Send Orders for Reprints to reprints@benthamscience.ae

604 *Current Topics in Medicinal Chemistry,* 2016, *16,* 604-615

Prediction and Dissection of Protein-RNA Interactions by Molecular Descriptors

Zhi-Ping Liu[1,*] and Luonan Chen[2,3,4,*]

Zhi-Ping Liu

[1]*Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, P.R. China;* [2]*Key Laboratory of Systems Biology, Collaborative Innovation Center of Cancer Medicine, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, P.R. China;* [3]*School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, P.R. China;* [4]*Collaborative Research Center for Innovative Mathematical Modelling, Institute of Industrial Science, University of Tokyo, Tokyo 153-8505, Japan*

Luonan Chen

Abstract: Protein-RNA interactions play crucial roles in numerous biological processes. However, detecting the interactions and binding sites between protein and RNA by traditional experiments is still time consuming and labor costing. Thus, it is of importance to develop bioinformatics methods for predicting protein-RNA interactions and binding sites. Accurate prediction of protein-RNA interactions and recognitions will highly benefit to decipher the interaction mechanisms between protein and RNA, as well as to improve the RNA-related protein engineering and drug design. In this work, we summarize the current bioinformatics strategies of predicting protein-RNA interactions and dissecting protein-RNA interaction mechanisms from local structure binding motifs. In particular, we focus on the feature-based machine learning methods, in which the molecular descriptors of protein and RNA are extracted and integrated as feature vectors of representing the interaction events and recognition residues. In addition, the available methods are classified and compared comprehensively. The molecular descriptors are expected to elucidate the binding mechanisms of protein-RNA interaction and reveal the functional implications from structural complementary perspective.

Keywords: Bioinformatics, Molecular descriptor, Prediction, Protein-RNA interaction, Protein-RNA recognition.

1. INTRODUCTION

The interactions between protein and RNA play crucial roles in many fundamental biological processes, such as alternative splicing [1], RNA interference [2], and RNA stability and degradation [3] in the post-transcriptional gene expression regulation [4]. Nowadays, various RNAs are gradually discovered to be key regulators of performing essential functions, such as microRNA (miRNA) [5] and long noncoding RNA (lncRNA) [6]. The techniques of next generation sequencing identified a number of novel lncRNAs, some of which have been shown to have significant impacts on human diseases [7, 8]. However, the specific functions of many RNAs are still mysteries. RNAs are always recognized to cooperate with particular proteins and form certain complexes so as to carry out paramount important activities. Without the interactions with proteins, these RNAs cannot correctly perform their critical roles [4].

*Address correspondence to these authors at the Key Laboratory of Systems Biology, Chinese Academy of Sciences, Shanghai, P.R. China;
Tel: 81-21-5492-0100; Fax: 81-21-5492-0120; E-mail: lnchen@sibs.ac.cn
and Department of Biomedical Engineering, Shandong University, Shandong, P.R. China; Tel: 86-531-8839-2280; Fax: 86-531-8839-2205;
E-mail: zpliu@sdu.edu.cn.

During the booming research interests of RNA functions, experimental techniques have correspondingly been developed for investigating protein-RNA interactions *in vivo*, such as RNA immunoprecipitation (RIP) and cross-linking immunoprecipitation (CLIP). RIP uses an antibody-based technique to precipitate specific RNA-binding proteins and their associated RNAs that can be identified by real-time PCR, microarray or deep sequencing (ChIP) [9]. The low resolution of RIP-ChIP limits the identification of actual binding sites and specificities. CLIP combines ultraviolet (UV) cross-linking with immunoprecipitation to map RNA-binding sites of a specific protein [10]. Joint with high-throughput sequencing termed as HITS-CLIP or CLIP-Seq [11], enhanced CLIP-based techniques are currently popular to identify genome-wide protein-RNA interactions. iCLIP is such a modified CLIP with individual-nucleotide resolution of binding sites [12]. It has successfully identified the protein-RNA interactions in various crucial biological processes [13, 14]. The improved methods demonstrate the promising sequencing technologies in detecting the protein-RNA interactions [11, 13, 15, 16].

The experimental approaches are usually time consuming and labor costly. Also, there are the throughput difficulties of identifying the protein-RNA interactions and the problems

1873-5294/16 $58.00+.00

on low resolutions of binding sites [10]. Therefore, bioinformatics methods of predicting protein-RNA interactions trained on the available knowledge provide alternative pipelines to identify the interactions between protein and RNA [17, 18]. The predicted interacting pairs can be further validated by *in vivo* experiments and used to design downstream experiments. More importantly, bioinformatics methods are often built on some assumption of protein-RNA interaction mechanism. The computational methods collect the data sources of protein-RNA interactions based on which the predictors are trained. The dissected propensities behind the protein-RNA interaction events, as well as the locations of sequence specificities in the protein-RNA recognition residues are extremely valuable to reveal their binding principle [19].

In the paper, we review the main computational strategies developed for predicting protein-RNA interactions and binding residues. We firstly illustrate them by a general machine learning framework. Secondly, we introduce the derived features of molecular descriptors, the encoding procedure and several learning algorithms. Then the available protein-RNA interaction resources and the existing bioinformatics methods of prediction are summarized respectively. We also briefly summarize the local structural motifs of protein-RNA interactions for dissecting their recognition structure basis. Last but not least, we present a discussion of future research directions, and conclude the achievements and outlooks of predicting protein-RNA interactions.

2. FRAMEWORK OF PREDICTION

Fig. (**1**) demonstrates the framework of predicting protein-RNA interactions. The available prediction methods can be categorized as two subgroups. First is the prediction of protein-RNA interaction events. Given a pair of protein and RNA, it is to determine whether they interact or not. Second is the prediction of binding sites. That is to predict the binding residues in the protein-RNA recognition. Generally speaking, an assumption underlying these prediction methods is that the principles of determining the interaction or non-interaction between one pair of protein and RNA can be generalized and extended to another pair of protein and RNA. In these methods, protein and RNA are represented and encoded by their numerical features respectively. A pair of protein and RNA is often formulated as a single feature vector. A classification algorithm is then utilized to build a predictor by distinguishing the interactions from non-interactions. After we train the predictor in these feature vectors of interactions and non-interactions, a decision is made for determining whether there exists an interaction between a predicting protein-RNA pair according to their encoded feature vector.

Although the two predictions focus on different objects, a similar strategy can be implemented for both of them. In the prediction of protein-RNA interactions, the strategy is to represent the predicting pair of protein and RNA into a feature vector (Fig. **1a**). By collecting the prior knowledge of protein-RNA interactions (with protein-RNA non-interactions) (Fig. **1c**), the protein-RNA pairs are represented into feature vectors individually. After labeling the feature vectors with binary signs of '1' for protein-RNA interactions and '0' for protein RNA non-interactions (Fig. **1d**), a machine learning algorithm is resorted to learn the correspondence between the interaction and the encoded features (Fig. **1e**). After a classifier is trained (Fig. **1b**), it can be used to predict if there is a new interaction or not in a given pair of protein and RNA according to the encoded feature vector as that in the training samples.

Often, a cross-validation procedure is implemented to evaluate the prediction performance in the training data before an independent prediction (Fig. **1g**). That is to divide the available training samples into several complementary subsets, then perform the learning on one subset (training set), and validate the prediction on the other subset (testing set). The prediction and validation steps are rotationally estimated in these subsets. Then, the prediction performance in terms of sensitivity, specificity, accuracy, F1-measure and the area under receiver operating characteristic curve can be achieved [17, 20].

The framework of predicting binding residues in protein-RNA recognition is similar, except that the encoding procedure is to represent each candidate residue into a feature vector. The knowledge of RNA-binding residues can be extracted in three-dimensional (3D) structure of protein-RNA complexes. After assigning the binary labels to these residues ('1' for binding residues and '0' for non-binding residues), the latter training and testing steps are very similar to that in the prediction of protein-RNA interaction events.

Currently, various molecular descriptors have been extracted as features to encode protein and RNA in the prediction of protein-RNA interactions and binding residues. Some physicochemical properties (e.g., amino acid type and atom number), biochemical features (e.g., hydrophobic indices of amino acids and nucleotides), evolutionary information (e.g., position-specific scoring matrix), self-defined propensities and statistical potentials, sequence-based conjoint triad features, and structure-based solution accessible area features are represented numerically as descriptors for protein and RNA respectively. The features reflect their properties from various aspects of protein-RNA interaction and recognition. They are the elements in the encoded feature vector.

