2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Computational Prediction of lncRNA-Protein Interactions using Machine learning

Muhammad Mushtaq , Hammad Naveed, Zoya Khalid

Abstract— **Long non-coding RNAs have generated much scientific interest because of their functional significance in regulating various biological processes and also their dysfunction has been implicated in disease progression. LncRNAs usually bind with proteins to perform their function. The experimental approaches for identifying these interactions are time taking and expensive. Lately, numerous method on predicting lncRNA-protein interactions have been reported yet, they all have some prevalent drawbacks that limit their prediction performance. In this research, we proposed a computational method based on a similarity scheme that integrates features derived from sequence and structure similarities. When compared with the state of the art, the proposed method has achieved highest performance with accuracy and F1 measure of 98.6% and 98.7% using XGBoost as classifier. Our results showed that by combining sequence and structure based features the lncRNA protein interactions can be better predicted and can also complement the experimental techniques for this task.**

Clinical Relevance— **The lncRNA-protein interactions play significant role in regulating various biological processes. This can help in providing early diagnosis and better treatment for cancer related diseases.**

I. Introduction

Recent studies have shown that only 2% of RNA transcripts are involved in the protein translation process. The remaining 98% that do not encode proteins were declared as transcriptional noise [1]. Long non-coding RNAs (lncRNAs) are a class of non-coding RNAs (ncRNAs) having more than 200 nucleotides, which are involved in various biological processes including regulation of gene expression, transcription, post-translational regulation, chromatin modification, and disease progression Hu et al. (2018). LncRNAs usually have independent regulatory components such as promoters and enhancers. LncRNAs also have complicated higher-order or secondary structures and are longer than other types of ncRNAs [2].

Currently, lncRNAs are gaining increasing scientific interest as they are involved in a variety of processes which predominantly include cell differentiation, apoptosis and cancer progression. The lncRNAs have to bind to a protein to perform its function. Numerous experimental methods were designed to discover lncRNA-protein interactions including RNA immunoprecipitation and mass spectrometry [1]. Such experimental techniques are costly and time-consuming, hence to complement such processes many computational methods have been proposed lately. NPInter database is built on the data produced from these high throughput experiments, hence is a widely used source for carrying out computational analysis.

Zhang et al. proposed the sequence based ensemble method "SFPEL-LPI" [3] that extracts sequence derived features of lncRNAs and proteins. The features are based on a similarity scheme derived from lncRNA-lncRNA and protein-protein similarities. The method also predicted novel interactions that were verified by the literature. The reported accuracy and F1-score of the model are 96% and 47%. Another study reported by Huan et al. proposed a lncRNA-protein interaction prediction model known as HLPI-Ensemble [1]. HLPI-Ensemble specifically predicts only human lncRNA-protein interactions. The model is an ensemble method that combines SVM, Random Forests and XGB models. The method introduces random pairing strategy for generating a negative dataset for the prediction task. The model is trained on sequence based features with the reported accuracy of 74.4% and f1-score of 77.8%. Xie et al proposed a method known as LPI-IBNRA to predict the lncRNA-protein interactions, by implementing bipartite network recommender algorithm. The feature set combines known lncRNA-protein interactions, protein-protein interactions and lncRNA expression similarity features. The method achieved 88.4% accuracy and 52.7% f1-score.

Another method proposed by [4] integrated the random walk and neighborhood regularized logistic matrix factorization algorithms into a semi-supervised model. This method did not use negative dataset for training rather uses known interactions to predict the unknown ones. The model combines know lncRNA-protein interactions, lncRNA similarity network and protein similarity network for predicting lncRNA-protein interactions. The method has achieved 90% accuracy with 65.2% f1-score.

