International Journal of
Molecular Sciences

MDPI

Article

Long Noncoding RNA and Protein Interactions: From Experimental Results to Computational Models Based on Network Methods

Hui Zhang [1], Yanchun Liang [1,2], Siyu Han [1], Cheng Peng [1] and Ying Li [1,*]

[1] College of Computer Science and Technology, Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China; huizhang16@mails.jlu.edu.cn (H.Z.); ycliang@jlu.edu.cn (Y.L.); hansy15@mails.jlu.edu.cn (S.H.); chengpengeace@gmail.com (C.P.)

[2] Zhuhai Laboratory of Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, Zhuhai College of Jilin University, Zhuhai 519041, China

* Correspondence: liying@jlu.edu.cn; Tel.: +86-135-0431-9660

Received: 27 January 2019; Accepted: 11 March 2019; Published: 14 March 2019

check for updates

Abstract: Non-coding RNAs with a length of more than 200 nucleotides are long non-coding RNAs (lncRNAs), which have gained tremendous attention in recent decades. Many studies have confirmed that lncRNAs have important influence in post-transcriptional gene regulation; for example, lncRNAs affect the stability and translation of splicing factor proteins. The mutations and malfunctions of lncRNAs are closely related to human disorders. As lncRNAs interact with a variety of proteins, predicting the interaction between lncRNAs and proteins is a significant way to depth exploration functions and enrich annotations of lncRNAs. Experimental approaches for lncRNA–protein interactions are expensive and time-consuming. Computational approaches to predict lncRNA–protein interactions can be grouped into two broad categories. The first category is based on sequence, structural information and physicochemical property. The second category is based on network method through fusing heterogeneous data to construct lncRNA related heterogeneous network. The network-based methods can capture the implicit feature information in the topological structure of related biological heterogeneous networks containing lncRNAs, which is often ignored by sequence-based methods. In this paper, we summarize and discuss the materials, interaction score calculation algorithms, advantages and disadvantages of state-of-the-art algorithms of lncRNA–protein interaction prediction based on network methods to assist researchers in selecting a suitable method for acquiring more dependable results. All the related different network data are also collected and processed in convenience of users, and are available at https://github.com/HAN-Siyu/APINet/.

Keywords: lncRNA–protein interaction prediction; computational model; biological network science; machine learning

1. Introduction

Long non-coding RNAs (lncRNAs) are non-protein-coding transcripts with a length of more than 200 nucleotides, which can regulate gene expression at different levels [1]. LncRNAs were first regarded as transcriptional noise, and later it was found that they can play an important role in cell division, differentiation, metabolism and other physiological processes [2–4]. With the development of biotechnology and the emergence of computational models, there is now a great deal of evidence suggesting that lncRNAs are significant in diverse mechanisms and are involved in almost the whole process of cells from one division to the next [5,6], such as in transcriptional and post-transcriptional regulation, epigenetic regulation, tissue development, the process of genome selective expression in time and space and apotheosis, metabolic processes, cell cycle control and morphological and structural changes in chromosomes [7–14]. More and more reports have indicated that lncRNAs participate energetically in various stages of gene expression, including as signals, decoys, scaffolds, and leaders [15]. Compared with the characteristics of protein coding genes, lncRNAs tend to be less conserved across species and often show low expression level and high tissue specificity, which make the research more challenging and have attracted the attention of scientists and given rise to considerable discussions in recent decades.

Similar to protein-coding genes and microRNAs, lncRNAs have also been found in the regulation of many human complex diseases, including various types of cancer. At present, there are many databases of lncRNA associated with diseases, such as LncRNADisease database [16] and Lnc2Cancer database [17], which can be used to collect many kinds of disease-related lncRNA. The LncRNADisease database contains nearly 2000 lncRNA–disease associations, and Lnc2Cancer database contains 1488 lncRNA–cancer associations. It further confirms that lncRNA is closely related to diseases, even cancer and prognosis regulation. Obviously, the number of annotated lncRNAs involved in these two databases is relatively small compared with the number of identified lncRNAs, and most of the functions of lncRNAs associated with diseases are unclear. It is worth mentioning that lncRNA–protein interaction is a very important mechanism of lncRNAs. To fully understand function or molecular mechanism of lncRNAs, it is necessary to mine interactions between lncRNAs and other molecules, especially lncRNA–protein interactions.

It is of great importance to identify lncRNA–protein interactions to gain a comprehensive and profound understanding of the potential functions encompassed in their molecular mechanisms. At present, the main methods for identifying lncRNA–protein interactions are based on experimental approaches and computational approaches. Several large-scale experimental approaches for lncRNA–protein interaction prediction include RNAcompete [18], RNP immunoprecipitation-microarray (RIP-Chip) [19], high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [20] and photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) [21]. There are also many effective methods for the analysis of experimental data, such as several methods for finding RNA motifs from crosslinking-immunprecipitation and high-throughput sequencing (CLIP-Seq) or other high-throughput experiments, such as BEAM (BEAr Motif finder) [22] and SMARIV (Sequence and Structure Motif enrichment Analysis for Ranked RNA daTa generated from In-Vivo binding experiments) [23]. NPInter database and StarBase database are built on these data, which are based on high-throughput experiments and have a certain degree of false positivity. True interactions may involve the integration of multi-source data, such as the STRING database containing PPI (protein–protein interactions), which integrates information from various sources, including experimental data, co-expression data, text mining, etc.

The methods of predicting lncRNA–protein interactions based on computational approaches are mainly divided into machine-based learning methods and network-based methods. The methods based on machine learning construct a classifier by fusing the features of sequence, structure and physicochemical properties, so as to form an interactive or non-interactive classification model. At present, the existing methods are RPISeq [24], de novo prediction [25], CatRAPID [26], LncPro [5], RPI-Pred [27] and rpiCOOL [28]. Random Forest, Nave Bayesian, Extended Nave Bayesian and SVM are the classifiers used in the above machine-based learning methods. There are also two methods to construct classification model based on deep learning: IPMiner [29] and lncADeep [30]. Current network-based approaches include LPBNI [31], fusing multiple protein–protein similarity networks (PPSNs) proposed by Zheng et al. [32], the method to predict lncRNA–protein interactions based on the relevance search method proposed by Yang et al. [33], LPIHN [34] and PLPIHS [35]. Compared to machine learning-based methods, network-based methods can accommodate more heterogeneous data, not only avoiding ignoring the external links between molecules, but also mining the hidden topological structure information in heterogeneous networks.

Nowadays, network science is being extensively used in biological and related fields. It provides many practical descriptions to characterize various biological systems [36] and the relationships between diseases and biological factors [37]. Network science is becoming more and more popular, and has achieved remarkable results in various fields of bioinformatics. Network science has also made rapid advances in disease gene prioritization [38], disease lncRNA prioritization [39–41], disease-related miRNA identification [42–48], disease metabolite prioritization [49] and drug–target interaction prediction [50–52]. In this paper, we focus on re-viewing network-based methods used for integrating heterogeneous data to predict lncRNA–protein interactions directly. The materials, interaction score calculation algorithms, and advantages and disadvantages of state-of-art algorithms of lncRNA–protein interaction prediction based on network methods are summarized and discussed to assist researchers in selecting a suitable method for acquiring more dependable results. This article is organized as follows. Section 2 summarizes the relevant databases used for analyzing lncRNA–protein interaction. Section 3 gives a brief introduction to experimental approaches and machine learning-based computational approaches for studying lncRNA–protein interaction. Section 4 systematically analyzes biological network-based computational models for lncRNA–protein interaction prediction. Section 5 includes the performance comparison of different network-based models for lncRNA-protein interaction prediction. And Section 6 briefly summarizes the discussion in this paper and looks forward to the future feasible methods.

2. A Brief Introduction to the Relevant Databases Used for Analyzing LncRNA–Protein Interactions

The various databases discussed in this article incorporate lncRNAs from different tissues and focus on lncRNAs as well as lncRNA-related interactions. Some of these databases are available at RNAcentral [53]. Although there is a great deal of overlapping sections among these databases, each database nonetheless offers considerable unique features. We present herein an overview of their respective contents and search features in order that researchers can get a quick glance of what each can offer. Then, we give a brief summary of the relevant databases mentioned in Table 1, including the name and website of the database and a brief description. We provide data information on all possible interactions between biomolecules that may be used in the research of lncRNA functions (which users can browse and download from https://github.com/HAN-Siyu/APINet/), that is, lncRNA–disease associations, lncRNA–lncRNA interactions, lncRNA–microRNA interactions, lncRNA–gene interactions, lncRNA–Gene Ontology (GO) interactions, microRNA–microRNA interactions, microRNA–disease associations, microRNA–gene interactions, microRNA–target interactions, gene–gene interactions, gene–metabolite interactions, metabolite–metabolite interactions, gene–GO interactions, gene–disease associations, gene–drug associations, metabolite–disease associations, drug–disease associations, drug–drug interactions, drug–side-effect interactions and and disease–disease interactions. The details of the data information are shown in Table 2. As some interaction data are integrated by multi-source data, in Table 2, we can see the types of these interactive data information, the number of sets of interaction data composed of several biological molecules and the sources of these data, which determine association data that can be used to construct heterogeneous networks, i.e., the composition of heterogeneous networks.

Table 1. Description of lncRNA relevant databases.

Database	Description	Availability
ncRNA database (Especially lncRNAs):		
NONCODE [54]	Comprehensive knowledge database of non-coding RNAs, including lncRNAs from 17 species, and predicted/validated lncRNA–disease relationships.	http://www.noncode.org
MNDR [55]	Database of ncRNA–disease associations in mammals.	http://www.rna-society.org/mndr
deepBase [56]	Database for identification, expression, evolution and function of small RNAs, lncRNAs and circular RNAs from deep-sequencing data.	http://rna.sysu.edu.cn/deepBase
NRED [57]	Database integrating annotated human and mouse ncRNA expression data from various resources.	http://nred.matticklab.com
ChIPBase [58]	Database on the transcriptional regulation of ncRNAs based on ChIP-sequencing data.	http://rna.sysu.edu.cn/chipbase
SomamiR [59]	Cancer somatic mutations with altering microRNA–ceRNA interactions.	http://compbio.uthsc.edu/SomamiR
LncRNA2Function [60]	Functional annotations and expression profiles (RNAseq) of human lncRNAs.	http://mlg.hit.edu.cn/lncrna2function
LincSNP [61]	A database containing human lncRNAs information about linking disease related SNPs.	http://bioinfo.hrbmu.edu.cn/LincSNP
LncRNA-SNP [62]	A database of SNPs in lncRNAs and their predicted effects in human and mouse.	http://bioinfo.life.hust.edu.cn/lncRNASNP
LNCipedia [63]	A database for annotated human lncRNA transcript sequences and structures.	http://www.lncipedia.org
ALDB [64]	A farm livestock lncRNA database.	http://res.xaut.edu.cn/aldb/index.jsp
lncRNAtor [65]	A database for functional investigation of lncRNAs that encompasses annotation, sequence analysis, gene expression, protein binding and phylogenetic conservation.	http://lncrnator.ewha.ac.kr
Co-LncRNA [66]	A web-sever containing effects of lncRNAs in GO functions and KEGG pathways based on co-expressed genes.	http://www.bio-bigdata.com/Co-LncRNA
Lnc2Cancer [17]	A database for experimentally validated associations between lncRNAs and cancers.	http://www.bio-bigdata.net/lnc2cancer
LncRNADisease [16]	A database for experimentally validated lncRNA-associated diseases.	http://www.cuilab.cn/lncrnadisease
lncRNAMap [67]	A map of putative regulatory functions in the long non-coding transcriptome.	http://lncRNAMap.mbc.nctu.edu.tw/
TANRIC [34]	A web-resource for interactive exploration of lncRNAs in cancer.	http://ibl.mdanderson.org/tanric/_design/basic/index.html
LncRNA ontology [64]	A web-resource for inferring lncRNA functions based on chroma-tin states and expression patterns.	http://www.bio-bigdata.com/lncrnaontology/
LNCediting [68]	A database for functional effects of RNA editing in lncRNAs.	http://bioinfo.life.hust.edu.cn/LNCediting/
LncBase [69]	A database of interactions between miRNAs and lncRNAs.	http://www.microrna.gr/LncBase
TF2LncRNA [70]	A Web-resource for the identification of common transcription factors for a list of lncRNA genes.	http://mlg.hit.edu.cn/tf2lncrna
LncSubpathway [71]	A web server for the identification of dysfunctional subpathways associated with risk lncRNAs.	http://www.bio-bigdata.com/lncSubpathway/
LncRNA2Target [72]	A database of differentially expressed genes after lncRNA knock-down or overexpression.	http://lncrna2target.org
LncReg [73]	A reference resource for lncRNA-associated regulatory networks.	http://bioinformatics.ustc.edu.cn/lncreg/
lncRNAdb [74]	An annotation database of eukaryotic lncRNAs.	http://www.lncrnadb.org/

Table 1. *Cont.*

Database	Description	Availability
	Database information on proteins or microRNAs that may be associated with lncRNAs:	
NPInter [75]	Database of noncoding RNA-associated interactions.	http://www.bioinfo.org/NPInter
PRIDB [76]	Comprehensive database of RNA–protein interfaces extracted from complexes in the PDB.	http://bindr.gdcb.iastate.edu/PRIDB
PDB [77]	A database of experimentally determined three-dimensional structures of proteins, nucleic acids and other biomolecules.	http://www.rcsb.org/
StarBase v 2.0 [78]	A database of experimentally supported interactions from RBPs, mRNAs, miRNAs, RNAs, proteins and so on.	http://starbase.sysu.edu.cn/
Nucleic acid database (NDB) [79]	A database about three-dimensional nucleic acid structures and their complexes, geometric data, structure information.	http://ndbserver.rutgers.edu/

Table 2. Details of interactions between biomolecules and the research of lncRNA functions.

