International Conference on Recent Trends in Data Science and its Applications
DOI: rp-9788770040723.073

Prediction of RNA Function in the Context of Linear Non-Coding RNAusing a Heterogeneous Network Approach

R.PanneerSelvi
Assistant Professor,Department of Computer Science and Engineering, Vel Tech RangarajanDr.Sagunthala R&D Institute of Science and Technology,
Chennai, Tamilnadu,India,
panneerselvir@veltech.edu.in

K Saravanan
Associate professor, Department of Information Technology, R.M.K Engineering college (Autonomous)
Kavaraipettai, GummidipondiTaluk, ThiruvallurDist,
Chennai 601206, Tamilnadu,
sara.shapnaasree@gmail.com

Kaushal Kumar
Department of Botany, School of Life Sciences, Mahatma Gandhi
Central University, Bihar,
India,kaushalkumar127@gmail.com

R.Sree Parimala
Associate Professor, Department of Mathematics, Sri Eshwar College of Engineering,
Coimbatore,Tamilnadu,India,sreeparimala.r@sece.ac.in

DevvretVerma
*Assistant Professor, Department of Biotechnology, Graphic Era Deemed To Be University,*Dehradun,Uttarakhand,India,248002,
devvret@geu.ac.in

A.V.S.Ram Prasad
Department of Mechanical Engineering,KoneruLakshmaiah Education Foundation,
Guntur--522502, Andhra Pradesh,India,
avsrp_me@kluniversity.in

Abstract—Linear non-coding RNA (lncRNA) has recently attracted a variety of research, critical attention to some intriguing new knowledge about its essential functions. Thousands of lncRNAs were discovered in absurdly short durations because of tools like RNA-Seq. Only a handful of them should be intrinsically depicted, although, due to lesser citation incidence. Wet lab tests to elucidate the roles of lncRNA are complex, laborious, and sometimes ridiculously expensive. This research explores the critical topic of generating molecular docking toward determining the essential functions of lncRNAs. The model provided here uses a measure based on a systematic review known as AvgSim on a heterogeneous network infrastructure to predict various lncRNA (HIN) RNA activities. The same framework is designed with proposed statistics on the lncRNA enzyme or activity relationships and information on the lncRNA co-expression and linkage proteins. The proposed methodology forecasts probable capabilities for 2,695 lncRNAs with a reliability of 73.68% out of 2,758 lncRNAs evaluated for investigation and was performed remarkably better than other methodologies like a random forest for an impartial training dataset. These same linked roles of two well-known lncRNAs are investigated inside as an example. The findings were confirmed by research observations identified in this review.

Keywords—Heterogeneous Intelligence Network; noncoding RNA; statistical approaches; Random forest approach

I. INTRODUCTION

They possess essential features like enzyme translating capability and structural conservation, which are also required for biological functions. Until a generation later, many quasi segments of RNA were thought to be 'garbage,' having no specific genes [1]. Recent evidence suggests that lncRNAs play a part in developing cellular mechanisms, including cell-type-specific expression, intracellular element distribution, and sickness correlation [2]. The activity of lncRNAs is engaged by these physiological and morphological attributes because it is no longer considered. Even though many lncRNAs have been discovered to date, the amount of adequately annotated lncRNAs is meager [3]. The asymmetry in lncRNA identification or categorization frequencies is responsible for the lack of functional awareness concerning lncRNAs. Wet labs' investigation to characterize lncRNA operations is costly, consequential, and exhausting. As a result, the computational intelligence approaches anticipating lncRNA activities are an urgent requirement in lncRNA exploration [4]. Consequently, a distinct field of study has evolved in which computer architecture is employed for lncRNA investigation. The web techniques rely more on the channel's implicit information than on the biological features of lncRNAs [5]. Consequently, such attempts gained popularity in a short amount of time. The cornerstone of infrastructure approaches is built on two assumptions. First is the "remorse by association" theory, which holds that genes that regulate biochemical reactions might co-express proteins involved in the same activity [6]. The second is that while executing a function, compounds communicate with others. As an outcome, while developing a model for predicting lncRNA activities, the relationships of lncRNAs, especially their co-expression, should be emphasized [7].