Suresh et al. presented the RPI-Pred (RNA-protein interaction predictor), an SVM based method, which integrates both sequence and structure-based features for

M. M. Author is with the Computational Biology Research Lab, Department of Computer Science, National University of Computer and Emerging Sciences, NUCES-FAST, Islamabad, Pakistan (e-mail: derawall@gmail.com).
H.N. Author is with the Computational Biology Research Lab, Department of Computer Science, National University of Computer and Emerging Sciences, NUCES-FAST, Islamabad, Pakistan (e-mail: hammad.naveed@nu.edu.pk).
Z.K Author is with the Computational Biology Research Lab, Department of Computer Science, National University of Computer and Emerging Sciences, NUCES-FAST, Islamabad, Pakistan (e-mail: zoya11khalid@gmail.com).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

predicting lncRNA-Protein interactions [5]. The protein structures are represented in terms of 16 structural fragments called protein blocks (PBs). For the RNA high order structure, RPI-Pred used the five classes of RNA secondary structures (RSS), namely, stem, hairpin, loop, bulges and an internal loop. The obtained PBs and RSS were combined with their corresponding amino acids and nucleotide sequences and passed to SVM for the classification of lncRNA and protein interactions. The accuracy of RPI-Pred on NPInter database is reported to be 86.9%. Pan et al proposed a fully sequence-based method, IPMiner [6] to predict the ncRNA-protein interactions. IPMiner extracts the raw sequence composition features from protein and RNA sequences and used them into a stacked auto-encoder to extract the high-level hidden features, then deploy them into a random forest to predict the interaction between ncRNA and protein. To integrate the different models' predictions the stacked ensemble technique is used. The reported accuracy by IPMiner is 95.2% on the NPInter dataset.

In addition to these methods, there are other computational methods based on deep learning algorithms. RPITER proposed by [7] is a hierarchical deep learning method for predicting RNA–protein interactions. For sequence coding, RPITER improved the method for coding the conjoint triad feature (CTF) by adding sequence and structure information into the coding vectors. RPITER used a convolution neural network (CNN) and a stacked auto encoder (SAE) architecture. The authors have compared their deep learning model with other reported methods on ncRNA-protein interaction predictions and the proposed model has outperformed with the achieved accuracy of 95.5%. Zhan et al proposed a method BGFE by employing the sequence-based features in a deep learning model Stacked Auto Encoder (SAE) combined with random forest (RF) classifier Zhan et al. (2019). The K-mers sparse matrix is used to represent the lncRNA sequences and Singular Value Decomposition (SVD) is used to extract the lncRNA sequence-based features. They have employed Position-Specific Scoring Matrix (PSSM) to extract protein sequence-based features and the bi-gram algorithm to get feature vectors from PSSMs. SAE is used to learn high-level hidden information and RF to classify lncRNAs-proteins interactions. They used 3 different datasets and reported the results, on RPI488 dataset the accuracy is 88.68%, on RPI1708 the accuracy is 96.0% and on RPI2241 the accuracy is 91.3%.

However, the reported methods have few limitations to address. First, the majority of the models used only positive dataset considering only the known interactions between proteins and lncRNAs which thus creates biased predictions. Secondly, the previous reported studies have considered all species data together to predict lncRNA-protein interactions however, the lncRNA sequence homology is very low so, one generalized model cannot be applicable for all species sequence data. Also, selecting suitable features for predicting lncRNA-protein interactions is challenging. The methods employing deep learning algorithms carry an inherent drawback of requiring more amount of data with more computing power as well. Keeping this in consideration, we have developed a computational method based on a similarity scheme that combines protein and lncRNA features derived from sequence similarities and structural similarities. Unlike previous machine learning based methods our method has utilized negative dataset and produced comparable results. Compared with the deep learning methods our method does not require high computation power to predict lncRNA-protein interactions.