Name	Samples	Interactions	Source
LncRNA–Disease	804 × 288	1454	LncRNADisease [16], Lnc2Cancer [17]
LncRNA–LncRNA	1114 × 1114	1,179,256	LFSCM [80]
LncRNA–microRNA	1127 × 277	10,198	StarBase v2.0 [78]
LncRNA–Gene	240 × 15,527	6186	LncRNA2Target [72]
LncRNA–GO	240 × 6428	3094	GeneRIF [81]
MicroRNA–MicroRNA	271 × 271	24,062	Zhong et al. [82]
MicroRNA–Disease	1080 × 592	11,835	HMDD [83], miR2Disease [84], miRCancer [85]
MicroRNA–Gene	495 × 15,527	135,852	miRTarBase [86]
MicroRNA–Target	495 × 15,527	135,852	miRTarBase [86]
Gene–Gene	16,785 × 16,785	1,515,370	Yao et al. [49]
Gene–Metabolite	12,342 × 3278	192,763	Yao et al. [49]
Metabolite–Metabolite	3764 × 3764	74,667	Yao et al. [49]
Gene–GO	15,527 × 6428	1,191,503	GO Annotation [87]
Gene–Disease	1715 × 1886	2603	DisGeNET [88]
Gene–Drug	155,275 × 8283	3760	DrugBank [89]
Metabolite–Disease	388 × 149	664	HMDB [90]
Drug–Disease	15,527 × 412	115,317	CTD [91]
Drug–Drug	8283 × 8283	453,436	DrugBank [89]
Drug–Side-effects	1430 × 5880	140,064	SIDER [92]
Disease–Disease	5080 × 5080	20,280,092	Yao et al. [49]

3. A Brief Introduction of Experimental Approaches and Computational Approaches Based on Machine Learning to Study LncRNA–Protein Interactions

3.1. LncRNA–Protein Interactions: From Experimental Approaches to Computational Models Based on High-Throughput Experiments

Several large-scale experimental approaches for lncRNA–protein interaction prediction include RNA immunoprecipitation (RIP) followed by mass spectrometry analysis, RNAcompete [18], RIP-Chip [19], high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) [20], and photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) [21]. Although these approaches can provide valuable data to construct a network of lncRNA–protein interactions, they are expensive and time-consuming, which are disadvantages that cannot be ignored. It is therefore urgent to put forward the computational approaches.

There are many effective methods for the analysis of experimental data, such as several methods for finding RNA motifs from CLIP-Seq or other high-throughput experiments, such as BEAM [22] and SMARIV [23]. BEAM is a method for structural motif discovery from a set of unaligned RNAs. Tested in various scenarios, BEAM is successful in retrieving structural motifs even in highly noisy datasets, such as those that can arise in CLIP-Seq or other high-throughput experiments. To solve the problem that the previous methods cannot provide information about protein structure preferences, the sequence and structure preferences of RNA-binding proteins can be inferred based on the feasibility of obtaining RNA structure information. SMARTIV is a novel computational tool for discovering combined sequence and structure binding motifs from in vivo RNA binding data relying on the sequences of the target sites, the ranking of their binding scores and their predicted secondary structures. The combined motifs are presented in a unified form, which is rich in information and easy for visual perception. These high-throughput experimental data can be used to predict the next step by developing machine learning methods. The quality of these models depends directly on the experimental data. At present, NPInter database and StarBase database are constructed from high throughput experimental data and are existing databases for lncRNA–protein interactions.

3.2. LncRNA–Protein Interactions: From Experimental Results to Computational Models Based on Machine Learning

Computational approaches for lncRNA–protein interaction prediction can be grouped into the following two ways of expressions. The first category is based on sequence and structural information and physicochemical properties, including RPISeq [24], de novo prediction [25], CatRAPID [26], LncPro [5], RPI-Pred [27], rpiCOOL [28], IPMiner [29] and lncADeep [30]. The second category is based on the fusion of heterogeneous data to construct a network, such as the lncRNA–protein bipartite network inference (LPBNI) method [31], fusing multiple protein–protein similarity networks (PPSNs) [32], the method to predict lncRNA–protein interactions based on the relevance search method proposed by Yang et al. [33], the prediction method of interactions between lncRNAs and proteins on heterogeneous networks (LPIHN) [34] and the predicting lncRNA–protein interactions using HeteSim scores (PLPIHS) method [35].

From the point of view of characteristics such as sequence information, various classical methods have been proposed. RPISeq [24] is proposed to predict RNA–protein interactions only using sequence information. The support vector machine (SVM) classifier and the random forest (RF) classifier, which are supervised machine learning algorithms, are used in the RPISeq. De novo prediction of RNA–protein interactions [25] also only considers sequence information. A set of known RNA–protein interactions is collected as gold-standard positives, where sequence-based features are extracted for each RNA–protein pair [25]. In the process of constructing the Bayes classifier, these effective features are used to train an RNA–protein interaction prediction model. CatRAPID [26] is proposed by using physicochemical properties, including the secondary structures of the molecules and their propensities for hydrogen bonding and van der Waals interactions. Encoding the protein–RNA pairs into feature vectors is the first step, followed by calculating the interaction score through the matrix computation. LncPro [5] is proposed to predict ncRNA–protein interactions by using Fisher's linear discriminant approach. The training features are not only from protein secondary structures and their propensities for hydrogen bonding and van der Waals interactions, but also from RNA secondary structures [93]. LncPro also requires the identification of a matrix and calculation of the interaction score to represent degree of interactions through matrix computation by a simple machine-learning model for matrix computation. RPI-Pred [27], a SVM-based machine-learning approach, is proposed by considering sequence features and combining the high-order structures of both proteins and RNAs. This interaction prediction considers protein blocks rather than classical three-state protein secondary structures. Five classes of RNA secondary structures are regarded as high-order structures. RpiCOOL [28] is a tool developed for detecting RNA–protein interactions in silico by using the RF classifier, which classifies RNA and protein based on whether there are interactions between them. The sequence composition and repetitive patterns are used as heterogeneous information of the protein and RNA, which is then used to encode feature vectors to express pairs between RNA and protein. IPMiner [29], a tool based on simple sequence composition features, integrates deep neural network and stacked ensembling classifiers to identify RNA–protein interactions. The extracted original features, SDA (stacked denoising autoencoder) and SDA-FT (SDA with fine tuning), are provided to the RF classifier, respectively. The outputs of these three classifiers, which are trained by a logistic regression mode, are integrated through superposition. These computational methods fill the broadening gap between raw and annotated data that has been generated as a result of the large amount of data obtained by high-throughput technologies. LncADeep is proposed to predict lncRNA-protein interactions based on deep neural networks, using both sequence and structure information.

With the development of computational approaches, experimental methods are now suffering the great disadvantage to predict lncRNA–protein interactions, such as high cost and long time. Intrinsic features of lncRNA and protein have increasingly interested the researchers. The advantage of intrinsic features has been demonstrated in the research of lncRNA identification. The methods of lncRNA–protein interaction prediction focus on intrinsic features of lncRNA and protein, such as sequence information, structure information, and physicochemical properties, including hydrogen-bond and van der Waals propensities. We analyzed the dataset of methods based on what kind of information they use, such as sequence, structure and physicochemical information. We also analyzed what machine learning algorithms are employed in the different methods. The comparison of each method is shown in Table 3. In this article, we give a more detailed introduction to each computational model for lncRNA–protein interaction prediction based on intrinsic features of lncRNA and protein. To make it easier for users to use these computation models, we have supplemented the availability network resources. We give more details about each computational method's availability, such as the web server or offline package for lncRNA–protein interaction prediction based on sequence and structural information and physicochemical properties.

Whereas the machine learning-based methods only consider the properties of the RNAs or proteins and neglect interactions between lncRNAs and proteins, the network-based methods pay more attention to this kind of interactions, which are implicated in the topologies between nodes in the heterogeneous networks of lncRNAs. When the sequence is too long or the randomness of structural information is predicted, the computational models based on machine learning will be affected to some extent.

4. Computational Models for LncRNA–Protein Interaction Prediction Based on Biological Networks

The previously described methods for predicting the interactions between lncRNAs and proteins more focus on the intrinsic features of lncRNAs and proteins but do not take the topological structures of biological networks associated with the lncRNAs into consideration. A biological network can apply to biological systems. Nowadays, network science is being used extensively in the biological and related fields. Network science provides many practical descriptions of biological systems and relationships between diseases and other biomolecules as biological factors [33]. Moreover, we could integrate known lncRNA–protein interaction networks, lncRNA–lncRNA similarity networks and PPI networks that were downloaded in the databases and fused by multiple PPSNs to construct heterogeneous networks and implement a model based on computing node similarity between networks to discover possible interactions between lncRNAs and proteins, such as random walk on heterogeneous networks and kinds of propagation algorithms that can discover potential associations. The overview is presented in Figure 1. We analyzed which heterogeneous data are selected by each method, how to fuse heterogeneous data to construct the network, and what methods are used to deal with heterogeneous networks to predict lncRNA–protein interactions. We analyzed the differences among the different network-based methods such as the datasets that are used in each method, how to fuse heterogeneous data to construct the network and algorithms for specific computation interactions. The differences of each network-based method are shown in Table 4. In this articl, we give a more detailed introduction to each computational model for lncRNA–protein interaction prediction based on biological networks.

Table 3. The comparison of each method by analyzing the differences in intrinsic features and classifiers.

		CatRAPID [26]	**RPISeq** [24]	**De novo** [25]	**LncPro** [5]	**RPI-Pred** [27]	**rpiCOOL** [28]	**IPMiner** [29]	**lncADeep** [30]
Feature	RNA Sequence		√	√	√	√	√	√	
	Protein Sequence		√	√	√	√	√	√	
	3D Structure(RNA)					√			
	3D Structure (protein)					√			
	The secondary structure (RNA)	√			√				
	The secondary structure(protein)				√				
	Hydrogen-Bonding Propensities	√			√				
	van der Waals' Propensities	√			√				
Classifier	Random Forest		√				√	√	
	Naive Bayesian			√					
	Extended NB			√					
	SVM		√			√			
	Fisher's linear				√				
	automatic encoder							√	
	deep neural network								√
	p-values	√							√
Web server or offline package		√	√		√	√	√	√	√

[1] http://s.tartaglialab.com/page/catrapid_group (web server); [2] http://pridb.gdcb.iastate.edu/RPISeq (web server); [3] http://bioinfo.bjmu.edu.cn/lncpro/ (offline package and web server); [4] http://ctsb.is.wfubmc.edu/projects/rpi-pred (web server); [5] http://biocool.ir/softs/rpicool.html (offline package); [6] https://github.com/xypan1232/IPMiner (offline package); [7] https://github.com/cyang235/LncADeep (offline package).

Table 4. Differences in each network-based methods.

Method		Dataset	Algorithm	AUC
LPBNI [31]	LPI	4870 lncRNA–protein interactions from NPInter database (2380 lncRNAs and 106 proteins)	Bipartite Network	0.8780
	PPI	×		
	LLI	×		
Yang et al. [33]	LPI	4883 lncRNA–protein interactions from NPInter database (1116 lncRNAs and 99 proteins)	A random walk model HeteSim	0.7972
	PPI	1608 protein–protein interactions from STRING database		
	LLI	×		
LPIHN [34]	LPI	10232 lncRNA–protein interactions from NPInter database (1113 lncRNAs and 99 proteins)	Random Walk with Restart	0.8839
	PPI	804 protein–protein interactions from STRING database		
	LLI	lncRNA expression similarity from NONCODE 4.0 database (1113 lncRNA expression profiles)		
Zheng et al. [32]	LPI	4467 lncRNA–protein interactions from NPInter database (1050 lncRNAs and 84 proteins)	SNF; A random walk model HeteSim	0.9068

	PPI	Sequence similarity from UniProt database;
		Functional annotation similarity from GO database;
		Protein domain similarity from Pfam database;
		STRING similarity from STRING database;		

	LLI	×		

PLPIHS [35]	LPI	lncRNA–protein interactions from GENCODE Release 24 (15941 lncRNAs and 20284 proteins)
		Co-expression data from COXPRESdb;
		Co-expression data from ArrayExpress and GEO;
		lncRNA–protein interactions from NPInter database;
			SVM; A random walk model HeteSim	**0.9678**

	PPI	Protein–protein interactions from STRING database		
	LLI	lncRNA co-expression similarity from NONCODE database (lncRNA expression profiles)		

Bold representation performs best in AUC values and we found that the performance of the method is better when the heterogeneous network is composed by more sources. When heterogeneous networks are constructed by the same sources, the performance will be better for the heterogeneous networks constructed by weighted networks. [1] https://github.com/USTC-HIlab/LPBNI (offline package); [2] https://github.com/cyang235/LncADeep (offline package); [3] lncRNA–protein interactions; [4] protein–protein interactions; [5] lncRNA–lncRNA interactions; [6] A relevance search based on random walk in heterogeneous network to evaluate the relevance between a pair of lncRNA and protein, and a large relevance score means a high possibility that the lncRNA and protein interacts [94]. [7] Similarity Network Fusion: It is a nonlinear message-passing based method that iteratively updates each network and makes it more and more similar to the other [95].

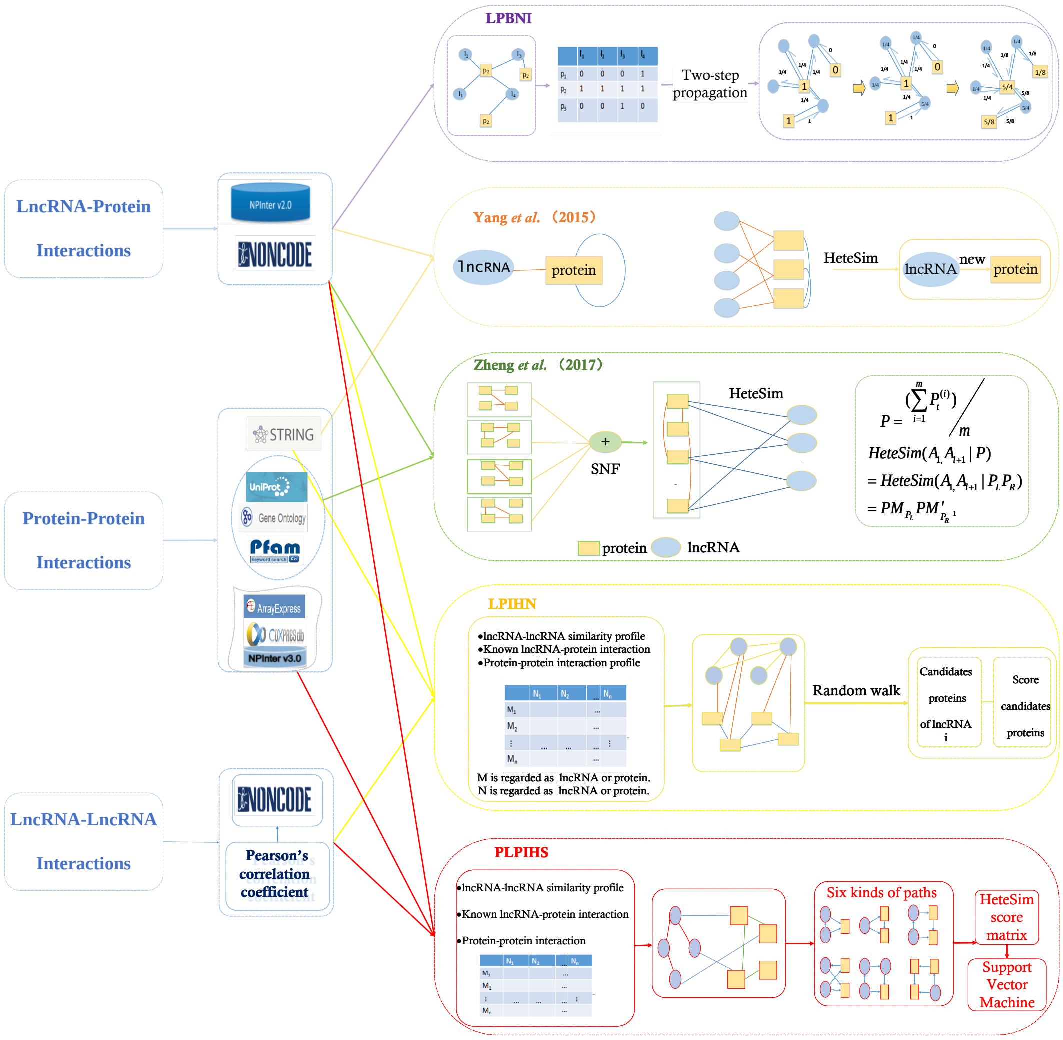

LncRNA-Protein Interactions
Protein-Protein Interactions
LncRNA-LncRNA Interactions
NPInter v2.0
NONCODE
STRING
UniProt
Gene Ontology
Pfam
ArrayExpress
NPInter v3.0
Pearson's correlation coefficient
LPBNI
Two-step propagation
Yang et al. (2015)
lncRNA
protein
HeteSim
new
Zheng et al. (2017)
SNF
+
$P = (\sum_{i=1}^{m} P_t^{(i)}) / m$
$HeteSim(A_1, A_{l+1} | P)$
$= HeteSim(A_1, A_{l+1} | P_L P_R)$
$= PM_{P_L} PM'_{P_R^{-1}}$
protein
lncRNA
LPIHN
•lncRNA-lncRNA similarity profile
•Known lncRNA-protein interaction
•Protein-protein interaction profile
M is regarded as lncRNA or protein.
N is regarded as lncRNA or protein.
Random walk
Candidates proteins of lncRNA i
Score candidates proteins
PLPIHS
•lncRNA-lncRNA similarity profile
•Known lncRNA-protein interaction
•Protein-protein interaction
Six kinds of paths
HeteSim score matrix
Support Vector Machine

