EMR-I) and the Indian Council of Medical Research (ICMR) (sanction no. RBMH/FW/2020/10), Government of India.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interest to declare.

ORCID

Zhumur Ghosh https://orcid.org/0000-0002-4877-7551

REFERENCES

1. Kung JT, Colognori D, Lee JT. Long noncoding RNAs: Past, present, and future. Genetics. 2013;193(3):651–69.
2. Sanchez Calle A, Kawamura Y, Yamamoto Y, Takeshita F, Ochiya T. Emerging roles of long non-coding RNA in cancer. Cancer Sci. 2018;109(7):2093–100.
3. Cheng C, Moore J, Greene C. Applications of bioinformatics to non-coding RNAs in the era of next-generation sequencing. Pac Symp Biocomput. 2014;412–6. https://doi.org/10.1142/9789814583220_0039
4. Nakaya HI, Amaral PP, Louro R, et al. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. Genome Biol. 2007;8(3):R43.
5. Liu M, Zhang H, Li Y, et al. HOTAIR, a long noncoding RNA, is a marker of abnormal cell cycle regulation in lung cancer. Cancer Sci. 2018;109(9):2717–33.
6. Wilusz JE, Sunwoo H, Spector DL. Long noncoding RNAs: Functional surprises from the RNA world. Genes Dev. 2009;23(13):1494–504.
7. He RZ, Luo DX, Mo YY. Emerging roles of lncRNAs in the post-transcriptional regulation in cancer. Genes Dis. 2019;6(1):6–15.
8. Froberg JE, Yang L, Lee JT. Guided by RNAs: X-inactivation as a model for lncRNA function. J Mol Biol. 2013;425(19):3698–706.
9. Aliperti V, Skonieczna J, Cerase A. Long non-coding RNA (lncRNA) roles in cell biology, neurodevelopment and neurological disorders. Noncoding RNA. 2021;7(2):7–16.
10. Hur K, Kim SH, Kim JM. Potential implications of long noncoding RNAs in autoimmune diseases. Immune Netw. 2019;19(1):e4.
11. Chen H, Shan G. The physiological function of long-noncoding RNAs. Noncoding RNA Res. 2020;5(4):178–84.
12. Szcześniak MW, Makałowska I. lncRNA-RNA interactions across the human transcriptome. PloS One. 2016;11(3):e0150353.
13. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol. 2021;22(2):96–118.
14. Peng L et al. Probing lncRNA-protein interactions: Data repositories, models, and algorithms. Front Genet. 2019;10:1346.
15. Zeng T, Ni H, Yu Y, et al. BACE1-AS prevents BACE1 mRNA degradation through the sequestration of BACE1-targeting miRNAs. J Chem Neuroanat. 2019;98:87–96.
16. Izaguirre DI, Zhu W, Hai T, Cheung HC, Krahe R, Cote GJ. PTBP1-dependent regulation of USP5 alternative RNA splicing plays a role in glioblastoma tumorigenesis. Mol Carcinog. 2012;51(11):895–906.
17. Zhu L, Wei Q, Qi Y, et al. PTB-AS, a novel natural antisense transcript, promotes glioma progression by improving PTBP1 mRNA stability with SND1. Mol Ther. 2019;27(9):1621–37.
18. Sun J, Wang X, Fu C, et al. Long noncoding RNA FGFR3-AS1 promotes osteosarcoma growth through regulating its natural antisense transcript FGFR3. Mol Biol Rep. 2016;43(5):427–36.
19. Li Z, Liu L, Jiang S, et al. LncExpDB: An expression database of human long non-coding RNAs. Nucleic Acids Res. 2021;49(D1):D962–8.
20. Cheng L, Wang P, Tian R, et al. LncRNA2Target v2.0: A comprehensive database for target genes of lncRNAs in human and mouse. Nucleic Acids Res. 2019;47(D1):D140–4.
21. Zhao H, Shi J, Zhang Y, et al. LncTarD: A manually-curated database of experimentally-supported functional lncRNA-target regulations in human diseases. Nucleic Acids Res. 2020;48(D1):D118–26.
22. Terai G, Iwakiri J, Kameda T, Hamada M, Asai K. Comprehensive prediction of lncRNA-RNA interactions in human transcriptome. BMC Genomics. 2016;17(1):12.
23. Yi Y, Zhao Y, Li C, et al. RAID v2.0: An updated resource of RNA-associated interactions across organisms. Nucleic Acids Res. 2017;45(D1):D115–8.
24. Gong J, Shao D, Xu K, et al. RISE: A database of RNA interactome from sequencing experiments. Nucleic Acids Res. 2018;46(D1):D194–201.
25. Lu Z, Zhang QC, Lee B, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. Cell. 2016;165(5):1267–79.
26. Nguyen TC, Cao X, Yu P, et al. Mapping RNA-RNA interactome and RNA structure in vivo by MARIO. Nat Commun. 2016;7:12023.
27. Gawronski AR, Uhl M, Zhang Y, et al. MechRNA: Prediction of lncRNA mechanisms from RNA-RNA and RNA-protein interactions. Bioinformatics. 2018;34(18):3101–10.
28. Mann M, Wright PR, Backofen R. IntaRNA 2.0: Enhanced and customizable prediction of RNA-RNA interactions. Nucleic Acids Res. 2017;45(W1):W435–9.
29. Antonov I, Marakhonov A, Zamkova M, Medvedeva Y. ASSA: Fast identification of statistically significant interactions between long RNAs. J Bioinform Comput Biol. 2018;16(1):1840001.
30. Fukunaga T, Iwakiri J, Ono Y, Hamada M. LncRRIsearch: A web server for lncRNA-RNA interaction prediction integrated with tissue-specific expression and subcellular localization data. Front Genet. 2019;10:462.
31. Li J, Ma W, Zeng P, et al. LncTar: A tool for predicting the RNA targets of long noncoding RNAs. Brief Bioinform. 2015;16(5):806–12.
32. Fukunaga T, Hamada M. RIblast: An ultrafast RNA-RNA interaction prediction system based on a seed-and-extension approach. Bioinformatics. 2017;33(17):2666–74.
33. Huang YA, Huang ZA, You ZH, et al. Predicting lncRNA-miRNA interaction via graph convolution auto-encoder. Front Genet. 2019;10:758.
34. Yang L, Li LP, Yi HC. DeepWalk based method to predict lncRNA-miRNA associations via lncRNA-miRNA-disease-protein-drug graph. BMC Bioinformatics. 2022;22(Suppl 12):621.
35. Li J et al. WAFNRLTG: A novel model for predicting LncRNA target genes based on weighted average fusion network representation learning method. Front Cell Dev Biol. 2021;9:820342.
36. Guo T, Xu C, Shi B, Xu C, Tao D. Learning from bad data via generation. Paper presented at Proceedings of the 33rd