3. MOLECULAR DESCRIPTORS AS FEATURES

The knowledge of protein-RNA interactions indicates some molecular descriptors are related to protein-RNA interactions [21-23]. These descriptors are extracted and combined together to transform protein and RNA into feature vectors. As shown in Fig. (**1**), one pair of protein and RNA is represented by one feature vector, in which the elements are the molecular descriptors of their diverse properties, from physicochemical characteristics (PC) to derived or defined properties of amino acids and nucleotides. The following subsections summarize some widely-used molecular descriptors in the predictions.

Physicochemical Characteristics

Firstly, the 20 amino acids and 4 nucleotides are those easily achievable sequence-based descriptors. Some particular residues are found to be enriched in the RNA-binding sites, such as arginine-rich motifs [21, 23]. Different types of

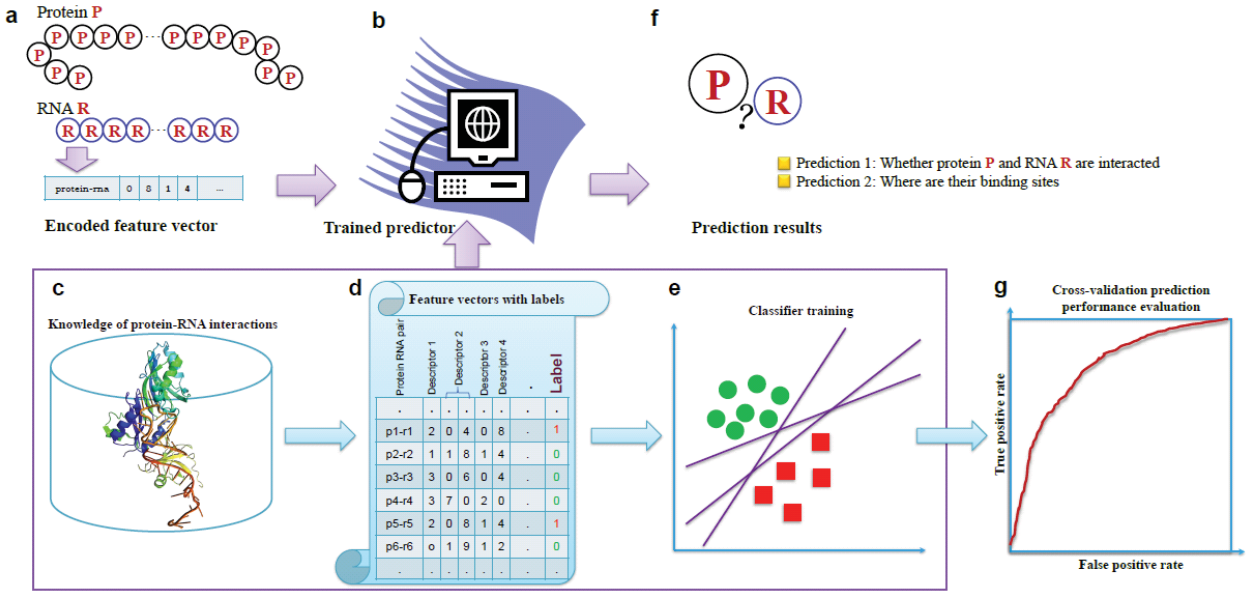

Fig. (1). The framework of predicting protein-RNA interactions. When predicting binding sites in the protein-RNA recognition, the encoding features in the learning and testing are the candidate residues, not the components in a pair of protein and RNA.

residues can be represented by different numbers. The single-sequence information (SSI) encodes the sequences into features [24, 25]. The number of atoms, number of electrostatic charge, and number of potential hydrogen bonds contained in amino acids and nucleotides are also often extracted as features [17].

Hydrophobicity

Hydrophobicity of protein surfaces is identified to be important in protein–RNA binding [17, 26]. The hydrophobic indices of amino acids are often identified as the descriptors of characterizing protein-RNA interaction [27].

Side Chain pKa Value

Protein pKa value measures its acid dissociation constant of a solution [28]. The side chain pKa value (pKa) is important in determining the pH-dependent characteristics of a protein [29]. The descriptor describes the environmental characteristic of RNA-binding, and thus it is often used in the prediction of protein-RNA interaction [25].

Position-Specific Scoring Matrix

The conservation or evolutionary information of sequences is important information of general interactions with RNA across protein complexes and species. Position-specific scoring matrix (PSSM) is a commonly-used representation of these patterns [30]. The protein-RNA interaction and their binding profiles are represented as the numerical descriptors. PSSM significantly facilitates the prediction of interactions between protein and RNA [31], as well as the discovery of binding motif [32].

Conjoint Triad Feature

The derived conjoint triad feature (CTF) represents protein and RNA by their percentages of conjoint triad residues in the sequences. The 20 types of amino acids are divided into 7 groups, i.e., {A, G, V}, {C}, {D, E}, {H, N, Q, W}, {I, L, F, P}, {Y, M, T, S} and {R, K} [33]. In such a way, each protein sequence is reduced to 7-letter alphabets. The feature is to represent the normalized frequency of the corresponding CTF in 3-mer of the 7-letter form [18]. Similarly, for RNA sequence, the *k*-mer (where *k* often equals to 3) frequency representation encodes the nucleotide percentages. The CTF descriptors characterize the binding sequence features of protein-RNA interaction [34].

Secondary Structure

Secondary structure (SS) describes the 3D form of local segments of protein and RNA [35]. RNA-binding is often located between α helix or loop and a groove of pocket, or between β sheet surface and unpaired RNA nucleotide bases [36, 37]. The SS descriptor of α helix, β sheet or loop provides the detailed local protein-RNA binding structure information [17].

Accessible Surface Area

Protein surfaces are identified as the main locations of protein-RNA interaction [19, 38]. The accessible surface area (ASA) is the protein surface area that is accessible to a solvent [35]. The descriptor defines the environment and exposure of protein surface where the RNA-binding events take place. The relative ASA is often implemented by dividing the ASA based on the accessible surface area of fully exposed amino acids [39].

Interaction Propensity

Beyond the physiochemical and biochemical descriptors of amino acids and nucleotides, some self-defined descriptors based on the ratios of different residue type in the entire dataset are often implemented as features of interaction pro-

pensity (IP). For instance, the residue propensity of each amino acid or nucleotide residue k (k = 1,…,20 for amino acids and k = 1,…,4 for nucleotides) is defined as the logarithm ratio of its percentage [40], i.e.,

$$P(k) = \log_2 \frac{\text{percentage of residue } k \text{ in contacts}}{\text{percentage of residue } k \text{ in entire dataset}}.$$

Specifically, the interface propensity $P(k)$ on protein surface can be calculated by

$$P(k) = \log_2 \frac{N_k^i / \sum_k N_k^i}{N_k^s / \sum_k N_k^s},$$

where N_k^i is the number of interface residues of type k, $\sum N_k^i$ is the total number of interface residues, N_k^s is the number of surface residues of type k, and $\sum_k N_k^s$ is the total number of surface residues [41, 42].

Based on relative entropy, we proposed a mutual IP between a residue triplet in protein and a nucleotide in RNA as follows [17]:

$$S(x,y) = \sum_{p,r} f_{p,r}(x,y) \log_2 \frac{f_{p,r}(x,y)}{f_p(x) f_r(y)},$$

where x denotes an amino acid triplet, y denotes an RNA nucleotide, (p,r) denotes a pair of protein-RNA interaction, $f_p(x)$ and $f_r(y)$ are the frequency of x and y respectively, and $f_{p,r}(x,y)$ is the frequency of triplet x interacting with nucleotide y in the pair (p,r). The ratio defines a descriptor which has been shown be crucial in the prediction of RNA binding residues [17].

Statistical Potentials

Statistical potentials describe the states of residue and atom by an energy function [43]. Various statistical potentials have been proposed to define the interactions among amino acid residues in protein folding [44]. The energy-based descriptor also estimates the interface propensity of protein-RNA binding. The statistical potential is often defined as

$$S(i,j) = -RT \ln \frac{N_{obs}(i,j)}{N_{\exp}(i,j)},$$

where R is the gas constant, T is the temperature, $N_{obs}(i,j)$ is the observed number of atomic pairs (i,j) within a distance, and $N_{\exp}(i,j)$ is the expected number of atomic pairs (i,j) in the same distance in the reference state [45].