Figure 1. Overview of five computational models for lncRNA–protein interaction prediction based on network method, including data collection and core algorithm. Illustration: The specific algorithm implementation of each method is represented by rectangular boxes with dotted lines of different colors, and the solid lines with different colors outside the rectangular boxes of dotted lines represent the data sources used by different methods. These colors are the same as the colors used by method names. In addition, the solid line color in the dotted rectangular frame is used to distinguish the interaction of lncRNA–lncRNA, protein–protein or lncRNA–protein.

4.1. LPBNI: A Bipartite Network-Based Method for the Prediction of LncRNA–Protein Interactions

Inspired by resource methods in dynamically allocated networks, Zhou et al. [96] proposed algorithms based on the propagation process of the LPBNI method. Li et al. [34] developed this method on the basis of an lncRNA–protein bipartite network to predict lncRNA–protein interactions. A graph G can be used to represent the lncRNA–protein interaction network. The structure of the bipartite network of lncRNA–protein is simply shown in graphic language, as shown in Figure 2. Finally, they chose to apply the propagation method on the constructed network and calculated the degree of lncRNA–protein interactions as a score. In the $G(L, P, E)$, the propagation matrix is used as W, where W_{ik} represents the information transferred from the p_k node to the p_i node, and the transmission of key information between two nodes represents the importance of nodes. For each lncRNA l_j, they defined $S_0(i) = s_{ij}, i \in \{1, 2, \ldots, m\}$ as the first information on protein P, where $s_{ij} = 1$ if p_i interacts with l_j; otherwise, $s_{i,j} = 0$. $S_L(l_j), j \in \{1, 2, \ldots, n\}$ represents the score on l_j after the first step of information propagation, which can be calculated as

$$S_L = \sum_{i=1}^{m} \frac{a_{ij} S_0(i)}{d(p_i)}. \tag{1}$$

In the formula above, $d(p_i) = \sum_{j=0}^{n} a_{ij}$ is the number of lncRNAs that interact with p_i.

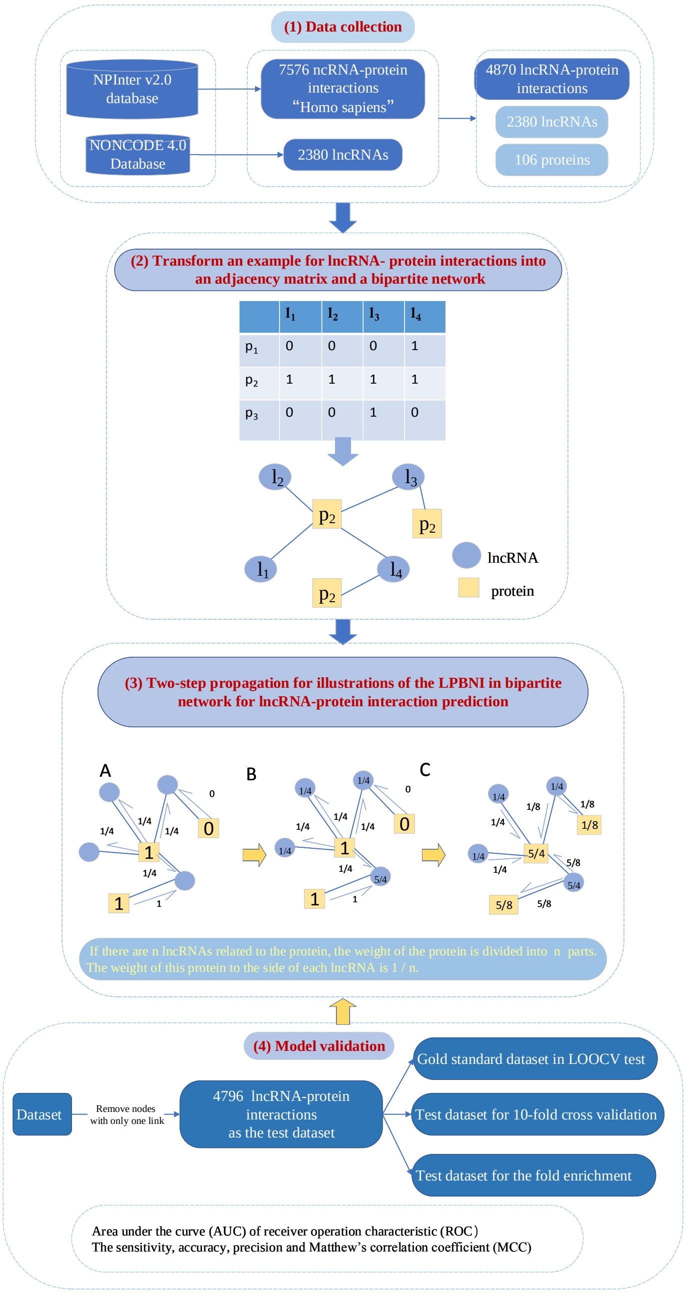

(1) Data collection
NPInter v2.0 database
NONCODE 4.0 Database
7576 ncRNA-protein interactions "Homo sapiens"
2380 lncRNAs
4870 lncRNA-protein interactions
2380 lncRNAs
106 proteins
(2) Transform an example for lncRNA- protein interactions into an adjacency matrix and a bipartite network
l1 l2 l3 l4
p1 0 0 0 1
p2 1 1 1 1
p3 0 0 1 0
l2 l3 p2 p2 l1 l4 p2
lncRNA
protein
(3) Two-step propagation for illustrations of the LPBNI in bipartite network for lncRNA-protein interaction prediction
A
1/4 1/4 1/4 0 0 1 1/4 1 1
B
1/4 1/4 1/4 1/4 1/4 0 0 1/4 1/4 1 1/4 5/4 1 1
C
1/4 1/4 1/8 1/8 1/8 1/4 1/4 5/4 5/8 1/4 5/4 5/8 5/8
If there are n lncRNAs related to the protein, the weight of the protein is divided into n parts. The weight of this protein to the side of each lncRNA is 1 / n.
(4) Model validation
Dataset
Remove nodes with only one link
4796 lncRNA-protein interactions as the test dataset
Gold standard dataset in LOOCV test
Test dataset for 10-fold cross validation
Test dataset for the fold enrichment
Area under the curve (AUC) of receiver operation characteristic (ROC)
The sensitivity, accuracy, precision and Matthew's correlation coefficient (MCC)

Figure 2. Framework of LPBNI mainly including four modules: (**1**) Data collection: the lncRNA–protein interaction network is from NPInter and NONCODE. (**2**) Bipartite network construction (a toy example in Figure 1). (**3**) Two-step propagation on the bipartite network: (**A**) The process of the initial information propagated from proteins to their direct neighbor lncRNAs. For example, the initial information of three proteins is 1, 1 and 0, respectively. (**B**) The score on red circles is the information of each lncRNA received from proteins. (**C**) The process of the information propagated from lncRNAs back to proteins. The score on blue hexagon in (**C**) is the final information of each protein after the two-step propagation. The red circles represent lncRNAs and the blue hexagons represent proteins. (**4**) Model validation based on leave one out cross validation (LOOCV), the area under the receiver operating characteristic curve (AUC) and Matthew's correlation coefficient (MCC).

In the next step, all the information in L propagates back to P. $S_F(p_i)$ is defined as the final information on protein p_i, signifying the interaction score of protein p_i with l_j. S_F can be defined as

$$S_F(i) = \sum_{j=1}^{n} \frac{a_{ij} S_L(l_j)}{d(l_j)} = \sum_{j=1}^{n} \frac{a_{ij}}{d(l_j)} \sum_{k=1}^{m} \frac{a_{kj} S_0(k)}{d(p_k)}, \tag{2}$$

where $d(l_j) = \sum_{l=0}^{m} a_{ij}$ is the number of proteins that interact with l_j. The final information S_F can be defined in the matrix as

$$\vec{S}_F = W\vec{S}_0, \tag{3}$$

where $\vec{S}_0$ is the column vector of S_0, and $\vec{S}_F$ is the final score of the lncRNA that users query after the two-step information propagation in the lncRNA–protein interaction network. S_F can be represented as

$$S_F(i) = \sum_{k=1}^{m} w_{ik} S_0(k), W_{ij} = \frac{1}{d(p_i)} \sum_{j=1}^{n} \frac{a_{ij} a_{kj}}{d(l_j)}. \tag{4}$$

After calculations, the protein sorted by the final score S_F for l_j is obtained. All the candidate proteins are ranked in decreasing order, and proteins with a high ranking are considered to interact with lncRNA l_j.

LOOCV was performed on the heterogenous network containing lncRNA–protein interactions, leaving only one sample for the test set at a time, and the other samples were used as the training set. Although the calculation was more complicated than other verification methods, the sample utilization rate was the highest. LOOCV aws used to evaluate the performance of the proposed method. In the course of the calculation, each lncRNA–protein pair was omitted as a test sample by changing the value in the adjacency matrix A to 0. The performance of LPBNI could be estimated by the ratio of its predicted interactions to the originally known lncRNA-protein. A receiver operating characteristic (ROC) curve was selected as a criterion to evaluate the LPBNI and random walk with restart methods. The propagation matrix W proposed in the LPBNI method is dependent on the adjacency matrix A of the bipartite network. When applying LOOCV, multiple values of W were obtained, owing to the change of A values during each step of the cross-validation. Consequently, the value of W was recalculated for each lncRNA–protein pair that was left out as a test sample. In addition, nodes that do not propagate information are not considered when evaluating the performance of the method, where nodes with fewer than two links are defined as nodes that do not propagate information in the process of cross-validation.

4.2. Fusing Multiple Protein–Protein Similarity Networks to Effectively Predict LncRNA–Protein Interactions

To improve the performance of lncRNA–protein interaction prediction, Zheng et al. [32] fused multiple PPSNs to construct a multilevel heterogeneous network. New lncRNA–protein interaction predictions are inferred by integrating the fused PPSNs with known lncRNA–protein interaction predictions (Figure 3). Protein sequences, protein domains, GO terms, and the STRING database are first used to construct four Protein–Protein Similarity Networks (PPSNs), following which the SNF algorithm [95] is employed to combine the four protein–protein similarity networks into a fused protein–protein similarity network. Then, a heterogeneous lncRNA–protein network is built including based on the fused protein–protein similarity network and the known lncRNA–protein interactions. Finally, the HeteSim algorithm [94] is used to infer new lncRNA–protein interaction predictions. Extensive experiments show that this approach outperforms not only the existing methods for predicting the lncRNA–protein interactions, but also performs well by using only one PPSN as a protein–protein interaction network without fusing four different aspects of the protein–protein similarity network into a protein–protein interaction network. After fusing all the four matrices, the area under the curve (AUC) value of 0.9068 indicates the best performance. This result shows that a more reliable and informative network can be obtained by fusing multiple matrices.

