Neurocomputing 206 (2016) 28–34

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Prediction of protein-RNA interactions using sequence and structure descriptors

Zhi-Ping Liu [a,*], Hongyu Miao [b,*]

[a] Department of Biomedical Engineering, School of Control Science and Engineering, Shandong University, Jinan, Shandong 250061, China
[b] Department of Biostatistics, School of Public Health, University of Texas Health Science Center, Houston, TX 77030, USA

ARTICLE INFO

Article history:
Received 8 August 2015
Received in revised form
28 October 2015
Accepted 28 November 2015
Available online 2 June 2016

Keywords:
Protein-RNA interaction
Sequence and structure descriptor
Interaction propensity
Random forest predictor

ABSTRACT

Protein-RNA interactions play critical roles in numerous biological processes such as posttranscriptional regulation and protein synthesis. However, experimental screening of protein-RNA interactions is usually laborious and time-consuming. It is therefore desirable to develop efficient bioinformatics methods to predict protein-RNA interactions, which can provide valuable hints for future experimental design and advance our understanding of the interaction mechanisms. In this study, we propose a novel method for predicting protein-RNA interactions based on both sequence and structure descriptors of protein and RNA (e.g., the sequence-based physicochemical features, the secondary and three-dimensional structure-based features). We train and compare several classifiers using these descriptors on several benchmark datasets, and the random forest method is selected to build an efficient predictor of protein-RNA interactions. We conduct further cross-validation and case studies, and the results clearly suggest the efficacy of the proposed method.

1. Introduction

Interactions between protein and RNA molecules are of paramount importance to a variety of fundamental biological processes, including pre-mRNA processing, post-transcriptional gene regulation and RNA degradation [1,2]. Although previous studies have identified that RNA-binding proteins can recognize the target RNA molecules via certain structural domains or motifs [3–7], our understanding of the principles and mechanisms of protein-RNA interactions is still limited. The development of experimental assays and next-generation sequencing techniques for detecting protein-RNA interactions (e.g., CLIP-Seq [8–10], mobility shift assay [11], RNA pull-down assay [12], and FISH/ISH col-localization [13]) certainly helps to advance our knowledge in this area. However, such experimental approaches have their throughput limitations and are usually time and labor consuming. Therefore, alternative methods such as computational predictions of protein-RNA interactions are desirable to expedite the exploration of millions of potential interactions and generate useful hypotheses and resources for further experimental validations and designs.

A number of previous studies have made attempts to computationally investigate the interaction patterns between proteins and RNAs. For instance, Engtangle [14] was proposed to identify the residue interactions at the interface between protein and RNA, and these interactions may result from hydrogen bond, electrostatic force, hydrophobicity, or van der Waals force. Kim et al. [15] proposed a new definition of residue-based propensity and statistically evaluated the pairing preferences of different amino acid residues in the RNA interface using 85 protein-RNA complexes. Perez-Cano and Fernandez-Recio [16] also proposed a propensity-based method called OPRA (Optimal Protein-RNA Area) to predict RNA-binding areas on proteins. Furthermore, using machine learning techniques such as support vector machine (SVM) and naïve Bayesian (NB), several bioinformatics tools, including BindN [17], RNABindR [18] and PPRint [19], have been implemented to predict RNA-binding residues in proteins. Recently, we also developed a more efficient algorithm named PRNA for RNA-binding site prediction [20], which considered the features of both protein and RNA. However, it should be stressed that all the methods described above focus on determining the RNA-binding residues based on the information from existing protein-RNA complexes, and thus cannot be directly applied to the prediction of protein-RNA interactions.

An increasing number of data on protein-RNA interactions have been generated and made publically available [21], which enables more systematic investigations of such interactions from a bioinformatics point of view. For example, Jones et al. [6] and Ellis et al. [7] computationally analyzed 32 and 89 protein-RNA complexes, respectively, to investigate the RNA-binding patterns of protein

* Corresponding authors.
E-mail addresses: zpliu@sdu.edu.cn (Z.-P. Liu), hongyu.miao@uth.tmc.edu (H. Miao).

http://dx.doi.org/10.1016/j.neucom.2015.11.105
0925-2312/

residues. However, the methods for predicting protein-RNA interactions are still lacking. Recently, Our group [22] and Muppirala et al. [23] have proposed approaches to meet the needs in methodology development in this area, both of which are solely based on sequence-derived features. Nevertheless, there exists strong evidence suggesting that both sequence and structure features of protein and RNA can significantly contribute to the protein-RNA binding events [3], it is therefore necessary and desirable to take both sequence and structure information into consideration to develop more accurate and efficient computational methods.

In this study, we integrate several important sequence-based and structure-based features to develop a new computational method for predicting protein-RNA interactions. More specifically, we consider six and four sequence-based descriptors for protein amino acid residues and RNA nucleotide bases, respectively. In addition, we extract the structure feature by calculating the mutual interaction propensity between amino acids and nucleotides in protein-RNA complexes. The RNA secondary structure is obtained by employing a folding algorithm and the paired nucleotide information are then encoded. All the features from one pair of protein and RNA are stacked into a feature vector, and the feature vectors associated with different protein-RNA pairs are transformed into vectors of the same length using the auto covariance technique. Using a cross-validation scheme, we compare and select the classifier with the best performance and build a predictor for protein-RNA interactions. We also evaluate the importance of the selected molecular descriptors or features to the prediction of protein-RNA interactions. Further case studies are conducted to illustrate the accuracy and usefulness of the proposed method for predicting the interactions between protein and microRNA (miRNA) or long non-coding RNA (lncRNA).

