 doi:10.20944/preprints202105.0240.v1

Review

A Survey of Current Resources to Study lncRNA-protein Interactions.

Melcy Philip [1], Tyrone Chen [1] and Sonika Tyagi [2,3]*

[1] School of Biological Sciences, Monash University, 25 Rainforest Walk, Clayton, VIC 3800, Australia.
[2] Department of Infectious Disease, Monash University, 85 Commercial Road, 3004, VIC, Australia
[3] Monash eResearch Centre, Monash University, 15 Innovation Walk, Clayton, VIC 3800, Australia
* Correspondence: sonika.tyagi@monash.edu

Abstract: Phenotypes are driven by regulated gene expression, which in turn are mediated by complex interactions between diverse biological molecules. Protein-DNA interactions such as histone and transcription factor binding are well studied, along with RNA-RNA interactions in short RNA silencing of genes. In contrast, lncRNA-protein interaction (LPI) mechanisms are comparatively unknown, likely driven by the difficulties in studying LPI. However, LPI are emerging as key interactions in epigenetic mechanism, playing a role in development and disease. Their importance is further highlighted by their conservation across kingdoms. Hence, interest in LPI research is increasing. We therefore review the current state of the art in lncRNA-protein interactions. We specifically surveyed recent computational methods and databases which researchers can exploit for LPI investigation. We discovered that algorithm development is heavily reliant on a few generic databases containing curated LPI information. We show that early methods predict LPI using molecular docking, have limited scope and are slow, creating a data processing bottleneck. Recently, machine learning has become the strategy of choice in LPI prediction, likely due to the rapid growth in machine learning infrastructure and expertise. While many of these methods have notable limitations, machine learning is expected to be the basis of modern LPI prediction algorithms.

Keywords: LPI, lncRNA, ncRNA, protein, transcriptomics, molecular docking, machine learning, deep learning, databases

1. Introduction

The introduction should briefly place the study in a broad context and highlight why it is important. Transcriptomics is the study of a complete set of RNA transcripts in a cell, measuring variable expression levels of the genome under different conditions. Modern transcriptomics is performed with high throughput sequencing to investigate the function of genes and biological pathways, commonly with bioinformatics methods applying differential gene expression analyses, splice site identification, transcript variant identification or determining alternative promoter usage for protein-coding transcripts [1]. However, these protein-coding transcripts only represent a small proportion of the transcriptome. A large proportion of the genome generates RNA transcripts which do not directly code for protein products [2]. These non-coding RNA (ncRNA) transcripts have been known to exist, but their properties make them difficult to characterize compared to coding transcripts. ncRNA can be divided into multiple categories based on function and length [3]. In this review, we specifically consider the long non-coding RNA (lncRNA) category of ncRNA and their interaction with proteins, an important functional mechanism of lncRNA.

LncRNA are very broadly defined as RNA transcripts exceeding 200 nucleotides (nt) in length without coding potential. Their length varies widely, ranging from hundreds to thousands of nucleotides [4]. LncRNA can act as a gene regulator, and like other epigenetic mechanisms are involved in numerous biological processes. They achieve their

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

regulatory function with their ability to interact with a wide range of biological molecules, such as other nucleic acids and proteins [5], as well as with small molecules [6]. Among their more direct modes of action are sequestering and releasing transcript to modulate gene expression, stabilizing transcript and binding to DNA to sterically hinder transcription initiation [7]. More indirectly, they can recruit proteins and other molecules to form a functional complex, or act as a scaffold for targeted chromatin formation [8].

An important layer of lncRNA-mediated gene regulation is LPI (lncRNA-protein interactions). We illustrate the importance of LPI in developmental and abiotic stress pathways with several examples encompassing multiple distinct species. In *Drosophila melanogaster*, regulatory networks mediated by LPI regulate key eye development [9] and dosage compensation pathways [10] mediated by RNA binding proteins. In the plant *Arabidopsis thaliana*, LPI controls alternative splicing within the nucleus by selectively displacing existing transcripts and subsequently altering root development [11 and12]. Response to abiotic stress is also governed by LPI, as shown by a lncRNA recruiting histone methylases to suppress *Arabidopsis thaliana* flowering during cold conditions [13]. *Dario renio* LPI are also observed to interface with transcription factors and other RNA-binding proteins during embryonic development, although their exact mechanism of action is not well known [14]. LPI also act as mediators of other epigenetic mechanisms, for instance as chromatin scaffolds to organize the three-dimensional structure of the genome in *Mus musculus* [15]

Due to the widespread involvement of LPI in epigenetics, dysregulation of certain LPI contributes to disease states, particularly cancers. Severity of a human pancreatic cancer phenotype is driven by a lncRNA-protein complex, which triggers a positive feedback loop of protein overexpression leading to poor patient outcomes [16]. Similarly, formation of a lncRNA-protein complex is associated with poorer prognosis in breast cancer [17], colon cancer [17] and lymphoma [18] by blocking phosphorylation sites, stabilizing other epigenetic factors, and through an unknown mechanism, respectively. Infectious diseases are also associated with LPI dysregulation, including COVID-19 [19, 20]. A more exhaustive list of known LPI-disease associations is available at the LncTarD database [21]. Despite the wealth of information on LPI-disease associations, their precise mechanism of action remains unknown. Therefore, insight into LPI will be valuable in complex disease research, potentially resulting in improved diagnosis and treatment procedures.

Multiple high-throughput laboratory assays were developed to investigate LPI, some of which will be briefly discussed in this review article. However, exhaustively performing an experimental validation for each individual LPI is not practical given their volume and variety. Hence, computational methods are necessary to screen these high throughput assays for potential LPI which can then be subsequently experimentally validated, similar to transcriptomics workflows for conventional protein-coding RNA [22]. A variety of these computational LPI predictors exist, each applying different strategies to achieve their goals, and are dependent on a few biological databases containing subsets of experimentally validated LPI. In this review, we will discuss recent bioinformatics resources for studying LPI, with an emphasis on software and databases.

2. LPI laboratory assays

Because of the biological importance of LPI, many laboratory assays were developed to identify these interactions. Two general categories of such assays exist, protein-centric assays and RNA-centric assays, which can capture either the cellular environment of a living cell or extracted biological material [23]. Protein-centric assays target the protein component of a LPI, while RNA-centric assays target the lncRNA component. Each method varies in sensitivity and specificity, has different prerequisites and has unique advantages as well as disadvantages. Comprehensively comparing and contrasting these laboratory assays is out of scope of this review, but we provide a high-level overview only

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

to give the computational methods discussed in this article some biological context. A more detailed overview of these assays can be found in a separate review article [23].

To discover proteins bound to RNA of interest (RNA-centric methods), IVT *(in vitro* transcribed) RNA can be tagged with biotin, and selectively bound to streptavidin for purification [24]. RaPID (RNA–protein interaction detection) [25] operates in a conceptually similar way to the previous method. IVT RNA can also be tagged with dyes and bound to protein microarrays, with fluorescence providing a quantitative output [26]. *In vivo,* cross-linking RNA with protein, either through formaldehyde or UV light, is used to identify LPI by purifying and extracting the RNA-bound proteins. CHART (capture hybridization analysis of RNA targets) [27], ChIRP (Chromatin isolation by RNA purification and capture hybridization analysis of RNA targets) [28], MS2-BioTRAP (MS2 in vivo biotin-tagged RAP) [29], PAIR (peptide-nucleic-acid-assisted identification of RBPs) [30], RAP (RNA affinity purification) [31] and TRIP (tandem RNA isolation procedure) [32] all use either of these cross-linking strategies.