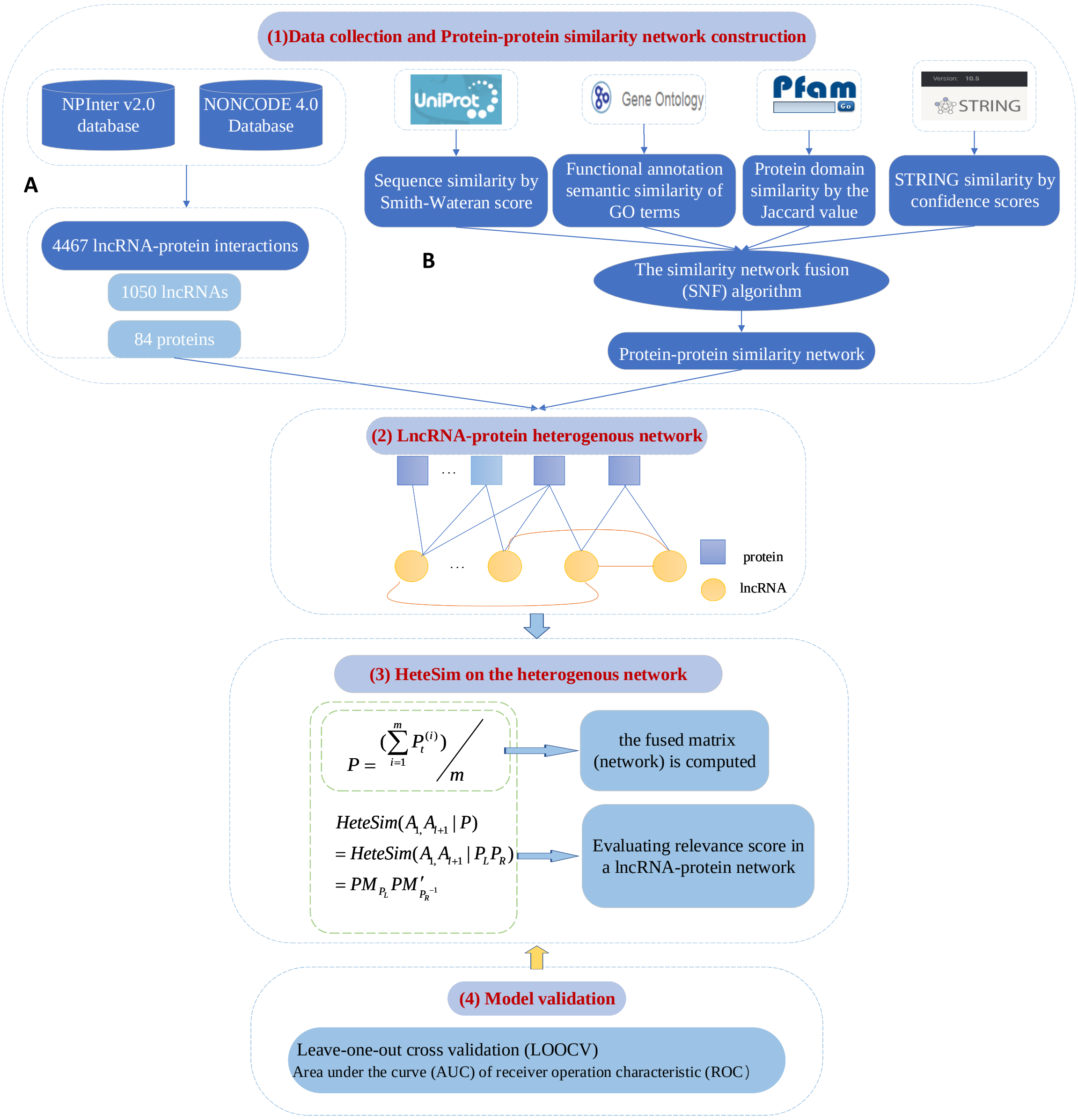

(1)Data collection and Protein-protein similarity network construction
NPInter v2.0 database
NONCODE 4.0 Database
A
4467 lncRNA-protein interactions
1050 lncRNAs
84 proteins
UniProt
Gene Ontology
Pfam
STRING
Sequence similarity by Smith-Wateran score
Functional annotation semantic similarity of GO terms
Protein domain similarity by the Jaccard value
STRING similarity by confidence scores
B
The similarity network fusion (SNF) algorithm
Protein-protein similarity network
(2) LncRNA-protein heterogenous network
protein
lncRNA
(3) HeteSim on the heterogenous network
$P = (\sum_{i=1}^{m} P_t^{(i)}) / m$
the fused matrix (network) is computed
$HeteSim(A_1, A_{l+1} | P) = HeteSim(A_1, A_{l+1} | P_L P_R) = PM_{P_L} PM'_{P_R^{-1}}$
Evaluating relevance score in a lncRNA-protein network
(4) Model validation
Leave-one-out cross validation (LOOCV)
Area under the curve (AUC) of receiver operation characteristic (ROC)

Figure 3. Framework of the proposed method by Zheng et al. [32] mainly containing four modules. (**1**) (**A**) Data collection: The lncRNA–protein network is from NPInter and NONCODE. The datasets from Uniprot, GO, Pfam and STRING database are collected for protein–protein similarity network construction. (**B**) Protein–protein similarity network construction: based on similarity network fusion (SNF) algorithm by integration of multi-resource information. (**2**) A heterogeneous network construction. (**3**) HeteSim computation on the heterogeneous network. (**4**) Model validation based on LOOCV and AUC.

The advantage of SNF algorithm is that it can obtain valuable information from a relatively small number of samples, and it has strong robustness in dealing with noise and data heterogeneity. It is a nonlinear method based on the typical nature of the complexity of the natural world based on message-passing. The nonlinear method is closer to the nature of the objective thing itself. It is one of the important methods to quantitatively study and understand complex knowledge. This method iteratively updates each network and makes it more and more similar to other networks. A protein similarity network can be represented as a graph $G = (V, E)$, where $V = \{v_1, v_2, \ldots, v_n\}$ represents a set of corresponding proteins in the network, and E represents a set of edges, each of which has a similarity weight. The authors denoted the corresponding similarity matrix as W, where $W(i, j)$ is the similarity between proteins v_i and v_j. They defined a full and sparse kernel on each matrix in order to compute the fused network from four protein similarity matrices. The full kernel is a normalized weight matrix $P = D^{-1}W$, where D is a diagonal matrix and $D(i,j) = \sum_j W(i,j)$. Because P involves self-similarities on the diagonal entries of W, a better form for avoiding numerical instability is as follows [96]:

$$P(i,j) = \begin{cases} \frac{W(i,j)}{2\sum_{k\neq i} W(i,k)}, & j \neq i \\ 1/2, & j = i. \end{cases} \tag{5}$$

Protein v_i's neighbors are denoted as N_i and use k nearest neighbors (kNN) to measure the local part as follows:

$$S(i,j) = \begin{cases} \frac{W(i,j)}{\sum_{k\in N_i} W(i,k)}, & j \in N_i \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

A protein has much better similarities to its neighbors than it has to remote proteins. Similarity based on graph diffusion principle can be propagated to remote proteins. Matrix P provides all the information of the PPSN, whereas S provides the local similarity information of the network. Then, iterative computation can occur as follows:

$$P_t^{(i)} = S^{(i)} \times \left(\frac{\sum_{k\neq i} P_{t-1}^{(k)}}{m-1} \right) \times \left(S^{(i)} \right)^T, i = 1,2,3,4, \tag{7}$$

where $P_t^{(i)}$ is the ith similarity matrix after $t(\geq 0)$ iterations, and $S^{(i)}$ is the kNN matrix of the similarity matrix or network. Following that, m is the number of PPSNs used. As S is the kNN matrix of P, it contains the most important information of P and also alleviates the noise effect of P. In each iteration, each similarity matrix can get more reliable information from other similarity matrices, at the same time, it will update its own matrix based on other similarity matrices. After t iterations, the fusion network can be replaced by a fusion matrix, which is defined as follows:

$$P = \left(\sum_{i=1}^{m} P_t^{(i)} \right) \Big/ m. \tag{8}$$

Note that the authors normalized matrix P_t after each iteration, each protein has a higher degree of similarity to itself in order to ensure that the matrix is in a full rank state than other proteins. With the known lncRNA–protein interactions and the fused PPSN, they built a lncRNA–protein heterogeneous network, on which a random walk model HeteSim was used to infer new lncRNA–protein interactions. HeteSim is used to evaluate the relevance between a lncRNA–protein pair, where a large relevance score means the lncRNA and protein have more interactions.

For this method, 15 settings made up of different combinations of the similarity matrices (Seqs, Pfam, GO, and STRING, respectively) were implemented. The path selection is very important since HeteSim is a path-constrained relevance measure. In the fusion work, the relevance path was chose as lncRNA-protein-protein, which was the same as that used in the work of Yang et al. [33]. With the proof of the experiment and more matrix merging, the AUC value becomes more ideal. For example, the AUC value of GO + Pfam + STRING is 0.9066, which is larger than the AUC value of GO + Pfam, GO + STRING and Pfam + STRING. When all four protein similarity matrices were fused, AUC achieved the best result of 0.9068. This shows that, with the increase of the number of fusion matrices, we could get more specific information of protein similar network. This multi-matrix fusion method is convenient to get more reliable and informative data representation.

4.3. Prediction of Interactions between lncRNA and Protein by Using Relevance Search in a Heterogeneous LncRNA–Protein Network

Yang et al. [33] tried to use the possible hidden information in the biological network topologies containing lncRNA layer networks. Thus, an algorithm named HeteSim is introduced to measure the relevance between lncRNAs and proteins on the basis of the heterogeneous lncRNA–protein network, which integrates the known lncRNA–protein interaction networks and PPSNs. Figure 4 shows a network model and the schema of the interaction network. The AUC of HeteSim for the lncRNA MALAT1 is 0.955. The performance results of network-assisted method confirm a difficult problem. It is difficult to break through the low conservatism of lncRNAs by traditional methods to predict the interactions between lncRNAs and proteins, which is a challenge to propose new methods to predict lncRNA–protein interactions, which generally uses information from intrinsic features of the RNA and protein alone. Their approach also demonstrates the tremendous value of the network-based approach in lncRNA-related fields, and has valuable implications for predicting interactions in heterogeneous networks constructed from biomolecules.

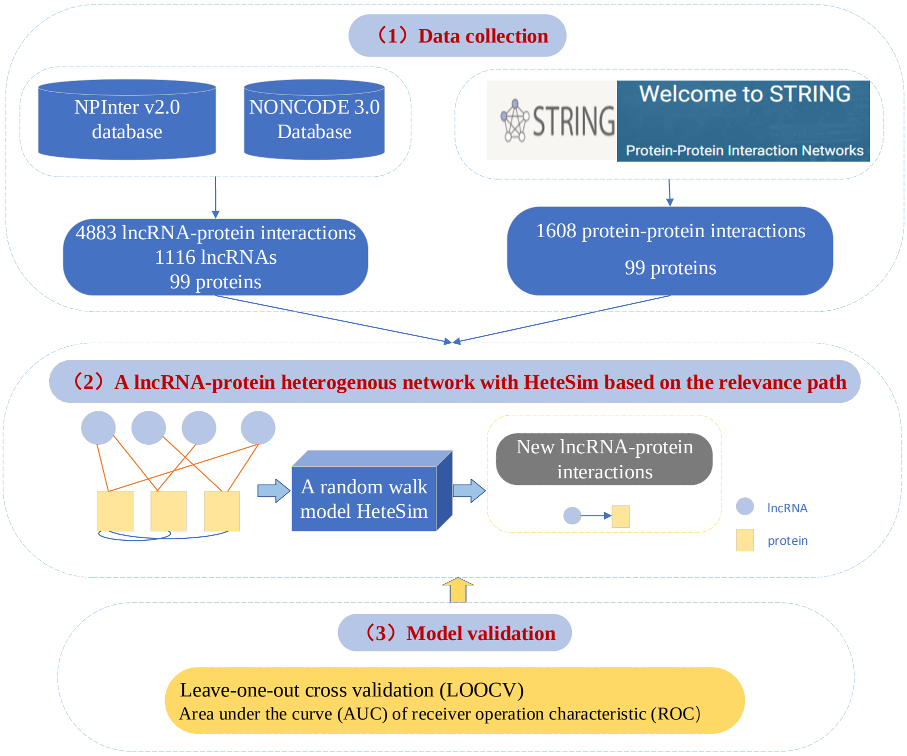

（1）Data collection
NPInter v2.0 database
NONCODE 3.0 Database
STRING
Welcome to STRING
Protein-Protein Interaction Networks
4883 lncRNA-protein interactions
1116 lncRNAs
99 proteins
1608 protein-protein interactions
99 proteins
（2）A lncRNA-protein heterogenous network with HeteSim based on the relevance path
A random walk model HeteSim
New lncRNA-protein interactions
lncRNA
protein
（3）Model validation
Leave-one-out cross validation (LOOCV)
Area under the curve (AUC) of receiver operation characteristic (ROC)

Figure 4. Pipeline of the method proposed by Yang et al. [33]. (**1**) Data collection: lncRNA–protein interactions from NPInter and NONCODE and protein–protein interactions from STRING database. (**2**) HeteSim computation based on relevance path of heterogenous network for lncRNA–protein interaction predictions. (**3**) Model validation based on LOOCV and AUC.

In the HeteSim algorithm [94], relevance paths are defined. A relevance path P, denoted as $A_1 \xrightarrow{R_1} \ldots \xrightarrow{R_l} A_{l+1}$, is a path defined over the schema $T_G = (A, R)$. A composite relation $R = R_1 \circ R_2 \circ \cdots \circ R_l$ between node types A_1 and A_{l+1} is revealed by the symbolization of the relevance path, where $\circ$ denotes the composition operator of relations. For a given relevance path $R = R_1 \circ R_2 \circ \cdots \circ R_l$, HeteSim can measure the similarity between two objects x and y ($x \in R_1.X$ and $y \in R_1.Y$) according to the relevance score:

$$
\begin{aligned}
&HeteSim(x,y\big|R_1 \circ R_2 \circ \cdots \circ R_{l-1} \circ R_l) = \frac{1}{\big|O(x|R_1)\big|\big|I(y|R_l)\big|}, \\
&\sum_{I(v|R_l)}^{O(x|R_1)} HeteSim(O_i(x|R_1), I_j(y|R_l)\big|R_1 \circ R_2 \circ \cdots \circ R_{l-1} \circ R_l).
\end{aligned} \tag{9}
$$

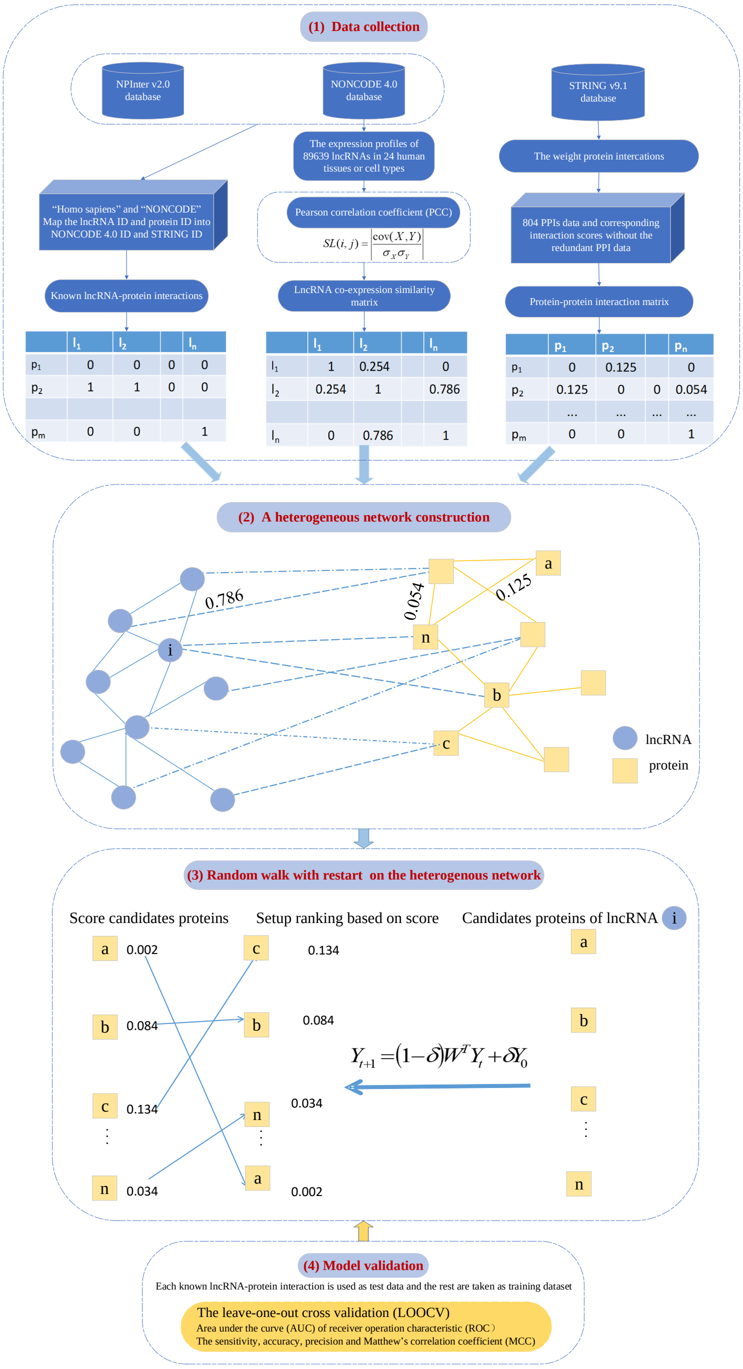
$O(x|R_1)$ represents the out-neighbors of x based on relation $R_1 \bigcup R_2$, and $I(y|R_l)$ represents the neighbors of y based on relation $R_{l-1} \circ R_l$. In fact, x and y can also be the same type according to the random walk model pair. For an arbitrary relevance path $P = A_1A_2 \cdots A_{l+1}$, the HeteSim relevance between any two objects $a \in A_1$ and $b \in A_{l+1}$ is the corresponding component in the score matrix named HeteSim $(A_1, A_{l+1}\big|P)$. Finally, the relatedness between A_1 and A_{l+1} in the relevance path $P = A_1A_2 \cdots A_{l+1}$ is defined as follows:

$$
\begin{aligned}
&HeteSim(A_1, A_{l+1}\big|P) \\
&= HeteSim(A_1, A_{l+1}\big|P_LP_R) \\
&= PM_{P_R} * PM'_{P_R^{-1}} \\
&= U_{A_1A_2} \cdots U_{A_{mid-1}M}V_{MA_{mid+1}} \cdots V_{A_lA_{l+1}} \\
&= U_{A_1A_2} \cdots U_{A_{mid-1}M}U'_{A_{mid+1}M} \cdots U'_{A_{l+1}A_l} \\
&= U_{A_1A_2} \cdots U_{A_{mid-1}M}(U_{A_{l+1}A_l} \cdots U_{A_{mid+1}M}).
\end{aligned} \tag{10}
$$