II. RELATED WORKS

The interacting links among lncRNA and enzyme are used as the primary component to create a channel in the number of existing methodologies for a statistical sense of lncRNA functions. Approximately 340 lncRNAs were designated categories based on the functionalities of nearby molecules in an encoding non-coding genetic co-expression system. This was among the first experiments at predicting lncRNA function. Molecular interactions were not examined in this study. Interacting protein data from the international lncRNA function and performance tool (link-GFP) into the co-expression system [8]. It identified 1,625 lncRNAs that had biological functions. However, none of the other approaches processed data from Next Generation Technology (NGS).In the available research, all internet techniques rely on the lncRNA - protein association as a critical criterion for developing a predictive model the capabilities of lncRNAs [9].As a consequence, these approaches could be used to forecast the activities of lncRNAs with known protein complexes. The lncRNA-

DOI: rp-9788770040723.073

lncRNA linkages should be considered in order to employ this same benefit of an infrastructure paradigm. Even with the lack of protein sequence relationships, the proposed study analyses lncRNA-lncRNA interactions and forecasts overall activities of lncRNAs.After the HIN is built, a coefficient of determination is used to determine relevant promoter regions meta-paths [10]. Along certain results were compared, the connectedness metric AvgSim is determined. The features for a Classification Algorithm that predicts lncRNA processes were formed by combining the AvgSim scores along multiple metapaths. In comparison to existing approaches that rely solely on lncRNA-protein interactions for forecasting, the proposed study makes advantage of HIN's meta-path based knowledge. The technique assigns available features to a maximum of 2,695 lncRNAs.The precision is proved statistically using involves attempting through a recent review research [11]. The outcomes of a research study of two well-studied lncRNAs are indeed confirmed.

III. Proposed Methods

The Heterogeneous LncRNA-Protein-Function Connection (HLPFN) is a network which connects five separate binding interactions: proteins connections, lncRNA co-expression, lncRNA operational connections, or lncRNA-protein interrelations and protein function connections [12]. To preserve the heterogeneity of vertices and edges, all connections are retained as independent pairwise models. Fig.1 depicts the multiple transitive vectors required to generate HIN. This same transitive index of subnetworks is constructed as follows.

NPInter is used to gather information on LncRNA-protein interactions. It discusses the interactions of ncRNAs with various macromolecules [13]. To extract enzymes, the lncRNAs are separated from the ncRNAs using NONCODE ID and the particular intervention is limited to 'RNA-protein.' Only 'Human Ancestors' can converse with others. The lncRNA-protein association system is managed as a data structure, with lncRNAs in the rows and enzymes in the columns. The borders are generated based on the obtained contextual information. The analysis of the structure MLP is employed in this work. This method is impossible because of the amount of potential conceptual grows exponentially as the duration of the conceptual expands, rendering the issue unsolvable. One option is to follow the regular patterns identified in HIN research [14].

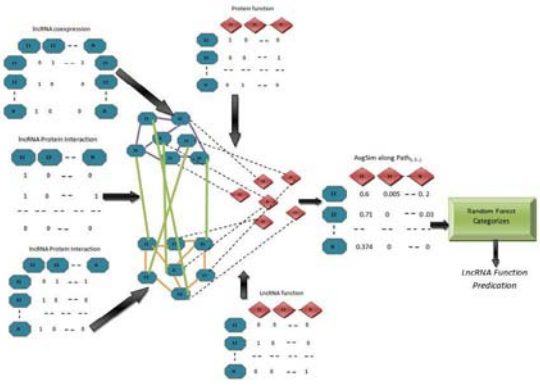

Fig.1. Determining lncRNA expression

$$K_{lp}(x,y) = \begin{cases} 1, if\ LncRNA_x, interacts\ \ with\ Proteins_y \\ 0, otherwise \end{cases}$$

In most cases, increasing overall conceptual duration endlessly in an HIN is pointless. Consequently, the chain of relationships between of expected completion elements might be very extensive in prolonged metapaths. As a consequence, inadvertently increasing conceptual duration is generally unproductive and produces outcomes that are far less accurate. As a consequence of this understanding, most HIN-based approaches specify a conceptual duration maximum limit in preparation. It only examine the meta-paths that are smaller than the criterion. It's critical to ensure that the findings aren't harmed by the removal of longer paths when setting such a threshold.If the findings are impacted, the threshold value should be reviewed, and this technique attempts to justify the maximum permissible setting process. A criterion limit can be established arbitrarily, statistically, or heuristically in most cases.