4. ENCODING PROCEDURE

The former molecular descriptors derived from the sequence and structure of protein and RNA enable us to encode them into feature vectors. By coordinating and concentrating the values into vectors, the two macromolecules are represented by features. The elements in the feature vector are the corresponding numerical values of the molecular descriptors. Different methods use different descriptors and the vector dimension corresponds to the number of combined descriptors. For example, if only using the CTF, the dimension of the feature vector for protein is 343, and 64 for RNA, respectively [18, 34]. Thus, one pair of protein-RNA is encoded into one feature vector with dimension 407. During the encoding procedure, a sliding-window technique [17, 33] is often employed to integrate the sequential information of neighbor residues into feature vectors.

5. MACHINE LEARNING ALGORITHMS

After encoding a pair of protein and RNA into a feature vector, we label it as '1' when they are interacted and '0' when they are not. Then, the prediction of protein-RNA interaction is formulated into a binary classification problem. Based on entire dataset, we often train a machine learning algorithm to be a classifier of distinguishing protein-RNA interactions from non-interactions. Classic algorithms such as neural network (NN) [46], naïve Bayes (NB) [18], support vector machine (SVM) [47], and random forest (RF) [48] are often implemented as predictors.

To evaluate the prediction accuracy, fivefold or tenfold cross-validations are often implemented. When the built predictor is tested to achieve good prediction performance, such as high sensitivity, specificity, accuracy and F1-measure [17, 49], it is used to predict an independent protein-RNA interaction by encoding the candidate pair of protein and RNA into a feature vector. The features and the parameters of machine learning classifier are gradually tuned to be optimal in the training. After inputting the encoded feature vector of the candidate pair, the trained predictor outputs a prediction score.

6. EXISTING SOURCES OF PROTEIN-RNA INTERACTIONS

The descriptors and the training datasets extensively relied on the existing resources of protein-RNA interactions. The knowledge of protein-RNA interactions provides the fundamental background of machine learning. With the development of structure biology and high-throughput technology, more and more protein-RNA interactions are deposited in databases. Table **1** lists some popular databases of documenting protein-RNA interactions and their binding sites.

Protein Data Bank (PDB) is a worldwide repository of information about the 3D structure of proteins and nucleic acids [50]. The source determines the physical interactions of protein-RNA interactions of their complexes. The database provides important data resources of detecting the interacting residues in protein-RNA binding. Many databases related to protein-RNA interaction events and recognition sites are built on PDB, such as NDB [51], PRD [52], PRIDB [53], RPintDB [34], NPIDB [54] and RBPDB [55]. These databases extract specific protein-RNA interactions and provides certain services of inquiring these interactions. For instance, NDB specifically contains the information about experimentally-determined nucleic acids and complex assemblies. It provides the categories and the other detailed information about the specificities of protein-RNA interaction [51]. The brief descriptions of the other databases are shown in Table **1**.

For detecting the genome-scale of protein-RNA interactions, there has been great development in the field of ribonomics [12, 13, 56]. RIP-ChIP [9] and CLIP [57] based

Table 1. Databases of protein-RNA interactions. They are ordered alphabetically.

Database	Brief Description	Website	Reference
BIPA	A biological interaction database for protein-nucleic acid	http://www-cryst.bioc.cam.ac.uk/bipa	[67]
CLIPdb	A CLIP-seq database for protein-RNA interactions	http://clipdb.ncrnalab.org	[60]
CLIPZ	A database of post-transcriptional regulatory elements of RNA-binding proteins built from CLIP data	http://www.clipz.unibas.ch	[58]
DBBP	A database of the hydrogen bonding interactions between proteins and nucleic acids	http://bclab.inha.ac.kr/dbbp	[68]
DoRiNA	A database of RNA interactions in post-transcriptional regulation	http://dorina.mdc-berlin.de/	[69]
NDB	NDB contains the experimentally-determined protein-nucleic acid complexes	http://ndbserver.rutgers.edu/	[51]
NPIDB	A nucleic acid–protein interaction database	http://npidb.belozersky.msu.ru	[54]
NPInter	NPInter documents functional interactions between ncRNAs and proteins	http://www.bioinfo.org/NPInter	[62]
PDB	A protein repository containing the three-dimensional structural data of protein-nucleic acid interactions	http://www.rcsb.org/pdb/	[50]
PRD	A protein-RNA interaction database	http://pri.hgc.jp/	[52]
PRID	A protein-RNA interaction database	http://www-bioc.rice.edu/~shamoo/prid.html	[70]
PRIDB	A comprehensive database of protein-RNA interfaces extracted from complexes in PDB	http://bindr.gdcb.iastate.edu/PRIDB	[53]
RAID	A comprehensive resource for human RNA-associated interaction	http://www.rna-society.org/raid/	[71]
RBPDB	A database of RNA-binding protein specificities	http://rbpdb.ccbr.utoronto.ca/index.php	[55]
RBPmap	A web server for mapping binding sites of RNA-binding proteins	http://rbpmap.technion.ac.il/	[65]
RPintDB	A database of known RNA-protein interactions	http://pridb.gdcb.iastate.edu/RPISeq/RPIntDB.html	[34]
RsiteDB	A database of protein binding pockets that interact with RNA nucleotide bases	http://bioinfo3d.cs.tau.ac.il/RsiteDB/	[64]
starBase	A database of the interactions between protein and RNA from high-throughput data	http://starbase.sysu.edu.cn/	[72]

techniques have generated high-throughput protein-RNA interactions and binding sites in a relatively high-resolution coverage. Some databases, such as CLIPZ [58], starBase [59] and CLIPdb [60], recorded these high-throughput protein-RNA interactions. With the research field about non-coding RNA (ncRNA) booms, some databases such as NPInter [61] particularly collect the protein-ncRNA interactions. NPInter collects ncRNA interactions from recent literature and related databases in various species. It provides a comprehensive database of protein-ncRNA interactions [62].

Protein-RNA recognition sites such as the binding residues in proteins provide the local structure basis of protein-RNA interaction. Based on the 3D structures in PDB and its generated databases, some databases focus on the interfaces of protein-RNA binding [63]. To dissect the mechanism of protein-RNA interaction, sequence domains, structural motifs and functional sites related to protein-RNA binding highlight the functional importance of the spatial environments and local structure patterns in protein-RNA interaction. This kind of database, such as RsiteDB [64], PRIDB [53] and RBPmap [65], is also available. RsiteDB deposits the binding pockets that interact with single-stranded RNA nucleotides [64]. The binding sites are derived from NDB and classified into several groups. Essentially, the descriptors in the prediction methods aim to characterize the interactions by sequence and structure features, and the binding residues coordinate the places where the binding events take place. They characterize the protein-RNA interactions from different scales and perspectives.

So far, few information is available for the non-interacting protein-RNA pairs. In the training step of building a predictor shown in Fig. (**1e**), the negative interaction samples are often generated randomly or according to an iterative reduce strategy [49, 66]. That is to say, the non-interacting protein-RNA pairs are generated computationally. If an coupled protein-RNA pair cannot find its interolog

in the knowledge, it will be used as the protein-RNA non-interaction for training [49]. The same encoding procedure is implemented as that in the positive interactions. Based on the available protein-RNA interactions and the constructed protein-RNA non-interactions, some machine learning algorithms have been proposed to predict protein-RNA interactions.