Based on the random walk model [37], P is divided into two equal path lengths P_L and P_R, where $P_L = A_1A_2 \cdots A_{mid-1}M$ and $P_R = MA_{mid+1} \cdots A_{l+1}$. Depending on whether the length of P is even or odd, the node type of M is impacted differently. If the length of P is even, M is the middle position node type, which could be one of A. Otherwise, it is just the defined middle type. P_R is equal to P_L^{-1}. The transition probability matrix of $A_i \rightarrow A_j$ denoted as $U_{A_iA_j}$ is the normalized matrix of the adjacent matrix $W_{A_iA_j}$ that contains the row vector, and the transition probability matrix of $A_i \rightarrow A_j$ denoted as $V_{A_iA_j}$ is the normalized matrix of $W_{A_iA_j}$ that contains the column vector. It easily proves that $V_{A_iA_j}$ is equal to $U'_{A_iA_j}$. Finally, the score between two objects is normalized to ensure that the correlation between the same objects is 1. Based on HeteSim algorithm in the heterogeneous network of lncRNA–protein, the lncRNA–protein-related pathway is considered. In this network, a group of data is randomly extracted from the measured data as a training dataset, and the rest of the data are used as the test dataset. The AUC of HeteSim achieved on the lncRNA–protein–protein path is 0.879.

4.4. LPIHN: LncRNA–Protein Interaction Prediction Based on Heterogeneous Network Models

Based on this assumption, interrelated lncRNAs tend to exhibit interaction patterns that have similarities with proteins. Li et al. [34] proposed the network-based computational method LPIHN for predicting new lncRNA–protein interactions. The LPIHN procedure is shown in Figure 5. A heterogeneous network is constructed, which is integrated by a similarity network containing lncRNA–lncRNA expression data, a lncRNA–protein interaction network and a PPI network. The similarity network containing lncRNA–lncRNA expression data is calculated by the Pearson's correlation coefficient [97–102] between the expression profiles of each lncRNA–lncRNA interaction. The lncRNA–protein interaction network is constructed from NPInter, by Shang et al. [103], who made a detailed and comprehensive analysis of it. The protein–protein interaction network is not a single source; it is based on computational prediction methods, and some of the interaction data are obtained through high-throughput experiments, from the STRING v9.1 database [104] to text mining, data obtained from the three weighted protein interaction degrees. Then, they walk randomly over the heterogeneous network to infer and predict the interaction between new lncRNAs and proteins.


(1) Data collection
NPInter v2.0 database
NONCODE 4.0 database
STRING v9.1 database
The expression profiles of 89639 lncRNAs in 24 human tissues or cell types
The weight protein intercations
"Homo sapiens" and "NONCODE" Map the lncRNA ID and protein ID into NONCODE 4.0 ID and STRING ID
Pearson correlation coefficient (PCC)
$SL(i,j)=\left|\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}\right|$
804 PPIs data and corresponding interaction scores without the redundant PPI data
Known lncRNA-protein interactions
LncRNA co-expression similarity matrix
Protein-protein interaction matrix
(2) A heterogeneous network construction
0.786
0.054
0.125
a
n
i
b
c
lncRNA
protein
(3) Random walk with restart on the heterogenous network
Score candidates proteins
Setup ranking based on score
Candidates proteins of lncRNA i
a 0.002
b 0.084
c 0.134
n 0.034
c 0.134
b 0.084
n 0.034
a 0.002
a
b
c
n
$Y_{t+1}=(1-\delta)W^TY_t+\delta Y_0$
(4) Model validation

	l_1	l_2		l_n
p_1	0	0	0	0
p_2	1	1	0	0
p_m	0	0		1

	l_1	l_2		l_n
l_1	1	0.254		0
l_2	0.254	1		0.786
l_n	0	0.786		1

	p_1	p_2		p_n
p_1	0	0.125		0
p_2	0.125	0	0	0.054
	…	…	…	…
p_m	0	0		1

Each known lncRNA-protein interaction is used as test data and the rest are taken as training dataset

The leave-one-out cross validation (LOOCV)
Area under the curve (AUC) of receiver operation characteristic (ROC）
The sensitivity, accuracy, precision and Matthew's correlation coefficient (MCC)

Figure 5. Pipeline of LPIHN, containing three modules: (**1**) Data collection: lncRNA–protein interactions from NPInter, protein–protein interactions from STRING database and lncRNA–lncRNA similarity network computed based on lncRNA expression profile from NONCODE. (**2**) A heterogeneous network construction. (**3**) LncRNA–protein interactions prediction based on the random walk with restart. A score is assigned to each candidate protein of a query lncRNA, by the random walk with restart on the heterogeneous network. The candidate proteins are ranked based on the scores. (**4**) Model validations based on LOOCV and AUC. For LPIHN, the lncRNA–lncRNA similarity network is calculated by using the lncRNA expression profiles based on the PCC of each pair of lncRNAs. The heterogeneous network is constructed by connecting the lncRNA–lncRNA similarity network and PPI network together with the known lncRNA–protein interaction network. Blue circles indicated lncRNAs, orange squares indicated proteins, blue edges indicated lncRNA–lncRNA similarities, orange edges indicated protein–protein interactions, and blue dotted edges indicated known lncRNA–protein interactions.

In the RWR procedure [37], an iterative walker starts at a source node with the first probability, and then it can either move to a randomly selected direct neighbor in the process of random walking or restart at a source node with probability δ in each step. Therefore, when random walks are completed on heterogeneous networks, researchers can determine the initial probability, transfer matrix, and restart probability. However, it is based on information provided by heterogeneous networks. During the process of predicting the potential proteins for lncRNA l_i, Y_0 represents the first probability of the walker starting at every node, where l_i and the proteins that are known to interact with l_i are assigned positive values, and the nodes that remain are assigned as zero. It means that the node where the random walk begins is l_i, or that the protein interacts with l_i. Y_i represents the relevance of l_i to all other nodes, where j represents the node and t represents the step. Y_{t+1} can be defined by the following equation:

$$Y_{t+1} = (1-\delta)W^T Y_t + \delta Y_0,$$

where $\delta \in (0,1)$ represents the restart probability of the random walk. W is the transition matrix and Y_0 is the first probability of the random walk. For a given lncRNA l_i, l_i is the seed node in the lncRNA network, the probability of vertex l_i is 1, and other elements in the lncRNA network are assigned as zero, which forms the first probability of the lncRNA network v_0. When protein p_j interacts with lncRNA l_i, p_j becomes the seed node in the protein network. The first probability vector of the protein network u_0 is formed by assigning equal probabilities to the protein seed nodes. For the heterogeneous network, the first probability is

$$Y_0 = \begin{bmatrix} (1-\beta)u_0 \\ \beta v_0 \end{bmatrix}. \tag{11}$$

The parameter $\beta \in (0,1)$ can decide whether to focus more on lncRNA networks or more on protein networks. When $\beta = 0.5$, failure to focus more on one side of a similar network means that the lncRNA–lncRNA similarity network and the PPI network are given the same weight. With $\beta < 0.5$, the random walk tended to return to the protein network. The transition matrix was defined in order to complete the random walk on the heterogeneous network. The authors defined $W = \begin{bmatrix} W_P & W_{PL} \\ W_{LP} & W_L \end{bmatrix}$ as the transition matrix, where W_P is the subnetwork transition matrix showing the probability of the random walker transiting between the protein and another protein in the random walking process. W_L between lncRNA and another lncRNA can be calculated in a similar way. W_{PL} represents the probability of the random walker transiting from the protein network to the lncRNA network, and W_{LP} represents the relationship of the lncRNA network to the protein network. In the process of transition, they defined γ as the probability of the random walker transiting from the protein network to the lncRNA network, where the reverse is also true. W, the probability of the random walker transiting from protein p_i to p_j, is defined as

$$W_P(i,j) = p(p_j|p_i) = \begin{cases} \frac{SP'(i,j)}{\sum_j SP'(i,j)}, & \sum_k I(i,k) = 0 \\ \frac{(1-\gamma)SP'(i,j)}{\sum_j SP'(i,j)}, & \text{otherwise,} \end{cases} \tag{12}$$

where $\sum_k I(i,k) = 0$ means that p_i can bind to multiple lncRNAs and at least one lncRNA, and can be transferred from p_i to a similar network of lncRNA–lncRNA at random. In this case, the probability with γ of p_i transferring to l_i can be further calculated. The probability of p_i transiting to p_j should multiply $1-\gamma$. The probability of transiting from lncRNA l_i to l_j can be defined as:

$$W_L(i,j) = p(l_j|l_i) = \begin{cases} \frac{SL(i,j)}{\sum_j SL(i,j)}, & \sum_k I(k,i) = 0 \\ \frac{(1-\gamma)SL(i,j)}{\sum_j SL(i,j)}, & \text{otherwise.} \end{cases} \tag{13}$$

The probability of transiting from protein p_i to lncRNA l_j is defined as

$$W_{PL}(i,j) = p(l_j|p_i) = \begin{cases} \frac{\gamma I(i,j)}{\sum_j I(i,j)}, & \sum_k I(i,k) \neq 0 \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where $\sum_k I(i,k) \neq 0$ means that p_i is bound to at least one lncRNA, and the walker can transit to the lncRNA–lncRNA network from p_i with probability γ; under that condition, one can further calculate the probability of p_i transiting to l_j. The probability of transiting from lncRNA l_i to protein p_j can be defined in a similar manner as

$$W_{LP}(i,j) = p(p_j|l_i) = \begin{cases} \frac{\gamma I(j,i)}{\sum_j I(i,j)}, & \sum_k I(k,i) \neq 0 \\ 0, & \text{otherwise}. \end{cases} \tag{15}$$

As the first probability Y_0 and the transition matrix W were defined, the RWR procedure [37] could be used for the heterogeneous network. After multiple iterations, the change between Y_t and Y_{t+1} was less than 10^{-10}, which meant that a stable probability $Y_\infty = [(1-\beta)u_\infty,\ \beta v_\infty]^T$ had been obtained.

The result of the LOOCV test showed that the approach could achieve 0.96 with an AUC value. Some predicted interactions between lncRNAs and proteins have been confirmed in recent research studies and databases, indicating the strong influence of LPIHN in predicting lncRNA–protein interactions. In each cross-validation experiment, the test dataset was associated with each known lncRNA–protein interaction, while the rest was used as a training dataset. The method has been successfully reconstructed and possible interactions have been evaluated. In particular, the authors use curves and fold enrichment to measure performance, and it is worth mentioning that the average-fold enrichment of all test data is also used to evaluate the model.

4.5. PLPIHS: Prediction of LncRNA–Protein Interactions Using HeteSim Scores Based on Heterogeneous Networks

Predicting the association between biological molecules based on biological networks has been widely used in many types of research, such as searching for gene sequencing of a disease [27] and predicting drug target interactions. Some of them have achieved good prediction results and good performance. Xiao et al. [35] proposed the PLPIHS method (Figure 6) to predict lncRNA–protein interactions using HeteSim scores and they used a path metric to calculate the interrelationship between nodes in heterogeneous networks. Zeng et al. [105] inferred the association between heterogeneous nodes by means of uniform and symmetric metrics of random paths, regardless of whether they are the same or different types according to the score. Because the relevance path captures the semantic information and also also restricts the wandering path, the score depends on the similarity measure of the path. A heterogeneous network is first constructed with an lncRNA–lncRNA similarity network, which uses the Pearson's correlation coefficient between the expression profiles of each pair of lncRNAs to calculate the lncRNA–protein association network downloaded from GENCODE Release 24 [106] and a PPI network obtained from the STRING v10.0 database [107]. Then, they used the HeteSim to measure the degree of interaction of each lncRNA–protein in the network and showed it in fractions. The SVM classifier is built on the basis of the scores of different paths.

LOOCV is carried out to evaluate the performance of PLPIHS [108]. The results show that the AUC of PLPIHS for the network cutoff value of 0.3 is 96.8%, which is higher than LPIHN. Similarly, PLPIHS outperforms other methods in the 0.5 network and 0.9 network as well. A total of 2000 lncRNA–protein associations from positive samples of the 0.9 network and 2000 interactions from the remaining negative samples of the 0.3 network are randomly selected to construct an independent test set to further conduct the performance evaluation. Using this independent test set, PLPIHS achieves an AUC value of 0.879.