To discover RNA bound to proteins of interest (protein-centric methods), exploiting cross-linking is also common. The largest group of protein-centric methods are CLIP (cross-linking immunoprecipitation) based methods [33]. Many variants of CLIP methods exist [34], and when paired with high throughput sequencing are capable of generating libraries of data for further analysis. RIP-seq (RNA Immunoprecipitation) [35] and TRIBE (targets of RNA-binding proteins identified by editing) [36] also belong to this category of protein-centric methods.

3. LncRNA - protein Resource Databases

Starbase [37], POSTAR [38], RAIN [39], RNAInter [40], NPInter [41], ATtRACT [42] and oRNAment [43] are examples of databases that contain information associated with lncRNA-protein interactions obtained by the previously discussed laboratory assays. Two broad classes exist, databases containing curated **lncRNA-protein interactions** and databases **containing RNA-binding motifs.**

Starbase, RNAInter, POSTAR, NPInter and RAIN all contain details of curated lncRNA-protein interactions, and many additional attributes (including functional annotation) associated with the interactions, derived from a combination of the laboratory assays discussed in the previous section [Table S1]. These are not limited exclusively to lncRNA, and contain various other interaction information, including interactions with other ncRNA, other nucleic acids and proteins [44,45,46]. Some contrasts between these databases are also observable from a species, usability and scope perspective, which will be discussed here. Starbase, POSTAR and RAIN contain LPI information from a small number (two to four) of species, while RNAInter and NPInter host a wide range of species. To improve usability, Starbase, RNAInter and RAIN feature third party tool integration to streamline bioinformatics workflows. In terms of scope, POSTAR and NPInter appear to be focused on disease phenotypes, providing disease association information, while Starbase, RNAInter and RAIN have a more generic focus.

ATtRACT and oRNAment databases contain details of RBP (RNA-binding protein) motifs. While not directly containing LPI, these can be applied to predict putative LPI and are a useful starting point or supplementary tool in screening for LPI.

All databases feature at least mouse and human datasets, likely due to their status as model organisms relevant to human disease, although some incorporate other model organisms as well. It is interesting to note that all databases feature advanced querying and search functions, likely reflecting the volume and complexity of LPI data. We have reviewed and compared them in Table S1 [Table S1]. In summary, we discovered that there

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

is a surprising lack of specialized LPI databases, with most databases featuring combinations of other nucleic acid and protein combinations.

4. LPI prediction algorithms

Most LPI prediction algorithms exploit these curated databases of prior LPI knowledge to tune their predictions. Computational strategies for LPI prediction can be divided into two high-level categories, molecular docking and machine learning. Lower-level subdivisions among the methods we surveyed are visualized in Figure 1, and include deep learning, tree-based methods, graph-based methods, similarity networks, image segmentation, matrix factorization and variants of the Fourier transform. Conventional molecular docking methods operate by finding the optimal configuration of a lncRNA-protein complex, and ranking the highest scoring configurations for further evaluation. Within the past decade, a large number of prediction algorithms based on machine learning have emerged. Most machine learning methods do not involve molecular docking simulations. Instead, they exploit known interactions between lncRNA and protein and/or biomolecular sequence information directly, although many also leverage known secondary structures to improve their performance Table [1,2]. As with the LPI databases, it is worth noting that none of these methods are tuned specifically for LPI prediction, and represent broader scopes of identifying combinations of nucleic acid-protein interaction.

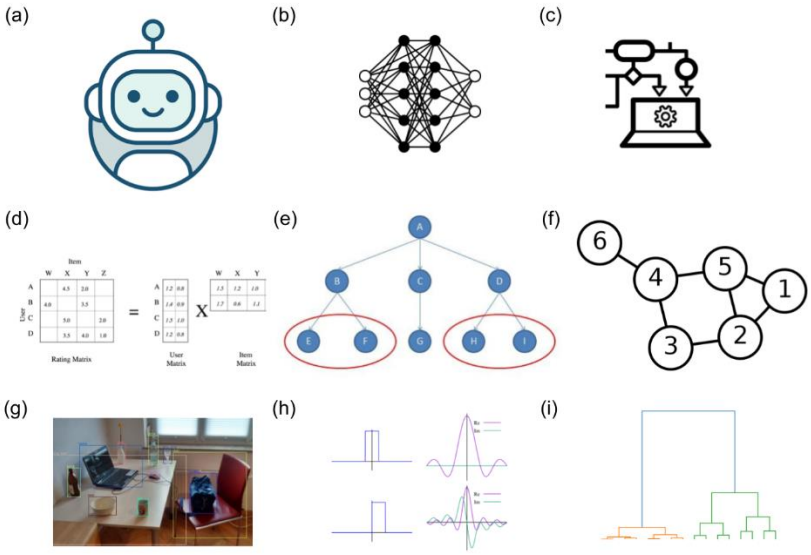

Figure 1. Visualization of the broad categories of strategies used for predicting lncRNA-protein interactions. (a) Machine learning, (b) deep learning, (c) ensemble learning, (d) matrix factorization, (e) similarity network analysis, (f) graph theory, (g) segmentation, (h) Fourier transform (in lncRNA-protein molecular docking simulations) and (i) hierarchical clustering. Training data is commonly higher-level features (e.g., structure, orientation) of lncRNA and proteins as well as the sequences recoded into tensors of varying dimensions.

5. Molecular docking approaches

Before the current ecosystem of machine learning algorithms was established, molecular docking was the dominant strategy used to predict and investigate LPI or RNA - protein interactions in general. By developing custom equations, which account for conformation and other steric properties, the likelihood of lncRNA-protein complex formation is scored. Low-level methodology does not vary significantly, with most methods applying a variant of the FFT (Fast Fourier transform) to extract features from three-dimensional molecule representations, template or optimizing for a minimal energy state. Key factors considered include docking pose, distance and area of interracial sites, energy-based criteria, and selection of the most structurally conserved docked complex [47]. Several methods also account for sequence homology or electrical charge between biological

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

molecules [48]. Hierarchical clustering to group complexes of interest is not uncommon. However, at a high-level these strategies are applied in different ways, and on different steric features. In many cases, a set of parameters must be specified by the user.

Most of the molecular docking methods we reviewed use methods which incorporate at least two of the previously discussed low-level methodologies [Table 1]. To provide some context for the building blocks of these more complex methods, we first present examples of methods that use an individual strategy, which include 3dRPC [49], HexServer [50], FireDOCK [51], HADDOCK [52] and PatchDOCK [53]. 3dRPC and HexServer are FFT-based methods. 3dRPC exploits the fact that LPI complexes have looser packing, and implements FFT on geometric complementarity and electrostatics with a custom scoring function. HexServer uses an FFT-based algorithm to exploit shape complementarity as a feature for optimization. Its key advantage is its reformulation of the conventional 3D search space to greatly boost the speed of the FFT, achieving results in seconds. Meanwhile, FireDOCK and HADDOCK optimize the minimum free energy of the lncRNA-protein complex. While FireDOCK and focuses on exploiting side chain information, HADDOCK leverages ambiguous interaction restraints, and is one of the few methods which can generalize to multi-body problems as well as other biomolecular interactions. Among molecular docking tools, PatchDOCK takes a more unconventional strategy by summarizing low-level geometric features into higher level features, and has some conceptual similarities to image segmentation. It is interesting to note that FireDOCK and PatchDOCK both complement each other, where PatchDOCK can feed output directly into FireDOCK.