An observational technique is used in this study. The connection of all life conceptual with real existing life linkages is repeatedly evaluated beginning at duration three. The repetition ends when the number of adversities linked metapaths equals half of the number of favorably connected metapaths. The duration criterion is determined by path length at this point. The next eight lines go into this technique in great depth. The prospective meta-paths to be included in the investigation are discovered in the first stage, which is an incremental method also on conceptual duration. Because of mission is to find a link among lncRNA to service, only meta-paths that begin with lncRNA or conclude with feature are taken into account. The repetition must begin with shortest conceptual conceivable: lf, with a duration of two. In reality, this route correlates to actual function in the HIN and illustrates the actual functional association of lncRNAs. In practice, the iteration begins with a three-length meta-path. Every repetition leads to formation of a vector containing an overall maximum count of the concept of that type among each combination of lncRNAs or processes.The next challenge is to decide the appropriate meta-paths. If the concept question is significant, connectivity features suggest that its number across l or f should be higher. This suggests that the lncRNA is quite likely to fulfil that role. In contrast, if the amount alternative conceptual is low, this same likelihood of such lncRNA performing that purpose is low. A comparison study of the conceptual column is done against the known relationship pattern in order to function this tendency.All conceptual in the meta-path vector that have a significant correlation to established of connections are thought necessary, while the rest are considered invalid. This same terminate requirement is written so that the repetition comes to a halt so when the amount of negatively correlated meta-paths exceeds half of number of positively associated meta-paths. As previously indicated, the repetition begins with a three-length meta-path and continues until the termination point is achieved. A termination condition is satisfied in this study at a conceptual duration of four, that is used as duration criterion.

DOI: rp-9788770040723.073

IV. Performance Evaluations

K-fold cross-validation idea to evaluate overall forecasting accuracy. The specimens TP or TN have accurately predicted advantages and disadvantages correspondingly. FP and FN reflect the amount of incorrectly anticipated positive and negative samples, respectively. It can be produced in the following manner: An indicator value for each connection is calculated using the possibilities supplied by the algorithm. This same forecast value is used to order related couples. True Positives are known relationships with a more excellent predictive value than a certain criterion value point. In contrast, True Negatives are unknown correlations with a lower predictive value than the threshold. False Positives are recognized relationships with a value just below the criterion, while False Negatives are undiscovered connections with a value above the criterion.

V. Results and Discussions

Assuming 73.68% for the model can predict novel capabilities for 2,695 lncRNAs. Different functions were expected for some of the lncRNAs. The factor influencing outcomes suggests that lncRNAs are directly implicated in physiological systems rather than cellular and molecular level capabilities. Many previously unknown lncRNA functions were predicted using this strategy. The functions come from the GO collaboration. The GO ontology tracks parent-child connections using GOBasic, GOSlim, and other operational groups. To comprehend any such activities provided by lncRNAs, various operational GO Terms are grouped according to their GOSlim category. Table 1 shows this category-by-category list. The figure displays a list of functional domains and the number of GO keywords within each classification. It demonstrates that lncRNAs have key roles in cellular mechanisms, growth, cellular component, and metabolism, molecular function. The CateGOrizer program was used to generate these outcomes.

