Computational Prediction of RNA-Protein Interactions

Hilal Kazan
Department of Computer Engineering, Faculty of Engineering, Antalya International University, Antalya, Turkey

Received Date: 2016-06-15 Accepted Date: 2016-10-10

Abstract

RNA-protein interactions play critical roles in diverse cellular processes including post-transcriptional regulation of gene expression and infection by pathogens. As such, characterization of RNA-protein interactions will lead to a better understanding of these mechanisms and associated diseases. Experimental methods to determine RNA-protein interactions remain tedious and expensive. An alternative strategy is to use computational methods to predict RNA-protein interactions. Here, we develop a random forest model that uses sequence information of an RNA-protein pair to determine whether they will interact or not. We evaluate our model with three diverse datasets including one dataset that has never been used for this purpose before. For the two other datasets, our model gives a better performance than existing methods. We also show that including features that represent the physico-chemical properties of the protein or RNA secondary structure. Altogether, these results show that RNA-protein interactions can be predicted accurately with computational models.

Keywords: RNA-protein interaction, machine-learning, random forests, physico-chemical properties, RNA secondary structure

1. Introduction

RNA-protein interactions have fundamental roles in several biological mechanisms such as RNA processing [1], gene expression control [2], protein synthesis [3,4], viral replication and pathogen resistance [5]. Aberrations in RNA-protein interactions have been linked to several diseases including neurodegenerative disorders and cancer [6]. As such, understanding the principles that govern RNA-protein interactions is critical.

RNA-protein recognition is more complex than DNA-protein recognition; because, in contrast to the B-form helical structure of the DNA, RNA molecules fold into an A-form helical structure whose major groove is less accessible for proteins. Therefore, base-specific interactions are mostly seen with the single-stranded regions of the RNA. Moreover, unlike DNA, RNA molecules can fold into complex and diverse structures, and this makes the RNA-protein recognition problem more challenging.

The most reliable approach to characterize RNA-protein interactions is to solve the three-dimensional structure using X-ray crystallography or NMR spectroscopy. However, these methods are time-consuming and expensive. Recently, a number of high throughput approaches have been developed to identify RNA-protein interactions *in vivo* and *in vitro*. RIP-Chip is an *in vivo* method that first immunoprecipitates the RNA-protein complexes, and then identifies them using the microarray technique [7]. CLIP and PAR-CLIP are *in vivo* methods based on UV crosslinking and immunoprecipitation [8, 9]. Due to its complex protocol, CLIP method has been applied to a small number of RBPs so far. On the other hand, RNAcompete is an *in vitro* array-based method that has been applied to a large number of RBPs [10].

Computational methods provide a less costly and more robust alternative to the experimental characterization of RNA-protein interactions. Pancaldi et al predicted RNA-protein interactions in budding yeast with a machine learning model that uses more than 100 gene and protein features including features related to RNA structure, translational features, expression levels and protein properties [11]. Wang et al proposed a naïve Bayes classifier to predict RNA-protein interactions [12]. They encode the protein sequences with a reduced alphabet that groups the amino acids into four classes. The feature set that they use includes the frequency of all 3-mers in protein sequences (64 features), the frequency of all 3-mers in RNA-sequences (64 features), and all possible combinations of 3-mers in proteins and RNAs (64*64 features). Since this results in a large number of features, they apply feature selection methods to reduce the parameter space. Suresh et al use both sequence and structure-based features in a support vector machine model to predict RNA-protein interactions [13]. Namely, they use the protein block representation derived from the crystal structure of the proteins. Also, they include the structural context of each RNA nucleotide in their features. Their results indicate that

*Corresponding author: Address: Department of Computer Engineering, Faculty of Engineering, Antalya International University, Antalya, Turkey. E-mail address: hilal.kazan@antalya.edu.tr, Phone: +902422450271

Doi:10.21541/apjes.92217

structure-based features improve the prediction accuracy over sequence-only models.

Here, we propose a new machine-learning method that uses simple sequence-based features in a random-forest model to predict RNA-protein interactions. We also assess the predictive accuracy of additional features that represent the physico-chemical properties of the protein and RNA secondary structure. Unlike many of the previous models, we tune the parameters of the random forest correctly using nested cross validation. We evaluate our model on three datasets from diverse organisms and show that it can predict RNA-protein interactions accurately. One of these datasets, RNAcompete is used in RNA-protein interaction prediction for the first time. For the other two datasets, our model gives a better a better performance than existing models that use a much larger feature set.