Methods implementing a mixture of these strategies include HDOCK [54], MPRDOCK [55], P3DOCK [56] and NPDOCK [57]. HDOCK integrates template-based modeling as well as ab initio free docking, with a scope that extends to both proteins and nucleic acids. In addition, the user may specify binding sites of interest directly. MPRDOCK exploits protein flexibility by applying FFT and considering sequence homology of the target of interest to generate a repertoire of structures for "ensemble docking". We note that in this specific context of MPRDOCK, "ensemble docking" refers to the library of proteins generated by MPRDOCK, and is distinct from "ensemble learning" in the machine learning section [65,66,67] where the output of multiple algorithms are aggregated to obtain a result. P3DOCK (http://www.rna-binding.com/P3DOCK/P3DOCK.html) integrates the previously discussed 3dRPC, as PRIME that leverages sequence as well as structural homology in addition to the features used by 3dRPC. P3DOCK's authors claim that by complementing free docking and template-based docking strategies in a hybrid approach, a more accurate classification is possible. Finally, NPDOCK does not use a hybrid or ensemble strategy, but chains multiple methods into a pipeline of tools, which implement mostly FFT-based methods.

Table 1. A comparison of molecular docking tools used to predict lncRNA-protein interactions. Important attributes of these molecular docking tools, including their effectiveness and a link to their corresponding server are listed

.Sl:No	Resource	Resource type	Comment	Weblink	Reference paper
1	P3DOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	Free docking and template-based docking strategies in a hybrid approach, results in an accurate classification	http://www.rna-binding.com/P3DOCK/P3DOCK.html	[56]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

2	HDOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	Integrates template-based modeling as well as ab initio free docking, with a scope that extends to both proteins and nucleic acids	http://hdock.phys.hust.edu.cn/	[54]
3	PATCHDOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	Low-level geometric features into higher level features, FireDOCK and PatchDOCK both complement each other, where PatchDOCK can feed output directly into FireDOCK.	https://bioinfo3d.cs.tau.ac.il/PatchDock/	[53]
4	FIREDOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	Focuses on exploiting side chain information, optimise the minimum free energy of the lncRNA-protein complex	http://bioinfo3d.cs.tau.ac.il/FireDock/	[51]
5	NPDOCK	Exclusively LncRNA - protein docking server, developed for nucleic acid docking only	Chains multiple methods into a pipeline of tools, which implement mostly FFT-based methods.	http://genesilico.pl/NPDock	[57]
6	HADDOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	It everages ambiguous interaction restraints, and it can generalise to multi-body problems as well as other biomolecular interactions,optimise	https://wenmr.science.uu.nl/haddock2.4/	[52]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

			the minimum free energy of the lncRNA-protein complex		
7	MPRDOCK	LncRNA - protein docking server (Adapted from conventional docking servers)	Implies protein flexibility by applying FFT and considering sequence homology of the target of interest to generate a repertoire of structures for “ensemble docking"	http://huanglab.phys.hust.edu.cn/mprdock/	[55]
8	Hexserver	LncRNA - protein docking server (Adapted from conventional docking servers)	FFT-based algorithm to exploit shape complementarity as a feature for optimisation	http://hex-server.loria.fr/	[50]

With the exception of one or two methods such as HexServer, many of these algorithms are computationally expensive and time-consuming (hours to days of real time) to run. Some methods like HexServer require advanced hardware such as GPUs and specialized software engineering tools. Biological molecules are complex and dynamic, with their wide range of possible conformations as well as orientations greatly increasing the search space for algorithms. The molecular docking community is mindful of this, and provides their software on publicly accessible and user-friendly web servers for users to run these programs remotely, although time remains a bottleneck for these workflows.

6. Machine learning approaches

Most modern lncRNA-protein interaction (LPI) prediction algorithms use machine learning, where large datasets with attributes of interest are passed to an algorithm [Table 2]. The algorithm then “learns” from the data, discovering patterns in the data with minimal human intervention such as user-defined equations. In the case of LPI, known LPI and their corresponding sequences as well as structures are used for training the prediction models. Their strategies can be divided into several broad categories, including graph methods, ensemble learning, matrix factorization and deep learning. Of these strategies, matrix factorization appears to be the most popular and is integrated into many other higher-level strategies. LPI are commonly formulated as similarity matrices, which can then be easily formulated as a matrix factorization problem. Broader strategies incorporating matrix factorization, such as ensemble learning and methods which leverage multimodal data appear to have consistently robust performance. Few deep learning models exist, but they both perform and generalize well in comparison to other methods, and are likely to become more popular as they have become in other areas of biology.

Matrix factorization is the most common way to formulate LPI for prediction algorithms, including LPI-FKLKRR (LncRNA-Protein Interaction Kernel Ridge Regression, based on Fast Kernel Learning) [58], LPI-KTASLP (Prediction of LncRNA-Protein

 doi:10.20944/preprints202105.0240.v1

Interaction by Semi-Supervised Link Learning With Multivariate Information) [59], LPI-NRLMF (lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization) [60], LPI-INBRA (Long non-coding RNA-Protein Interaction Prediction based on Improved Bipartite Network Recommender Algorithm) [61] and LPI-BNPRA (Long non-coding RNA-Protein Interaction bipartite network projection recommended algorithm) [62]. These methods share a common theme of formulating lncRNA-protein interactions as a matrix factorization problem and using them in broader strategies such as multiple kernel learning or recommender algorithms. Known structural features are often used together with sequence features. In the special case of LPI-FKLKRR, matrices are reformulated into kernels for direct optimization with kernel ridge regression, increasing performance in the common scenario of class imbalance.

Some graph-based methods for LPI prediction are PBLPI (path-based lncRNA-protein interaction) [63] and PLPIHS (Predicting lncRNA-Protein Interactions using HeteSim Scores) [64]. PBLPI takes into account both functional and semantic similarity between proteins, while PLPIHS uses a custom distance metric to unify co-expression, lncRNA-protein interactions and protein-protein interaction scores to construct a network which is then provided to a SVM classifier. Performance is improved by preserving information regarding the biological network, taking into account lncRNA-protein interactions similar to the target.