TABLE 1: LncRNAs' Important Features

Go Class ID	Definition	Count	Go Class ID	Definition	Count
GO:0007034	biological process	2647	GO:00014680	kinase activity	24
GO:0007152	metabolism	868	GO:0068159	plasma membrane	20
GO:0004674	molecular_function	494	GO:0004164	signal transducer activity	23
GO:0008271	development	354	GO:008536	behavior	23
GO:0006572	cellular component	369	GO:0005819	nucleoplasm	16
GO:0018040	cell organization & biogenesis	288	GO:0005056	receptor binding	16
GO:0004628	cell	389	GO:00016487	viral life cycle	16
GO:0008159	cell communication	330	GO:006156	protein kinase activity DNA binding	15
GO:0008164	signal transduction	297	GO:0004822	regulation of gene expression & epigenetic	15
GO:0017536	protein metabolism	234	GO:0007952	response to biotic stimulus	15
GO:0007819	transport biosynthesis	319	GO:0006651	RNA binding	18
GO:0008056	binding	293	GO0007463	mitochondrion	18
GO:0003487	cell differentiation	267	GO:00030997	enzyme regulator activity	14
GO:0040156	catalytic activity	204	GO-0008738	peptidase activity	12
GO:0004822	response to stress	299	GO:0005513	cell growth	12
GO:0006952	morphogenesis	273	CO:0008620	nucleotide binding	12
GO:0009551	protein modification	256	GO:0008642	endoplasmic reticulum	11
GO0008463	organelle organization & biogenesis	208	GO-0008813	endosome	10
GO:0005997	cytoplasm	244	GO0008463	nuclear chromosome	9
GO-0006738	protein binding	237	GO:0006997	Golgi apparatus	9
GO:0004813	lipid metabolism	97	GO-0007738	vacuole transcription factor activity	8
GO:0005620	cell cycle	83	GO:0005513	lysosome	9
GO:0008042	ion transport	98	GO:0005620	nuclease activity	10
GO-0007813	response to endogenous stimulus	61	GO:0005042	structural molecule activity	7
GO:0008719	response to external stimulus	77	GO-0005813	nucleolus	6
GO:0009604	hydrolase activity	67	GO:0019997	secondary metabolism	5
GO:0018780	catabolism	69			
GO:0006057	DNA metabolism	75			
GO:0005256	transferase activity	70			
GO.0015741	reproduction	69			

To use a different experimental dataset, lncRNA2GO-55, accessible for downloadable in NeuraNetL2GO, we compare our estimate to two state-of-the-art algorithms, LncRNA2Function and NeuraNetL2GO. The result is evaluated the using F-score, accuracy, and recall measures, and our approach is shown to be the greatest. The benefits of performance assessment are shown in Fig.2. In terms of availability, the proposed methodology has also demonstrated satisfying performance (see Table 2).

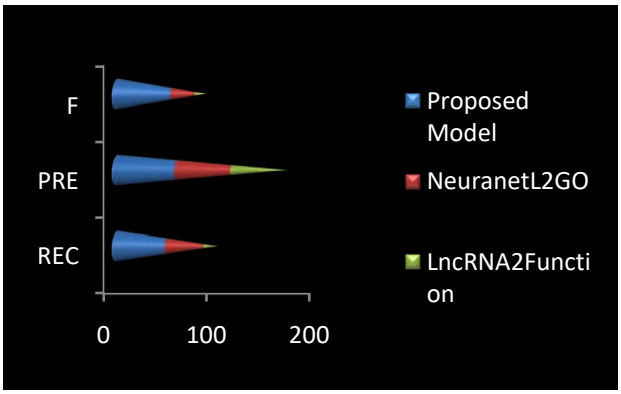

Fig.2. Performance measures

TABLE 2. Comparison of proposed and existing systems

Model	Annotated	Coverage percentage
Proposed Model	52	88.03
NeuranetL2GO	47	94.10
LncRNA2Function	15	29.55

A consideration of the perceived significance of the meta-paths as RF classifier characteristics is offered here. Statistical assurance is also provided that paths proposed by the appropriate path selection procedure are significant. The final section demonstrates the validity by using four as the maximum criterion for the conceptual duration. Furthermore, the appropriate meta-paths chosen are relevant in terms of biological scraps of evidence. The meaning of the unnecessary abstract, lplf, is that one lncRNA connects with an enzyme, which binds with another lncRNA to execute a purpose. According to existing research, enzymes perform operations immediately, while lncRNAs are used relatively infrequently to generate functionalities. Therefore, the conceptual interpretations of all essential meta paths identified either by automated selection procedure or on either extreme are closely related to scientifically confirming lncRNA-function pathways.