II. Methods

A. Dataset

The total of 8162 lncRNA-protein interactions were downloaded from NPInter [8] database that contains experimentally confirmed lncRNA-protein interactions. The sequences of lncRNAs were extracted from NONCODE V3 [9] database, while the protein sequences were downloaded from Uniprot Consortium (2018). We kept only those lncRNAs whose nctype is NONCODE and lncRNA, inter-class is binding and organism is Homo Sapiens because the focus of the study is predicting the human lncRNA-protein interactions. After removing the duplicate entries, the dataset has reduced to 3951 lncRNA-protein interactions containing 1625 unique lncRNAs and 26 proteins. As NPInter database only provides the positive dataset comprising of interactions data, for creating a negative dataset we used the random pairing strategy as mentioned in [1] [10] [11].

B. LncRNA-Protein Interactions

We denoted the number of proteins with P and the number of lncRNAs with L and L∈P as a feature matrix of interactions between lncRNAs and proteins. The interaction among lncRNA L_i and protein P_i could be denoted as follow:

$$I(p_i, l_i) = \{1 \text{ if } p_i \text{ interact with } l_i, 0 \text{ otherwise.}$$

C. Local Pairwise Sequence Similarity

Protein-protein and lncRNA-lncRNA local pairwise sequence similarities were calculated using the smith waterman algorithm (SW) [12]. As the length of the sequences is varied, we used the normalized smith waterman score between protein p and p^{-} by using the following equation as provided in [13]. The authors divide the SW score of two proteins by their geometric mean of the self-alignment SW score. same procedure is applied for lncRNAs to normalize the sequence similarity score

$$S(p, p^{-}) = \mathrm{SW}(p,p^{-}) / \sqrt{SW(p,p)} \ \sqrt{SW(p,p^{-})}$$

D. Extracting Kmers

Normally, the term k-mer refers to all possible length k sub-strings that are present in a string. In computational genomics, k-mers refer to all possible (length k) sub-sequences from a sequence read obtained from DNA , RNA or a protein sequence. To extract the k-mer features we have used k=1,2 and 3 for protein and K=1,2,3 and 4 for lncRNA to generate overlapping k-mers For proteins total extracted features for

K=3 are $20^3 = 8000$ while in RNA the total computed features for K=4 are $4^4 = 256$

E. PC-PseAAC-General

PseAAC-General provides the pseudo amino acid composition for protein sequences. Pseudo amino acid composition incorporates the global sequence order information of protein sequences into their feature vectors via the physicochemical properties of their constituent amino acids and have been widely used in bioinformatics tasks. There are different modes of PseAAC available in web server Pse-in-One [14] we used PC-PseAAC-General which contains additional features of gene ontology and functional domain mode. PC-PseAAC-General creates the protein feature vectors by combining the amino acid composition and global sequence-order effects. We have selected the parameters with lambda value 10 and weight factor 0.5.

F. PC-PseDNC-General

The abbreviation of PseDAC is pseudo deoxyribonucleic acid compositions for DNA/RNA sequences. There are various modes of PseDAC available in Pse-in-One [14] tool, we used PC-PseDNC-General to generate the feature vector for lncRNA. PseDNC uses the six local RNA structural properties of dinucleotides and using PC-PseDNC-General mode that is based on the properties of dinucleotides, we generated parallel correlation components feature vector of 26 features using 6 physicochemical properties (Rise (RNA), Tilt (RNA), Twist (RNA), Slide (RNA), Shift (RNA),Roll (RNA)), and parameters with lambda =10 and weight factor = 0.5.

G. Protein structure similarity

For computing protein structure similarity we have computed the pairwise local structural alignment using the ProBis Konc and [15] tool, between 26 distinct proteins in filtered NPInter V2 dataset. ProBis algorithm uses the complete protein surfaces, motifs or protein binding sites to align the two protein structures. ProBis computes the pairwise alignment of entire protein structure or selected binding sites. It also enables the fast database searches for similar protein binding sites, it can search similar binding sites in different protein 3D structures without prior knowledge of their locations.

H. lncRNAs structure similarity

To calculate the pairwise structure alignment of LncRNA secondary structure we used RNAforester HHochsmann M [14] tool distributed in Vienna RNA package Lorenz. RNAforester calculates¨ the pairwise alignment of RNA secondary structures. It computes tree and forest alignment using local similarity algorithm. The paired and unpaired bases of RNA secondary structure can be represented as a forest.

