Received: 14 March 2025 | Revised: 6 June 2025 | Accepted: 29 October 2025 | Published online: 13 November 2025

RESEARCH ARTICLE

Medinformatics
2025, Vol. 00(00) 1–9
DOI: 10.47852/bonviewMEDIN52025688

Identifying lncRNA–Disease Association Based on Dormant Factor Model

Iyappan Ramalakshmi Oviya[1,*], **Guruprasath Manika Rameshbabu**[1], **Shree Prasad Muthukrishnan**[1], **Balu Bhasuran**[2], **Tharun Kaarthik Gunasekaran Kumutha**[1], **Sudeesh Kumar Venkatesh**[1] and **Shalini Deena Dhayalan**[1]

[1]*Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, India*

[2] *College of Communication and Information, Florida State University, USA*

Abstract: Long non-coding RNA (lncRNA), once thought to be a noisy gene, has been discovered to engage in a range of biological processes. The association between lncRNA and diseases can be explored by studying its evolutionary conservation. Recent research has demonstrated the role of lncRNA in a number of human disorders like cancer. Computational approaches for identifying the association between lncRNA and diseases are necessary to establish its role. This research aims at creating lncRNA–disease associations based on dormant factor models using lncRNAs–miRNAs and miRNA–disease association data. Similarity functions were statistically analyzed to build the interaction matrices, followed by dormant feature extraction, where the lncRNA–disease association is ranked, with cancer being at the top. The obtained results showed that the current study had a higher accuracy with an area under the curve value of 0.934 for the computed association matrix. These findings highlight the potential of computational models in uncovering novel lncRNA–disease associations effectively.

Keywords: lncRNA, miRNA, dormant, evolution, association, similarity matrices

1. Introduction

The non-coding portion of the genome, which was ignored once, has been gaining huge interest for its vital roles in cellular regulation and disease pathogenesis. Among the non-coding genes, long non-coding RNAs (lncRNAs) seem to be highly noticeable. LncRNAs are a large and diverse class of RNA transcripts, which are longer than 200 nucleotides in length and are not translated into proteins [1]. They were mentioned as transcriptional noise or genomic junk because the main function of RNA is to evolve DNA into proteins, whereas the RNAs that are not involved in protein synthesis were considered nonfunctional. The later studies in the genome provide evidence that lncRNA plays a decisive role in modulating gene expression, chromatin structure, and diverse cellular processes [2, 3]. The dysregulation of lncRNAs has been widely noted in several human diseases such as cancer [4], cardiovascular disorders, neurodegenerative diseases, and immune disorders.

The interest in lncRNAs and their disease associations started in the early genomic studies revealed that protein-coding genes constitute only a small fraction of the genome while the remaining large portion is transcribed into non-coding RNAs. As the interest in understanding the roles and functions of non-coding.

As RNAs increased, the research on profiling the transcriptomes of various cell types and tissues using the high-throughput RNA sequencing technology [5, 6] showed extensive lncRNA transcription across the genome, proving their significance as biologically meaningful entities rather than transcriptional noise. As the regulatory functions of lncRNAs were recognized, it led to the discovery of their involvement in diseases. Meanwhile, microRNAs (miRNAs) are small non-coding RNAs that regulate gene expression post-transcriptionally, leading to mRNA degradation. These RNA molecules play critical roles in diverse biological processes, such as development, differentiation, apoptosis, and metabolism, but are also associated with various diseases [7, 8].

The discovery of these associations would help identify the underlying reason for the disease's development and progression by taking into account its molecular mechanisms. Understanding the contribution of dysregulated lncRNAs to diseases can create novel pathways to targeted treatments. For instance, modulating the lncRNA expression and restoring cellular homeostasis with the help of lncRNA inhibition using antisense oligonucleotides, small molecule inhibitors, or gene editing technology offer a possibility for therapeutic targets for patients. The disease-associated lncRNAs can serve as diagnostic biomarkers for the detection of early disease, for prognosis assessment to estimate the risk, and for monitoring the treatment response [9]. Integrating disease

[*] **Corresponding author:** I R Oviya, Department of Computer Science and Engineering, Amrita School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India. Email: ir_oviya@ch.amrita.edu.

associations would enable personalized treatments based on the patient's genomic profiles and help optimize the outcomes.

After understanding the functional aspect of lncRNAs in terms of their diversity and functions, it is apparent that their regulatory action includes essential processes in cells and therefore impacts physiology in healthy situations and pathogenesis of numerous diseases. As explained in [10], the realization of the fact that protein-coding genes constitute only a small fraction of the genome has fueled growing knowledge on the roles played by the ubiquitous non-coding RNAs, especially lncRNAs. The advent of high-throughput RNA sequencing technology has been key to revealing widespread transcription of lncRNAs across the genome, solidifying their status as biologically significant molecules and not background transcriptional noise [9]. It is this technology that underpins the data-driven methodology used in this proposed strategy since it has the backing of large-scale lncRNA, miRNA, and disease association datasets.

In addition, the therapeutic target and biomarker value of lncRNAs in diagnostics is the prime focus driving research in the field [10–12]. Dysregulation of lncRNAs in disease, as highlighted in most studies, holds promise for the early detection of disease, prognosis, and monitoring of the response to treatment [10]. The computational method outlined here will attempt to aid this cause by accurately identifying disease-related lncRNAs, with a special focus on cancer. The inclusion of lncRNA–miRNA and miRNA–disease relationships in their model exposes the complexities of the regulatory networks the lncRNAs are a part of, as noted in [11, 13, 14]. The experimental strategy employed in [13], such as lncRNA knockdown and pull-down assays, is equally vital in validating the computationally inferred interactions and the distinct molecular pathways through which the lncRNAs act in disease contexts. The combination of robust computational models with swift experimental verification is critical in an attempt to convert the shared body of knowledge regarding lncRNA biology into clinical benefits.

Recently, various computational models have been structured for predicting the associations between lncRNAs and diseases [15]. In this research, this study aims to construct a predictive model that can accurately find the disease-associated lncRNAs. Understanding the interactions between lncRNAs, miRNAs, and disease would lead to the development of biomedical research. Here, we address the challenge of identifying lncRNA–disease associations by developing a computational framework based on a novel dormant factor model (DFM). Building on previous research that highlighted the regulatory significance of lncRNAs and their associations with diseases, we integrate lncRNA–miRNA and miRNA–disease interaction data to construct a predictive model.