7. EXISTING METHODS OF PREDICTING PROTEIN-RNA INTERACTIONS

Several methods have been developed for predicting protein-RNA interactions by molecular descriptors. For the available protein-RNA interactions documented in the databases listed in Table **1**, protein and RNA are encoded into a feature vector as illustrated in Fig. (**1**). The training data are those feature vectors with labels of interaction and non-interaction individually. A classification algorithm such as SVM is implemented to learn the discriminant features of protein-RNA interactions and non-interactions. After training, the predictor can be used to predict novel protein-RNA interactions.

catRAPID is a method of integrating multiple descriptors, i.e., secondary structure (SS), hydrogen bonding (HB), interaction propensity (IP), polarity (PO), and van der Waals (VW), to predict protein associations with lncRNAs [73]. Only based on sequence descriptors, we proposed a prediction method, ncRNAP, which performs a feature selection procedure to choose the distinct features in the CTF descriptors [18]. ncRNAP focuses on the prediction of ncRNA-protein interactions. We predicted protein-ncRNA interactions by an extended NB classifier and validated the predictions by wet experiments. Specifically, we first selected effective descriptor features and compressed the feature vector dimension for reducing the computational complexity. The selected features provide biological insights and allow relatively transparent prediction. Then an extended NB classifier was constructed to infer protein-ncRNA interactions. We also conducted ncRNA pull-down experiments and identified interacting proteins of sbRNA CeN72 in *C. elegans* [18]. Similarly, RPIseq provides a prediction method based on SVM/RF algorithm by the CTF descriptors [34]. By integrating sequence and structure features, i.e., protein block (PB), secondary structure (SS) and conjoint triad feature (CTF), RPI-Pred provides an SVM-based predictor of RNA-protein interaction [74]. In yeast, Pancaldi and Bahler provided an SVM/RF-based classifier for predicting the interacting protein-mRNA pairs with several feature classes, i.e., protein properties, such as physical features and gene ontology associations; mRNA properties, such as UTR characteristics, RNA structure, translational features, and expression levels [75].

8. EXISTING METHODS OF PREDICTING PROTEIN-RNA BINDING SITES

Compared to the prediction of protein-RNA interactions, relatively more methods have been proposed to predict protein-RNA binding sites because the 3D structures in PDB/NDB [51] are already available for learning before the availability of high-throughput data about protein-RNA interactions [10].

The definition of RNA-binding sites can be easily categorized as distance-based and energy-based [23]. In the distance definition, when the distance between two bases of protein amino acid and RNA nucleotide is close enough to a threshold, such as 3Å. The two residues are defined as interacting residues or the RNA-binding sites. In the energy-based definition, when a residue is within van der Waals contact or hydrogen binding distance to a RNA, it is defined as a RNA-binding residue. The binding sites are the local structures of binding residues. Apparently, the non-RNA-binding residue is the rest residues. The positive and negative RNA-binding sites are then easily composed. Compared to the non-binding residues, there are usually fewer residues identified as binding residues. Note that the imbalanced data classification problem can be tackled by random forest [77]. The negative samples are different from those in the prediction of protein-RNA interaction events, in which they are mainly constructed by randomly coupling (few methods also use the physical contact distance in the protein-RNA complexes as the cutoff to distinguish the protein-RNA interactions from non-interactions [73]). By using a similar strategy in the prediction of protein-RNA interactions, various methods for predicting RNA binding residues in proteins or in RNAs have been developed. Table **3** lists some representatives of them.

Jeong *et al.* proposed a NN-based method for predicting RNA-binding residues with amino acid types (AA) and secondary structures (SS) [46]. Then, they improved their predictor by the weighted PSSM [78]. BindN [25] develops an SVM-based tool for predicting RNA (and DNA) binding sites in proteins by single sequence information (SSI) with three sequence features, i.e., side chain pKa value (pKa), hydrophobicity (HP) index and molecular mass (MM). RNABindR [79] presents a naïve Bayes method to predict RNA-binding sites by the molecular descriptors of interface propensity (IP), relative accessible surface area (ASA), sequence entropy (SE), hydrophobicity (HP), secondary structure (SS), and electrostatic potentials (EP). In a feature integration manner, we proposed a method named PRNA [17] to predict the RNA-binding residues in proteins by combining various descriptors, i.e., mutual interaction propensity (IP), sequence and structure-based features. Specifically, the mutual IP represents a binding specificity of a protein residue to the interacting RNA nucleotide by considering its two-side neighborhood in a protein residue triplet. In addition, the sequence and structure based features are combined together to discriminate the interaction of amino acids with RNA. RNA-binding residues in proteins are then predicted by implementing a well-built RF classifier. PRNA is shown to be able to detect the annotated protein-RNA interaction sites with a high accuracy by outperforming the existing methods. It is also compared with the other machine learning algorithms, such as SVM, NN and NB, the better prediction performance provides more evidence of its efficiency and advantage [17].

9. STRUCTURE BASIS OF PROTEIN-RNA INTERACTIONS

After using the molecular descriptors to transform the proteins and RNAs into feature vectors, the accurate prediction of protein-RNA interaction events and recognition sites

Table 2. Available methods of predicting protein-RNA interactions.

Method	Descriptor	Classifier	Website	Reference
catRAPID	SS, HB, VW, HP, PO, IP	classified by discriminant power (based on IP)	http://s.tartaglialab.com/page/catrapid_group	[73]
lncPro	SS, HB, VW	Fisher's linear discriminant method	http://cmbi.bjmu.edu.cn/lncpro	[76]
ncRNAP	CTF	NB	http://doc.aporc.org/wiki/NCRNAP	[18]
Pancaldi and Bahler	Several feature classes	SVM and RF	NA	[75]
PRNAinter	PC, HP, pKa, PSSM, SS, ASA, IP	RF	http://doc.aporc.org/wiki/PRNAinter	[17]
RPIseq	CTF	SVM and RF	http://pridb.gdcb.iastate.edu/RPISeq/	[34]
RPI-Pred	PB, SS, CTF	SVM	http://ctsb.is.wfubmc.edu/projects/rpi-pred	[74]

Table 3. Methods of predicting protein-RNA binding residues.

Method	Descriptor	Classifier	Website	Reference
BindN	SSI, pKa, HP, MM	SVM	http://bioinfo.ggc.org/bindn/	[25]
Choi and Han	SL, SSI, MM, IP	SVM	http://bclab.inha.ac.kr/primer/	[80]
Jeong *et al.*	AA, SS, PSSM	NN	NA	[46, 78]
KYG	IP, PSMSP	Classified by propensity score (based on IP)	http://cib.cf.ocha.ac.jp/KYG/	[81]
PiRaNhA	PSSM, IP, ASA, HP	SVM	http://www.bioinformatics.sussex.ac.uk/PIRANHA	[26]
PPRInt	PSSM	SVM	http://www.imtech.res.in/raghava/pprint/	[82]
PRBR	SS, PSSM, PO, HP	RF	http://www.cbi.seu.edu.cn/PRBR/	[83]
PRINTR	PSSM, SSI, ASA, SS	SVM	http://210.42.106.80/printr/	[84]
PRNA	PC, HP, pKa, PSSM, SS, ASA, IP	RF	http://doc.aporc.org/wiki/PRNA	[17]
RISP	PSSM	SVM	http://grc.seu.edu.cn/RISP.	[85]
RNABindR	IP, ASA, SE, HP, SS, EP	NB and NN	http://bindr.gdcb.iastate.edu/RNABindR	[79]
RNAProB	PSSM	SVM	NA	[31]

becomes binary classification problems. Machine learning algorithms learned the rules of determining protein-RNA interactions in the training data and then generalized the principles into the testing data. The rules are often not transparent and cannot be explicitly described. From the mechanism perspective, they are highly related to the latent codes of determining protein-RNA interactions Table **2**. However, it is very difficult to identify a general principle underlying protein-RNA interactions [4, 86]. Obviously, the importance evaluation of molecular descriptors in the predictions will benefit to identify the key features of affecting the prediction performance. The feature evaluation provides a direct way of identifying the crucial factors in determining protein-RNA interactions. In our method ncRNAP [18], we selected the distinct features, which are significantly different in the interactions from the non-interactions. In our method PRNA [17], the importance of each feature has been evaluated by reducing prediction accuracy individually. We identified the mutual IP descriptor has improved the prediction accuracy. The descriptor is also different in binding residues from non-binding residues. This indicates the importance of local structural complementary in binding RNA.

In protein-RNA recognition, some common sequence patterns have been identified, such as RNA-recognition domains and motifs [21, 23, 36, 86, 87] (Table **2**). They provide the structure basis of protein-RNA interaction and recognition. In [19], we provided a systematic analysis of these RNA-binding domains and pockets on protein surfaces. Table **4** lists some of the major Pfam [88] domains and their families in RNA-binding proteins. The domain peptides

Table 4. Pfam superfamilies and domains of RNA-binding proteins [19]. The proteins are those representative proteins containing the domains.