2. Materials and methods

2.1. Datasets

We compile four datasets (called PDBInter, NPInter, RPI2241 and RPI369, respectively) on protein-RNA interactions from representative public databases. First, we download the documented protein-RNA complexes from the PDB database [21], and obtain 896 complexes. For simplicity, we only consider the complexes with a protein sequence length of 15–200 amino acid residues and an RNA sequence length of 5–200 nucleotides. Using BLASTclust [24], we remove the homologous proteins and RNAs according to a sequence similarity of 25% and 60%, respectively. Such thresholds are widely used to remove the homologous protein-RNA pairs that will result in biases in predicting interactions [4,20]. After that, we obtain 158 protein-RNA interaction pairs in multiple species, and this dataset is denoted as PDBInter. Second, we retrieve data from the NPInter database [25], which contains experimentally-verified interactions between proteins and non-coding RNAs in six organisms [25]. A total of 11,441 pairs of non-redundant protein-RNA interactions are obtained using the similar procedure described above for PDB, and this dataset is named as NPInter. Third, we collect data from the RPISeq database [23]. This database contains two sets of data: one consists of 2241 experimentally-validated protein-RNA pairs in various organisms, and the other one only consists of 369 non-redundant pairs selected from the former dataset [23]. We denote the two datasets from RPISeq as RPI2241 and RPI369, respectively. Note that in all these databases, there are not explicit data on non-interacting pairs so we have to generate them by randomly pairing proteins and RNAs and avoiding all known interactions, which is necessary and critical to classifier training. Also, for fair evaluation, we suggest to make the number of non-interacting pairs comparable with the number of known protein-RNA interactions in each dataset. All the datasets and computer codes are available at our website http://doc.aporc.org/wiki/PRNAinter.

2.2. Descriptors of protein-RNA interactions

To quantitatively characterize complex molecular interactions, we consider molecular descriptors (i.e., features) of protein and RNA that are potentially important from the physicochemical, constitutional, or topological point of view [23,26]. Several sequence and structure descriptors have been shown to be likely to play crucial roles in predicting protein-RNA interactions, including number of atoms [14], electrostatic charges [14], number of potential hydrogen bonds [14,27], hydrophobicity [17,26], evolutionary conservation score [28], pKa value [17], mass value [17], and the mutual interaction propensity [7,16,20,26]. Therefore, we consider the descriptors above as representative features and incorporate them into our prediction strategy in this study. More specifically, for an amino acid residue, six sequence-based physicochemical descriptors are extracted, including the number of atoms (Atom), the number of electrostatic charges (Charge), the number of potential hydrogen bonds (Bond), the hydrophobicity, the conservation score, and the pKa value. The hydrophobicity of an amino acid residue is quantified using the hydrophobic index (Hydro) proposed by Kabsch and Sander [29]; the conservation score of a residue is retrieved from the corresponding position-specific scoring matrix (PSSM), which is generated using PSI-BLAST [24] by searching the non-redundant protein sequences in the UniProt database [30]; and the pKa value of an amino acid residue is readily available in Lehninger et al. [31] and the details can be found in Nakamura [32]. For each of the four types of RNA nucleotides, we consider three sequence-based descriptors, including the number of atoms (AtomR), the mass value (Mass) and the pKa value (pKaR).

For structure-based descriptors, we consider the mutual interaction propensity between protein and RNA residues. First, we use Entangle [14] to identify the binding sites in protein-RNA complexes. If the central amino acid residue of a triplet interacts with a nucleotide, we define the mutual interaction propensity (MIP) between a residue triplet in protein and a nucleotide in RNA as follows [20]:

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

where $f_{p,r}(x,y)=\frac{N_{p,r}(x,y)}{\sum_{x,y}N_{p,r}(x,y)}$, $f_p(x)=\frac{N_p(x)}{\sum_x N_p(x)}$ and $f_r(y)=\frac{N_r(y)}{\sum_y N_r(y)}$. In addition, (p,r) denotes a protein-RNA interaction pair, x denotes an amino acid residue triplet, y denotes an RNA nucleotide, $N_p(x)$ is the number of residue triplets in the protein, $N_r(y)$ is the number of nucleotide residues in the RNA, and $N_{p,r}(x,y)$ is the number of x binding to y in the protein-RNA pair. According to the definitions above, one can tell that $f_p(x)$ 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 a pair (p,r). By definition, the MIP between x and y is calculated over an entire dataset of protein-RNA complexes. We derive the MIPs between amino acid triplets and nucleotides from the whole non-redundant dataset of protein-RNA complexes in PDBInter. For a given pair of protein and RNA, if their three-dimensional complexes are not available, the MIPs can be calculated by searching their correspondences in these derived MIPs from PDBInter. When there is no hit for a querying pair of amino acid triplet and nucleotide, the MIP value is simply denoted as 0. The secondary structure (Second) of RNA plays an important role in mediating protein-RNA interactions [33], so we also extract and encode this descriptor into the feature vector. RNAfold in