I. Model Building

The sequence and structure-based features were trained on SVM and XGB classifiers. Features were trained both separately and combined to check the model performance. For SVM the model was trained on both sequence and structure features separately and combined with different hyperparameters to obtain the highest performance of the model. The best results were obtained with kernel = linear, C=2 and gamma = 10. Four different models of XGB were trained on different features combinations (sequence based and structure based). The best performance was obtained using the n-estimator 77 and 100, learning rate 0.1, max depth 5 and objective function binary: logistic. We have used stratified 5 fold cross validation to evaluate the performance of the models.

III. Results

A. Model Performance

To evaluate our classification models we have implemented stratified 5 fold cross validation and computed the Accuracy, Precision, Recall, and F-1 score. The whole dataset is randomly divided into two parts 20% test set and 80 % training set. Further, the training set is divided into five equal folds, each time 4 folds are used for training and one fold is used as a validation set. Both sequence and structural features were trained on SVM and XGB classifiers. The features were tested separately and combined to evaluate the model performance. The results obtained showed significant increase of the performance measure by adding structure-based features. The results are tabulated in Table 1. By combining sequence and structure similarity features the XGB has achieved 98.68% accuracy and 98.71% F1-Score.

B. Comparison with the state of the art

The proposed model is compared with the existing methods RPITER [6] and IPMiner [7]. RPITER is a deep learning-based hierarchical model which predicts ncRNA-Protein interactions by integrating the sequence and structural information of lncRNAs and proteins, IPMiner is a fully sequence based method that extracts the raw sequence composition features from protein and RNA sequences and deploy them into a random forest to predict the interaction between ncRNA and protein. We have trained the two methods on our NPInter filtered dataset and evaluated the performance of the methods. On RPITER method, the classification model has achieved accuracy of 95.9%, precision and recall of 96.9 % and F1-score of 97.8% which is better as compared than the original datasets used in this study. IPMINER achieved 95.7% accuracy and 95.6% precision, recall and F1 score. The results are tabulated in Table 1.

Both these methods are deep learning based which requires high computational power and comparatively large amount of training data for obtaining good results. Our method has tackled various limitations of the previous methods by adding the negative dataset for better prediction and building a human specific model for lncRNA-protein prediction. Moreover, we have incorporated structure based features which are deemed to increase the model efficiency. We believe that the proposed method has achieved comparable results with the state of the art and can also be computed with limited resources.

TABLE I. RESULTS OBTAINED ON BOTH SEQUENCE AND STRUCTURE BASED FEATURES

Method	**Acc%**	**Pre%**	**Rec%**	**F1%**
Proposed Method	98.6	97.7	99.6	98.7
REPITER Peng et al. (2019)	95.5	96.9	96.9	97.8
IPMiner Pan et al. (2016)	95.7	95.6	95.6	95.6

TABLE II. COMPARISON WITH THE STATE OF THE ART

Model	***LncRNA Features***	***Protein Features***	***Accuracy***	***F1-Score***
XGB	All features	All features	98.68	98.71
XGB	Sequence features	Sequence features	74.85	76.14
SVM	All features	All features	70.83	73.04
SVM	Sequence features	Sequence features	72.16	74.0

IV. CONCLUSION

Long non-coding RNAs (lncRNAs) play a significant functional role in regulating various biological processes and disease progression, by interacting with specific proteins. The experimental approaches for predicting such interactions are very time consuming and costly. We aimed to complement this, by developing a computational method that predicts protein-lncRNA interactions in a speedy and cost-effective manner. Our computational method combines the features derived from the sequence and structure of proteins and lncRNAs. The model is designed specifically for human lncRNA-protein interactions unlike the previous approaches. SVM and XGBoost models were trained on sequence and structure based features separately as well as combined by applying 5 fold cross validation. The results with XGBoost has outperformed the other classification models by achieving 98.6% accuracy and 98.7 F1-score. We have tried to implement a simpler model that caters all the required features and can be executed in less time and resources. Although, we have achieved very significant results, still the method has few limitations which includes the limited availability of known lncRNA-protein interactions data and the unavailability of protein structure data for deploying structure based features for model execution.