Clan ID	Description	Domain ID	Description	Protein
CL0007	K-Homology (KH) domain Superfamily	PF00013	KH domain	1EC6:A;2ANN:A;3AEV:B
CL0027	RNA dependent RNA polymerase	PF30561	RNA dependent RNA polymerase	1UVL:A;2E9T:A;3BSO:A
CL0039	HIGH-signature proteins, UspA, and PP-ATPase	PF38198	tRNA synthetases class I (I, L, M and V)	1FFY:A;1GAX:A;2BTE:A;2CSX:A
CL0040	Class II aminoacyl-tRNA and Biotin synthetases	PF01409	tRNA synthetases class II core domain (F)	1EIY:B;2DU3:A;2DU4:A;2IY5:A
CL0055	Positive stranded ssRNA viruses coat protein	PF02247	Large coat protein	1BMV:2
CL0063	FAD/NAD(P)-binding Rossmann fold Superfamily	PF05958	tRNA (Uracil-5-)-methyltransferase	2BH2:A;3BT7:A
CL0101	Pelota - RNA ribose binding superfamily	PF01248	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family	1E7K:A;1SDS:A;1T0K:B
CL0178	PUA/ASCH superfamily	PF01472	PUA domain	1J2B:A;1R3E:A;1ZE2:B;2RFK:A
CL0196	DSRM-like clan	PF00035	Double-stranded RNA binding motif	1DI2:A;1RC7:A;3ADI:A;3ADL:A
CL0219	Ribonuclease H-like superfamily	PF02171	Piwi domain	1YTU:A;2F8S:A;3F73:A
CL0221	RRM-like clan	PF00076	RNA recognition motif. (a.k.a. RRM, RBD, or RNP domain)	1B7F:A;1CVJ:A;1ZH5:A;2G4B:A;3NNH:A
CL0383	Phenylalanine- and lysidine-tRNA synthetase domain superfamily	PF03483	B3/4 domain	1EIY:B
CL0527	Sm (Small RNA binding protein domain)	PF01423	LSM domain	1KQ2:A;1M8V:A;3AHU:A
CL0537	CCCH-zinc finger	PF00642	Zinc finger C-x8-C-x5-C-x3-H type (and similar)	3D2S:A
CL0539	RNase III domain-like superfamily	PF14622	Ribonuclease-III-like	1RC7:A

recognize the major groove or loop of the RNA and they provide a structure platform of interacting with RNA. For instance, KH domain is a major family with $\alpha\beta$ motifs that bind to RNA, which contains ~70 amino acids in length [89]. The RNA strand is bound between its first and second α helices, and between its two major loops, the GXXG loop and variable loop with 3 to 60 residues long [86]. Often, KH only binds to 4 bases. Double-stranded RNA binding motif (dsRBM) is another domain type containing an $\alpha\beta\beta\alpha$ fold. The contacts are located between its loops 2 and 4 of the fold in the nearby major grooves [4, 86, 89].

RNA also has structural motifs of binding to specific proteins, such as tetraloop, kink-turn, loop-E, π-turn, Ω-turn and S2-turn [90]. These RNA structural motifs are also shown important roles from the RNA side in protein-RNA interaction [91]. Many of the structure details of the protein-RNA binding specificity and determination are still not very clear [4, 92, 93]. When most types of the 3D structures of RNA-binding proteins are crystallized and categorized with the information of the binding domains information, we will hopefully get better understanding of the RNA-binding structure basis.

As for RNA-binding pockets on protein surfaces, we clustered them into several major groups by a computational procedure based on pocket similarity network [19, 94]. Fig. (**2**) shows the major patterns of local binding pockets in seven selective non-redundant protein groups. The local structures demonstrate the binding grooves of recognizing RNA. As shown in Fig. (**2**), we identified some RNA-binding structural motifs in the form of pockets on protein surfaces (format: PDB_ID:Chain_ID:Pocket_ID). Most of the RNA-binding pockets are contained in the domains of these RNA-binding proteins [19]. This indicates the RNA-binding domains fold into certain local tertiary structures to facilitate the binding with RNA. These pockets provide the structural shapes needed in protein-RNA recognitions [38] (Table **4**).

DISCUSSION AND CONCLUSION

As more and more crucial functions of RNA are revealed in recent years [95], predicting and analyzing protein-RNA interactions have been hot research topics, especially in the field of post-transcriptional gene regulation [96-98]. Beyond the crystallized 3D structure of protein-RNA complexes

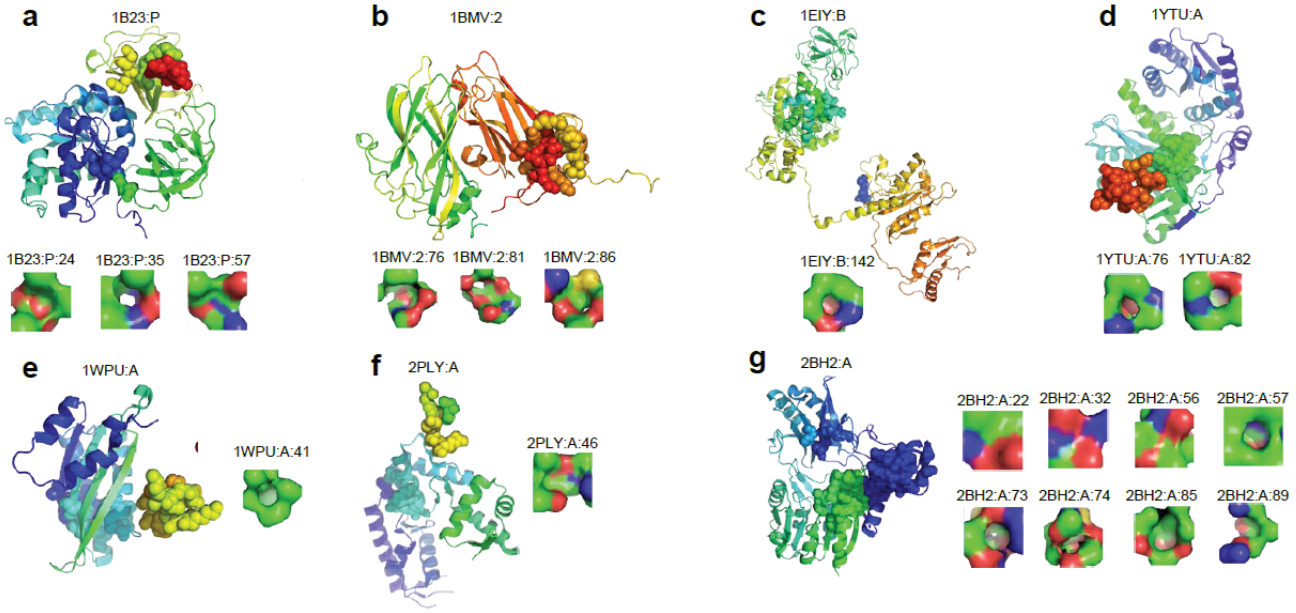

Fig. (2). Seven major types of RNA-binding pockets and their locations on protein surfaces.

documented in PDB, the high-throughput techniques as CLIP has revolutionized the detection of protein-RNA interactions [57]. Bioinformatics methods provide the precious opportunities of identifying novel protein-RNA interactions and dissecting the mechanisms related to their binding. The growing data sources of protein-RNA interactions accelerate the development of more reliable predictions. When more protein-RNA interactions are identified in new experiments, more samples and knowledge can be integrated to train the predictor. It is expected to develop more accurate methods of predicting protein-RNA interactions at both molecular coupling level and atomic binding level.

As to the existing methods, different predictors have their advantages in different datasets or perspectives. Collaborative prediction projects have received promising achievements in protein structure and gene regulatory network [99, 100]. It is expected to build benchmarked datasets or international challenges for evaluating the prediction strategies. Several future directions will be expected to improve the prediction of protein-RNA interaction and recognition. In the feature representation, the molecular descriptors encode the protein and RNA into feature vectors. Introducing more powerful molecular descriptors of protein-RNA interaction and recognition, not only the isolated descriptors of protein and RNA, will improve the prediction accuracy, such as mutual IP [17] and the knowledge-based scoring functions [101]. In these descriptors, identifying the most important features related to protein-RNA interaction will reduce the feature dimension and computational complexity. It will definitely improve the prediction power with more complicated and ensemble learning algorithms. Moreover, the selected features will provide new insights into dissecting the mechanism of protein-RNA interaction.