Examples of hybrid and ensemble learning approaches are IRWNRLPI (Integrating Random Walk and Neighborhood Regularized Logistic Matrix Factorization for lncRNA-Protein Interaction Prediction) [65], SFPEL-LPI (sequence-based feature projection ensemble learning method) [66], HLPI-Ensemble (human lncRNA-protein interactions ensemble) [67], GPLPI (graph predict lncRNA-protein interaction) [68] and LPI-BLS (predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier) [69]. IRWNRPLI uses lncRNA-protein interactions and lncRNA/protein sequence similarity as input into a hybrid approach of random walk and neighborhood regularized logistic matrix factorization. Being an integrative model, it appears to be robust, although its accuracy varies on different biological systems. Ensemble approaches PMKDN, SFPEL-LPI, HLPI-Ensemble and LPI-BLS are all robust against noise due to their ensemble strategy incorporating multiple approaches, and are capable of discovering new LPI. LPI-BLS in particular stands out for its unconventional flat network architecture and aggregation strategy. However, we note that HLPI-Ensemble is specifically intended for human LPI only. GPLPI uses both sequence features and known secondary structures to train a graph-based neural network. In addition, by using an ensemble of features including evolutionary information, GPLPI's effectiveness was increased. An important distinction between these two methods is that GPLPI is trained on known plant lncRNA, and plant non-coding RNA have different properties (some ncRNA lose function even with 1-2 nucleotide changes) to that of animal non-coding RNA [70]. For this model to be effective on non-plant organisms, retraining is likely necessary but viable due to the relatively higher volume of data associated with animals, in particular humans [67].

Only a few deep learning approaches exist, DeepBind [70], LPI-CNNCP (lncRNA–protein interactions convolutional neural network copy-padding trick) [71] and DeepLPI (deep lncRNA-protein interactions) [72]. DeepBind was one of the first applications of deep learning to predict nucleic acid-protein binding, and is applicable to LPI. By reformulating the classical position weight matrix [73] as a convolutional kernel, it operates on raw sequence data to provide a simple prediction score for a nucleic acid-protein interaction [74]. LPI-CNNCP uses only lncRNA and protein sequence data recorded as k-mers as input into a CNN but achieves good results. It is also interesting to note that it appears to be one of the few models that are effective across different species. Meanwhile, DeepLPI feeds co-expression, sequence and structural data to a neural network optimized by a conditional random field. Using protein isoform data makes DeepLPI the only method to date

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

with the ability to predict lncRNA interaction with different protein isoforms. Furthermore, its flexibility allows it to be extended to other biomolecular interactions such as miRNA.

Other methods used to predict LPI that do not fall into a specific category include LPI-SKF (lncRNA-protein interaction similarity kernel fusion) [75], PMKDN (projection-based neighborhood non-negative matrix decomposition model) [76] and LPI-MiRNA [77]. LPI-SKF uses an integrative approach where verified lncRNA-protein interactions are used to build a network, and similarity kernel fusion is used to integrate protein and lncRNA similarity scores before applying manifold learning. PMKDN uses multiple features from lncRNA (nucleotide composition, expression levels) and protein (amino acid subcategories) to build a similarity matrix for similarity network fusion with a nearest neighbor's approach. Both these methods are robust against noise and capable of interaction discovery, but like most methods that express LPI as similarity matrices, they make a strong assumption that sequence homology correlates with interactivity, which may not hold in all cases. LPI-MiRNA takes a unique approach, exploiting miRNA as an intermediate unit of lncRNA-protein binding, and uses this in a network-based approach. While this gives LPI-MiRNA the ability to operate on datasets without prior knowledge of lncRNA interactions, a different limitation is introduced of relying on known miRNA-lncRNA and miRNA-protein interactions. An assumption is also made that miRNAs which interact with both lncRNA and a protein would also form LPI, which may not always hold. Nevertheless, this method was shown to be effective.

lncPro [78] and catRAPID [79] are older methods but are featured in this manuscript because of their historical significance. lncPro was one of the first published machine learning LPI prediction algorithms, and many LPI algorithms resemble it. Higher-level features are extracted from lncRNA and protein sequence, which are then recorded as vectors as input into their model. Although the authors noted limitations associated with data availability and computational complexity at the time, this method became a template for many other machine learning methods, including those discussed in this manuscript. catRAPID does not apply machine learning, but instead constructs an interaction matrix from known secondary structure and other molecular features. A major limitation of this approach is its reliance on obsolete genomic data, which is expected to reduce prediction accuracy.

However, it is important to note that the scope of most LPI prediction algorithms are limited. Not all methods can predict interactions for novel lncRNA or proteins, and few methods generalist across species [62,69,71]. This is partly due to the limited availability of curated training data, with a small number of samples mostly from human or mouse present in a few databases [66,67,69]. LPI prediction for different protein isoforms is also not an active area of prediction algorithm development, with only one method having this functionality. Another limitation observed is that some methods exploit sequence similarity as an intermediate metric for LPI prediction, particularly methods which formulate LPI as similarity matrices. While this appears to be effective within the specific training datasets used by each study, this implicit assumption of similar sequence homology correlating to interactivity may not always hold, especially across different species [80, 81]. At the same time, we consider that small nucleotide changes in biological molecules can cause major functional changes, which can potentially cause improperly trained prediction algorithms to produce misleading results [82].

We also note the limited accessibility of many of these machine learning methods. Among the methods reviewed that were published within the last five years, many do not make their source code publicly available and/or are written in proprietary programming languages such as MATLAB [83]. This restricts reproducibility and prevents usage of more than half of the methods we reviewed [Table 2]. At least, partly because of the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

computational complexity required, machine learning methods which are well suited to resolving non-linear variables in high dimensional data have recently become a focus of the LPI field. Although, computational methods that integrates the identification and functional annotation of LPI are not yet developed or established, which leaves a void that has to be filled.

In contrast to published molecular docking algorithms, only a few methods provide active web servers for convenient use by the community, further raising the barrier for usability by biologists.

Table 2. A comparison of machine learning algorithms used to predict lncRNA-protein interactions. Important attributes of these machine learning algorithms, including their scope, strategies, training data, effectiveness and reproducibility are listed. More than half of these methods are not reproducible as their source code is proprietary or not available. A few methods provide web interfaces for users to enter their own data

Sl:no	Resource	Scope	Comment	Strategy	Problem formulation	Model training data	Web-link/source code	Reference paper
1	LPI-FKLKRR (LncRNA-Protein Interaction Kernel Ridge Regression, based on Fast Kernel Learning)	prediction	Effective in datasets with imbalanced classes.	Kernel Ridge Regression	Similarity matrices formulated as kernels	lncRNA-protein interactions, lncRNA expression, protein ontology, lncRNA sequence, protein sequence	https://github.com/6gbluewind/LPI_FKLKRR	[58]
2	LPI-KTASLP (Prediction of LncRNA-Protein Interaction by Semi-Supervised Link Learning With Multivariate Information)	prediction, discovery	Effective in datasets with imbalanced classes.	Multiple Kernel Learning	Similarity matrices formulated as kernels	lncRNA-protein interactions, lncRNA expression, lncRNA sequence	https://github.com/6gbluewind/LPI_KTASLP	[59]
3	LPI-NRLMF (lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization)	prediction, discovery	Prediction bias is expected due to the sparsity of the training dataset.	Matrix factorisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[60]
4	LPI-INBRA (Long non-coding RNA-Protein Interaction Prediction based on Improved Bipartite Network Recommender Algorithm)	prediction	Robust against false positives.	Matrix factorisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[61]
5	LPI-BNPRA (Long non-coding RNA-Protein Interaction	prediction	Effective in humans and	Bipartite network recommendation	Similarity matrices	lncRNA-protein interactions, lncRNA	NA	[62]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