2. Material and Methods

2.1 Data Collection

We treat the problem of RNA-protein interaction prediction as a binary classification task. Namely given two strings that correspond to the sequences of the protein and the RNA, we have to determine whether the pair will interact or not. We used three diverse datasets to evaluate our model: (i) RNAcompete dataset (ii) NDB-PRIDB dataset (iii) Pancaldi dataset.

RNAcompete is a high-throughput microarray based binding assay to detect the binding preferences of RNA-binding proteins *in vitro*. Namely, the protein of interest is incubated with a large pool of short RNA sequences and the bound fraction is identified with a microarray containing probes that are complementary to the RNA sequences in the pool. In the end, the intensities of the probes correspond to an estimate of the binding affinity of the protein to each of the RNA sequences in the pool. The custom designed RNA pool contains more than 240,000 short RNAs (30-41 nucleotides) where each 9-mer (i.e., subsequence of length 9) appears at least 16 times and each 7-mer appears at least 155 times. To summarize the binding preferences of an RBP, a score is calculated for each 7-mer. First, all probes that have negative normalized values are assigned zero. Then, the score for a 7-mer is calculated by taking the trimmed mean (ignoring the top and bottom 5% quartile) of the intensities of the probes that contain that 7-mer.

We downloaded the 7-mer scores of 66 human RBPs from the supplementary website (http://hugheslab.ccbr.utoronto.ca/supplementary-data/RNAcompete_eukarya/). We identified the top and bottom scoring 10 7-mers for each RBP. Then, we formed the dataset by pairing the sequence of the protein with each of the 7-mers and labeling 1 or 0 based on whether it is a high scoring or low scoring 7-mer, respectively. In total, this dataset includes 1320 RNA-protein interactions. When preparing the training and test sets, we ensured that the top and bottom 7-mers of the same protein are considered completely in either the training set or the test set to avoid any bias. In other words, we split the training and test sets based on protein IDs.

We downloaded the NDB-PRIDB dataset from the supplementary data of Suresh et al paper. This dataset is compiled from RNA-protein complexes available in Nucleic Acid Database (NDB) [14] and the protein-RNA interface database (PRIDB) [15]. A distance cutoff of 3.4 Angstrom is used to define the positive and negative pairs. Namely, if any of the atoms in the protein is within distance <= 3.4 Angstrom to any of the atoms in the RNA molecule this pair is classified as interacting. Otherwise, the pair is labeled as non-interacting. Final dataset includes 1807 positive pairs (including 1807 protein and 1078 RNA chains) and 1436 negative pairs (including 1436 protein and 493 RNA chains).

Lastly, we also downloaded the dataset prepared by Pancaldi et al. The positive instances of this dataset include the 5166 mRNA-protein interactions detected by RIP-chip experiments performed in yeast [16] These interactions are shuffled to form the negative set. We should note that the average length of the RNAs (i.e., 1334 nts) is much longer compared to the previous two datasets.

2.2. Feature Compilation

We compiled features to represent a protein-RNA pair. For protein sequences, we counted the number of times each 3-mer (amino acid sequence of length 3) appears. To reduce the number of features, we encoded the protein sequences using a smaller alphabet that classifies the amino acids into 7 groups according to their dipole moments and the volume of their side chain: {A, G, V}, {I,L,F,P},{Y,M,T,S},{H,N,Q,W},{R,K},{D,E},{C} [17]. Similarly, for RNA sequences, we counted the frequency of each 3-mer. As such, the feature vector of a protein-RNA pair contains 407 features (343 for proteins and 64 for RNA sequences). Hereafter we call this feature set "sequence-only feature set".

We also tried including features that represent the physico-chemical properties of the aminoacids (downloaded from http://www.bmrb.wisc.edu/). These additional features include Chou-Fassman helix and sheet propensity values to represent the structure of the protein. We also included pKa value for free amino acid carboxylate, pKa value for free amino acid

amine and pKa value for amino acid side chain as these are important in determining the pH-dependent characteristics of the protein. Next, there are features that represent the number of carbon, hydrogen, nitrogen, oxygen and sulfur atoms in each amino acid, and hydrophobicity. Finally, we included features related to the accessible surface area of the protein. The number of features for each protein-RNA pair is 421 (14 features for physico-chemical properties and 407 for sequence-only features). Hereafter, we call this feature set "physico-chemical feature set".