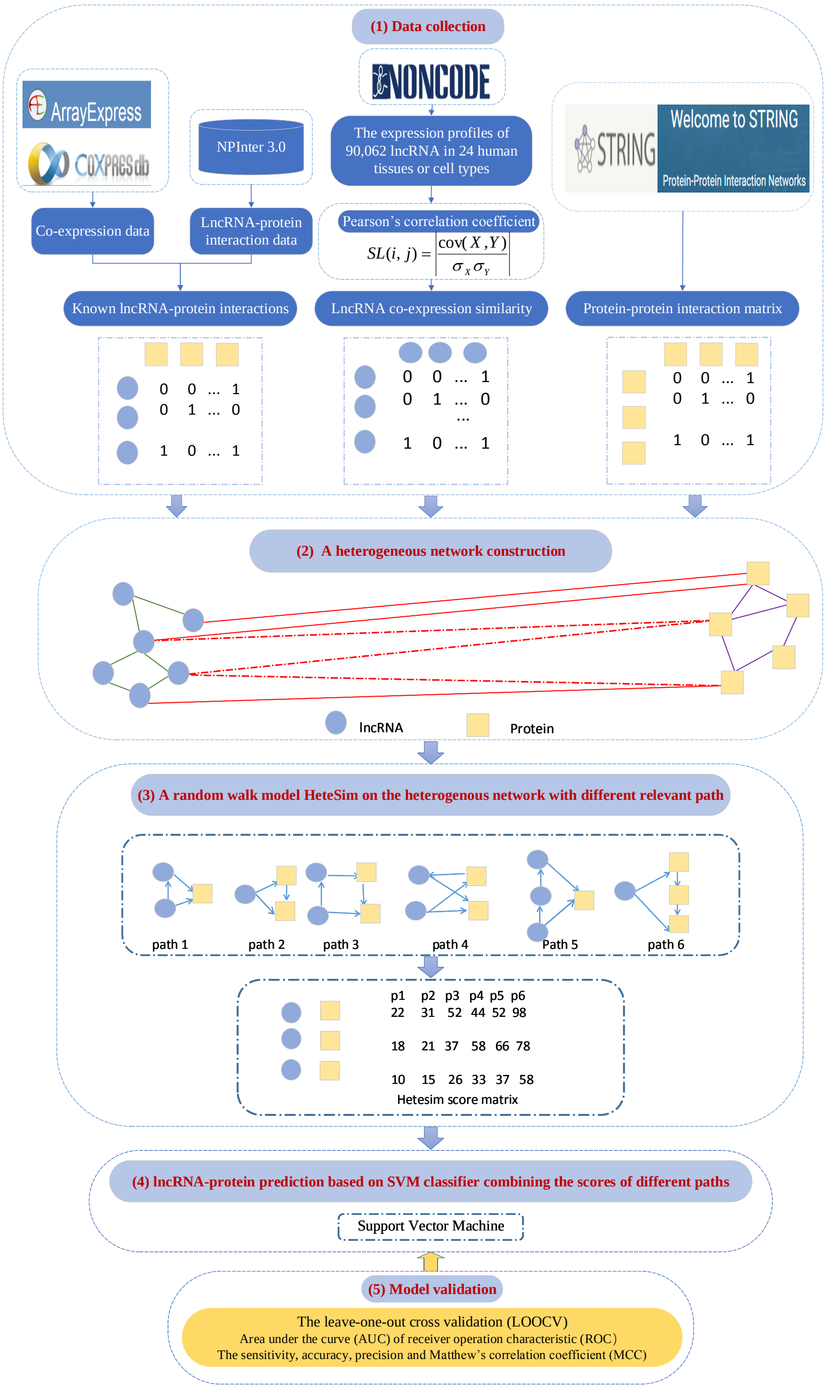

(1) Data collection
ArrayExpress
COXPRESdb
NPInter 3.0
NONCODE
The expression profiles of 90,062 lncRNA in 24 human tissues or cell types
Welcome to STRING
STRING
Protein-Protein Interaction Networks
Co-expression data
LncRNA-protein interaction data
Pearson's correlation coefficient
$SL(i,j)=\left|\frac{\mathrm{cov}(X,Y)}{\sigma_X\sigma_Y}\right|$
Known lncRNA-protein interactions
LncRNA co-expression similarity
Protein-protein interaction matrix
(2) A heterogeneous network construction
lncRNA
Protein
(3) A random walk model HeteSim on the heterogenous network with different relevant path
path 1
path 2
path 3
path 4
Path 5
path 6
p1 p2 p3 p4 p5 p6
22 31 52 44 52 98
18 21 37 58 66 78
10 15 26 33 37 58
Hetesim score matrix
(4) lncRNA-protein prediction based on SVM classifier combining the scores of different paths
Support Vector Machine
(5) Model validation
The leave-one-out cross validation (LOOCV)
Area under the curve (AUC) of receiver operation characteristic (ROC)
The sensitivity, accuracy, precision and Matthew's correlation coefficient (MCC)

Figure 6. Flowchart of PLPIHS, including four modules: (**1**) Data collection. (**2**) Heterogeneous network construction consisting of a lncRNA–lncRNA similarity network, a lncRNA–protein interaction network and a protein–protein interaction network. (**3**) HeteSim measure is used to calculate a score for each lncRNA–protein pair in each path. (**4**) LncRNA–protein prediction based on SVM classifier combining the scores of different paths. (**5**) Model validations based on LOOCV, AUC and MCC.

5. Results of Comparisons of the Network-Based Models for Predicting LncRNA–Protein Interactions

To compare the network-based methods, the fusion of heterogeneous data and performance evaluation were analyzed. All of the above-described methods used LOOCV to validate their respective performances. The test results of the network-based methods are shown in Table 5.

Table 5. Differences in evaluation measures by the network-based methods.

Method	Measure for the Evaluation			Test Dataset	Measurement or Illustration							
	LOOCV	Precision Versus	Fold Enrichment		AUC	SPE	ACC	PRE	MCC	REC	F1-Score	SEN
LPBNI [31]	√		√	4870 lncRNA–protein interactions from NPInter v2.0	0.878	0.99 0.95	0.873 0.880	0.852 0.681	0.449 0.534	–	–	0.288 0.532
Zheng et al. [32]	√			4467 lncRPIs, including 1050 lncRNAs and 84 proteins	AUC values of 15 settings: Seqs-0.8565, Pfam-0.8459, GO-0.8584, STRING-0.7972; Seqs+Pfam-0.8689, Seqs+GO-0.8626, Seqs+STRING-0.8762, Pfam+GO-0.8677, Pfam+STRING-0.8977, and GO+STRING-0.8814; Seqs+Pfam+GO-0.8704, Seqs+Pfam+STRING-0.9023, Seqs+GO+STRING-0.8904, Pfam+GO+STRING-0.9066; Seqs+Pfam+GO+STRING-0.9068.							
Yang et al. [33]	√			MALAT1 with all 99 proteins AK0951949 with all 99 proteins	0.955 0.973	–	–	–	–	–	–	–
LPIHN [34]	√	√	√	The test dataset is the interaction of each known lncRNA–protein, and the rest is used as training dataset.	0.96	√	√	√	√	√		√
PLPIHS [35]	√			The remaining positive samples found in the 0.9 network had 2000 lncRNA–protein interactions and the same number of negative interactions in the 0.3 network	0.879	–	√	√	√		√	√

LOOCV, leave-one-out cross validation; AUC, area under the curve; SPE, specificity; ACC, accuracy; PRE, precision; MCC, Matthew's correlation coefficient; REC, recall; SEN, sensitivity; OMIM, Online Mendelian Inheritance in Man compendium.

Yang et al. [33] proposed that the relevance path is the same as fusing multiple PPSNs. They extracted MALAT1 and AK0951949, respectively, with all 99 proteins as two experimental datasets. Known interactions data between two lncRNAs and their protein chaperones are considered as positive samples, while negative samples are new pairs of lncRNA–protein interactions that have not been experimentally verified. From the ROC curves of the prediction results, the AUC is 0.955 for MALAT1 with all 99 proteins and 0.973 for AK0951949 with all 99 proteins.

LPBNI obtained 4870 lncRNA–protein interactions data from NPInter 2.0. The method used the propagation matrix and the lncRNA–protein interaction networks to set the test sample. First, the test sample is set according to the interaction pair of each lncRNA–protein in the adjacent matrix, leaving a node and setting one at the zero corresponding value of the adjacent matrix. In this process, some nodes will be deleted during the evaluation process. Considering that these nodes do not have more than two connection nodes, it is considered that there is no information dissemination between them. Compared with random walk with restart, it is clear that LPBNI showed the highest true positive rate in each false positive rate, and the AUC value was 0.878.

PLPIHS selected data samples in different cutoff values of networks and obtained 2000 positive samples from 0.9 network and randomly selected 2000 negative samples from 0.3 network. PLPIHS calculated the AUC in different network cutoff values (0.3 and 0.9), where that for the 0.3 network was 0.968, which was higher than the value calculated by LPIHN. To verify that PLPIHS has better performance, the authors select the same number of positive and negative samples from different cutoff values of the network, respectively, and use this random selection to construct independent test datasets. Compared with the values generated by LPIHN, the AUC value of PLPIHS was 0.879. The accuracy, sensitivity, precision, Matthew's correlation coefficient, and F1-Score were also chosen as measurements to evaluate performance.

Fusing multiple PPSNs to effectively predict lncRNA–protein interactions was from the perspective of a fusion protein. The best relevance path was lncRNA–protein–protein according to HeteSim. The fusion matrix is an effective means for users to get more reliable and richer information matrix or network. The best AUC value was 0.9068 with Go+Pfam, Go+String, and Pfam+String. The AUC values of the 15 settings implemented in the paper by Zheng et al. [32] are shown in Table 5, which included using only one similarity matrix, fusing two similarity matrices, fusing three similarity matrices, and fusing all four similarity matrices.

In the LPIHN model, the determination of test datasets is consistent with other interaction prediction methods, leaving a cross-validation method. This model used not only LOOCV but also precision versus recall curves and fold enrichment to measure the performance, whereas the average fold enrichment of all test data was used for assessment. The LOOCV results showed that LPIHN obtained an AUC of 0.96. When more attention was paid to the predicted first 4870 lncRNA–protein interactions, 802 of the predicted LPIHN interactions were within this ranking.

To better understand the performance of network-based computational models to predict lncRNA–protein interactions, we divided the heterogeneous network into three cases according to the source of components: (1) only the lncRNA–protein interaction network; (2) the network combining the interactions of lncRNA–protein and protein–protein; and (3) the network with the integration of the interactions of lncRNA–protein, protein–protein and lncRNA–lncRNA. For each case, different methods were validated with the same set of test datasets, and the performances are compared by AUC in Figure 7. LPBNI (green) used leave-one-out cross validation on 4796-lncRNA–protein interaction network. The method proposed by Yang et al. [33] and method (orange) by Zheng et al. [32] used leave-one-out cross validation on 4467 lncRNA–protein interaction networks. The remaining two methods (blue) used leave-one-out cross validation on the dataset which 2000 lncRNA–protein interactions from network of PLPIHS with cutoff of 0.9 were extracted as positive samples, 2000 negative samples were randomly selected in 0.3 network. The gold set containing 185 lncRNA–protein interactions downloaded from NPInter database. In Figure 7, different colors represent different network types, and the same color bar graphs represent the verification results under the same set of data. In Figure 7, the performance of the method is better when the heterogeneous network is composed by more sources. When heterogeneous networks are constructed by the same sources, the performance will be better for the heterogeneous networks constructed by weighted networks. (The implication of more data here can be illustrated by the interactions of lncRNA–lncRNA. The interactions of lncRNA–lncRNA can be considered from many perspectives. It can be calculated from expression profile data, sequence alignment or experiment.) For example, the method proposed by Yang et al. and method (orange) by Zheng et al. both integrated the interactions of lncRNA–protein and protein–protein to construct a heterogeneous network, and both methods were based on the relevant path of HeteSim random walk in the heterogeneous network. However, for protein–protein interaction networks construction, Yang et al. only considered the protein–protein interactions from STRING database. Zheng et al. considered not only the protein–protein interactions from STRING database, but also the sequence similarity, functional annotation semantic similarity and protein domain similarity protein–protein interactions constructed based on. The method (orange) by Zheng et al. with AUC 0.9068 is better than the method proposed by Yang et al. with AUC 0.7972.

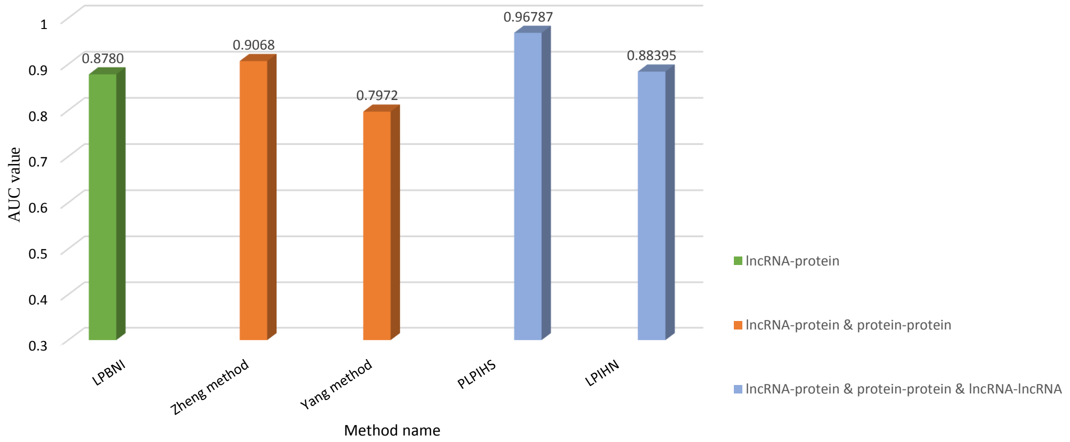


Figure 7. The AUC value of five methods under at three different levels of heterogeneous networks. Different colors represent different network cases, and the same color bar graphs represent the verification results on the same set of data.

6. Discussion

Prediction of the interactions between lncRNAs and proteins is a very important step for research about lncRNAs. Based on the results of lncRNA–protein interactions, the functions as well as the associated diseases of lncRNAs can be inferred. The lncRNA–protein interaction is a very significant molecular mechanism. Computational approaches to predict lncRNA–protein interactions can be grouped into two broad categories. The first category is based on intrinsic features of the lncRNAs and proteins, including the sequence, structural information, and physicochemical property. The second category is based on the fusion of heterogeneous data to construct a network.

Whereas the sequence-based methods only consider the properties of the RNA and neglect the internal relationship between the lncRNAs and proteins, the network-based methods have paid more attention to this kind of internal relationship. The main advantage of a network-based computational model is that it can predict lncRNA–protein interactions with heterogeneous data. Predictions using the intrinsic features of sequences alone may lead to more false-positive interaction pairs than that obtained using a network-based method. Unavoidably, the network-based computational model can have some disadvantages. The prediction of the network-based computational model can be affected when it is carried out in the case of finite interactions. When there are no interaction data, the network-based computational model cannot be used to predict interactions.

New lncRNA–protein interactions are predicted more effectively by using several kinds of heterogeneous data sources. As the study of proteins becomes ever more comprehensive, the proposed effective computational models for predicting lncRNA–protein interactions from heterogeneous biological data can benefit our understanding of more comprehensive annotations for lncRNAs.