	bipartite network projection recommended algorithm)		closely related species.			sequence, protein sequence		
6	PBLPI (path-based lncRNA-protein interaction)	prediction, discovery	Prediction accuracy limited due to technical limitations.	Graph	Similarity matrices	lncRNA-protein interactions, Protein semantic similarity, lncRNA functional similarity, Gaussian interaction profile kernel similarity, Integrated similarity for lncRNAs and proteins	NA	[63]
7	PLPIHS (Predicting lncRNA-Protein Interactions using HeteSim Scores)	prediction, discovery	Performance is improved by preserving information regarding the biological network, taking into account lncRNA-protein interactions similar to the target.	Graph	Similarity matrices	Co-expression data of lncRNA-protein pairs, lncRNA-protein interaction data	NA	[64]
8	IRWNRLPI (Integrating Random Walk and Neighborhood Regularized Logistic Matrix Factorization for lncRNA-Protein Interaction Prediction)	prediction	Robust due to hybrid approach, but known to be unstable.	Hybrid: random walk, neighborhood regularised logistic matrix factorisation algorithm	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	NA	[65]
9	SFPEL-LPI (sequence-based feature projection ensemble learning method)	prediction, discovery	Multi-modal approach boosts	Ensemble: graph Laplacian regularisation	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, protein sequence	http://www.bio-in-fotech.cn/SSFLM-LPI/	[66]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

			prediction accuracy.					
10	HLPI-Ensemble (human lncRNA-protein interactions ensemble)	prediction	Scope restricted to human.	Ensemble: Support Vector Machines (SVM), Random Forests (RF) and Extreme Gradient Boosting (XGB)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, lncRNA features, protein sequence, protein features	http://ccsipb.lnu.edu.cn/hlpiensemble/	[67]
11	GPLPI (graph predict lncRNA-protein interaction)	prediction	Scope restricted to plants.	Deep learning, Ensemble learning, Graph attention LSTM-autoencoder	Recoded sequence and structure vectors	lncRNA sequences, protein sequences, structural features from predicted secondary structures from lncRNA and protein sequences.	https://github.com/Mjwl/GPLPI	[68]
12	LPI-BLS (predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier)	prediction	Flat network architecture boosts speed and accuracy. Effective in several model organisms.	Ensemble: Broad learning system (flat neural network)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, lncRNA features, protein sequence, protein features	https://github.com/NWPU-903PR/LPI_BLS	[69]
13	LPI-CNNCP (lncRNA–protein interactions convolutional neural network copy-padding trick)	prediction	Can be extended to predict other biomolecular interactions, effective across different species.	Deep learning (Convolutional Neural Network)	Recoded feature vectors	lncRNA-protein interactions, lncRNA sequence, protein sequence	https://github.com/NWPU-903PR/LPI-CNNCP	[70]
14	DeepLPI (deep lncRNA-protein interactions)	prediction, discovery	Can be extended to other biomolecular interactions, unique	Deep learning (embedding, convolution, LSTM)	Recoded feature tensors	lncRNA-protein interactions, lncRNA sequence, lncRNA structure, protein	https://github.com/dls03/DeepLPI	[72]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

			capability to predict lncRNA interaction with different protein isoforms.			sequence, protein structure		
15	LPI-SKF (lncRNA-protein interaction similarity kernel fusion)	prediction, discovery	Aggregating multiple similarities increases robustness against noise.	Similarity Kernel Fusion, Manifold learning	Similarity matrices	lncRNA-protein interactions, pairwise similarities for lncRNAs, pairwise similarities for proteins	https://github.com/zyk2118216069/LPI-SKF	[75]
16	PMKDN (projection-based neighborhood non-negative matrix decomposition model)	prediction	Strategy avoids overfitting and sparsity issues, allowing more generalisability to different datasets.	Neighborhood regularised matrix factorisation algorithm	Similarity matrices	lncRNA-protein interactions, lncRNA sequence, lncRNA expression, protein sequence, protein annotation	NA	[76]
17	LPI-MiRNA	prediction, discovery	Can operate on datasets without prior knowledge of lncRNA interactions but relies on known miRNA-lncRNA and miRNA-protein interactions.	Heterogeneous network model	Similarity matrices	lncRNA–miRNA interactions, protein–miRNAs interactions	https://github.com/zyk2118216069/LncRNA-protein-interactions-prediction	[77]
18	lncPro	prediction	Training dataset limited,	Fourier transform, matrix factorisation	Recoded feature tensors	lncRNA-protein interactions, lncRNA	http://cmbi.bjmu.edu.cn/lncpro	[78]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 May 2021 doi:10.20944/preprints202105.0240.v1

			effective on short sequences.			sequence, lncRNA features, protein sequence, protein features		
19	catRAPID	prediction	Visualization is available, prediction accuracy may be limited by reliance on very old lncRNA annotations.	Discrete Fourier transform	lncRNA and protein secondary structure, hydrogen bonding, van der Waals forces	NA	http://s.tartaglialab.com/page/catrapid_group	[79]
20	3dRPC	prediction	Effective on well-characterised molecules, may have lower accuracy if this is not the case.	Fast Fourier transform, Root Mean Square Deviation	conformations of nucleotide-amino-acid pairs	NA	http://biophy.hust.edu.cn/3dRPC	[49]
21	DeepBind	prediction	Effective, generalisable across species, but more effective at predicting protein-DNA binding than protein-RNA binding.	Deep learning (Convolutional Neural Network)	Recoded feature tensors	lncRNA-protein interactions, lncRNA sequence, protein sequence	http://tools.genes.toronto.edu/deepbind/	[70]

7. Conclusions

LPI forms a unique layer of gene regulation across many species, and a growing interest in the field has resulted in the creation and expansion of curated databases as well as LPI prediction algorithms. Here, we are reviewing some of the established (older than five years) and recent (within the last five years) LPI prediction approaches as well as databases. We note four important points. First, there has been a clear and recent shift from conventional molecular docking algorithms to machine learning methods, which attempts the direct prediction of LPI from biomolecular sequence identity and higher-level features. This shift to machine learning is observable across different fields of biology and

 doi:10.20944/preprints202105.0240.v1

is likely to continue with the rising availability of computational infrastructure and machine learning expertise. Secondly, these methods are heavily dependent on a set of curated data across several databases. Across these databases, a lack of universal standardization complicates data merging [84], preventing the community from unlocking the full potential of LPI data, in contrast to conventional transcriptomics databases such as SRA [85], EBI [86] and DDBJ [87]. This is in part due to the diversity of assays used to capture the LPI information, as well as the scope of the databases, which may subsequently bias the machine learning algorithms developed on these data. Third, there is a distinct lack of methods and databases which are specifically designed for LPI's unique properties, with most having a generic scope despite LPI's biological significance. Finally, it is concerning that more than half of the recent machine learning methods we surveyed are not reproducible or usable due to the absence of their source code. However, LPI acts as an important but less-studied regulatory layer and understanding them will provide key context to deepen our understanding of biological systems.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/ Table S1: LncRNA-protein data repositories (Table-1 S1). Seven databases, four with LPI information and three with RNA motif information are surveyed. Each database holds information on at least one combination of nucleic acid and protein interaction. The number of species each database contains varies widely, from 4-154. Every database contains at least human and mouse data, and has been updated within the past five years.