The structural motifs in protein-RNA interaction give the recognition details between protein [19] and RNA [102]. The biologically meaningful residues in structures generate the binding spatial patterns and active sites of performing certain roles. It is meaningful to determine the functions underlying the protein-RNA interactions. The structural motifs can be directly used for the target design of ligand binding in RNAi drugs [103-105]. In the prediction of protein-RNA interactions, novel descriptors derived from the identified structural motifs which directly extract the structural complementary patterns will be an interesting research direction in predicting protein-RNA interactions.

In conclusion, we provided a review of the current state-of-art of bioinformatics methods of predicting protein-RNA interactions. The widely-used molecular descriptors and machine learning algorithms, available datasets of depositing protein-RNA interactions, and existing methods of predicting protein-RNA interaction events and recognition sites are summarized and commented, respectively. In addition to costly experiments *in vivo*, computational predictions of protein-RNA interactions provide promising alternatives of identifying protein-RNA interactions and recognizing their binding specificities. The analysis of molecular descriptors, the identification of structural motifs, and the functional annotations to binding sites highlight the downstream of deciphering the mechanism of protein-RNA interaction and the following biomedical applications. Moreover, accurate identification of protein-RNA interactions will also benefit the research works on network biomarkers [106-108] and dynamical network biomarkers [109, 110] with the consideration of ncRNAs on biology and medicine.

LIST OF ABBREVIATIONS

3D	=	Three-dimension
ASA	=	Accessible surface area
ChIP	=	Chromatin immunoprecipitation
CLIP	=	Cross-linking immunoprecipitation

CTF = Conjoint triad feature
dsRBM = Double-stranded RNA binding motif
EP = Electrostatic potentials
HB = Hydrogen bonding
HITS-CLIP = High-throughput sequencing by CLIP
HP = Hydrophobic indices
IP = Interaction propensity
lncRNA = Long non-coding RNA
miRNA = MicroRNA
MM = Molecular mass
NA = Not available
NB = Naïve Bayes
ncRNA = Non-coding RNA
NN = Neural network
PB = Protein block
PC = Physicochemical characteristics
pKa = Side chain pKa value
PO = Polarity
PSMSP = Position-specific multiple sequence profiles
PSSM = Position-specific scoring matrix
RF = Random forest
RIP = RNA immunoprecipitation
RNA = Ribonucleic acid
SE = Sequence entropy
SL = Sequence length
SP = Statistical potentials
SS = Secondary structure
SSI = Single-sequence information
SVM = Support vector machine
UTR = Untranslated regions
UV = Ultraviolet
VW = Van der Waals

CONFLICT OF INTEREST

The authors confirm that this article content has no conflict of interest.

ACKNOWLEDGEMENTS

This work was partially supported by the Strategic Priority Research Program of Chinese Academy of Sciences under Grant No. XDB13040700; National Program on Key Basic Research Project under Grant No. 2014CB910504; National Natural Science Foundation of China (NSFC) under Grant Nos. 61572287, 31100949, 61134013 and 91439103; the Fundamental Research Funds of Shandong University under Grant No. 2014TB006; Shandong Provincial Natural Science Foundation under Grant No. ZR2015FQ001; and the Scientific Research Foundation for the Returned Overseas Chinese Scholars, Ministry of Education of China.

REFERENCES

[1] Wahl, M. C.; Will, C. L.; Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **2009,** *136*(4), 701-718.
[2] Hannon, G. J. RNA interference. *Nature,* **2002,** *418*(6894), 244-251.
[3] Houseley, J.; Tollervey, D. The many pathways of RNA degradation. *Cell,* **2009,** *136*(4), 763-776.
[4] Lunde, B. M.; Moore, C.; Varani, G. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **2007,** *8*(6), 479-490.
[5] Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell,* **2004,** *116*(2), 281-297.
[6] Kim, E. D.; Sung, S. Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.,* **2011,** *17*(1), 16-21.
[7] Pefanis, E.; Wang, J.; Rothschild, G.; Lim, J.; Chao, J.; Rabadan, R.; Economides, A. N.; Basu, U. Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. *Nature,* **2014,** *514*(7522), 389-393.
[8] Pefanis, E.; Wang, J.; Rothschild, G.; Lim, J.; Kazadi, D.; Sun, J.; Federation, A.; Chao, J.; Elliott, O.; Liu, Z. P.; Economides, A. N.; Bradner, J. E.; Rabadan, R.; Basu, U. RNA exosome-regulated long non-coding RNA transcription controls super-enhancer activity. *Cell,* **2015,** *161*(4), 774-789.
[9] Townley-Tilson, W. H.; Pendergrass, S. A.; Marzluff, W. F.; Whitfield, M. L. Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein. *RNA,* **2006,** *12*(10), 1853-1867.
[10] Milek, M.; Wyler, E.; Landthaler, M. Transcriptome-wide analysis of protein-RNA interactions using high-throughput sequencing. *Semin. Cell Dev. Biol.,* **2012,** *23*(2), 206-212.
[11] Zhang, C.; Darnell, R. B. Mapping *in vivo* protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol.,* **2011,** *29*(7), 607-614.
[12] Huppertz, I.; Attig, J.; D'Ambrogio, A.; Easton, L. E.; Sibley, C. R.; Sugimoto, Y.; Tajnik, M.; Konig, J.; Ule, J. iCLIP: protein-RNA interactions at nucleotide resolution. *Methods,* **2014,** *65*(3), 274-287.
[13] Konig, J.; Zarnack, K.; Rot, G.; Curk, T.; Kayikci, M.; Zupan, B.; Turner, D. J.; Luscombe, N. M.; Ule, J. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.,* **2010,** *17*(7), 909-915.
[14] Tollervey, J. R.; Curk, T.; Rogelj, B.; Briese, M.; Cereda, M.; Kayikci, M.; Konig, J.; Hortobagyi, T.; Nishimura, A. L.; Zupunski, V.; Patani, R.; Chandran, S.; Rot, G.; Zupan, B.; Shaw, C. E.; Ule, J. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nat. Neurosci.,* **2011,** *14*(4), 452-458.
[15] Hafner, M.; Landthaler, M.; Burger, L.; Khorshid, M.; Hausser, J.; Berninger, P.; Rothballer, A.; Ascano, M., Jr.; Jungkamp, A. C.; Munschauer, M.; Ulrich, A.; Wardle, G. S.; Dewell, S.; Zavolan, M.; Tuschl, T. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell,* **2010,** *141*(1), 129-141.
[16] 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,* **2011,** *44*(4), 667-678.
[17] Liu, Z. P.; Wu, L. Y.; Wang, Y.; Zhang, X. S.; Chen, L. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics,* **2010,** *26*(13), 1616-1622.
[18] Wang, Y.; Chen, X.; Liu, Z. P.; Huang, Q.; Xu, D.; Zhang, X. S.; Chen, R.; Chen, L. De novo prediction of RNA-protein interactions from sequence information. *Mol. Biosyst.,* **2013,** *9*(1), 133-142.
[19] Liu, Z. P. Systematic identification of local structure binding motifs in protein-RNA recognition. *Proc. 8th Inter. Conf. Syst. Biol.* **2014**, 74-80.
[20] Metz, C. E. Basic principles of ROC analysis. *Semin. Nucl. Med.,* **1978,** *8*(4), 283-298.
[21] Jones, S.; Daley, D. T.; Luscombe, N. M.; Berman, H. M.; Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.,* **2001,** *29*(4), 943-954.
[22] Ellis, J. J.; Broom, M.; Jones, S. Protein-RNA interactions: structural analysis and functional classes. *Proteins,* **2007,** *66*(4), 903-911.

[23] Allers, J.; Shamoo, Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.,* **2001,** *311*(1), 75-86.

[24] Han, L. Y.; Cai, C. Z.; Lo, S. L.; Chung, M. C.; Chen, Y. Z. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA,* **2004,** *10*(3), 355-368.

[25] Wang, L.; Brown, S. J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.,* **2006,** *34* (Web Server issue), W243-248.