2. Research Methodology

2.1. Dataset and preprocessing

For the mentioned methodology, this research requires a dataset that correlates lncRNA, miRNA, and disease, and the dataset required is created by combining three datasets. The first dataset used in this research gives the lncRNA–disease interaction, which is obtained from LncRNADisease v3.0 [16]. This supports lncRNA–disease associations, with a total of 25,440 entries, including 6,066 lncRNAs, 484 diseases, and 13,191 interactions between lncRNAs and diseases. The second dataset provides an interaction between lncRNA and miRNA, obtained from lncRNASNP v3 [17]. This dataset provides a comprehensive repository of single nucleotide polymorphisms and somatic mutations in lncRNAs and their impacts on lncRNA structure and function. This is also supported by lncRNA–miRNA interaction, which has identified 45,774,338 lncRNA–miRNA pairs. Finally, the RNADisease v4.0 [18] dataset is selected to provide this research with RNA–disease interaction. This only takes miRNA-type RNA from this dataset, and it provides a comprehensive and concise data resource of RNA–disease associations, which contains a total of 3,428,58 RNA–disease entries covering 18 RNA types, 117 species, and 4,090 diseases.

Combining the above datasets in the same manner, such as with the inner join, a final dataset containing lncRNA, miRNA, and disease as features will be obtained [19]. First, the lncRNA–miRNA and miRNA–disease interactions are combined by taking common miRNA in both datasets, which results in a dataset with three features (lncRNA, miRNA, disease), which contains 9,376,657 records with 7554 lncRNAs, 239 miRNAs, and 514 diseases. The set of lncRNA can be denoted as $L = [l_1, l_2, \ldots, l_\mathrm{n}]$ with length l, then the set of miRNA $M = [m_1, m_2, \ldots, m_\mathrm{n}]$ with the size of m, and finally, the set of disease $D = [d_1, d_2, \ldots, d_\mathrm{n}]$ with the size of d.

2.2. Construct Features Integrated Network

Features Integrated Network is a graph that interconnects the features and represents them in sets of nodes and relationships between them. Mathematically, it is represented as $G = (N, E)$, where $N \in$ is the set of features and $E \in$ is the set of relationships between features. In our case, $N \in (L \cup M \cup D)$ and $E \in$ represents the interaction between features.

2.3. Construct feature relation matrix

This step creates matrices that represent the relation between any two features from G. Let us begin by creating three separate graphs (LM_g, MD_g, LD_g) from G; that is, the first graph has the relation between lncRNA and miRNA, followed by miRNA and disease, and finally lncRNA and disease. From the individual graph, let's create the matrices LM, MD, and LD of size $1 \times m$, $m \times d$, and $1 \times d$. Then each element of the matrix is defined as

$$LM[i][j] = \begin{cases} 1 \text{ if } \operatorname{edge}(\mathrm{L}_i, M_j) \in G \\ 0 \qquad \mathit{otherwise} \end{cases}$$

$$MD[i][j] = \begin{cases} 1 \text{ if } \operatorname{edge}(\mathrm{M}_i, D_j) \in G \\ 0 \qquad \mathit{otherwise} \end{cases}$$

$$LD[i][j] = \begin{cases} 1 \text{ if } \operatorname{edge}(L_i, D_j) \in G \\ 0 \qquad \mathit{otherwise} \end{cases}$$

2.4. Building interaction within the feature

Functional similarity within the features is constructed. Functional similarity represents how a feature relates to another feature by finding similarities by combining Cosine similarity and Jaccard similarity. Cosine similarity is a function that measures the similarity between two vectors based on their direction [19–21].

$$\mathit{CosineSimilarity}(A, B) = CS(A, B) = \frac{A.B}{|A||B|} \tag{1}$$

Here, $A = < a_1, a_2, \ldots, a_\mathrm{n} >$ and $B = < b_1, b_2, \ldots, b_\mathrm{n} >$ are the vectors, and A, $B \in R$. The Jaccard similarity provides similarities between two sets based on the intersection and union (15).

$$JaccardSimilarity(A, B) = JS(A, B) = \frac{A \cap B}{A \cup B} \tag{2}$$

Here, $A = (a_1, a_2, \ldots, a_n)$ and $B = (b_1, b_2, \ldots, b_n)$ are sets, and $A, B \in R$.

This research utilizes two similarity measures to overcome the disadvantages of one another; here, Cosine similarity can lead to higher sparsity in the result. So, to overcome this issue, this research uses Jaccard similarity to reduce the sparsity. Finally, the Integrated Similarity Function (ISF) would be

$$ISF(A, B) = \begin{cases} CS(A, B) & \text{if} CS(A, B) = 1 \\ \frac{CS(A, B) + JS(A + B)}{2} & \text{if} CS(A, B) = 0 \end{cases}$$

2.4.1. Interaction within lncRNA

LM is used for obtaining the interaction within lncRNAs. The row in the matrix represents the association between lncRNA and all miRNAs in the dataset. Every row in the matrix acts as a vector that represents the lncRNA. The matrix ISF_l has size $l \times l$, where $ISF_l[i][j]$ represents the interaction between lncRNAs $L[i]$ and $L[j]$.

$$ISF_l[i][j] = \sum_{i=0}^{l} \sum_{j=0}^{l} ISF(L[i], L[j]) \tag{3}$$

2.4.2. Interaction within disease

MD is used for obtaining the interaction within lncRNAs. The column in the matrix represents the association between the disease and all miRNAs in the dataset. Every column in the matrix acts as the vector that represents the lncRNA matrix ISF_d of size $(d \times d)$, where $ISF_d[i][j]$ represents the interaction between diseases $D[i]$ and $D[j]$.

$$ISF_d[i][j] = \sum_{i=0}^{d} \sum_{j=0}^{d} ISF(D[i], D[j]) \tag{4}$$

2.5. Dormant feature extraction from preliminary lncRNA–disease association

This research uses matrix decomposition to extract dormant features from preliminary lncRNA–disease, where ψ_p is the matrix that is the initial lncRNA–disease, and this is defined as

$$\psi_p = LM \times MD \tag{5}$$

By utilizing the ψ_p, two dormant feature matrices X and Y with sizes of $l \times k$ and $d \times k$, respectively, are created, where k is the number of dormant features. And by multiplying these two matrices, the final dormant matrix ψ with rows related to lncRNA and columns to diseases is obtained.