Author Contributions: Conceptualization, ST; methodology, MP, TC, ST.; formal analysis, MP, TC, ST; resources, ST; data curation, MP, TC.; writing—original draft preparation, MP, TC.; writing—review and editing, MP, TC, ST.; visualization, TC, MP; supervision, ST.; project administration, ST.; funding acquisition, ST. All authors have read and agreed to the published version of the manuscript.

Funding: S. T acknowledges the AISRF EMCR Fellowship by the Australian Academy of Science and Australian Women Research Success Grant at Monash University. T. C received funding from the Australian Government Research Training Program Scholarship and Monash Faculty of Science Dean's Postgraduate Research Scholarship.

Informed Consent Statement: "Not applicable."

Acknowledgments: We thank Yashpal Ramakrishnaiah for his proofreading and feedback on this article. The authors thank the HPC team at Monash eResearch Centre for their continuous personnel support. This work was supported by the \href{www.massive.org.au}{MASSIVE HPC facility}. \href{https://www.monash.edu/indigenous-australians/about-us/recognising-traditional-owners}

We acknowledge and pay respects to the Elders and Traditional Owners of the land on which our 4 Australian campuses stand.

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

Appendix A

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data is shown in the main text can be added here if brief, or as Supplementary data. Mathematical proofs of results not central to the paper can be added as an appendix.

References

1. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457. Published 2017 May 18. doi:10.1371/journal.pcbi.1005457
2. Guo JC, Fang SS, Wu Y, et al. CNIT: a fast and accurate web tool for identifying protein-coding and long non-coding transcripts based on intrinsic sequence composition. Nucleic Acids Res. 2019;47(W1):W516-W522. doi:10.1093/nar/gkz400
3. Wang C, Wang L, Ding Y, Lu X, Zhang G, Yang J, Zheng H, Wang H, Jiang Y, Xu L. LncRNA Structural Characteristics in Epigenetic Regulation. International Journal of Molecular Sciences. 2017; 18(12):2659. https://doi.org/10.3390/ijms18122659

4. Kazimierczyk M, Kasprowicz MK, Kasprzyk ME, Wrzesinski J. Human Long Noncoding RNA Interactome: Detection, Characterization and Function. Int J Mol Sci. 2020;21(3):1027. Published 2020 Feb 4. doi:10.3390/ijms21031027
5. Jalali S, Bhartiya D, Lalwani MK, Sivasubbu S, Scaria V (2013) Systematic Transcriptome Wide Analysis of lncRNA-miRNA Interactions. PLoS ONE 8(2): e53823. https://doi.org/10.1371/journal.pone.0053823
6. Kazimierczyk M, Kasprowicz MK, Kasprzyk ME, Wrzesinski J. Human Long Noncoding RNA Interactome: Detection, Characterization and Function. International Journal of Molecular Sciences. 2020; 21(3):1027. https://doi.org/10.3390/ijms21031027
7. Li, J., Chen, Y., Xu, X. et al. HNRNPK maintains epidermal progenitor function through transcription of proliferation genes and degrading differentiation promoting mRNAs. Nat Commun 10, 4198 (2019). https://doi.org/10.1038/s41467-019-12238-x
8. Fang Y, Fullwood MJ. Roles, Functions, and Mechanisms of Long Non-coding RNAs in Cancer. Genomics Proteomics Bioinformatics. 2016 Feb;14(1):42-54. doi: 10.1016/j.gpb.2015.09.006. Epub 2016 Feb 12. PMID: 26883671; PMCID: PMC4792843.
9. Lo Piccolo L, Mochizuki H, Nagai Y. The lncRNA hsrω regulates arginine dimethylation of human FUS to cause its proteasomal degradation in Drosophila. J Cell Sci. 2019;132(20):jcs236836. Published 2019 Oct 23. doi:10.1242/jcs.236836
10. Militti C, Maenner S, Becker PB, Gebauer F. UNR facilitates the interaction of MLE with the lncRNA roX2 during Drosophila dosage compensation. Nat Commun. 2014;5:4762. Published 2014 Aug 27. doi:10.1038/ncomms5762
11. Bardou F, Ariel F, Simpson CG, et al. Long noncoding RNA modulates alternative splicing regulators in Arabidopsis. Dev Cell. 2014;30(2):166-176. doi:10.1016/j.devcel.2014.06.017
12. Rigo R, Bazin J, Romero-Barrios N, et al. The Arabidopsis lncRNA ASCO modulates the transcriptome through interaction with splicing factors. EMBO Rep. 2020;21(5):e48977. doi:10.15252/embr.201948977
13. Zhao X, Li J, Lian B, Gu H, Li Y, Qi Y. Global identification of Arabidopsis lncRNAs reveals the regulation of MAF4 by a natural antisense RNA. Nat Commun. 2018;9(1):5056. Published 2018 Nov 29. doi:10.1038/s41467-018-07500-7
14. Huang C, Zhu B, Leng D, Ge W, Zhang XD. Long noncoding RNAs implicated in embryonic development in Ybx1 knockout zebrafish. FEBS Open Bio. 2021;11(4):1259-1276. doi:10.1002/2211-5463.13057
15. Zhao T, Cai M, Liu M, et al. lncRNA 5430416N02Rik Promotes the Proliferation of Mouse Embryonic Stem Cells by Activating Mid1 Expression through 3D Chromatin Architecture. Stem Cell Reports. 2020;14(3):493-505. doi:10.1016/j.stemcr.2020.02.002
16. Li N, Yang G, Luo L, et al. lncRNA THAP9-AS1 Promotes Pancreatic Ductal Adenocarcinoma Growth and Leads to a Poor Clinical Outcome via Sponging miR-484 and Interacting with YAP. Clin Cancer Res. 2020;26(7):1736-1748. doi:10.1158/1078-0432.CCR-19-0674
17. Liu B, Sun L, Liu Q, et al. A cytoplasmic NF-κB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis. Cancer Cell. 2015;27(3):370-381. doi:10.1016/j.ccell.2015.02.004
18. Kim SH, Kim SH, Yang WI, Kim SJ, Yoon SO. Association of the long non-coding RNA MALAT1 with the polycomb repressive complex pathway in T and NK cell lymphoma. Oncotarget. 2017;8(19):31305-31317. doi:10.18632/oncotarget.15453
19. Turjya RR, Khan MA, Mir Md Khademul Islam AB. Perversely expressed long noncoding RNAs can alter host response and viral proliferation in SARS-CoV-2 infection. Future Virol. 2020;15(9):577-593. doi:10.2217/fvl-2020-0188
20. Laha S, Saha C, Dutta S, et al. In silico analysis of altered expression of long non-coding RNA in SARS-CoV-2 infected cells and their possible regulation by STAT1, STAT3 and interferon regulatory factors. Heliyon. 2021;7(3):e06395. doi:10.1016/j.heliyon.2021.e06395
21. Zhao H, Shi J, Zhang Y, et al. LncTarD: a manually-curated database of experimentally-supported functional lncRNA-target regulations in human diseases. Nucleic Acids Res. 2020;48(D1):D118-D126. doi:10.1093/nar/gkz985
22. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007
23. Ramanathan M, Porter DF, Khavari PA. Methods to study RNA-protein interactions [published correction appears in Nat Methods. 2019 Apr;16(4):351]. Nat Methods. 2019;16(3):225-234. doi:10.1038/s41592-019-0330-1
24. Faoro, C. & Ataide, S. F. Ribonomic approaches to study the RNA-binding proteome. FEBS Lett. 588, 3649–3664 (2014)
25. Ramanathan, M., Majzoub, K., Rao, D. et al. RNA–protein interaction detection in living cells. Nat Methods 15, 207–212 (2018). https://doi.org/10.1038/nmeth.4601
26. Kretz, M. et al. Control of somatic tissue differentiation by the long non-coding RNA TINCR. Nature 493, 231–235 (2013).
27. Simon, M. D. et al. Te genomic binding sites of a noncoding RNA. Proc. Natl Acad. Sci. USA 108, 20497–20502 (2011).
28. Chu, C., Qu, K., Zhong, F. L., Artandi, S. E. & Chang, H. Y. Genomic maps of long noncoding RNA occupancy reveal principles of RNA-chromatin interactions. Mol. Cell 44, 667–678 (2011).
29. Tsai, B. P., Wang, X., Huang, L. & Waterman, M. L. Quantitative profiling of in vivo–assembled RNA-protein complexes using a novel integrated proteomic approach. Mol. Cell. Proteomics 10, M110.007385 (2011).
30. Zeng, F. et al. A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. Nat. Protoc. 1, 920–927 (2006).
31. McHugh, C. A. & Guttman, M. RAP-MS: a method to identify proteins that interact directly with a specific RNA molecule in cells. Methods Mol. Biol. 1649, 473–488 (2018).
32. Matia-González, A. M., Iadevaia, V. & Gerber, A. P. A versatile tandem RNA isolation procedure to capture in vivo formed mRNA-protein complexes. Methods 118–119, 93–100 (2017).
33. Ule, J. et al. CLIP identifies Nova-regulated RNA networks in the brain. Science 302, 1212–1215 (2003).
34. Kim, B. & Kim, V. N. fCLIP-seq for transcriptomic footprinting of dsRNA-binding proteins: lessons from DROSHA. Methods 152, 3–11 (2019).
35. Nicholson, C. O., Friedersdorf, M. & Keene, J. D. Quantifying RNA binding sites transcriptome-wide using DO-RIP-seq. RNA 23, 32–46 (2017).