[26] Spriggs, R. V.; Murakami, Y.; Nakamura, H.; Jones, S. Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics,* **2009,** *25*(12), 1492-1497.

[27] Sweet, R. M.; Eisenberg, D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J. Mol. Biol.,* **1983,** *171*(4), 479-488.

[28] Nakamura, H. Roles of electrostatic interaction in proteins. *Q. Rev. Biophys.,* **1996,** *29*(1), 1-90.

[29] Gibas, C. J.; Subramaniam, S. Explicit solvent models in protein pKa calculations. *Biophys. J.,* **1996,** *71*(1), 138-147.

[30] Stormo, G. D.; Schneider, T. D.; Gold, L.; Ehrenfeucht, A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res.,* **1982,** *10*(9), 2997-3011.

[31] Cheng, C. W.; Su, E. C.; Hwang, J. K.; Sung, T. Y.; Hsu, W. L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics,* **2008,** *9*(Suppl 12), S6.

[32] Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **2000,** *16*(1), 16-23.

[33] Shen, J.; Zhang, J.; Luo, X.; Zhu, W.; Yu, K.; Chen, K.; Li, Y.; Jiang, H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. USA,* **2007,** *104*(11), 4337-4341.

[34] Muppirala, U. K.; Honavar, V. G.; Dobbs, D. Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics,* **2011,** *12*, 489.

[35] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers,* **1983,** *22*(12), 2577-2637.

[36] Draper, D. E. Themes in RNA-protein recognition. *J. Mol. Biol.,* **1999,** *293*(2), 255-270.

[37] Chen, Y.; Varani, G. Protein families and RNA recognition. *FEBS J.,* **2005,** *272*(9), 2088-2097.

[38] Liu, Z. P.; Wu, L. Y.; Wang, Y.; Zhang, X. S.; Chen, L. Bridging protein local structures and protein functions. *Amino Acids* **2008,** *35*(3), 627-650.

[39] Rost, B.; Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins,* **1994,** *20*(3), 216-226.

[40] Terribilini, M.; Lee, J. H.; Yan, C.; Jernigan, R. L.; Honavar, V.; Dobbs, D. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA,* **2006,** *12*(8), 1450-1462.

[41] Han, K.; Nepal, C. PRI-Modeler: extracting RNA structural elements from PDB files of protein-RNA complexes. *FEBS Lett.,* **2007,** *581*(9), 1881-1890.

[42] Perez-Cano, L.; Fernandez-Recio, J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins,* **2009,** *78*(1), 25-35.

[43] Skolnick, J. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.,* **2006,** *16*(2), 166-171.

[44] Buchete, N. V.; Straub, J. E.; Thirumalai, D. Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.,* **2004,** *14*(2), 225-232.

[45] Zhou, H.; Zhou, Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.,* **2002,** *11*(11), 2714-2726.

[46] Jeong, E.; Chung, I. F.; Miyano, S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.,* **2004,** *15*(1), 105-116.

[47] Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.,* **1995,** *20*(3), 273-297.

[48] Breiman, L. Random forests. *Mach. Learn.,* **2001,** *45*(1), 5-32.

[49] Liu, Z. P.; Chen, L. Proteome-wide prediction of protein-protein interactions from high-throughput data. *Protein Cell,* **2012,** *3*(7), 508-520.

[50] Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.,* **2000,** *28*(1), 235-242.

[51] Coimbatore Narayanan, B.; 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.,* **2013,** *42*(Database issue), D114-122.

[52] Fujimori, S.; Hino, K.; Saito, A.; Miyano, S.; Miyamoto-Sato, E. PRD: A protein-RNA interaction database. *Bioinformation,* **2012,** *8*(15), 729-730.

[53] 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.,* **2010,** *39*(Database issue), D277-282.

[54] Kirsanov, D. D.; Zanegina, O. N.; Aksianov, E. A.; Spirin, S. A.; Karyagina, A. S.; Alexeevski, A. V. NPIDB: Nucleic acid-Protein Interaction DataBase. *Nucleic Acids Res.,* **2012,** *41*(Database issue), D517-523.

[55] Cook, K. B.; Kazan, H.; Zuberi, K.; Morris, Q.; Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.,* **2010,** *39*(Database issue), D301-308.

[56] Hogan, D. J.; Riordan, D. P.; Gerber, A. P.; Herschlag, D.; Brown, P. O. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.,* **2008,** *6*(10), e255.

[57] Chi, S. W.; Zang, J. B.; Mele, A.; Darnell, R. B. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature,* **2009,** *460*(7254), 479-486.

[58] Khorshid, M.; Rodak, C.; Zavolan, M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.,* **2010,** *39*(Database issue), D245-252.

[59] Yang, J. H.; Li, J. H.; Shao, P.; Zhou, H.; Chen, Y. Q.; Qu, L. H. starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.,* **2010,** *39*(Database issue), D202-209.

[60] Yang, Y. C.; Di, C.; Hu, B.; Zhou, M.; Liu, Y.; Song, N.; Li, Y.; Umetsu, J.; Lu, Z. CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics,* **2015,** *16*(1), 51.

[61] Wu, T.; Wang, J.; Liu, C.; Zhang, Y.; Shi, B.; Zhu, X.; Zhang, Z.; Skogerbo, G.; Chen, L.; Lu, H.; Zhao, Y.; Chen, R. NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Res.,* **2006,** *34*(Database issue), D150-152.

[62] Yuan, J.; Wu, W.; Xie, C.; Zhao, G.; Zhao, Y.; Chen, R. NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res.,* **2013,** *42*(Database issue), D104-108.

[63] Bahadur, R. P.; Zacharias, M.; Janin, J. Dissecting protein-RNA recognition sites. *Nucleic Acids Res.,* **2008,** *36*(8), 2705-2716.

[64] Shulman-Peleg, A.; Nussinov, R.; Wolfson, H. J. RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.,* **2009,** *37*(Database issue), D369-373.

[65] Paz, I.; Kosti, I.; Ares, M., Jr.; Cline, M.; Mandel-Gutfreund, Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Res.,* **2014,** *42*(Web Server issue), W361-367.

[66] Smialowski, P.; Pagel, P.; Wong, P.; Brauner, B.; Dunger, I.; Fobo, G.; Frishman, G.; Montrone, C.; Rattei, T.; Frishman, D.; Ruepp, A. The Negatome database: a reference set of non-interacting protein pairs. *Nucleic Acids Res.,* **2009,** *38*(Database issue), D540-544.

[67] Lee, S.; Blundell, T. L. BIPA: a database for protein-nucleic acid interaction in 3D structures. *Bioinformatics,* **2009,** *25*(12), 1559-1560.

[68] Park, B.; Kim, H.; Han, K. DBBP: database of binding pairs in protein-nucleic acid interactions. *BMC Bioinformatics,* **2014,** *15*(Suppl 15), S5.

[69] Blin, K.; Dieterich, C.; Wurmus, R.; Rajewsky, N.; Landthaler, M.; Akalin, A. DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.,* **2014,** *43*(Database issue), D160-167.

[70] Morozova, N.; Allers, J.; Myers, J.; Shamoo, Y. Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics,* **2006,** *22*(22), 2746-2752.

[71] Zhang, X.; Wu, D.; Chen, L.; Li, X.; Yang, J.; Fan, D.; Dong, T.; Liu, M.; Tan, P.; Xu, J.; Yi, Y.; Wang, Y.; Zou, H.; Hu, Y.; Fan, K.; Kang, J.; Huang, Y.; Miao, Z.; Bi, M.; Jin, N.; Li, K.; Wang, D.