Currently, there is very limited information on the interaction between lncRNAs and proteins, but computational methods can provide us with a large number of interaction pairs that can be further regarded as valuable material for the inference of lncRNA functions. First, a great deal of lncRNA–protein interactions can be provided by computational models based on intrinsic features. Second, since predictions using the intrinsic features of sequences alone may lead to some false-positive interaction pairs, computational models based on biological networks can be chosen to obtain more reliable predictions. In the future, a deep-learning-based framework can be considered to optimize the sequence-based and network-based computational models. Hopefully, long-short-term memory models can be employed to build a more advanced framework to build classifiers and achieve a more reliable classification model. We also can integrate machine learning with ab initio computing and network representation learning methods, and apply them to the prediction model of relationships between biological macromolecules. The interactions between lncRNAs and other molecules may enrich the functional annotations of lncRNAs. First, researchers can extract the characteristics of the molecules themselves by machine learning algorithm, and then they can use the appropriate algorithm in network representation learning to represent the feature vectors of relationships between nodes in heterogeneous networks. In this way, researchers can not only understand the internal features of molecules, but also not ignore the hidden topological information between molecules. This will overcome the weakness of most current research methods which only consider ab initio prediction or network-based methods.

Author Contributions: Conceptualization, Y.-C.L.; methodology, H.Z., C.P., S.H. and Y.L.; validation, H.Z.; formal analysis, Y.L.; writing—original draft preparation, H.Z.; writing—review and editing, Y.L., C.P. and S.H.; supervision, Y.-C.L.; and funding acquisition, Y.L. and Y.-C.L.

Funding: This work was supported by the National Natural Science Foundation of China (61472158, 61402194, and 71774154), Natural Science Foundation of Jilin Province (20180101331JC and 20180101050JC), Zhuhai Premier-Discipline Enhancement Scheme, Guangdong Premier Key-Discipline Enhancement Scheme, and Graduate Innovation Fund of Jilin University (2017124).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Nakaya, H.I.; Amaral, P.P.; Louro, R.; Lopes, A.; Fachel, A.A.; Moreira, Y.B.; El-Jundi, T.A.; da Silva, A.M.; Reis, E.M.; Verjovski-Almeida, S. 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*, R43. [CrossRef] [PubMed]
2. Guttman, M.; Amit, I.; Garber, M.; French, C.; Lin, M.F.; Feldser, D.; Huarte, M.; Zuk, O.; Carey, B.W.; Cassady, J.P.; et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* **2009**, *458*, 223–227. [CrossRef] [PubMed]
3. Guttman, M.; Garber, M.; Levin, J.Z.; Donaghey, J.; Robinson, J.; Adiconis, X.; Fan, L.; Koziol, M.J.; Gnirke, A.; Nusbaum, C.; et al. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.* **2010**, *28*, 503–510. [CrossRef] [PubMed]
4. Wang, K.C.; Chang, H.Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **2011**, *43*, 904–914. [CrossRef] [PubMed]
5. Lu, Q.; Ren, S.; Lu, M.; Zhang, Y.; Zhu, D.; Zhang, X.; Li, T. Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genom.* **2013**, *14*. [CrossRef]
6. Zhao, T.; Xu, J.; Liu, L.; Bai, J.; Xu, C.; Xiao, Y.; Li, X.; Zhang, L. Identification of cancer-related lncRNAs through integrating genome, regulome and transcriptome features. *Mol. BioSyst.* **2015**, *11*, 126–136. [CrossRef]
7. Wilusz, J.E.; Sunwoo, H.; Spector, D.L. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.* **2009**, *23*, 1494–1504. [CrossRef] [PubMed]

8. Managadze, D.; Rogozin, I.B.; Chernikova, D.; Shabalina, S.A.; Koonin, E.V. Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol. Evol.* **2011**, *3*, 1390–1404. [CrossRef]
9. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921. [CrossRef]
10. Qureshi, I.A.; Mattick, J.S.; Mehler, M.F. Long non-coding RNAs in nervous system function and disease. *Brain Res.* **2010**, *1338*, 20–35. [CrossRef]
11. Liao, Q.; Liu, C.; Yuan, X.; Kang, S.; Miao, R.; Xiao, H.; Zhao, G.; Luo, H.; Bu, D.; Zhao, H.; et al. Large-scale prediction of long non-coding RNA functions in a coding-non-coding gene co-expression network. *Nucleic Acids Res.* **2011**, *39*, 3864–3878. [CrossRef] [PubMed]
12. Moran, V.A.; Perera, R.J.; Khalil, A.M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res.* **2012**, *40*, 6391–6400. [CrossRef] [PubMed]
13. Zhu, J.J.; Fu, H.J.; Wu, Y.G.; Zheng, X.F. Function of lncRNAs and approaches to lncRNA-protein interactions. *Sci. China Life Sci.* **2013**, *56*, 876–885. [CrossRef] [PubMed]
14. Mercer, T.R.; Dinger, M.E.; Mattick, J.S. Insights into functions. *Nat. Rev. Genet.* **2009**, *10*, 155–159. [CrossRef] [PubMed]
15. Li, X.; Wu, Z.; Fu, X.; Han, W. LncRNAs: Insights into their function and mechanics in underlying disorders. *Mutat. Res./Rev. Mutat. Res.* **2014**, *762*, 1–21. [CrossRef] [PubMed]
16. Chen, G.; Wang, Z.; Wang, D.; Qiu, C.; Liu, M.; Chen, X.; Zhang, Q.; Yan, G.; Cui, Q. LncRNADisease: A database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.* **2013**, *41*, 983–986. [CrossRef] [PubMed]
17. Ning, S.; Zhang, J.; Wang, P.; Zhi, H.; Wang, J.; Liu, Y.; Gao, Y.; Guo, M.; Yue, M.; Wang, L.; Li, X. Lnc2Cancer: A manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res.* **2016**, *44*, D980–D985. [CrossRef]
18. Ray, D.; Kazan, H.; Chan, E.T.; Castillo, L.P.; Chaudhry, S.; Talukder, S.; Blencowe, B.J.; Morris, Q.; Hughes, T.R. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nat. Biotechnol.* **2009**, *27*, 667–670. [CrossRef]
19. Keene, J.D.; Komisarow, J.M.; Friedersdorf, M.B. RIP-Chip: The isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protoc.* **2006**, *1*, 302–307. [CrossRef] [PubMed]
20. Licatalosi, D.D.; Mele, A.; Fak, J.J.; Ule, J.; Kayikci, M.; Chi, S.W.; Clark, T.A.; Schweitzer, A.C.; Blume, J.E.; Wang, X.; et al. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **2008**, *456*, 464–469. [CrossRef]
21. Hafner, M.; Landthaler, M.; Burger, L.; Khorshid, M.; Hausser, J.; Berninger, P.; Rothballer, A.; Ascano, M.; Jungkamp, A.C.; Munschauer, M.; et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell* **2010**, *141*, 129–141. [CrossRef]
22. Pietrosanto, M.; Mattei, E.; Helmer-citterich, M. A novel method for the identification of conserved structural patterns in RNA: From small scale to high-throughput applications. *Nucleic Acids Res.* **2016**, *44*, 8600–8609. [CrossRef]
23. Polishchuk, M.; Paz, I.; Kohen, R.; Mesika, R.; Yakhini, Z. A combined sequence and structure based method for discovering enriched motifs in RNA from in vivo binding data. *Methods* **2017**, *118–119*, 73–81. [CrossRef]
24. Muppirala, U.K.; Honavar, V.G.; Dobbs, D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinform.* **2011**, *12*, 489. [CrossRef]
25. Wang, Y.; Chen, X.; Liu, Z.P.; Huang, Q.; Wang, Y.; Xu, D.; Zhang, X.S.; Chen, R.; Chen, L. De novo prediction of RNA–protein interactions from sequence information. *Mol. BioSyst.* **2013**, *9*, 133–142. [CrossRef]
26. Bellucci, M.; Agostini, F.; Masin, M.; Tartaglia, G.G. Predicting protein associations with long noncoding RNAs. *Nat. Methods* **2011**, *8*, 444–445. [CrossRef]
27. Suresh, V.; Liu, L.; Adjeroh, D.; Zhou, X. RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res.* **2015**, *43*, 1370–1379. [CrossRef]
28. Akbaripour-Elahabad, M.; Zahiri, J.; Rafeh, R.; Eslami, M.; Azari, M. rpiCOOL: A tool for in silico RNA-protein interaction detection using random forest. *J. Theor. Biol.* **2016**, *402*, 1–8. [CrossRef]
29. Pan, X.; Fan, Y.X.; Yan, J.; Shen, H.B. IPMiner: Hidden ncRNA-protein interaction sequential pattern mining with stacked autoencoder for accurate computational prediction. *BMC Genom.* **2016**, *17*, 1–14. [CrossRef]

30. Yang, C.; Yang, L.; Zhou, M.; Xie, H.; Zhang, C.; Wang, M.D.; Zhu, H. LncADeep: An ab initio lncRNA identification and functional annotation tool based on deep learning. *Bioinformatics* **2018**, *34*, 3825–3834. [CrossRef]
31. Ge, M.; Li, A.; Wang, M. A Bipartite Network-based Method for Prediction of Long Non-coding RNA-protein Interactions. *Genom. Proteom. Bioinform.* **2016**, *14*, 62–71. [CrossRef]
32. Zheng, X.; Wang, Y.; Tian, K.; Zhou, J.; Guan, J.; Luo, L.; Zhou, S. Fusing multiple protein-protein similarity networks to effectively predict lncRNA-protein interactions. *BMC Bioinform.* **2017**, *18*, 420. [CrossRef]
33. Yang, J.; Li, A.; Ge, M.; Wang, M. Prediction of interactions between lncRNA and protein by using relevance search in a heterogeneous lncRNA-protein network. In Proceedings of the 2015 34th Chinese Control Conference (CCC), Hangzhou, China, 28–30 July 2015; pp. 8540–8544. [CrossRef]
34. Li, A.; Ge, M.; Zhang, Y.; Peng, C.; Wang, M. Predicting long noncoding RNA and protein interactions using heterogeneous network model. *BioMed Res. Int.* **2015**, *2015*, 1–11. [CrossRef]
35. Xiao, Y.; Zhang, J.; Deng, L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. *Sci. Rep.* **2017**, *7*, 1–12. [CrossRef]
36. Emmert-Streib, F.; Tripathi, S.; Simoes, R.d.M.; Hawwa, A.F.; Dehmer, M. The human disease network. *Syst. Biomed.* **2013**, *1*, 20–28. [CrossRef]
37. Bauer, S.; Horn, D.; Robinson, P.N. Walking the interactome for prioritization of candidate disease genes. *AJHG* **2008**, *82*, 949–958. [CrossRef]
38. Barabási, A.L.; Gulbahce, N.; Loscalzo, J. Network medicine: A network-based approach to human disease. *Nat. Rev. Genet.* **2011**, *12*, 56–68. [CrossRef]
39. Chen, X.; Yan, C.C.; Zhang, X.; You, Z.H. Long non-coding RNAs and complex diseases: From experimental results to computational models. *Brief. Bioinform.* **2016**, *18*, bbw060. [CrossRef]
40. Chen, X.; Yan, G.Y. Novel human lncRNA-disease association inference based on lncRNA expression profiles. *Bioinformatics* **2013**, *29*, 2617–2624. [CrossRef]
41. Chen, X.; Sun, Y.z.; Guan, N.n.; Qu, J.; Huang, Z.a.; Zhu, Z.x.; Li, J.q. Computational models for lncRNA function prediction and functional similarity calculation. *Brief. Functional Genom.* **2019**, *18*, 58–82. [CrossRef]
42. Liu, Y.; Zeng, X.; He, Z.; Zou, Q. Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2017**, *14*, 905–915. [CrossRef]
43. Zeng, X.; Liu, L.; Lu, L. Prediction of potential disease-associated microRNAs using structural perturbation method. *Bioinformatics* **2018**, *34*, 2425–2432. [CrossRef] [PubMed]
44. Zeng, X.; Zhang, X.; Zou, Q. Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks. *Brief. Bioinform.* **2016**, *17*, 193–203. [CrossRef]
45. Chen, X.; Wang, L.; Qu, J.; Guan, N.n.; Li, J.-Q. Predicting miRNA-disease association based on inductive matrix completion. *Bioinformatics* **2018**, *34*, 4256–4265. [CrossRef] [PubMed]
46. Chen, X.; Xie, D.; Wang, L.; Zhao, Q.; You, Z.h.; Liu, H. Systems biology BNPMDA: Bipartite network projection for MiRNA-disease association prediction. *Bioinformatics* **2018**, *34*, 3178–3186. [CrossRef]
47. Chen, X.; Yin, J.; Qu, J.; Huang, L. MDHGI: Matrix Decomposition and Heterogeneous Graph Inference for miRNA-disease association prediction. *PLoS Comput. Biol.* **2018**, 1–25.[CrossRef]
48. Chen, X.; Huang, L. LRSSLMDA: Laplacian regularized sparse subspace learning for MiRNA-disease association prediction. *PLoS Comput. Biol.* **2017**, 1–29. [CrossRef]
49. Yao, Q.; Xu, Y.; Yang, H.; Shang, D.; Zhang, C.; Zhang, Y.; Sun, Z.; Shi, X.; Feng, L.; Han, J.; Su, F.; Li, C.; Li, X. Global prioritization of disease candidate metabolites based on a multi-omics composite network. *Sci. Rep.* **2015**, *5*, 1–14. [CrossRef]
50. Chen, X.; Yan, C.C.; Zhang, X.; Zhang, X.; Dai, F. Drug-target interaction prediction: Databases, web servers and computational models. *Brief. Bioinform.* **2016**, *17*, 696–712. [CrossRef]
51. Chen, X.; Guan, N.n.; Sun, Y.z.; Li, J.Q.; Qu, J. MicroRNA-small molecule association identification: From experimental results to computational models. *Brief. Bioinform.* **2018**, *16*, 1–15. [CrossRef]
52. Qu, J.; Chen, X.; Sun, Y.Z.; Li, J.Q.; Ming, Z. Inferring potential small molecule-miRNA association based on triple layer heterogeneous network. *J. Cheminform.* **2018**, *10*, 30. [CrossRef] [PubMed]
53. Consortium, T.R. RNAcentral: An international database of ncRNA sequences. *Nucleic Acids Res.* **2015**, *43*, D123–D129. [CrossRef]

