IOP Conf. Series: Earth and Environmental Science **69** (2017) 012056 doi:10.1088/1755-1315/69/1/012056

iRPIS-PseNNC: identifying RNA-protein interaction sites by incorporating the position-specific dinucleotide propensity into ensemble random forest approach

Long Li, Guojin Chen and Tingdu Jin

School of Mechanical Engineering, Hangzhou Dianzi University, HangZhou, 310018 China

Abstract. As the pile of RNA-Protein complexes sequences mounted, in order to overcome time-consuming problem of the traditional identify RNA-Protein interaction sites (RPIS) method, it is urgent need develop intelligent recognition approach for quickly and reliable recognition of the RNA-Protein interaction sites (RPIS). To settle the question, we developed a new method named iRPIS-PseNNC, in which each sample is a nineteen nucleotides segment that for positive the centre of the segments is RPIS and for negative the segments centre is non-RPIS, and the sample was obtained by sliding window. The RNA sample was formulated by combining the dipeptide position-specific propensity into random forest approach, and by using the random sampling to balance the training dataset. According the voting system, we combine eleven random forest together to construct an ensemble classifier. It is shown that via the rigorous cross validations that the new predictor "iRPIS-PseNNC" achieved very high percentage of accuracy than any other existing algorithms in this field, indicating that the iRPIS-PseNNC predictor will be an effective tool for prediction RNA-Protein interaction sites.

1 Introduction

The interaction between RNA and proteins play a critical role in many essential biochemical process, RNA-protein interaction(RPI) is involved in the regulation of many gene expression, human gene diseases, viral replication and transcription, and X-chromosome inactivation. Therefore, knowledge of RNA-protein interaction sites(RPIs) is of great importance for drug development and basic research.

Over the past decades, with the rapid development of the method to exploration and discover RNA biology. The method for exploring RNA-protein interactions can be divided into two general categories: 'experimental techniques 'and 'computational methods'. Although experimental studies on RPI such as RIP-Chip HITS-CLIP, RNA EMSA, SELEX, RNA compete, CLIP, oligonucleotide-targeted RNase H protection assays, PAR-CLIP, Ribonomics and Ribotrap. Unfortunately, those methods are very expensive time-consuming.

In fact, during the last decade, many computational methods for efficiently determinate the RPI without using experiments have been developed. For instance, Lu et al.[1] developed a method called lncPro which is by encoding RNA and protein sequence into numeric vectors and using matrix multiplication to score. By using the most common used machine learning methods SVM or RF that Muppirala et al.[2] proposed a new predictor named "RPISeq" to predict RNA-protein interaction. Perdrizet et al.[3] by using the physicochemical properties of the interface from the RNA-protein structures. Besides, several previous studies for prediction RNA-protein interaction sites in protein sequence.

However, only a few limit methods for predicting RNA-protein interaction in RNA sequence. Particularly, as the pile of RNA-Protein complexes sequences mounted in the post gen are, it is urgent need develop intelligent recognition approach for quickly and reliable recognition of the RNA-Protein interaction sites (RPIS).

Although the pervious works have important action in encouraging development in this area, which still exists the following disadvantages. The main limits are: (1) Ignore the information of the sequence order or the influence of the adjacent residues, therefore, pervious predictor accuracy be limited; (2) the prediction of RNA-protein interaction sites is a huge imbalance two kinds recognition problem, the reason is non-interaction sites of a RNA sequence is great than the number of interaction sites, and hence lead to identify deviations; (3) can further improve the classifier accuracy by introducing the most advance machine learning techniques.

In our work, we seek try to develop a new machine learning predictor called iRPIS-PseNNC for identifying the RNA-protein interaction sites(RPIS) in RNA sequence from the aforementioned three aspects. Accordingly, to construct the new machine learning predictor and avoiding those issues, including three critical factors. First, in order to ensure the reliable and stringent benchmark dataset, we adopt the published dataset; and then, based on sequence couple information that a discriminative statistic sequence represent method is proposed to descriptor our samples, and due to the dataset is extremely unbalanced, therefore, by using random sampling to optimize the imbalances training data sets that the result of minimizing the prediction error; finally, by incorporating the two factors mentioned into random forest to construct our predictor, and use the cross-validation to test the performance of the new classifier. Below, we will address those factors step by step.