REFERENCES

[1] Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., et al. (2018). Hlpi-ensemble: Prediction of human lncrna-protein interactions based on ensemble strategy. RNA Biology 15, 797–806. doi:10.1080/15476286.2018.1457935. PMID: 29583068

[2] Xie, G., Wu, C., Sun, Y., Fan, Z., and Liu, J. (2019). Lpi-ibnra: Long non-coding rna-protein interaction prediction based on improved bipartite network recommender algorithm. Frontiers in Genetics 10, 343. doi:10.3389/fgene.2019.00343

[3] Zhang, W., Yue, X., Tang, G., Wu, W., Huang, F., and Zhang, X. (2018). Sfpel-lpi: Sequence-based feature projection ensemble learning for predicting lncrna-protein interactions. PLOS Computational Biology 14, 1–21. doi:10.1371/journal.pcbi.1006616

[4] Zhao, Q., Zhang, Y., Hu, H., Ren, G., Zhang, W., and Liu, H. (2018). Irwnrlpi: Integrating random walk and neighborhood regularized logistic matrix factorization for lncrna-protein interaction prediction. Frontiers in Genetics 9, 239. doi:10.3389/fgene.2018.00239

[5] Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Research 43, 1370–1379. doi:10.1093/nar/gkv020

[6] Pan, X., Fan, Y., Yan, J., and Shen, H.-B. (2016). Ipminer: Hidden ncrna-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. BMC Genomics 17, 582. doi:10.1186/s12864-016-2931-8

[7] Peng, C., Han, S., Zhang, H., and Li, Y. (2019). Rpiter: A hierarchical deep learning framework for ncrna–protein interaction prediction. International Journal of Molecular Sciences 20, 1070. doi:10.3390/ijms20051070

[8] Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2013). NPInter v2.0: an updated database of ncRNA interactions. Nucleic Acids Research 42, D104–D108. doi:10.1093/nar/gkt1057

[9] Bu, D., Yu, K., Sun, S., Xie, C., Skogerbø, G., Miao, R., et al. (2011). NONCODE v3.0: integrative annotation of long noncoding RNAs. Nucleic Acids Research 40, D210–D215. doi:10.1093/nar/gkr1175

[10] Suresh, V., Liu, L., Adjeroh, D., and Zhou, X. (2015). RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. Nucleic Acids Research 43, 1370–1379. doi:10.1093/nar/gkv020

[11] Wekesa, J. S., Luan, Y., Chen, M., and Meng, J. (2019). A hybrid prediction method for plant lncrna-protein interaction. Cells 8, 521. doi:10.3390/cells8060521

[12] Smith, T. F., Waterman, M. S., and Burks, C. (1985). The statistical distribution of nucleic acid similarities. Nucleic Acids Research 13, 645–656. doi:10.1093/nar/13.2.645

[13] Bleakley, K. and Yamanishi, Y. (2009). Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics (Oxford, England) 25, 2397–403. doi:10.1093/bioinformatics/btp433

[14] Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., and Chou, K.-C. (2015). Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Research 43, W65–W71. doi:10.1093/nar/gkv458

[15] Konc, J. and Janezic, D. (2010). ProBiS: a web server for detection of structurally similar protein bindingˇ sites. Nucleic Acids Research 38, W436–W440. doi:10.1093/nar/gkq479

[16] Hoechsmann M, Toeller T, Giegerich R and Kurtz S, (2003) Local Similarity of RNA Secondary Structures, Proc. of the IEEE Bioinformatics Conference (CSB 2003), 159-168