The steps to create the final dormant matrix ψ are as follows:

Step 1: Create 2 matrices $X_{l\times k}$ and $Y_{d\times k}$ randomly

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots \\ x_{2,1} & x_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}_{l\times k} \quad Y = \begin{bmatrix} y_{1,1} & y_{1,2} & \cdots \\ y_{2,1} & y_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix}_{d\times k}$$

Step 2: Initialize ψ for the proceeding steps

$$\psi = X \times Y^T \tag{6}$$

Step 3: Optimize the psi and loss function LS, using gradient descent on X and Y, defined as

$$LS(X, Y) = \sum_{(i,j)\in k} (\psi - \psi_p)^2 + \frac{\lambda}{2} \sum_i ||X_i|| + \frac{\lambda}{2} \sum_i ||Y_j||^2 \tag{7}$$

where $\|X_i\|$ and $\|Y_j\|$ are the norms of X_i and Y_j, and λ is the regularization rate,

$$\begin{aligned} \frac{\partial LS}{\partial X_j} &= \frac{\sum_{(i,j)\in k} (\psi - \psi_p)^2 + \lambda \sum_j ||X_j||}{\partial X_j} \\ &= \sum_j 2(\psi_p - \psi) Y_j + \lambda X_i \end{aligned} \tag{8}$$

$$\begin{aligned} \frac{\partial LS}{\partial X_j} &= \frac{\sum_{(i,j)\in k} (\psi - \psi_p)^2 + \lambda \sum_j ||X_j||}{\partial X_j} \\ &= \sum_i 2(\psi_p - \psi) X_i + \lambda Y_j \end{aligned} \tag{9}$$

Update X_i and Y_j using gradient descent recursively

$$X_i = X_i - \alpha \frac{\partial LS}{\partial X_i}$$

$$X_i = X_i - \alpha \sum_j 2(\psi_p - \psi) Y_j + \lambda X_i \tag{10}$$

$$Y_j = Y_j - \alpha \frac{\partial LS}{\partial Y_i}$$

$$Y_i = Y_i - \alpha \sum_i 2(\psi_p - \psi) X_i + \lambda Y_j \tag{11}$$

In this research, $\alpha = 2 \times 10^{-6}$ and $\lambda = 4 \times 10^{-5}$ and iterated 213 times.

2.6. Dormant feature extraction from preliminary lncRNA–disease association

In this step, interaction with the feature and dormant feature is combined to provide the lncRNA–disease association. Initially, the dormant feature matrix (ψ) with the feature interaction matrices (IFS_l, IFS_d) is combined to get Dormant Feature Projection Matrices (DFP_l, DFP_d), and finally, Dormant Feature Projection and ψ are combined to get Dormant Feature Association Matrix (DFM)

$$DFD_l[i] = \frac{\sum_{j=0}^{l} (ISF_l[i][j] \times \psi[j])}{||ISF_l[i]||} \tag{12}$$

Here, $DFP_l[i]$ is the row vector of the DFP_l matrix, and $\psi[j]$ is the row vector of the ψ matrix

$$DFP_d[j] = \frac{\sum_{i=0}^{d} (ISF_d[i][j] \times \psi[i])}{||ISF_d[j]||} \tag{13}$$

Here, $DFP_d[j]$ is the column vector of the DFP_d matrix, and $\psi[i]$ is the column vector of the ψ matrix. Finally, combine the Dormant Feature Projection matrices to get the DFM. For integrating the two matrices in the final step, the parameter ω was adjusted to an optimal value ($\omega = 0.5$). $DFM_l[i]$ represents the row vectors of the DFM_l, and $DFM_d[j]$ represents the column vectors of DFM_d.

$$DFM[i][j] = \frac{\omega DFP_l[i][j] + (1-\omega)DFP_d[i][j]}{||DFP_l[i]|| + ||DFP_d[j]||} \tag{14}$$

After concluding the maximum frequency range to be 3, by manual inspection of the range, values below 3 are adjusted by subtracting a small fraction, while values above 3 are adjusted by adding a smaller fraction. This ensures that values are shifted within a controlled range, avoiding extreme outliers. Finally, the values are scaled to the range of 0 to 1, and by this normalization process, the algorithmic values in Table 1 are standardized, and a balanced distribution is maintained. Characterization of lncRNAs is important to understand their functionality in various diseases. However, limited knowledge of their contribution makes us explore their disease association statistically in the present work. This association highlights their evolutionary importance to be explored further experimentally.

This research utilizes various evaluation metrics to assess model performance. The metrics include precision, which determines the accuracy of positive predictions by measuring the proportion of true positives (TP) among all positively predicted instances.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{15}$$

Recall evaluates the model's sensitivity, indicating how effectively it identifies all positive cases within the dataset.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{16}$$

Accuracy serves as a metric, representing the percentage of all correctly predicted instances, both positive and negative, across the entire dataset.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{17}$$

The F1-score combines precision and recall into a single measure to provide a balanced assessment,

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{18}$$

especially valuable in cases of class imbalance. Additionally, ROC-AUC (receiver operating characteristic–area under the curve) is utilized to assess the model's ability to differentiate between positive and negative classes across various threshold levels.

The PR-AUC (precision-recall area under the curve) is also applied, particularly beneficial in imbalanced data scenarios, to evaluate the precision-recall trade-off.

3. Results

In the present study, the association matrix is built after collecting and integrating three large-scale public datasets to construct our lncRNA–miRNA–disease framework. From LncRNADisease v3.0, we obtained 25,440 entries with 13,191 validated lncRNA–disease interactions encompassing 6,066 lncRNAs and 484 diseases. The lncRNASNP v3 resource contributed 45,774,338 lncRNA–miRNA pairs across 7,554 lncRNAs and 239 miRNAs, while RNADisease v4.0 (filtered to miRNAs) provided 3,428,058 RNA–disease entries covering 18 RNA types, 117 species, and 4,090 diseases. An inner join on the common miRNA identifiers yielded a final dataset of 9,376,657 records linking 7,554 lncRNAs, 239 miRNAs, and 514 diseases.