2 Materials and methods

2.1 Figures and tables

In order to develop a new classifier, the most important first step is to construct a reliable standard data sets to train and test the new classifier. However, once the dataset has an error, the classifier accuracy will be unreliable and meaningless. In this study, the data set RNA-208 of the RNA-protein interaction RNA chain was derived from Panwar et al.[4], which contains 10198 interacting sites and 36384 non-interacting sites. In the previous studies, sliding widow approach has been are commonly used to research PTMs sites and HIV protease cleavage sites. In this study, we also use sliding window approach to study RNA-protein interaction nucleotides sites. According to use the sliding window method that the window size is given by $[-\xi,+\xi]$. Its width is $(2\xi+1)$, where ξ is an integer. In order to more clear description, we adopt the below method describe the RNA sample

$$R_{\xi} = N_{-\xi}N_{-(\xi-1)}\cdots N_{-2}N_{-1}N_0N_1N_2\cdots N_{(\xi-1)}N \qquad (1)$$

where $\mathbb{N}_0$ at the sample center is the target nucleotide, $N_{-\xi}$ is the ξth upstream residues, the -ξth downstream nucleotide, show as figure 1. The $(2\xi+1)-tuple$ peptides R_{ξ} can be divided into the two different parts

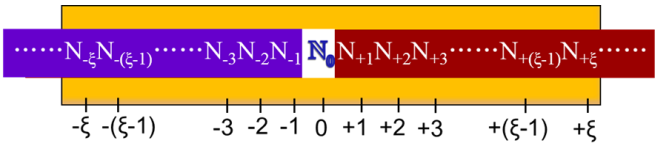
$$R_{\xi}(N)=\begin{cases} \mathrm{R}_{\xi}^{+} & if\ its\ center\ is\ \mathrm{int}eraction\ site \\ \mathrm{R}_{\xi}^{-} & otherwise \end{cases} \qquad (2)$$

According to the literature review [5], It is not necessary to divided a set of dataset into a training data and a testing data, if using the jack knife test or cross-validation to test the new classifier. Therefore, in this study, the data sets can be given as

$$\mathbb{I}_{\xi} = \mathbb{I}_{\xi}^{+} \bigcup \mathbb{I}_{\xi}^{-} \qquad (3)$$

where $\mathbb{I}_{\xi}^{+}$ only consists of positive samples R_{ξ}^{+}, i.e., the RPIS RNA segments; $\mathbb{I}_{\xi}^{-}$ only consists of negative samples R_{ξ}^{-}, i.e., the non-RPIS RNA segments; and $\bigcup$ represents "union" in the set theory.

Figure 1: The illustration shows the length sample of (2ξ+1) with interaction site at the centre.



The sample segment length $R_\xi(N)$ is $(2\xi+1)$, difference of values ξ will have different length samples in the this study, can be expression as

$$\mathbb{I}_\xi = \begin{cases} 17 \text{ residues} & when\ \xi = 8 \\ 19\ residues & when\ \xi = 9 \\ 21\ residues & when\ \xi = 10 \\ \vdots & \end{cases} \tag{4}$$

The detail process of construction $\mathbb{I}_\xi$ are described as below. If the length of the segment downstream or the length of the segment upstream was short of ξ, the missing parts were completed by the way of circulation. When the sample centre is annotated experimental RNA-protein sites remark the sample as positive and added into $\mathbb{I}_\xi^+$; otherwise, when the sample centre is annotated non RNA-protein sites remark it as negative and add it into $\mathbb{I}_\xi^-$. Therefore, in this study, all the samples were obtained through the same screen process as above mentioned. Before take the dataset to train and test, the most important thing is to exclude those self-conflict data, which the data are both exit in RPISsubset $\mathbb{I}_\xi^+$ and non-RPIS subset $\mathbb{I}_\xi^-$. According to these procedures, and using $\xi = 8,9,10$ as the half width of the window size, we construct four dataset; they are $\mathbb{I}_{\xi=17}, \mathbb{I}_{\xi=19}$, $\mathbb{I}_{\xi=20}$ and $\mathbb{I}_{\xi=21}$, respectively. However, based on the precendent experimental, when $\xi = 9$ that the RNA segments were consisted of $9\times2+1=19$ nucleotides residues, the results accuracy on the dataset were very high. Therefore, we choose the benchmark dataset $\mathbb{I}_{\xi=19}$ as training and testing the new predictor.

2.2 Representation of RNA Segment Samples

Based on the mentioned above process, the samples can be formulated as

$$\mathrm{R} = N_1 N_2 N_3 \cdots N_i \cdots N_L \tag{5}$$

Where

$$N_i \in \{\mathrm{A}(adenine), C(cyto\sin e), G(guanine), U(uracil)\}$$
$$(i = 1,2,3,\cdots,L)$$

where N_i is the nucleotide residue at the sequence position $(i = 1,2,3,\cdots,L = 19)$. Further, a RNA sequence feature vector can be expressed by its by its nucleotides composition (NC), s

$$R = [f(A), f(C), f(G), f(U)] \tag{6}$$

where $f(A)$, $f(C)$, $f(G)$, and $f(U)$ standard frequencies that adenine (A), cytosine (C), guanine (G), and guanine (U) appear in sample sequence; the symbol T means the transpose operator. As show Eq.6, however, if a RNA sample is represented by using NC, then the relationship between the sequence of residues would be fully lost. When the dinucleotide composition (DNC) are used to represent the RNA sequence, rather than the residues frequencies as show in Eq.6, the RNA sample as given by:

$$\begin{aligned} R &= [f(AA), f(AC), f(AG), \cdots, f(UU)] \\ &= [f_1^2, f_2^2, f_3^2, \cdots, f_{16}^2] \end{aligned} \tag{7}$$