36. McMahon, A. C. et al. TRIBE: hijacking an RNA-editing enzyme to identify cell-specific targets of RNA-binding proteins. Cell 165, 742–753 (2016).
37. Quinodoz S, Guttman M. Long noncoding RNAs: an emerging link between gene regulation and nuclear organization. Trends Cell Biol. 2014;24(11):651-663. doi:10.1016/j.tcb.2014.08.009
38. Ulitsky I. Interactions between short and long noncoding RNAs. FEBS Lett. 2018;592(17):2874-2883. doi:10.1002/1873-3468.13085
39. Ramakrishnaiah Y, Kuhlmann L, Tyagi S. Towards a comprehensive pipeline to identify and functionally annotate long noncoding RNA (lncRNA). Comput Biol Med. 2020;127:104028. doi:10.1016/j.compbiomed.2020.104028
40. Li JH, Liu S, Zhou H, Qu LH, Yang JH. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(Database issue):D92-D97. doi:10.1093/nar/gkt1248
41. Hu B, Yang YT, Huang Y, Zhu Y, Lu ZJ. POSTAR: a platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. Nucleic Acids Res. 2017;45(D1):D104-D114. doi:10.1093/nar/gkw888
42. Junge A, Refsgaard JC, Garde C, et al. RAIN: RNA-protein Association and Interaction Networks [published correction appears in Database (Oxford). 2017 Jan 1;2017(1):]. Database (Oxford). 2017;2017:baw167. Published 2017 Jan 10. doi:10.1093/database/baw167
43. Lin Y, Liu T, Cui T, et al. RNAInter in 2020: RNA interactome repository with increased coverage and annotation. Nucleic Acids Res. 2020;48(D1):D189-D197. doi:10.1093/nar/gkz804
44. Teng X, Chen X, Xue H, et al. NPInter v4.0: an integrated database of ncRNA interactions. Nucleic Acids Res. 2020;48(D1):D160-D165. doi:10.1093/nar/gkz969
45. Giudice G, Sánchez-Cabo F, Torroja C, Lara-Pezzi E. ATtRACT-a database of RNA-binding proteins and associated motifs. Database (Oxford). 2016;2016:baw035. Published 2016 Apr 7. doi:10.1093/database/baw035
46. Benoit Bouvrette LP, Bovaird S, Blanchette M, Lécuyer E. oRNAment: a database of putative RNA binding protein target sites in the transcriptomes of model species. Nucleic Acids Res. 2020;48(D1):D166-D173. doi:10.1093/nar/gkz986
47. Meng XY, Zhang HX, Mezei M, Cui M. Molecular docking: a powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011;7(2):146-157. doi:10.2174/157340911795677602.
48. Suravajhala R, Gupta S, Kumar N, Suravajhala P. Deciphering LncRNA-protein interactions using docking complexes [published online ahead of print, 2020 Dec 7]. J Biomol Struct Dyn. 2020;1-8. doi:10.1080/07391102.2020.1850354
49. Huang Y, Li H, Xiao Y. 3dRPC: a web server for 3D RNA-protein structure prediction. Bioinformatics. 2018;34(7):1238-1240. doi:10.1093/bioinformatics/btx742
50. Ghoorah AW, Devignes MD, Smaïl-Tabbone M, Ritchie DW. Protein docking using case-based reasoning. Proteins. 2013;81(12):2150-2158. doi:10.1002/prot.24433
51. Andrusier N, Nussinov R, Wolfson HJ. FireDock: fast interaction refinement in molecular docking. Proteins. 2007;69(1):139-159. doi:10.1002/prot.21495
52. van Zundert GCP, Rodrigues JPGLM, Trellet M, et al. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. J Mol Biol. 2016;428(4):720-725. doi:10.1016/j.jmb.2015.09.014
53. Duhovny D., Nussinov R., Wolfson H.J. (2002) Efficient Unbound Docking of Rigid Molecules. In: Guigó R., Gusfield D. (eds) Algorithms in Bioinformatics. WABI 2002. Lecture Notes in Computer Science, vol 2452. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45784-4_14
54. Yan Y, Tao H, He J, Huang SY. The HDOCK server for integrated protein-protein docking. Nat Protoc. 2020;15(5):1829-1852. doi:10.1038/s41596-020-0312-x
55. He J, Tao H, Huang SY. Protein-ensemble-RNA docking by efficient consideration of protein flexibility through homology models. Bioinformatics. 2019;35(23):4994-5002. doi:10.1093/bioinformatics/btz388
56. Zheng J, Hong X, Xie J, Tong X, Liu S. P3DOCK: a protein-RNA docking web server based on template-based and template-free docking. Bioinformatics. 2020;36(1):96-103. doi:10.1093/bioinformatics/btz478.
57. Tuszynska I, Magnus M, Jonak K, Dawson W, Bujnicki JM. NPDock: a web server for protein-nucleic acid docking. Nucleic Acids Res. 2015;43(W1):W425-W430. doi:10.1093/nar/gkv493.
58. Shen, C.; Ding, Y.; Tang, J.; Guo, F. Multivariate information fusion with fast kernel learning to kernel ridge regression in predicting lncrna-protein interactions. Frontiers in genetics 2019,139, 716.142.
59. Shen, C.; Ding, Y.; Tang, J.; Jiang, L.; Guo, F. LPI-KTASLP: prediction of lncRNA-protein interaction by semi-supervised link learning with multivariate information. IEEE Access 162019,7, 13486–13496.173.
60. Liu, H.; Ren, G.; Hu, H.; Zhang, L.; Ai, H.; Zhang, W.; Zhao, Q. LPI-NRLMF: lncRNA-protein interaction prediction by neighborhood regularized logistic matrix factorization. Oncotarget 192017,8, 103975.204.
61. Xie, G.; Wu, C.; Sun, Y.; Fan, Z.; Liu, J. Lpi-ibnra: Long non-coding rna-protein interaction prediction based on improved bipartite network recommender algorithm.Frontiers in genetics 222019, 10, 343.235.
62. Zhao, Q.; Yu, H.; Ming, Z.; Hu, H.; Ren, G.; Liu, H. The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. Molecular Therapy-Nucleic Acids 2018,13, 464–471.266.
63. Hu, H.; Zhang, L.; Ai, H.; Zhang, H.; Fan, Y.; Zhao, Q.; Liu, H. HLPI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy.RNA biology 2018, 3115, 797–806.328.
64. Xiao, Y.; Zhang, J.; Deng, L. Prediction of lncRNA-protein interactions using HeteSim scores based on heterogeneous networks. Scientific reports 2017, 7, 1–12.349.
65. Zhao, Q.; Zhang, Y.; Hu, H.; Ren, G.; Zhang, W.; Liu, H. IRWNRLPI: integrating random walk and neighborhood regularized logistic matrix factorization for lncRNA-protein interaction prediction.Frontiers in genetics 2018, 9, 239. Version 2