The lncRNA–disease database was processed into 3,882,756 different associations and 514 unique diseases. Figure 3 shows the graph representing the top 1,000 associations of lncRNA disease where each of the unique diseases is and lncRNAs are the nodes in the graph, and the edges represent the association between them. The top-ranked pair was lncRNA NONHSAT007829.2 with cancer (raw score 329.6302, normalized to 1.000), followed by

Table 1. Top associations between lncRNA and diseases

Ranking	lncRNA	Disease	Algorithmic Value	Normalized Values
1	NONHSAT007829.2	Cancer	329.63024	1.0
2	NONHSAT013732.2	Cancer	199.861441	0.961026
3	NONHSAT007829.2	Vascular diseases	147.514495	0.945304
4	NONHSAT017055.2	Cancer	134.41244	0.941369
5	NONHSAT000708.2	Cancer	133.70011	0.941155
6	NONHSAT007829.2	Neurodegenerative	128.761035	0.939672
7	NONHSAT007829.2	Carcinoma	126.0211	0.938849
8	NONHSAT017054.2	Cancer	121.956649	0.937628
9	NONHSAT007829.2	Brain disease	116.639702	0.936031
10	NONHSAT015532.2	Cancer	116.610551	0.936023
11	NONHSAT007829.2	Leukemia	106.862075	0.933095
12	NONHSAT008215.2	Cancer	106.452600	0.93297
13	NONHSAT004149.2	Cancer	103.154762	0.931981
14	NONHSAT005620.2	Cancer	99.311731	0.930827
15	NONHSAT012854.2	Cancer	97.947774	0.930417
16	NONHSAT007829.2	Diabetes mellitus	97.70052	0.930343
17	NONHSAT011344.2	Cancer	95.503554	0.929683
18	NONHSAT016265.2	Cancer	91.808802	0.92857
19	NONHSAT013732.2	Vascular diseases	90.309014	0.928123
20	NONHSAT002404.2	Cancer	89.287225	0.927816

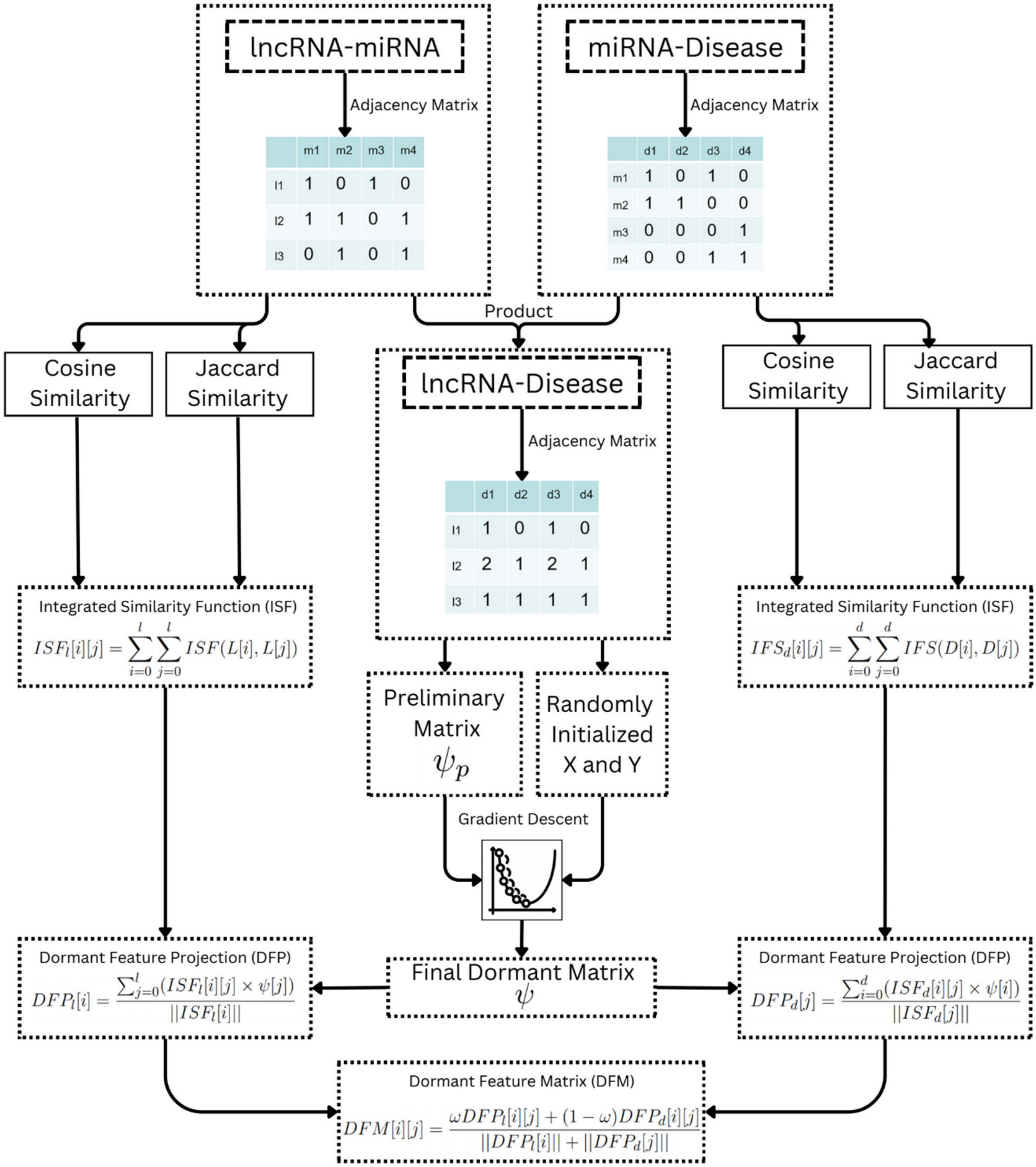
lncRNA-miRNA
Adjacency Matrix
m1 m2 m3 m4
l1 1 0 1 0
l2 1 1 0 1
l3 0 1 0 1
miRNA-Disease
Adjacency Matrix
d1 d2 d3 d4
m1 1 0 1 0
m2 1 1 0 0
m3 0 0 0 1
m4 0 0 1 1
Product
Cosine Similarity
Jaccard Similarity
lncRNA-Disease
Adjacency Matrix
d1 d2 d3 d4
l1 1 0 1 0
l2 2 1 2 1
l3 1 1 1 1
Cosine Similarity
Jaccard Similarity
Integrated Similarity Function (ISF)
$ISF_l[i][j] = \sum_{i=0}^{l}\sum_{j=0}^{l} ISF(L[i], L[j])$
Integrated Similarity Function (ISF)
$IFS_d[i][j] = \sum_{i=0}^{d}\sum_{j=0}^{d} IFS(D[i], D[j])$
Preliminary Matrix ψ_p
Randomly Initialized X and Y
Gradient Descent
Dormant Feature Projection (DFP)
$DFP_l[i] = \frac{\sum_{j=0}^{l}(ISF_l[i][j] \times \psi[j])}{||ISF_l[i]||}$
Final Dormant Matrix ψ
Dormant Feature Projection (DFP)
$DFP_d[j] = \frac{\sum_{i=0}^{d}(ISF_d[i][j] \times \psi[i])}{||ISF_d[j]||}$
Dormant Feature Matrix (DFM)
$DFM[i][j] = \frac{\omega DFP_l[i][j] + (1-\omega)DFP_d[i][j]}{||DFP_l[i]|| + ||DFP_d[j]||}$