Lastly, we also assessed the effect of RNA secondary structure by including features that take into account the structure context of RNA nucleotides. Namely, we predicted the secondary structure of RNA sequences using an existing computational method called RNAplfold [18]. RNAplfold employs local folding where RNA is folded locally in a sliding window approach and predictions are averaged over all windows. For each position, probabilities of being in the following five structural contexts are given as output: paired context, hairpin loop, internal loop, multiloop, external loop. We chose this method as a previous study showed that local folding is more accurate than global folding [19]. Also, in agreement with the fact that RNAs can fold into multiple diverse structures; RNAplfold takes into account the ensemble of all possible structures rather than predicting only the minimum free energy structure. We extended the RNA features to also include secondary structure information. For instance, instead of counting the frequency of AAA, we summed the average probability of AAA to be in paired context (i.e. AAA-P), the average probability of AAA to be in hairpin loop (i.e., AAA-H) etc. As such, for each 3-mer we included five features. In total, this feature set contains 663 features (320 features from RNA sequence and 343 features from protein sequence). Hereafter, we call this feature set "RNA structure feature set".

2.3. Random Forest

Random forest consists of a collection of decision trees [20]. Random forest is fit using two layers of randomness. First layer of randomness is due to the fact that a set of bootstrapped samples is used to fit each tree. The second layer of randomness is related to the selection of the feature to split a node. In standard decision trees, each node is split based on the best feature where best is defined based on a cost function. In random forest, first, a random set of features is sampled for each node. The split is then determined based on the best feature among the sampled features. In summary, each tree is trained with a random sample of data points and each split is based on a random sample of features. Random forest model also estimates the importance of a feature by assessing the increase in prediction error when the values of that feature are shuffled. We used the random forest implementation in the *scikit-learn* package (Python) to conduct our experiments.

We chose the parameter *n_estimators* (number of trees in the forest) using nested cross validation (10-fold). We tried the values 50, 100, and 500 for *n_estimators. We* set the *max_features* (the number of features to sample when splitting a node) as the square root of the number of features.

2.4. Performance Evaluation

We employed 10-fold cross-validation procedure to evaluate the performance of the random forest models. We used the following performance metrics:

$$\text{Accuracy} = (TP + TN) / N$$
$$\text{Precision} = TP / (TP + FP)$$
$$\text{Recall} = TP / (TP + FN)$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FN is the number of false negatives and N is the total number of data points. Additionally, we calculated the area under the ROC curve (AU-ROC) that corresponds to the expected proportion of positive data points ranked before a randomly chosen negative data point. We plot the interpolated ROC curve of 10 curves that correspond to the results on 10 CV folds.

3. Results

We evaluated our model with three datasets.

3.1. RNAcompete Dataset

The table below shows the average cross-validation results of our model on RNAcompete dataset using two different feature sets. In sequence-only model we only used the features that are based on the sequence of the RNA and protein sequences. This model predicted the interacting RNAs for RBPs successfully (AU-ROC: 0.90, Table 1). In physico-chemical model, in addition to the sequence features we also included features that represent the physico-chemical properties of the amino acids in the protein. We found that including physico-chemical features increased the recall slightly but decreased accuracy, precision and AU-ROC (Table 1 and Figure 1). This is likely to be due to the small size of the dataset. We could not include RNA secondary structure related features as each RNA sequence (i.e. 7-mer) appears in multiple probes and each probe has a distinct RNA secondary structure. We also note that RNA related features rank higher than protein related features in terms of importance values estimated by the random forest model.