This designer's large variety of differences is O, where n is the highest amount of lncRNAs/proteins/functions in the investigation, m or g are the amount and duration of meta-paths, and one or fare the number lncRNAs and activities in the study, correspondingly. It's worth noting that the amount and duration of concepts used in the research significantly impact the total intricacy. Furthermore, these are the variables that we can control by accepting or dismissing conceptual based on predetermined duration criteria.

According to a ROC curve analysis, the gain in outcome for more extensive conceptual was inappropriate to the processing effort required to analyze a more significant number of abstracts of more extensive durations. Fig.3 illustrates the ROC curves for tests with three, four, or five distances. The meta-path experimental with duration four

DOI: rp-9788770040723.073

performs much superior to one with three. The duration of five conceptual research, on the other hand, does not enhance the outcome to the extent that the increased processing complexity necessary to analyze individual concepts could be accounted for compensated. Furthermore, setting the upper boundary of conceptual duration to four for the particular data set seems to be the best option.

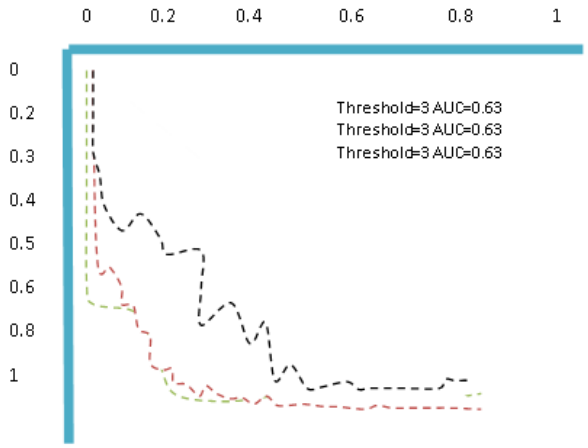

Fig.3.ROC curve for proposed system

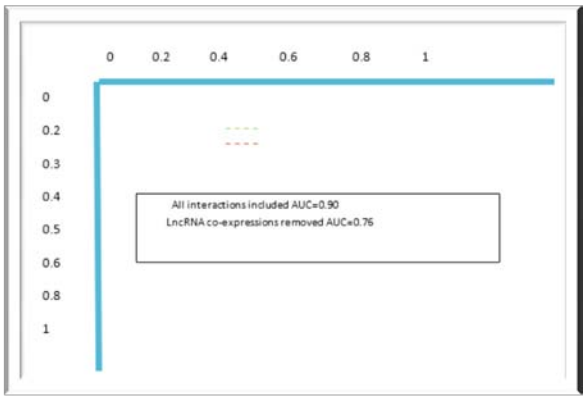

Fig.4.Impact of lncRNA co-expression information

The inclusion of lncRNA co-expression data into the network is among the innovative features of this article. This knowledge aids in the identification of lncRNAs that are structurally comparable. The relationship among lncRNAs was assessed using the co-expression patterns of lncRNAs in 24 organs. Experimental evidence indicates that lncRNAs have tissue-specific production, but that their position has a significant impact on their activity. As a consequence, tissue-specific co-expression features of lncRNAs could help determine important biological lncRNAs.In light of this, it is reasonable to assume that incorporating lncRNA co-expression information has aided in improving predictive performance. An ROC curve examination of the factor influencing outcome in a network with and without lncRNA co-expression data supports this hypothesis. Fig.4 depicts overall results of the analysis. It reveals that when the lncRNA co-expression system was removed from HIN, the AUC dropped consistently.