Figure 1. Dormant factor model flow chart

NONHSAT013732.2, which was also associated with cancer and was observed with most lncRNAs, showing 383 associations. Other diseases such as vascular diseases, neurodegenerative diseases, and brain diseases also have a notable number of associations. Since computational models have the ability to identify the most promising lncRNAs associated with human diseases, we have focused on predicting the possible correlations between lncRNAs and diseases. This research establishes a computational model by examining the known lncRNA–disease correlations and applying the dormant feature extraction technique along with gradient descent to obtain the association score of every lncRNA–disease pair. The association scores are ranked in Table 1 and give us insight to experimentally validate the lncRNA for the disease.

From these data, three bipartite interaction matrices were constructed: an LM matrix (7,554 × 239) for lncRNA–miRNA links, an MD matrix (239 × 514) for miRNA–disease associations, and an initially zeroed LD matrix (7,554 × 514) for direct lncRNA–disease pairs. Multiplying LM by MD produced the preliminary association matrix ψ_p, quantifying indirect lncRNA–disease connectivity via

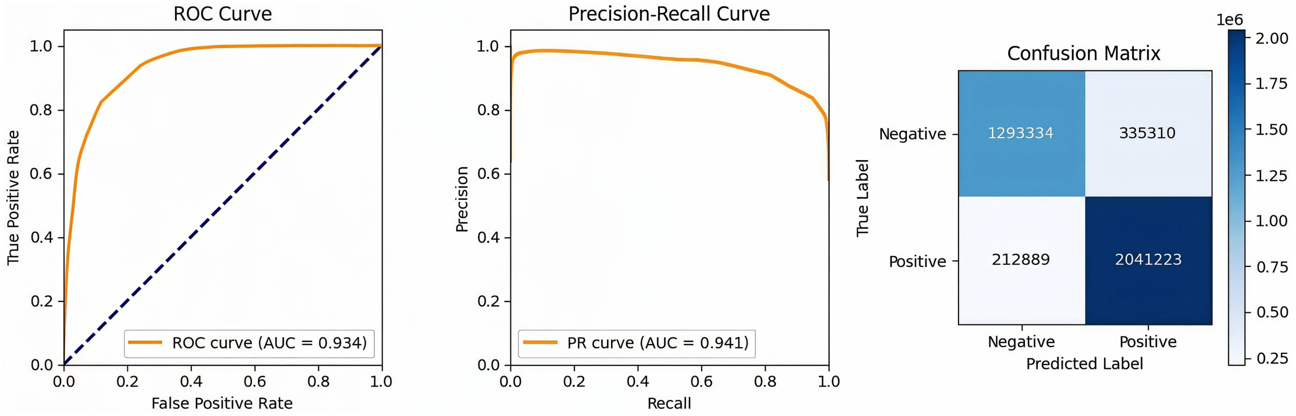

Figure 2. Evaluation of current DFM algorithm

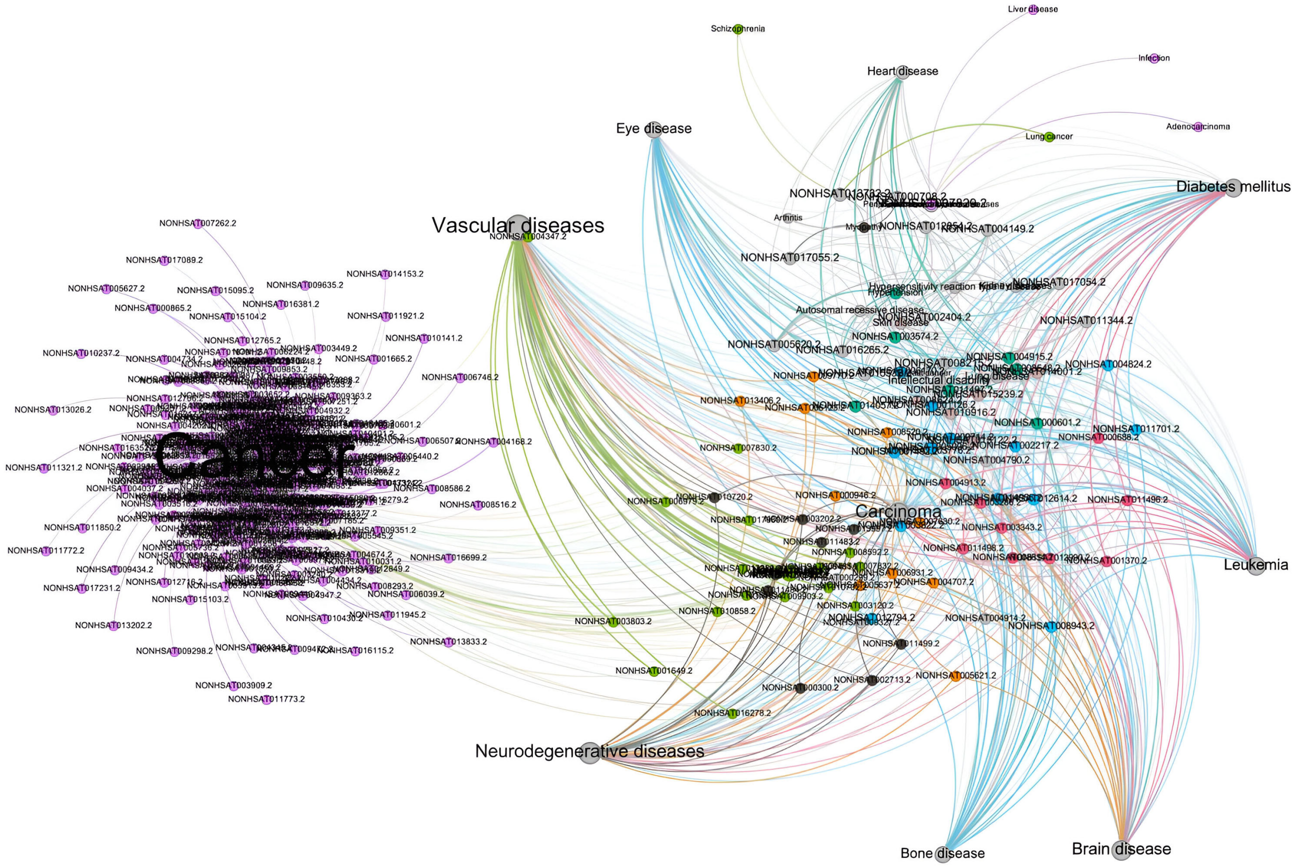

Figure 3. Top 1,000 lncRNA–disease associations represented in a graph

shared miRNAs. To extract latent association patterns, we initialized two feature matrices X (7,554 × k) and Y (514 × k) at random and then optimized them by minimizing a squared-error loss with ℓ_2 regularization ($\lambda = 4 \times 10^{-5}$) via alternating gradient descent (learning rate $\alpha = 2 \times 10^{-6}$ over 213 iterations). The product $X \times Y^{\mathrm{T}}$ formed the dormant feature matrix ψ, which we combined—using a weighting factor $\omega = 0.5$—with projections based on lncRNA and disease similarity to produce the final dormant feature model (DFM) association matrix.

Performance evaluation on held-out known associations yielded robust metrics: ROC-AUC of 0.934, PR-AUC of 0.941, accuracy of 0.859, precision of 0.859, recall of 0.906, and an F1-score of 0.882. These results demonstrate the model's strong ability to distinguish true lncRNA–disease links from spurious ones in a highly sparse search space.

4. Discussions

The DFM demonstrates strong predictive power for identifying lncRNA–disease associations within a complex, heterogeneous network. Its ROC-AUC of 0.934 and PR-AUC of 0.941 indicate that the model both discriminates true positive links with high confidence and maintains precision even as recall increases—an essential feature when only a small fraction of all possible lncRNA–disease pairs are experimentally validated. The balanced precision (0.859) and recall (0.906) yield an F1-score of 0.882, underscoring the model's ability to prioritize likely associations without incurring excessive false positives. And other works [24, 25] for the same association computation are carried