international conference on neural information processing systems. 2019, Curran Associates Inc. Article 543.
37. Zhang J, le TD, Liu L, Li J. Inferring and analyzing module-specific lncRNA-mRNA causal regulatory networks in human cancer. Brief Bioinform. 2019;20(4):1403–19.
38. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 2007;35-(Database issue):D61–5.
39. Teng X, Chen X, Xue H, et al. NPInter v4.0: An integrated database of ncRNA interactions. Nucleic Acids Res. 2020;48(D1): D160–5.
40. Yates AD, Achuthan P, Akanni W, et al. Ensembl 2020. Nucleic Acids Res. 2020;48(D1):D682–8.
41. Pearson K. LIII. On lines and planes of closest fit to systems of points in space. London Edinburgh Dublin Philosophic Magaz J Sci. 1901;2(11):559–72.
42. Hotelling H. Analysis of a complex of statistical variables into principal components. J Educ Psychol. 1933;24:417–41.
43. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
44. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Paper presented at Proceedings of the fifth annual workshop on computational learning theory. 1992, Association for Computing Machinery:Pittsburgh, Pennsylvania, USA. pp. 144–152.
45. Cortes C, Vapnik V. Support-vector networks. Machine Learn. 1995;20(3):273–97.
46. Schölkopf B, Williamson R, Smola A, Shawe-Taylor J, Platt J. Support vector method for novelty detection. Paper presented at: Proceedings of the 12th international conference on neural information processing systems. 1999, MIT Press:Denver, CO. pp. 582–588.
47. Liu FT, Ting KM, Zhou ZH. Isolation Forest. in 2008 Eighth IEEE International Conference on Data Mining. 2008.
48. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B. Scikit-learn: Machine learning in python. J Mach Learn Res. 2011;12:2825–30.
49. Thorndike RL. Who belongs in the family? Psychometrika. 1953;18(4):267–76.
50. Suresh V, Liu L, Adjeroh D, Zhou X. RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Res. 2015;43(3):1370–9.
51. Song J, Liu G, Wang R, Sun L, Zhang P. A novel method for predicting RNA-interacting residues in proteins using a combination of feature-based and sequence template-based methods. Biotechnol Biotechnol Equip. 2019;33(1):1138–49.
52. Seemann SE, Richter AS, Gesell T, Backofen R, Gorodkin J. PETcofold: Predicting conserved interactions and structures of two multiple alignments of RNA sequences. Bioinformatics. 2011;27(2):211–9.
53. Li AX, Marz M, Qin J, Reidys CM. RNA-RNA interaction prediction based on multiple sequence alignments. Bioinformatics. 2011;27(4):456–63.
54. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1): 81–106.
55. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inf Theory. 1967;13(1):21–7.
56. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
57. Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In NIPS. 2017.
58. Pan QH, Fan YH, Wang YZ, Li DM, Hu CE, Li RX. Long noncoding RNA NNT-AS1 functions as an oncogene in breast cancer via repressing ZFP36 expression. J Biol Regul Homeost Agents. 2020;34(3):795–805.
59. Qu Y, Wang Y, Wang P, Lin N, Yan X, Li Y. Overexpression of long noncoding RNA HOXA-AS2 predicts an adverse prognosis and promotes tumorigenesis via SOX4/PI3K/AKT pathway in acute myeloid leukemia. Cell Biol Int. 2020;44(8):1745–59.
60. Wang X, Zhang J, Wang Y. Long noncoding RNA GAS5-AS1 suppresses growth and metastasis of cervical cancer by increasing GAS5 stability. Am J Transl Res. 2019;11(8):4909–21.
61. Liu XH, Liu ZL, Sun M, Liu J, Wang ZX, De W. The long noncoding RNA HOTAIR indicates a poor prognosis and promotes metastasis in non-small cell lung cancer. BMC Cancer. 2013; 13:464.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Das G, Das T, Parida S, Ghosh Z. LncRTPred: Predicting RNA–RNA mode of interaction mediated by lncRNA. IUBMB Life. 2023. https://doi.org/10.1002/iub.2778