Table 1. Results of the random forest model on RNAcompete dataset

Model	Precision	Recall	Accuracy	AU-ROC
sequence only	0.883	0.711	0.809	0.904
physico-chemical	0.777	0.727	0.753	0.820

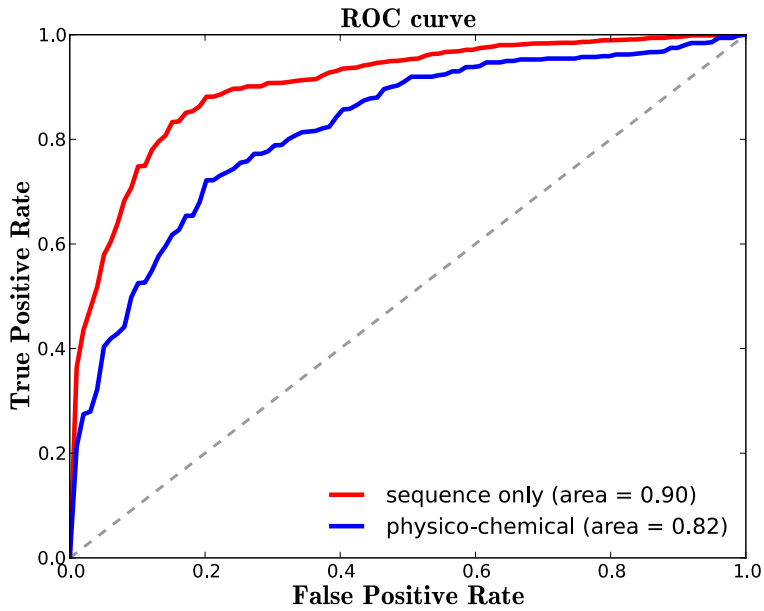

Figure 1. ROC curve of the models

3.2. NDB-PRIDB Dataset

Next, we evaluated our model with the dataset compiled from NDB and PRIDB datasets. Our sequence-only model has a remarkable performance with an AU-ROC of 0.983 (Table 2). Again, we observe no significant improvement when physico-chemical features or RNA structure features are included. The recently proposed RPI-Pred model by Suresh et al achieves an AU-ROC of 0.97 on the same dataset even though they use several additional features related to the structure of the protein and the RNA. Here our model that only uses sequence-based features achieves a better performance. We also note that the most important features are related to protein sequence according to the random forest model. For instance, the most important feature is the frequency of aminoacids that are from the groups {A, G, V}, {R,K}, {I,L,F,P}, respectively. This finding is in line with the previous literature which identified arginine and phenylalanine as enriched in the RNA binding sites [21].

3.3. Pancaldi Dataset

Lastly, we evaluated our model with the dataset prepared by Pancaldi et al. The model developed by Pancaldi et al gives an average accuracy of 0.69 and average AU-ROC of 0.77 on the same dataset. However, they use a large feature set (i.e., 120 features) that includes features based on diverse properties of protein and RNA such as the localization of the protein, gene ontology class of the protein, physical properties of the protein, expression level, ribosome density and structure of the mRNA etc. Their model cannot be generalized to other datasets specifically from other organisms as these features are difficult to compile. Our model based on simple sequence-based features result in a much better performance (AUROC: 0.80 vs 0.77) than Pancaldi et al (Table 3). Additionally, we see a slight improvement when we include the physico-chemical features. On the other hand, including RNA structure information decreases the performance. When we look at the feature importance values estimated by the random forest model, we observe RNA-derived features are found to be more important than protein-derived features similar to our results in RNAcompete dataset.

Table 2. Results of the random forest model on NDB-PRIDB dataset

sequence only	0.941	0.973	0.950	0.983
physico-chemical	0.937	0.973	0.947	0.984
RNA structure	0.935	0.969	0.944	0.984

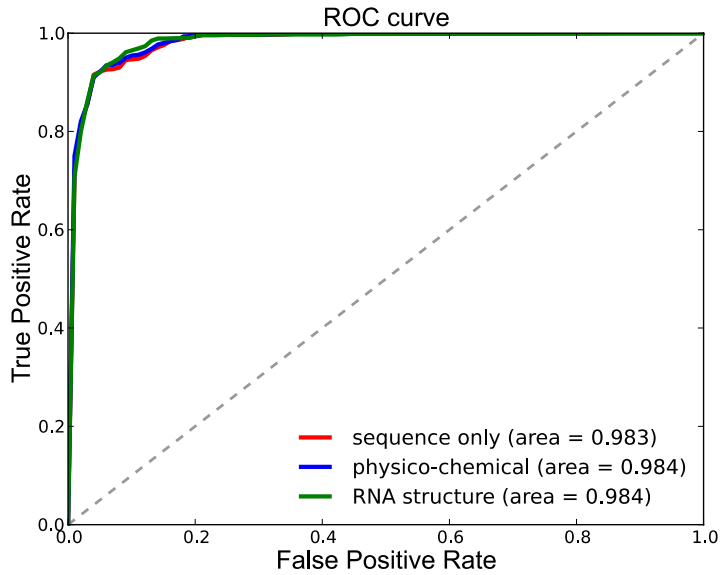

Figure 2. ROC curve of the models

Table 3. Results of the random forest model on Pancaldi dataset

Model	Precision	Recall	Accuracy	AU-ROC
sequence only	0.719	0.728	0.722	0.802
physico-chemical	0.722	0.729	0.724	0.803
RNA structure	0.713	0.714	0.714	0.792