out, in which the principal component analysis is used to construct primary feature vectors from the Gaussian interaction profile kernel of lncRNAs and functional similarity of illnesses, respectively. The interaction profile of a new lncRNA is determined by comparing it with the interaction profiles of its neighbors. Finally, it uses the principal features from the created feature matrices to finish the association matrix based on the inductive matrix completion framework. Also, another lncRNA–disease association model (ENCFLDA), proposed by Bo Wang et al. [26], incorporates collaborative filtering, matrix decomposition, and an elastic network. This approach predicts the relationship between an unknown lncRNA and disease using the available lncRNA–miRNA association data and miRNA–disease association data. The matrix is updated using matrix decomposition in conjunction with the elastic network, and the final prediction matrix is then obtained through collaborative filtering. The lesser-known disease and missing values issue in the current study's dataset was handled with a similar process to ENCFLDA, which was helpful in increasing the dormant features for the study. After incorporating more techniques, DFM outperforms ENCFLDA by 0.02.

To find lncRNA–miRNA and lncRNA–disease interactions, the matrices had to be combined. Xiao-xin Du et al. [25] proposed an effective method in DMWNN, which inspired the use of a technique based on the closest neighbor idea. In the model that merged possible associations, a novel similarity technique was employed. The suggested algorithm's superiority over the existing Cosine, Pearson, and Jaccard similarity algorithms was demonstrated by experimental validation. Lastly, using a fresh dataset to execute predictions, the suggested model's dependability was verified, producing an AUC of 0.92. This was DMWNN's highest score, while the current study's AUC overcame it significantly.

The resulting DFM represented the final lncRNA–disease association matrix, and the model's performance was evaluated using several key metrics. Notably, the study reported a high ROC-AUC of 0.934, PR-AUC of 0.941, accuracy of 0.859, precision of 0.859, recall of 0.906, and F1-score of 0.882.

The excellent results above demonstrate that the proposed methodology could accurately capture the complex interaction between lncRNAs, miRNAs, and diseases. The high values for ROC-AUC and PR- AUC indicate that the model secured an appropriate trade-off for the true positive rate and false positive rate and precision and recall on computing associations between lncRNAs and diseases for the given dataset.

The high accuracy, precision, recall, and F1-score imply that it is able to generate any association matrix for a given dataset. Accuracy is how well the model can make correct associations, while precision and recall both try to measure the performance of true positive associations. A good way to combine these two measures into just one value is that of the F1-score, which indicates the DFM's overall association-computing power.