54. Zhao, Y.; Li, H.; Fang, S.; Kang, Y.; Wu, W.; Hao, Y.; Li, Z.; Bu, D.; Sun, N.; Zhang, M.Q.; Chen, R. NONCODE 2016: An informative and valuable data source of long non-coding RNAs. *Nucleic Acids Res.* **2016**, *44*, D203–D208. [CrossRef]
55. Cui, T.; Zhang, L.; Huang, Y.; Yi, Y.; Tan, P.; Zhao, Y.; Hu, Y.; Xu, L.; Li, E.; Wang, D. MNDR v2.0: An updated resource of ncRNA-disease associations in mammals. *Nucleic Acids Res.* **2018**, *46*, 371–374. [CrossRef] [PubMed]
56. Zheng, L.l.; Li, J.h.; Wu, J.; Sun, W.j.; Liu, S.; Wang, Z.l.; Zhou, H. deepBase v2.0: Identification, expression, evolution and function of small RNAs, LncRNAs and circular RNAs from deep-sequencing data. *Nucleic Acids Res.* **2016**, *44*, 196–202. [CrossRef]
57. Dinger, M.E.; Pang, K.C.; Mercer, T.R.; Crowe, M.L.; Grimmond, S.M.; Mattick, J.S. NRED: A database of long noncoding RNA expression. *Nucleic Acids Res.* **2009**, *37*, 122–126. [CrossRef]
58. Zhou, K.R.; Liu, S.; Sun, W.j.; Zheng, L.l.; Zhou, H.; Yang, J.h.; Qu, L.h. ChIPBase v2.0: Decoding transcriptional regulatory networks of non-coding RNAs and protein-coding genes from ChIP-seq data. *Nucleic Acids Res.* **2017**, *45*, 43–50. [CrossRef]
59. Bhattacharya, A.; Ziebarth, J.D.; Cui, Y. SomamiR: A database for somatic mutations impacting microRNA function in cancer. *Nucleic Acids Res.* **2013**, *41*, 977–982. [CrossRef]
60. Jiang, Q.; Ma, R.; Wang, J.; Wu, X.; Jin, S.; Peng, J.; Tan, R.; Zhang, T.; Li, Y.; Wang, Y. LncRNA2Function: A comprehensive resource for functional investigation of human lncRNAs based on RNA-seq data. *BMC Genom.* **2015**, *16*, 1–11. [CrossRef]
61. Ning, S.; Yue, M.; Wang, P.; Liu, Y.; Zhi, H.; Zhang, Y.; Zhang, J.; Gao, Y.; Guo, M.; Zhou, D.; et al. LincSNP 2.0: An updated database for linking disease-associated SNPs to human long non-coding RNAs and their TFBSs. *Nucleic Acids Res.* **2017**, *45*, 74–78. [CrossRef]
62. Gong, J.; Liu, W.; Zhang, J.; Miao, X.; Guo, A.Y. lncRNASNP: A database of SNPs in lncRNAs and their potential functions in human and mouse. *Nucleic Acids Res.* **2015**, *43*, 181–186. [CrossRef] [PubMed]
63. Volders, P.J.; Helsens, K.; Wang, X.; Menten, B.; Martens, L.; Gevaert, K.; Vandesompele, J.; Mestdagh, P. LNCipedia: A database for annotated human IncRNA transcript sequences and structures. *Nucleic Acids Res.* **2013**, *41*, 246–251. [CrossRef] [PubMed]
64. Li, A.; Zhang, J.; Zhou, Z.; Wang, L.; Liu, Y.; Liu, Y. ALDB: A domestic-animal long noncoding RNA database. *PLoS ONE* **2015**, *10*, e0124003 [CrossRef]
65. Park, C.; Yu, N.; Choi, I.; Kim, W.; Lee, S. Databases and ontologies lncRNAtor: A comprehensive resource for functional investigation of long non-coding RNAs. *Bioinformatics* **2014**, *30*, 2480–2485. [CrossRef] [PubMed]
66. Zhao, Z.; Bai, J.; Wu, A.; Wang, Y.; Zhang, J.; Wang, Z.; Li, Y.; Xu, J.; Li, X. Co-LncRNA: Investigating the lncRNA combinatorial effects in GO annotations and KEGG pathways based on human RNA-Seq data. *Database* **2015**, *2015*, 1–7. [CrossRef] [PubMed]
67. Chan, W.L.; Huang, H.D.; Chang, J.G. lncRNAMap: A map of putative regulatory functions in the long non-coding transcriptome. *Comput. Biol. Chem.* **2014**, *50*, 41–49. [CrossRef]
68. Gong, J.; Liu, C.; Liu, W.; Xiang, Y.; Diao, L.; Guo, A.y.; Han, L. LNCediting: A database for functional effects of RNA editing in lncRNAs. *Nucleic Acids Res.* **2017**, *45*, 79–84. [CrossRef]
69. Paraskevopoulou, M.D.; Georgakilas, G.; Kostoulas, N.; Reczko, M.; Maragkakis, M.; Dalamagas, T.M.; Hatzigeorgiou, A.G. DIANA-LncBase: Experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. *Nucleic Acids Res.* **2013**, *41*, 239–245. [CrossRef]
70. Jiang, Q.; Wang, J.; Wang, Y.; Ma, R.; Wu, X.; Li, Y. TF2LncRNA: Identifying common transcription factors for a list of lncRNA genes from ChIP-Seq data. *BioMed Res. Int.* **2014**, *2014*. [CrossRef]
71. Xu, Y.; Li, F.; Wu, T.; Xu, Y.; Yang, H.; Dong, Q. LncSubpathway: A novel approach for identifying dysfunctional subpathways associated with risk lncRNAs by integrating lncRNA and mRNA expression profiles and pathway topologies. *Oncotarget* **2017**, *8*, 15453–15469. [CrossRef]
72. Jiang, Q.; Wang, J.; Wu, X.; Ma, R.; Zhang, T.; Jin, S.; Han, Z.; Tan, R.; Peng, J.; Liu, G.; Li, Y.; Wang, Y. LncRNA2Target: A database for differentially expressed genes after IncRNA knockdown or overexpression. *Nucleic Acids Res.* **2015**, *43*, D193–D196. [CrossRef]
73. Zhou, Z.; Shen, Y.; Khan, M.R.; Li, A. Original article LncReg: A reference resource for lncRNA-associated regulatory networks. *Database* **2015**, *2015*, 1–7. [CrossRef] [PubMed]

74. Quek, X.C.; Thomson, D.W.; Maag, J.L.V.; Bartonicek, N.; Signal, B.; Clark, M.B.; Gloss, B.S.; Dinger, M.E. lncRNAdb v2.0: Expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.* **2015**, *43*, D168–D173. [CrossRef] [PubMed]
75. Yuan, J.; Wu, W.; Xie, C.; Zhao, G.; Zhao, Y.; Chen, R. NPInter v2.0: An updated database of ncRNA interactions. *Nucleic Acids Res.* **2014**, *42*, 104–108. [CrossRef] [PubMed]
76. Lewis, B.A.; Walia, R.R.; Terribilini, M.; Ferguson, J.; Zheng, C.; Honavar, V.; Dobbs, D. PRIDB: A protein-RNA interface database. *Nucleic Acids Res.* **2011**, *39*, 277–282. [CrossRef] [PubMed]
77. Sussman, J.L.; Lin, D.; Jiang, J.; Manning, N.O.; Prilusky, J.; Ritter, O.; Abola, E.E. Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **1998**, *54*, 1078–1084. [CrossRef]
78. Li, J.H.; Liu, S.; Zhou, H.; Qu, L.H.; Yang, J.H. starBase v2.0: Decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res.* **2014**, *42*, D92–D97. [CrossRef]
79. Narayanan, B.C.; Westbrook, J.; Ghosh, S.; Petrov, A.I.; Sweeney, B.; Zirbel, C.L.; Leontis, N.B.; Berman, H.M. The nucleic acid database: New features and capabilities. *Nucleic Acids Res.* **2014**, *42*, 114–122. [CrossRef]
80. Chen, X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. *Sci. Rep.* **2015**, *5*, 1–12. [CrossRef]
81. Jimeno-Yepes, A.J.; Sticco, J.C.; Mork, J.G.; Aronson, A.R. GeneRIF indexing: Sentence selection based on machine learning. *BMC Bioinform.* **2013**, *14*, 1–11. [CrossRef]
82. Zhong, Y.; Xuan, P.; Wang, X.; Zhang, T.; Li, J.; Liu, Y.; Zhang, W. A non-negative matrix factorization based method for predicting disease-associated miRNAs in miRNA-disease bilayer network. *Bioinformatics* **2018**, *34*, 267–277. [CrossRef] [PubMed]
83. Li, Y.; Qiu, C.; Tu, J.; Geng, B.; Yang, J.; Jiang, T.; Cui, Q. HMDD v2.0: A database for experimentally supported human microRNA and disease associations. *Nucleic Acids Res.* **2014**, *42*, 1070–1074. [CrossRef]
84. Jiang, Q.; Wang, Y.; Hao, Y.; Juan, L.; Teng, M.; Zhang, X.; Li, M.; Wang, G.; Liu, Y. miR2Disease: A manually curated database for microRNA deregulation in human disease. *Nucleic Acids Res.* **2009**, *37*, 98–104. [CrossRef]
85. Xie, B.; Ding, Q.; Han, H.; Wu, D. MiRCancer: A microRNA-cancer association database constructed by text mining on literature. *Bioinformatics* **2013**, *29*, 638–644. [CrossRef]
86. Chou, C.H.; Chang, N.W.; Shrestha, S.; Hsu, S.D.; Lin, Y.L.; Lee, W.H.; Yang, C.D.; Hong, H.C.; Wei, T.Y.; Tu, S.J.; et al. miRTarBase 2016: Updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res.* **2016**, *44*, D239–D247. [CrossRef] [PubMed]
87. Dimmer, E.C.; Huntley, R.P.; Alam-Faruque, Y.; Sawford, T.; O'Donovan, C.; Martin, M.J.; Bely, B.; Browne, P.; Chan, W.M.; Eberhardt, R.; et al. The UniProt-GO annotation database in 2011. *Nucleic Acids Res.* **2012**, *40*, 565–570. [CrossRef] [PubMed]
88. Piñero, J.; Bravo, Á.; Queralt-Rosinach, N.; Gutiérrez-Sacristán, A.; Deu-Pons, J.; Centeno, E.; García-García, J.; Sanz, F.; Furlong, L.I. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **2017**, *45*, D833–D839. [CrossRef]
89. Law, V.; Knox, C.; Djoumbou, Y.; Jewison, T.; Guo, A.C.; Liu, Y.; MacIejewski, A.; Arndt, D.; Wilson, M.; Neveu, V.; et al. DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **2014**, *42*, 1091–1097. [CrossRef]
90. Wishart, D.S.; Jewison, T.; Guo, A.C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; et al. HMDB 3.0—The human metabolome database in 2013. *Nucleic Acids Res.* **2013**, *41*, 801–807. [CrossRef] [PubMed]
91. Mattingly, C.J.; Colby, G.T.; Forrest, J.N.; Boyer, J.L. The Comparative Toxicogenomics Database (CTD). *Environ. Health Perspect.* **2003**. [CrossRef]
92. Kuhn, M.; Letunic, I.; Jensen, L.J.; Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **2016**, *44*, D1075–D1079. [CrossRef] [PubMed]
93. Gruber, A.R.; Lorenz, R.; Bernhart, S.H.; Neuböck, R.; Hofacker, I.L. The Vienna RNA websuite. *Nucleic Acids Res.* **2008**, *36*, 70–74. [CrossRef]
94. Shi, C.; Kong, X.; Huang, Y.; Yu, P.S.; Wu, B. HeteSim: A general framework for relevance measure in heterogeneous networks. *IEEE Trans. Knowl. Data Eng.* **2014**, *26*, 2479–2492. [CrossRef]

95. Wang, B.; Mezlini, A.M.; Demir, F.; Fiume, M.; Tu, Z.; Brudno, M.; Haibe-Kains, B.; Goldenberg, A. Similarity network fusion for aggregating data types on a genomic scale. *Nat. Methods* **2014**, *11*, 333–337. [CrossRef]
96. Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.C. Bipartite network projection and personal recommendation. *Phys. Rev. E* **2007**, *76*, 1–7. [CrossRef]
97. Chen, X.; Ba, Y.; Ma, L.; Cai, X.; Yin, Y.; Wang, K.; Guo, J.; Zhang, Y.; Chen, J.; Guo, X.; et al. Characterization of microRNAs in serum: A novel class of biomarkers for diagnosis of cancer and other diseases. *Cell Res.* **2008**, *18*, 997–1006. [CrossRef]
98. Wang, F.; Zheng, Z.; Guo, J.; Ding, X. Correlation and quantitation of microRNA aberrant expression in tissues and sera from patients with breast tumor. *Gynecol. Oncol.* **2010**, *119*, 586–593. [CrossRef]
99. Ganegoda, G.U.; Wang, J.X.; Wu, F.X.; Li, M. Prioritization of candidate genes based on disease similarity and protein's proximity in PPI networks. In Proceedings of the 2013 IEEE International Conference on Bioinformatics and Biomedicine, IEEE BIBM 2013, Shanghai, China, 18–21 December 2013; pp. 103–108. [CrossRef]
100. Tang, X.; Wang, J.; Zhong, J.; Pan, Y. Predicting essential proteins basedon weighted degree centrality. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2014**, *11*, 407–418. [CrossRef]
101. Li, M.; Zheng, R.; Zhang, H.; Wang, J.; Pan, Y. Effective identification of essential proteins based on priori knowledge, network topology and gene expressions. *Methods* **2014**, *67*, 325–333. [CrossRef]
102. Li, M.; Zhang, H.; Wang, J.x.; Pan, Y. A new essential protein discovery method based on the integration of protein-protein interaction and gene expression data. *BMC Syst. Biol.* **2012**, *6*, 15. [CrossRef]
103. Shang, D.; Yang, H.; Xu, Y.; Yao, Q.; Zhou, W.; Shi, X.; Han, J.; Su, F.; Su, B.; Zhang, C.; et al. A global view of network of lncRNAs and their binding proteins. *Mol. BioSyst.* **2015**, *11*, 656–663. [CrossRef] [PubMed]
104. Franceschini, A.; Szklarczyk, D.; Frankild, S.; Kuhn, M.; Simonovic, M.; Roth, A.; Lin, J.; Minguez, P.; Bork, P.; von Mering, C.; et al. STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **2012**, *41*, D808–D815. [CrossRef] [PubMed]
105. Zeng, X.; Liao, Y.; Liu, Y.; Zou, Q. Prediction and validation of disease genes using HeteSim scores. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2017**, *14*, 687–695. [CrossRef] [PubMed]
106. Derrien, T.; Johnson, R.; Bussotti, G.; Tanzer, A.; Djebali, S.; Tilgner, H.; Guernec, G.; Martin, D.; Merkel, A.; Knowles, D.G.; et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. *Genome Res.* **2012**, 1775–1789. [CrossRef]
107. Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K.P.; et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **2015**, *43*, D447–D452. [CrossRef]
108. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection a study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montreal, QC, Canada, 20–25 August 1995.