Where $f_1^2 = f(AA)$ stands for the standard frequency of AA in the samples, $f_2^2 = f(AC)$ stands for the AC, $f_3^2 = f(AG)$ is AG, and so forth. According to do this way, we can only merge the local sequence-order information between the most adjacent nucleotides, but certainly does not contain it's the global or long-range sequence order or pattern information.

Based on the construct of other components, the sample information can be represented by the general formula as follow

$$R = [\Psi_1, \Psi_2, \cdots, \Psi\mu, \cdots, \Psi_\Omega]^T \quad (8)$$

where the subscript Ω it's value as same as the components $\Psi\mu = (\mu = 1, 2, \cdots, \Omega)$ will rest with the required information how extract from the sample information from sequence *R*. Next, the "position-specific dinucleotide propensity matrix" will be used to define the Ω elements in Eq.8 in this study.

The construct sample length in the benchmark dataset $\mathbb{I}_\xi$ is 19, and the number of possible different dinucleotides is $4 \times 4 = 16$ in an RNA sequence and the number of dinucleotides sub-site positions will be $19 - 1 = 18$ on the sequence of Eq.1. Therefore, the RNA sample R of Eq.1 can be formulated with a $16 \times (L-1)$ position-specific dinucleotide property matrix as given by

$$R = \begin{bmatrix} N_{1,1} & N_{1,2} & \cdots & N_{1,19} \\ N_{2,1} & N_{2,1} & \cdots & N_{2,19} \\ \cdots & \cdots & \cdots & \cdots \\ N_{16,1} & N_{16,1} & \cdots & N_{16,19} \end{bmatrix} \quad (9)$$

where

$$\mathrm{R}_{i,j} = R^+(\Delta_i \mid j) - R^-(\Delta_i \mid j) \quad (10)$$

and

$$\Delta_1 = AA, \Delta_2 = AC, \Delta_3 = AG, \cdots, \Delta_1 = TT \quad (11)$$

In Eq.8, $R^+(\Delta_i \mid j)$ is the frequency that the Δ_i dinucleotide occur in the *j-th* sub site on the sequence of Eq.1 and it can be easily obtained from the bench data dataset $\mathbb{I}_\xi^+$; while $R^-(\Delta_i \mid j)$ represents the contingent probability, but from the benchadata $\mathbb{I}_\xi^-$. In this progress, each predicted RNA-protein interaction sites (or non RNA-protein interaction sites) sample's own information is removed.

Therefore, the RNA sample of Eq.1 can be defined via Eq.8, and the dimension of Eq.8 Ω=18 and its μ-th component can be formulated as

$$\Psi_\mu = \begin{cases} N_{1,\mu} & when\ N_\mu N_{\mu+1} = AA \\ N_{2,\mu} & when\ N_\mu N_{\mu+1} = AC \\ N_{3,\mu} & when\ N_\mu N_{\mu+1} = AG \\ \quad\vdots & \\ N_{16,\mu} & when\ N_\mu N_{\mu+1} = TT \end{cases} \quad (1 \le \mu \le 19) \quad (12)$$

2.3 Ensemble Random Forest Algorithm

The random forest (RF) algorithm is a very effective algorithm, which has obtained a lot of satisfactory results in many fields, such as bioinformatics computing.

However, it should be noted that the current case that negative sample number is greater than the positive samples, but most of the classification (including random forest) is used to balance issue. To solve the problem, an asymmetric bootstrap approach was adopted as elaborated. According to bootstrap approach, we random selected non RNA-protein interaction sites equal number of the RNA-protein interaction sites, and then, wo do eleven times. Therefore, in this study, we can formulate with eleven different predictor modes, and which are used to train the predictor, and all of them are independent. Finally, we obtained eleven independent different classifiers, and used them to identify RNA-protein interaction sites, which can be given by

$$RPIS\ \text{predictor} = \mathbb{RF}(\mathrm{k}), k = 1, 2, 3, \cdots, 11 \quad (13)$$

Where $\mathbb{RF}(\mathrm{k})$ is one of the RF classifiers based on the eleven different modes.