The progress of lncRNA research is marked by excellence and methodology applications, both experimental and computational. The paper demonstrates the evolution of the maturity of computational methods to lncRNA–disease association prediction from outside correlative analysis to the incorporation of network-based quantifications and latent factor extraction. Its performance also relies on the quality and integrality of the underlying data, for example, lncRNA expression profiles, miRNA–target interactions, and disease–gene associations, typically acquired from big-scale efforts such as TCGA [8, 14].

However, the change is from prediction alone to more biological insight, and that calls for scientists to obtain solid experimental evidence. Scientists in [11] have come up with protocols that establish a minimal toolkit for probing the expression, localization, and molecular interactions of predicted lncRNAs by computational predictions. As an illustration, if the DFM strongly predicts a correlation between an lncRNA and cancer, scientists can utilize lncRNA in situ hybridization to quantify its level of expression in cancer tissue and utilize lncRNA immunoprecipitation to determine its protein interaction partners, which may be implicated in cancer-related pathways.

The computationally predicted functional activities can be independently confirmed by overexpression or knockdown of lncRNAs [10, 11]. By monitoring the response of the treatment to the characteristics of the cancer cells, for example, the proliferation, migration, or drug resistance, absolute knowledge of the function of the lncRNA in the disease can be gained. Integration with prediction and functional assays is required to avoid being the victim of guilt by association and to identify lncRNAs that act as active oncogenic drivers or cancer suppressors and not bystanders [10, 13].

Besides, resources such as lncRNAfunc [14] are required as part of the process of placing the predictions presented in this study into perspective because they give comprehensive information regarding experimentally validated lncRNA interactions, functional regulation, and their relationship with disease. A complementarity between lncRNA–disease interaction predictions and the information given by lncRNAfunc would enable scientists to maximize the integrity of their predictions and plan better experimental strategies. Ongoing innovation and adjustment of computational and experimental resources are the pillars of the breakthrough in the development of lncRNA research as it pertains to clinical diagnosis and therapy of cancer.

5. Conclusion

lncRNAs play multifaceted roles in evolution by influencing gene regulation, contributing to species-specific traits, and potentially serving as sources for new protein-coding genes. Their diverse functions and evolutionary dynamics underscore their significance in the complexity of genomic regulation and their role in diseases. In the present study, the DFM is used to identify lncRNA–disease associations with an accuracy of 93%, which, otherwise, is challenging due to research lacunae. Potential disease associations for newly discovered lncRNAs can be further justified with lab experiments. Beyond its strong predictive performance, DFM offers several practical advantages. Its scalable architecture handles millions of candidate associations without excessive computational demands, and its flexible parameterization (latent dimension, learning rate, regularization, and similarity weighting) allows adaptation to new data types, such as tissue-specific expression profiles, epigenomic modifications, or temporal regulatory snapshots. By integrating both latent factors and explicit similarity projections, the model captures both global network structure and localized feature relationships, leading to more nuanced association scoring. Additionally, the current DFM implementation assumes context-agnostic associations. To enhance the accuracy of the computational model, more robust and validated datasets are required for training, and the ranked associations must be experimentally verified. The present method can be made more robust with additional datasets. In summary, the DFM represents a significant advance in the computational prediction of lncRNA–disease associations. By combining robust latent-feature extraction with similarity-informed projections, it delivers high-accuracy predictions at scale, offering a powerful tool for researchers to uncover novel lncRNA functions in human health and disease. As the volume and diversity of non-coding RNA datasets continue to grow, DFM's adaptable framework poises it to remain a valuable resource for driving experimental discovery, biomarker development, and ultimately, the translation of lncRNA biology into clinical diagnostics and therapeutics.

Ethical Statement

The study does not require any kind of ethical approval as it completely relies on the use of computational techniques and models. This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The code and scripts supporting the findings of this study are openly available in the GitHub Repository. Additionally, the datasets used in this study were obtained from the following publicly accessible sources: LncRNADisease 3.0, LncRNASNP 3, and RNA Disease database. All data used are publicly available and were accessed in accordance with the respective database terms of use.

Author Contribution Statement

Iyappan Ramalakshmi Oviya: Conceptualization, Investigation, Resources, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Guruprasath Manika Rameshbabu:** Conceptualization, Methodology, Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing. **Shree Prasad Muthukrishnan:** Conceptualization, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Balu Bhasuran:** Conceptualization, Methodology, Validation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision. **Tharun Kaarthik Gunasekaran Kumutha:** Investigation, Writing – original draft. **Sudeesh Kumar Venkatesh:** Investigation, Writing – original draft. **Shalini Deena Dhayalan:** Investigation, Resources, Writing – original draft.

References

[1] Nemeth, K., Bayraktar, R., Ferracin, M., & Calin, G. A. (2024). Non-coding RNAs in disease: From mechanisms to therapeutics. *Nature Reviews Genetics*, *25*(3), 211–232. https://doi.org/10.1038/s41576-023-00662-1

[2] Wilusz, J. E., Sunwoo, H., & Spector, D. L. (2009). Long noncoding RNAs: Functional surprises from the RNA world. *Genes & Development*, *23*(13), 1494–1504. http://www.genesdev.org/cgi/doi/10.1101/gad.1800909

[3] Chen, X. (2015). Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Scientific Reports*, *5*(1), 13186. https://doi.org/10.1038/srep13186

[4] Fang, Y., & Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics, Proteomics & Bioinformatics*, *14*(1), 42–54. https://doi.org/10.1016/j.gpb.2015.09.006

[5] Camacho, C. V., Choudhari, R., & Gadad, S. S. (2018). Long noncoding RNAs and cancer: An overview. *Steroids*, *133*, 93–95. https://doi.org/10.1016/j.steroids.2017.12.012

[6] Choudhari, R., Sedano, M. J., Harrison, A. L., Subramani, R., Lin, K. Y., Ramos, E. I., . . . , & Gadad, S. S. (2020). Long noncoding RNAs in cancer: From discovery to therapeutic targets. *Advances in Clinical Chemistry*, *95*, 105–147. https://doi.org/10.1016/bs.acc.2019.08.003

[7] D'Agostino, N., Li, W., & Wang, D. (2022). High-throughput transcriptomics. *Scientific Reports*, *12*(1), 20313. https://doi.org/10.1038/s41598-022-23985-1