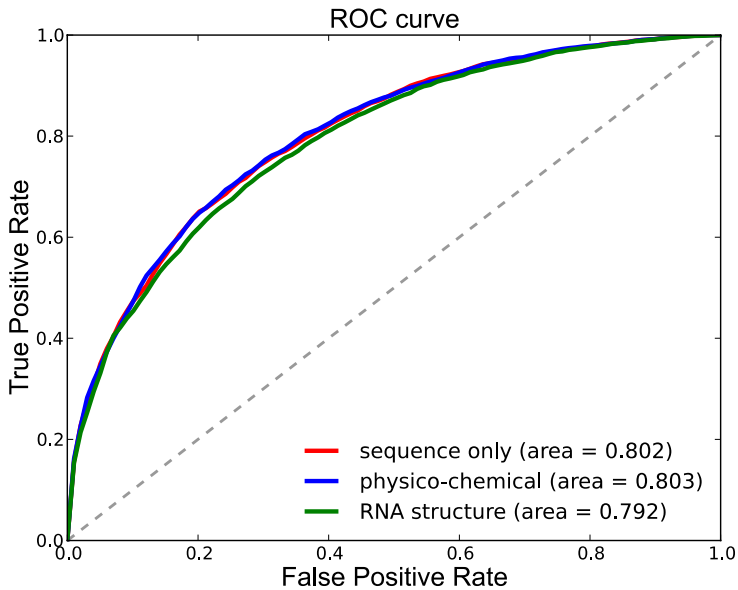

Figure 3. ROC curve of the models

3.4. Comparison with Other Approaches

Here we compare our random forest model that uses 3-mer frequencies with a reduced aminoacid alphabet with other statistical methods. We only performed the experiments with sequence-only model as including physico-chemical features or RNA secondary structure features do not result in a great change of performance. First, we show that using 1-mer and 2-mer frequencies with the full set of aminoacid alphabet reduces the performance. The decrease in performance was more pronounced for RNAcompete and Pancaldi datasets with small changes for the NDB-PRIDB dataset (Table 4). Next, we evaluated the effect of feature selection by only using the features that appear in the top 20 percentile when p-values obtained from chi-square test are sorted in increasing order (using Python's sklearn.feature_selection.chi2 method). Table 5 shows the results for all three datasets. We observe that using all the features gives a better performance.

Lastly, we tried other machine learning approaches: decision trees, naïve Bayes and SVM using scikit-learn package in Python. We tried both the linear and the RBF kernel for SVM model, which we name SVM-linear and SVM-RBF hereafter. Hyper parameters of the models were selected with nested cross validation (using GridSearchCV function). Table 6 shows the predictive performance of these approaches for the three datasets.

Table 4. Comparison of models that use 1-mer and 2-mer frequencies

	RNAcompete		NDB-PRIDB		Pancaldi	
Model	Accuracy	AU-ROC	Accuracy	AU-ROC	Accuracy	AU-ROC
1-mer	0.759	0.844	0.944	0.983	0.657	0.716
2-mer	0.800	0.896	0.944	0.982	0.659	0.720
our model	0.809	0.904	0.950	0.983	0.722	0.802

Table 5. The effect of applying feature selection on predictive performance

	RNAcompete		NDB-PRIDB		Pancaldi	
Model	Accuracy	AU-ROC	Accuracy	AU-ROC	Accuracy	AU-ROC
selected features	0.807	0.889	0.935	0.972	0.647	0.700
our model	0.809	0.904	0.950	0.983	0.722	0.802