VI. Conclusions

The current research topic of fast and effective protein identification of lncRNAs was inspired by the mounting evidence for operational characters played by lncRNAs in biological and cellular processes. Because the wet-lab technique for operationally annotating lncRNAs is costly and complex, algorithmic approaches have become very interesting lately. The research presented here forecasts the roles of lncRNAs analysis of protein behavior data and co-expression information. While previous techniques for identifying lncRNAs primarily focus on protein contact, this method considers lncRNA co-expression patterns and correlation to established benefits and associated proteins. More importantly, the technique can effectively examine the importance of lncRNA-function pairs in a HIN by associating functions with lncRNAs even if there is no protein connection. The model has a predictive performance of 74 percent altogether.

References

[1] G. Yu, Y. Wang, J. Wang, C. Domeniconi, M. Guo, and X. Zhang,“Attributed heterogeneous network fusion via collaborative matrix tri-factorization,” Information Fusion, vol. 63, pp. 153-65, 2020.

[2] G. Yu, K. Wang, C. Domeniconi, M. Guo, and J. Wang,“Isoform function prediction based on bi-random walks on a heterogeneous network,” Bioinformatics, vol. 36, no. 1, pp. 303-10, 2020.

[3] Y.K. Zhou, Z.A. Shen, H. Yu, T. Luo, Y. Gao, and P.F. Du,“Predicting lncRNA–protein interactions with miRNAs as mediators in a heterogeneous network model,” Frontiers in genetics, vol. 10, p.1341, 2020.

[4] T.P. Latchoumi, T.P. Ezhilarasi, andK. Balamurugan, “Bio-inspired weighed quantum particle swarm optimization and smooth support vector machine ensembles for identification of abnormalities in medical data,” SN Applied Sciences, vol. 1, no. 10, pp. 1-10, 2019.

[5] Y. Fan, J. Cui, and Q. Zhu,“Heterogeneous graph inference based on similarity network fusion for predicting lncRNA–miRNA interaction,” Rsc Advances, vol. 10, no. 20, pp. 11634-42, 2020.

[6] X.J. Lei, C. Bian, and Y. Pan,“Predicting CircRNA-disease associations based on improved weighted biased meta-structure,” Journal of Computer Science and Technology, vol. 36, no. 2, pp. 288-98, 2021.

[7] Gomathy, V., Janarthanan, K., Al-Turjman, F., Sitharthan, R., Rajesh, M., Vengatesan, K., &Reshma, T. P. (2021). Investigating the spread of coronavirus disease via edge-AI and air pollution correlation. ACM Transactions on Internet Technology, 21(4), 1-10.

[8] M. Ghanbari, and U. Ohler,“Deep neural networks for interpreting RNA-binding protein target preferences,” Genome research, vol. 30, no. 2, pp. 214-26, 2020.

[9] B. Du, Z. Liu, and F. Luo,“Deep multi-scale attention network for RNA-binding proteins prediction,” Information Sciences, vol. 582, pp. 287-301, 2022.

[10] Rajesh, M., &Sitharthan, R. (2022). Introduction to the special section on cyber-physical system for autonomous process control in industry 5.0. Computers and Electrical Engineering, 104, 108481.

[11] A. Kitaygorodsky, E. Jin, and Y. Shen, “Predicting localized affinity of RNA binding proteins to transcripts with convolutional neural networks,” bioRxiv, 2021.

[12] C.Bhuvaneshwari, and A.Manjunathan, “Reimbursement of sensor nodes and path optimization”, Materials Today: Proceedings, vol. 45, pp.1547-1551, 2021.

[13] R.Dineshkumar, P.Chinniah, S.Jothimani, N.Manikandan, A.Manjunathan, and M.Dhanalakshmi, “Genomics FANET Recruiting Protocol in Crop Yield Areas UAV”, Annals of the Romanian Society for Cell Biology, pp. 1515-1522, 2021.

[14] M. Ramkumar, A. Lakshmi, M.P.Rajasekaran, and A.Manjunathan, “MultiscaleLaplacian graph kernel features combined with tree deep convolutional neural network for the detection of ECG arrhythmia”, Biomedical Signal Processing and Control, vol. 76, p. 103639, 2022.