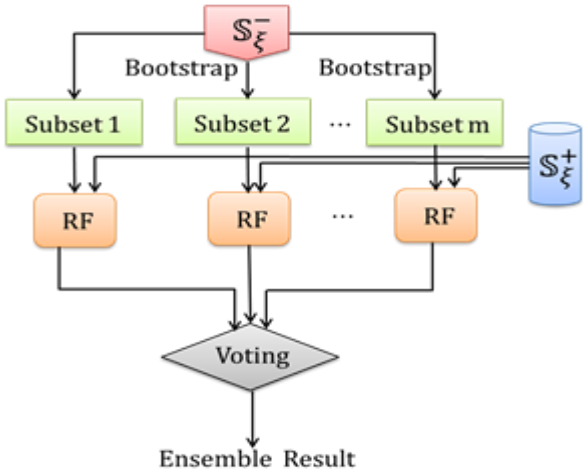
But for the moment, the key problem is to find a balance method which can combine the eleven predictor result that can make the result obtain maximize quality. According to the previous research,

we adopt the ensemble learning to fuse many different predictors which can improve the success accuracy for identifying protein subcellular location and protein quaternary structural attribute. Therefore, in this study, we also adopt the ensemble learning by using voting system to vote the eleven different predictor $\mathbb{RF}(\mathrm{k}), k=1,2,3,\cdots,11$ can be given as

$$\mathbb{RF}^{E}=\mathbb{RF}(1)\forall\cdots\forall\mathbb{RF}(11)=\forall_{k=1}^{11}\mathbb{RF}(\mathrm{k}) \qquad (14)$$

where $\mathbb{RF}^{E}$ is the ensemble learning predictor, and the symbol $\forall$ for the merge process. And the details of the voting system, can obtained from Eqs.30-35 in [5], where has been described clearly. Of cause, for quickly to know it, we give picture in below Figure 2, which show the voting process of the eleven RF predictor.

Figure 2: An overview of the sub-ensemble classifier. The number of the interaction sites is much less than that of non-interaction sites



3 Result and discussion

3.1 Comparison with the Existing Methods

By using four different properties index to e evaluate the predictor performance, which listed in Table 1, which are achieved by the iRPIS-PseNNC predictor via the 5-fold cross-validation and in order to more clear describe our predictor we also compare the existing method that the result also list in Table 1, we can observe the following: (1)our predictor iRPIS-PseNNC achieved very high percentage of accuracy than any other existing algorithms in this field[4]; (2) It is also applicable to other three index, which made it clear that identifying a new predictor for identifying RNA-protein interaction sites in RNA sequence, not only can produce higher prediction accuracy, but also has a more stable and higher sensitivity and specificity.

Predictor	ACC	MCC	Sn	Sp	AUC
rnapin[a]	0.8392	0.62	0.8482	0.8462	0.832
iRPIS-PseNNC	0.9436	0.8896	0.9088	0.9792	0.939

[a]Results reported by Panwar et al.[4].

Table 1. Comparison of the iRPIS-PseNNC with the other existing methods via the 5-fold cross-validation

According to the values list in Table 1, from the table, the new predictor iRPIS-PseNNC iRPIS-PseNNC achieved very high percentage of accuracy proposed in this paper remarkably the existing predictor, particularly in the below index Acc and MCC; ACC means the predictor have higher accuracy, and MCC means the predictor is more stability.

We also use the ROC curve[6,7] to evaluate the predictor performance, please see the Table 1, the value of AUC which is the area under ROC curve, our new predictor iRPIS-PseNNC have high value than exiting predictor.

4 Conclusion

To timely acquire the information of the RNA-Protein interaction sites in genomes sequence is important for in-depth studying gene expression, function and developing new drug. In this work, we proposed a new method for the prediction of the RNA-Protein interaction sites of genomes by combining the dipeptide position-specific propensity into random forest approach, and by using the random sampling to balance the training dataset. The results were very promising, which anticipated that our predictor may also be used for many other genome analysis problems.

Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (No. 51675148).

References

[1] Lu Q, Ren S, Ming L, Yong Z, Zhu D, et al. Computational prediction of associations between long non-coding RNAs and proteins. Bmc Genomics, **14**, pp. 1-10 (2013).
[2] Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-Protein Interactions Using Only Sequence Information. Bmc Bioinformatics, **12**, pp. 1-11, (2011).
[3] Parisien M, Wang X, Lamphear C, Fierke CA, Maheshwari KC, et al. Discovering RNA-Protein Interactome by Using Chemical Context Profiling of the RNA-Protein Interface. Cell Reports, **3**, pp. 1703–1713, (2013).
[4] Bharat P, Gajendra P S R. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. Genomics, **105**, pp. 197–203 (2015).
[5] Chou KC, Shen HB. Recent progress in protein subcellular location prediction. Anal Biochem, **370**, pp. 1-16, (2007).
[6] Fawcett T. ROC graphs: Notes and practical considerations for researchers. Machine Learning, **31**, pp. 1-38, (2004).
[7] LDavis J, Goadrich M. The relationship between Precision-Recall and ROC curves. ACM. pp. 233-240, (2006) .