Table 6. Comparison of the performance of other classification techniques

	RNAcompete		NDB-PRIDB		Pancaldi	
Model	Accuracy	AU-ROC	Accuracy	AU-ROC	Accuracy	AU-ROC
naïve Bayes	0.761	0.829	0.636	0.824	0.430	0.414
decision trees	0.698	0.698	0.910	0.913	0.608	0.608
SVM-linear	0.750	0.844	0.922	0.942	0.476	0.469
SVM-RBF	0.693	0.877	0.935	0.974	0.667	0.726
our model	0.809	0.904	0.950	0.983	0.722	0.802

We see that the naïve bayes and SVM models perform much better than the decision tree model for RNAcompete dataset. For NDB-PRIDB dataset, we observe that SVM-RBF model achieves a performance that is almost as good as our model. SVM with the linear kernel has slightly lower performance than SVM-RBF, whereas decision trees and naïve Bayes models perform worse. Lastly, for Pancaldi dataset, all models had considerably lower performance than our proposed model. In summary, for all the datasets, our random forest model has the best performance when compared to decision tree, naïve Bayes and SVM models.

4. Conclusion and Discussion

RNA-protein interactions play crucial roles in every step of RNA metabolism. Experimental methods to identify RNA-protein interactions are still time-consuming and expensive. As such, computational models are needed to fill this gap. In this study, we have developed a random forest model to predict whether a given RNA-protein pair will interact or not. We prepared different types of features: (i) features that are based on only the sequence content of the protein-RNA pair, (ii) features that represent the physico-chemical properties of the aminoacids in the protein, (iii) features that represent the secondary structure of the RNA. We evaluated our model with three diverse datasets. The first dataset is based on the RNAcompete outputs. This dataset is used for the first time for predicting RNA-protein interactions. Our model achieves an average accuracy of 0.82 and an average AU-ROC of 0.90. This result is quite promising as it indicates that the binding preferences of RBPs with no experimental data can be predicted accurately. Next, we ran our model with a dataset that is derived from the crystal structures of protein-RNA complexes in the NDB database. Again, our model achieved a dramatic performance with 0.983 average AU-ROC value. A recently proposed model named RPI-Pred that includes several additional features derived from the experimental structures of protein and RNA gives a lower performance (AU-ROC: 0.97). Our model is more advantageous as we can apply our model to any dataset even when the crystal structures of the protein-RNA complexes are unknown. Lastly, we predicted protein-mRNA interactions in yeast with an average accuracy of 0.72. Pancaldi et al achieves an average accuracy of 0.69 on the same dataset. Moreover, they use more than 100 complex features that are difficult to compile for datasets from other organisms than yeast. We should also note that our model differs from all the previous approaches in that the parameters of the model are set correctly using nested cross validation.

We performed a series of experiments to confirm that our proposed model gives optimal performance. We first showed that 3-mer frequencies with the reduced aminoacid alphabet gives better performance than lower order frequencies. Additionally, we observed that applying feature selection results in no gain of performance. However, more advanced techniques of feature selection must be applied before definite conclusions. Lastly, result of running naïve bayes, decision trees and SVM classifiers with the same datasets showed that our model performs much better than the other methods. The low performance of naïve bayes could be due to the assumption that features are independent. For all datasets, SVM-RBF has a higher AU-ROC than SVM-linear indicating that the data is not linearly separable. This high performance could be explained by the fact that the random forest model can represent higher order interactions among the features and it is less prone to overfitting. Altogether, these results indicate that using only sequence-based features can achieve good accuracy in predicting RNA-protein interactions.

We observed that including RNA secondary structure features did not result in improved performance. One reason could be the inaccuracy in predicting RNA secondary structure. Also, many of the RNA-binding proteins that appear in the three datasets might no have a strong preference for RNA secondary structure. Ray et al showed that motifs inferred from *in vitro* data could still be useful to predict *in vivo* binding [10]. However, in this study, we do not utilize the existing specific motif models for RBPs. Rather, we would like to infer general rules that can predict RNA-protein interactions only using the sequence of the protein and the sequence. As such, our model can still be used to pinpoint candidate RNA sequences that could be bound by the RBP of interest *in vivo* in the absence of any existing motif model. As a next step, experimental data on RNA secondary structure (e.g. [22]) can be utilized to determine the exact location of the binding sites within the RNA sequence.

References

[1] G. Varani and K. Nagai, "RNA recognition by RNP proteins during RNA processing", *Annu. Rev. Biophys. Biomol. Struct.*, vol. 27, pp. 407-445, 1998.