RAID: a comprehensive resource for human RNA-associated (RNA-RNA/RNA-protein) interaction. *RNA,* **2014,** *20*(7), 989-993.
[72] 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.,* **2013,** *42*(Database issue), D92-97.
[73] Bellucci, M.; Agostini, F.; Masin, M.; Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat. Methods,* **2011,** *8*(6), 444-445.
[74] Suresh, V.; Liu, L.; Adjeroh, D.; Zhou, X. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic Acids Res.,* **2015,** *43*(3), 1370-1379.
[75] Pancaldi, V.; Bahler, J. In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res.,* **2011,** *39*(14), 5826-5836.
[76] 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*, 651.
[77] Chen, C., Liaw, A., and Breiman, L. Using random forest to learn imbalanced data. *Technical Report 666, University of California, Berkeley* **2004**.
[78] Priami, C.; Cardelli, L.; Emmott, S.; Jeong, E.; Miyano, S. A Weighted Profile Based Method for Protein-RNA Interacting Residue Prediction. In *Transactions on Computational Systems Biology IV*, Springer Berlin Heidelberg: **2006**, *3939*, 123-139.
[79] Terribilini, M.; Sander, J. D.; Lee, J. H.; Zaback, P.; Jernigan, R. L.; Honavar, V.; Dobbs, D. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.,* **2007,** *35*(Web Server issue), W578-584.
[80] Choi, S.; Han, K. Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Comput. Biol. Med.,* **2013,** *43*(11), 1687-1697.
[81] Kim, O. T.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.,* **2006,** *34*(22), 6450-6460.
[82] Kumar, M.; Gromiha, M. M.; Raghava, G. P. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins,* **2008,** *71*(1), 189-194.
[83] Ma, X.; Guo, J.; Wu, J.; Liu, H.; Yu, J.; Xie, J.; Sun, X. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins,* **2011,** *79*(4), 1230-1239.
[84] Wang, Y.; Xue, Z.; Shen, G.; Xu, J. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids,* **2008,** *35*(2), 295-302.
[85] Tong, J.; Jiang, P.; Lu, Z. H. RISP: a web-based server for prediction of RNA-binding sites in proteins. *Comput. Methods Programs Biomed.,* **2008,** *90*(2), 148-153.
[86] Glisovic, T.; Bachorik, J. L.; Yong, J.; Dreyfuss, G. RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.,* **2008,** *582*(14), 1977-1986.
[87] Hall, K. B. RNA-protein interactions. *Curr. Opin. Struct. Biol.,* **2002,** *12*(3), 283-288.
[88] Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L.; Tate, J.; Punta, M. Pfam: the protein families database. *Nucleic Acids Res.,* **2014,** *42*(Database issue), D222-230.
[89] Valverde, R.; Edwards, L.; Regan, L. Structure and function of KH domains. *FEBS J.* **2008,** *275*(11), 2712-2726.
[90] Ciriello, G.; Gallina, C.; Guerra, C. Analysis of interactions between ribosomal proteins and RNA structural motifs. *BMC Bioinformatics,* **2010,** *11*(Suppl 1), S41.
[91] Apostolico, A.; Ciriello, G.; Guerra, C.; Heitsch, C. E.; Hsiao, C.; Williams, L. D. Finding 3D motifs in ribosomal RNA structures. *Nucleic Acids Res.,* **2009,** *37*(4), e29.
[92] Thapar, R.; Denmon, A. P.; Nikonowicz, E. P. Recognition modes of RNA tetraloops and tetraloop-like motifs by RNA-binding proteins. *Wiley Interdiscip. Rev. RNA,* **2013,** *5*(1), 49-67.
[93] Cook, K. B.; Hughes, T. R.; Morris, Q. D., High-throughput characterization of protein-RNA interactions. *Brief Funct. Genomics,* **2015,** *14* (1), 74-89.
[94] Liu, Z. P.; Wu, L. Y.; Wang, Y.; Zhang, X. S.; Chen, L. Analysis of protein surface patterns by pocket similarity network. *Protein Pept. Lett.,* **2008,** *15*(5), 448-455.
[95] Keene, J. D. RNA regulons: coordination of post-transcriptional events. *Nat. Rev. Genet.,* **2007,** *8*(7), 533-543.
[96] Liu, Z. P.; Wu, H.; Zhu, J.; Miao, H. Systematic identification of transcriptional and post-transcriptional regulations in human respiratory epithelial cells during influenza A virus infection. *BMC Bioinformatics,* **2014,** *15*, 336.
[97] Joshi, A.; Beck, Y.; Michoel, T. Post-transcriptional regulatory networks play a key role in noise reduction that is conserved from micro-organisms to mammals. *FEBS J.,* **2012,** *279*(18), 3501-3512.
[98] Mittal, N.; Scherrer, T.; Gerber, A. P.; Janga, S. C. Interplay between posttranscriptional and posttranslational interactions of RNA-binding proteins. *J. Mol. Biol.,* **2011,** *409*(3), 466-479.
[99] Cooper, S.; Khatib, F.; Treuille, A.; Barbero, J.; Lee, J.; Beenen, M.; Leaver-Fay, A.; Baker, D.; Popovic, Z.; Players, F. Predicting protein structures with a multiplayer online game. *Nature,* **2010,** *466*(7307), 756-760.
[100] Marbach, D.; Prill, R. J.; Schaffter, T.; Mattiussi, C.; Floreano, D.; Stolovitzky, G. Revealing strengths and weaknesses of methods for gene network inference. *Proc. Natl. Acad. Sci. USA,* **2010,** *107*(14), 6286-6291.
[101] Huang, S. Y.; Zou, X. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. *Nucleic Acids Res.,* **2014,** *42*(7), e55.
[102] Goodarzi, H.; Najafabadi, H. S.; Oikonomou, P.; Greco, T. M.; Fish, L.; Salavati, R.; Cristea, I. M.; Tavazoie, S. Systematic discovery of structural elements governing stability of mammalian messenger RNAs. *Nature,* **2012,** *485*(7397), 264-268.
[103] Ray, D.; Kazan, H.; Cook, K. B.; Weirauch, M. T.; Najafabadi, H. S.; Li, X.; Gueroussov, S.; Albu, M.; Zheng, H.; Yang, A.; Na, H.; Irimia, M.; Matzat, L. H.; Dale, R. K.; Smith, S. A.; Yarosh, C. A.; Kelly, S. M.; Nabet, B.; Mecenas, D.; Li, W.; Laishram, R. S.; Qiao, M.; Lipshitz, H. D.; Piano, F.; Corbett, A. H.; Carstens, R. P.; Frey, B. J.; Anderson, R. A.; Lynch, K. W.; Penalva, L. O.; Lei, E. P.; Fraser, A. G.; Blencowe, B. J.; Morris, Q. D.; Hughes, T. R. A compendium of RNA-binding motifs for decoding gene regulation. *Nature,* **2013,** *499*(7457), 172-177.
[104] Ray, D.; Kazan, H.; Chan, E. T.; Pena Castillo, L.; 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*(7), 667-670.
[105] Coelho, T.; Adams, D.; Silva, A.; Lozeron, P.; Hawkins, P. N.; Mant, T.; Perez, J.; Chiesa, J.; Warrington, S.; Tranter, E.; Munisamy, M.; Falzone, R.; Harrop, J.; Cehelsky, J.; Bettencourt, B. R.; Geissler, M.; Butler, J. S.; Sehgal, A.; Meyers, R. E.; Chen, Q.; Borland, T.; Hutabarat, R. M.; Clausen, V. A.; Alvarez, R.; Fitzgerald, K.; Gamba-Vitalo, C.; Nochur, S. V.; Vaishnaw, A. K.; Sah, D. W.; Gollob, J. A.; Suhr, O. B. Safety and efficacy of RNAi therapy for transthyretin amyloidosis. *N. Engl. J. Med.,* **2013,** *369*(9), 819-829.
[106] Liu, Z. P.; Wang, Y.; Zhang, X. S.; Chen, L. Network-based analysis of complex diseases. *IET Syst. Biol.,* **2012,** *6*(1), 22-33.
[107] Liu, X.; Liu, Z. P.; Zhao, X. M.; Chen, L. Identifying disease genes and module biomarkers by differential interactions. *J. Am. Med. Inform. Assoc.,* **2012,** *19*(2), 241-248.
[108] He, D.; Liu, Z. P.; Chen, L. Identification of dysfunctional modules and disease genes in congenital heart disease by a network-based approach. *BMC Genomics,* **2011,** *12*, 592.
[109] Chen, L.; Liu, R.; Liu, Z. P.; Li, M.; Aihara, K. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Sci. Rep.,* **2012,** *2*, 342.
[110] Liu, R.; Li, M.; Liu, Z. P.; Wu, J.; Chen, L.; Aihara, K. Identifying critical transitions and their leading biomolecular networks in complex diseases. *Sci. Rep.,* **2012,** *2*, 813.

Received: March 26, 2015 **Revised: June 25, 2015** **Accepted: June 26, 2015**