66. Zhang W, Yue X, Tang G, Wu W, Huang F, Zhang X. SFPEL-LPI: Sequence-based feature projection ensemble learning for predicting LncRNA-protein interactions. PLoS Comput Biol. 2018;14(12):e1006616. Published 2018 Dec 11. doi:10.1371/journal.pcbi.1006616
67. Hu H, Zhang L, Ai H, et al. HLPI-Ensemble: Prediction of human lncRNA-protein interactions based on ensemble strategy. RNA Biol. 2018;15(6):797-806. doi:10.1080/15476286.2018.1457935
68. Wekesa JS, Meng J, Luan Y. A deep learning model for plant lncRNA-protein interaction prediction with graph attention. Mol Genet Genomics. 2020 Sep;295(5):1091-1102. doi: 10.1007/s00438-020-01682-w. Epub 2020 May 15. PMID: 32409904.
69. Fan X, Zhang S. LPI-BLS: Predicting lncRNA–protein interactions with a broad learning system-based stacked ensemble classifier. Neurocomputing. 2019;370:88–93. doi: 10.1016/j.neucom.2019.08.084
70. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831-838. doi:10.1038/nbt.3300
71. Zhang, S.W.; Zhang, X.X.; Fan, X.N.; Li, W.N. LPI-CNNCP: Prediction of lncRNA-protein interactions by using convolutional neural network with the copy-padding trick. Analytical Biochemistry 2020, 601, 113767. 4714.
72. Shaw, D.; Chen, H.; Xie, M.; Jiang, T. DeepLPI: a multimodal deep learning method for predicting the interactions between lncRNAs and protein isoforms. BMC bioinformatics 2021, 4922, 1–22.5015.
73. Ma, Y.; He, T.; Jiang, X. Projection-based neighborhood non-negative matrix factorization for lncRNA-protein interaction prediction. Frontiers in genetics 2019, 10, 1148.5417.
74. Zhou, Y.K.; Shen, Z.A.; Yu, H.; Luo, T.; Gao, Y.; Du, P.F. Predicting lncRNA–protein interactions with miRNAs as mediators in a heterogeneous network model. Frontiers in genetics 562020, 10, 1341.5718.
75. Zhou YK, Hu J, Shen ZA, Zhang WY, Du PF. LPI-SKF: Predicting lncRNA-Protein Interactions Using Similarity Kernel Fusions. Front Genet. 2020;11:615144. Published 2020 Dec 9. doi:10.3389/fgene.2020.615144
76. Ma Y, He T, Jiang X. Projection-Based Neighborhood Non-Negative Matrix Factorization for lncRNA-Protein Interaction Prediction. Front Genet. 2019;10:1148. Published 2019 Nov 20. doi:10.3389/fgene.2019.01148
77. Zhou YK, Shen ZA, Yu H, Luo T, Gao Y, Du PF. Predicting lncRNA-Protein Interactions With miRNAs as Mediators in a Heterogeneous Network Model. Front Genet. 2020;10:1341. Published 2020 Jan 22. doi:10.3389/fgene.2019.01341
78. 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 genomics 2013, 14, 1–10.5919.
79. Agostini, F.; Zanzoni, A.; Klus, P.; Marchese, D.; Cirillo, D.; Tartaglia, G.G. catRAPIDomics: a web server for large-scale prediction of protein–RNA interactions. Bioinformatics 2013, 29, 2928–2930
80. Jacq, C.; Miller, J.; Brownlee, G. A pseudogene structure in 5S DNA of Xenopus laevis. Cell 691977, 12, 109–120.7023.
81. Lou, W.; Ding, B.; Fu, P. Pseudogene-derived lncRNAs and their miRNA sponging mechanism in human cancer. Frontiers in cell and developmental biology 2020, 8.7224.
82. Denning, G.M.; Anderson, M.P.; Amara, J.F.; Marshall, J.; Smith, A.E.; Welsh, M.J. Processing of mutant cystic fibrosis transmembrane conductance regulator is temperature-sensitive. Nature 1992, 358, 761–764.7525.
83. MATLAB.version 7.10.0 (R2010a); The MathWorks Inc.: Natick, Massachusetts, 2010
84. Ramakrishnaiah Y, Kuhlmann L, Tyagi S, linc2function: A deep learning model to identify and assign function to long noncoding RNA (lncRNA)bioRxiv, https://doi.org/10.1101/2021.01.29.428785
85. Leinonen R, Sugawara H, Shumway M; International Nucleotide Sequence Database Collaboration. The sequence read archive. Nucleic Acids Res. 2011;39(Database issue):D19-D21. doi:10.1093/nar/gkq1019
86. RNAcentral Consortium. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. Nucleic Acids Res. 2021;49(D1):D212-D220. doi:10.1093/nar/gkaa921
87. Ogasawara O, Kodama Y, Mashima J, Kosuge T, Fujisawa T. DDBJ Database updates and computational infrastructure enhancement. Nucleic Acids Res. 2020;48(D1):D45-D50. doi:10.1093/nar/gkz982