[2] T. Glisovic, J.L. Bachorik, J. Yong and G. Dreyfuss, "RNA-binding proteins and post-transcriptional gene regulation.", *FEBS letters*, vol. 582 , no14, pp. 1977–1986, June 2008.

[3] D. Moras, "Aminoacyl-tRNA synthetases.", *Curr. Opin. Struct. Biol.*, vol 2, pp. 138-142., 1992.

[4] P.B. Moore "The three-dimensional structure of the ribosome and its components.", *Annu. Rev. Biophys. Biomol. Struct.*, vol. 27, pp. 35–58, 1998.

[5] B. Tian, P. C. Bevilacqua, A. Diegelman-Parente and M. B. Mathews, "The double-stranded-RNA-binding motif: interference and much more.", *Nat. Rev. Mol. Cell Biol.* vol. 5, pp.1013-1023, 2004.

[6] K.E. Lukong, K.W. Chang, E.W. Khandjian and S. Richard. "RNA-binding proteins in human genetic disease." *Trends Genet.* vol. 24, no. 8 pp. 416-425, 2008.

[7] J.D. Keene, J.M. Komisarow and M.B. Friedersdorf. "RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts.", *Nat Proc,* vol. 1 no.1, pp:302-307, 2006.

[8] J. Ule, K. Jensen, A. Mele, and, R. B. Darnell. "CLIP: A method for identifying protein-RNA interaction sites in living cells". *Methods,* vol. 37, pp. 376–386, 2005.

[9] M. Hafner *et al.,* "Transcriptome-wide identification of RNA-Binding protein and MicroRNA target sites by PAR-CLIP," *Cell,* vol. 141, no. 1, pp. 129–141, Apr. 2010.

[10] D. Ray *et al.,* "A compendium of RNA-binding motifs for decoding gene regulation." *Nature*, vol. 499, pp.172– 177, 2013.

[11] V. Pancaldi and J Bahler (2011) In silico characterization and prediction of global protein-mRNA interactions in yeast. Nucleic Acids Research. vol. 39, no. 14, pp. 5286-5836. 2011.

[12] Y. Wang, X. Chen, Z. Liu, Q. Huang, D. Xu, and X. Zhang, "De novo prediction of RNA-protein interactions from sequence information," *Molecular BioSystems.*, vol. 9, no. 1, pp. 133–42, Nov. 2012.

[13] V. Suresh, L. Liu, D. Adjeroh, and X. Zhou, "RPI-Pred: Predicting ncRNA-protein interaction using sequence and structural information," *Nucleic Acids Research,* vol. 43, no. 3, Feb. 2015.

[14] Q. Lu *et al.*, "Computational prediction of associations between long non-coding RNAs and proteins," *BMC Genomics*, vol. 14, no. 1, p. 651, 2013.

[15] B. Lewis *et al.*, "PRIDB: A Protein-RNA interface database," *Nucleic Acids Research.*, vol. 39, Nov. 2010.

[16] D. J. Hogan, D. P. Riordan, A. P. Gerber, D. Herschlag, and P. O. Brown, "Diverse RNA-Binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system," *PLoS Biology*, vol. 6, no. 10, p. e255, Oct. 2008.

[17] J. Shen *et al.*, "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America.*, vol. 104, no. 11, pp. 4337–41, Mar. 2007.

[18] S. Bernhart, I. Hofacker, and P. Stadler, "Local RNA base pairing probabilities in large sequences," *Bioinformatics (Oxford, England).*, vol. 22, no. 5, pp. 614–5, Dec. 2005.

[19] S. J. Lange *et al.*, "Global or local? Predicting secondary structure and accessibility in mRNAs," *Nucleic Acids Research*, vol. 40, no. 12, p. 5215-5226, Feb. 2012.

[20] Breiman, L. "Random forests." *Machine Learning*. vol. 45, pp. 5-32, 2001.

[21] S. Jones, D. Daley, N. Luscombe, H. Berman, and J. Thornton, "Protein-RNA interactions: A structural analysis," *Nucleic Acids Research.*, vol. 29, no. 4, pp. 943–54, Feb. 2001.

[22] RC Spitale *et al.* (2015) "Structural imprints in vivo decode RNA regulatory mechanisms." *Nature*, 519, 486–490.