[8] Taniue, K., & Akimitsu, N. (2021). The functions and unique features of LncRNAs in cancer development and tumorigenesis. *International Journal of Molecular Sciences*, *22*(2), 632. https://doi.org/10.3390/ijms22020632

[9] Wang, W., & Chen, H. (2023). Predicting miRNA-disease associations based on lncRNA–miRNA interactions and graph convolution networks. *Briefings in Bioinformatics*, *24*(1), bbac495. https://doi.org/10.1093/bib/bbac495

[10] Winkler, L., & Dimitrova, N. (2022). A mechanistic view of long noncoding RNAs in cancer. *Wiley Interdisciplinary Reviews: RNA*, *13*(3), e1699. https://doi.org/10.1002/wrna.1699

[11] Feng, Y., Hu, X., Zhang, Y., Zhang, D., Li, C., & Zhang, L. (2014). Methods for the study of long noncoding RNA in cancer cell signaling. In *Cancer Cell Signaling: Methods and Protocols*, 115–143. https://doi.org/10.1007/978-1-4939-0856-1_10

[12] Torres-Bustamante, M. I., Vazquez-Urrutia, J. R., Solorzano-Ibarra, F., & Ortiz-Lazareno, P. C. (2024). The role of miRNAs to detect progression, stratify, and predict relevant clinical outcomes in bladder cancer. *International Journal of Molecular Sciences*, *25*(4), 2178. https://doi.org/10.3390/ijms25042178

[13] Olivero, C. E., & Dimitrova, N. (2020). Identification and characterization of functional long noncoding RNAs in cancer. *The FASEB Journal*, *34*(12), 15630–15646. https://doi.org/10.1096/fj.202001951R

[14] Yang, M., Lu, H., Liu, J., Wu, S., Kim, P., & Zhou, X. (2022). lncRNAfunc: A knowledgebase of lncRNA function in human cancer. *Nucleic Acids Research*, *50*(D1), D1295–D1306. https://doi.org/10.1093/nar/gkab1035

[15] Du, X. X., Liu, Y., Wang, B., & Zhang, J. F. (2022). lncRNA–disease association prediction method based on the nearest neighbor matrix completion model. *Scientific Reports*, *12*(1), 21653. https://doi.org/10.1038/s41598-022-25730-0

[16] Lin, X., Lu, Y., Zhang, C., Cui, Q., Tang, Y. D., Ji, X., & Cui, C. (2024). LncRNADisease v3. 0: An updated database of long non-coding RNA-associated diseases. *Nucleic Acids Research*, *52*(D1), D1365–D1369. https://doi.org/10.1093/nar/gkad828

[17] Yang, Y., Wang, D., Miao, Y. R., Wu, X., Luo, H., Cao, W., . . . , & Gong, J. (2023). lncRNASNP v3: An updated database for functional variants in long non-coding RNAs. *Nucleic Acids Research*, *51*(D1), D192–D198. https://doi.org/10.1093/nar/gkac981

[18] Chen, J., Lin, J., Hu, Y., Ye, M., Yao, L., Wu, L., . . . , & Wang, D. (2023). RNADisease v4. 0: An updated resource of RNA-associated diseases, providing RNA-disease analysis, enrichment and prediction. *Nucleic Acids Research*, *51*(D1), D1397–D1404. https://doi.org/10.1093/nar/gkac814

[19] Wang, B., Liu, R., Zheng, X., Du, X., & Wang, Z. (2022). lncRNA-disease association prediction based on matrix

decomposition of elastic network and collaborative filtering. *Scientific Reports*, *12*(1), 12700. https://doi.org/10.1038/s41598-022-16594-5

[20] Sheng, N., Wang, Y., Huang, L., Gao, L., Cao, Y., Xie, X., & Fu, Y. (2023). Multi-task prediction-based graph contrastive learning for inferring the relationship among lncRNAs, miRNAs and diseases. *Briefings in Bioinformatics*, *24*(5), bbad276. https://doi.org/10.1093/bib/bbad276

[21] Chen, Q., Qiu, J., Lan, W., & Cao, J. (2025). Similarity-guided graph contrastive learning for lncRNA-disease association prediction. *Journal of Molecular Biology*, *437*(6), 168609. https://doi.org/10.1016/j.jmb.2024.168609

[22] Sumathipala, M., Maiorino, E., Weiss, S. T., & Sharma, A. (2019). Network diffusion approach to predict lncRNA disease associations using multi-type biological networks: LION. *Frontiers in Physiology*, *10*, 888. https://doi.org/10.3389/fphys.2019.00888

[23] Chung, N. C., Miasojedow, B., Startek, M., & Gambin, A. (2019). Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data. *BMC Bioinformatics*, *20*(Suppl 15), 644. https://doi.org/10.1186/s12859-019-3118-5

[24] Lu, C., Yang, M., Luo, F., Wu, F. X., Li, M., Pan, Y., . . . , & Wang, J. (2018). Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics*, *34*(19), 3357–3364. https://doi.org/10.1093/bioinformatics/bty327

[25] Guo, Z. H., You, Z. H., Wang, Y. B., Yi, H. C., & Chen, Z. H. (2019). A learning-based method for LncRNA-disease association identification combing similarity information and rotation forest. *IScience*, *19*, 786–795. https://doi.org/10.1016/j.isci.2019.08.030

[26] Yao, D., Zhan, X., Zhan, X., Kwoh, C. K., Li, P., & Wang, J. (2020). A random forest based computational model for predicting novel lncRNA-disease associations. *BMC Bioinformatics*, *21*, 1–18. https://doi.org/10.1186/s12859-020-3458-1

[27] Wang, B., Liu, R., Zheng, X., Du, X., & Wang, Z. (2022). lncRNA-disease association prediction based on matrix decomposition of elastic network and collaborative filtering. *Scientific Reports*, *12*(1), 12700. https://doi.org/10.1038/s41598-022-16594-5.

How to Cite: Oviya, I. R., Rameshbabu, G. M., Muthukrishnan, S. P., Bhasuran, B., Kumutha, T. K.G., Venkatesh, S. K., & Dhayalan, S. D. (2025). Identifying lncRNA–Disease Association Based on Dormant Factor Model. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN52025688