Advances in Computer Science and Information Technology (ACSIT)
p-ISSN: 2393-9907; e-ISSN: 2393-9915; Volume 4, Issue 2; January-March, 2017, pp. 106-110

http://www.krishisanskriti.org/Publication.html

Predicting RNA-Protein Interaction using Machine Learning Approach

Chandan Pandey, Rokkam Sandeep, Aikansh Priyam and Sitanshu Sekhar Sahu

Birla Institute of Technology, Mesra
E-mail: thechandan94@gmail.com, sandeep2331996@gmail.com, aikanshshs@gmail.com, sssahu@bitmesra.ac.in

Abstract—*RNA-Protein interactions play an important role in the various cellular processes. In this paper machine learning approach is employed to predict the interaction between Protein and RNA using various sequence based information of RNA and proteins. It has been shown that the conjoint ternion method provides better accuracy of 89.67% and an MCC value of .79 in training set. Further, in an independent test, it provides an accuracy of 83.23%.*

1. INTRODUCTION

Ribonucleic Acid (RNA) a polymeric molecule which plays an essential role for various biological processes of our body. It is a long chain of four nucleotides called guanine, uracil, adenine and cytosine assembled together in a specific manner. Proteins are molecules which play a prominent role in various metabolic activities and molecular transportation. It is a long chain of 20 essential amino acids. RNA-Protein interactions play an important role in various cellular processes such as protein synthesis, sequence encoding RNA transfer, and gene regulation at the transcriptional and post-transcriptional levels, RNA splicing and various other processes .These RNA binding proteins play an important role in gene expression and regulation [1]. So there is a need for studying these RNA protein interactions in an elaborate way.

RNA-Protein interactions prediction gives way for the discovery of medicines for various diseases and methods for modification of gene expression and better understanding of RNA-Protein recognition. There are various methods established to study the RNA-Protein interactions experimentally but these methods are time consuming and costly and allow us to study only one or less number of interactions at a time [2]. Due to these disadvantages of the experimental methods development of computational methods took place which enables us to find the RNA-Protein interactions with less effort in terms of money and times and also enables the experimental methods to focus on these specific interactions. These computational methods are based on the sequential and structural information of the RNA and protein sequences. For now there are various methods developed to find the RNA protein binding interfaces but not a lot for predicting the specific RNA for a known RNA binding protein or a protein for non coded RNA. There are other type of computational methods which take into account the primary, secondary or tertiary-structure information which include StructNB which uses Naive Bayes classifier[3], residue singular interface propensity, residue interface doublet propensity [4] RNA-Protein interface predictor (PRIP) and OPRA (Optimal Protein-RNA Area). Due to less availability of structurally characterized RNA-Protein complexes computational methods based on sequential information are needed to be developed to predict RNA-Protein interactions [5] which take into account various parameters like hydrophobicity, electrostatics, side chain environment, residue interface propensity, residue accessibility etc [6]. In this study we have explained six different types of features AAC, DIPEP, Conjoint Triad, PseAAC and new feature extraction techniques are being developed namely Conjoint Ternion and Conjoint Dyad.

2. MATERIALS AND METHODS

2.1 Dataset

In this study, the dataset used for training the model was obtained from RNA-Protein Interaction Prediction (RPISeq) [5]. This benchmark dataset contains 2241 interacting protein and RNA pairs, obtained from 943 RNA-protein complexes from PRIDB (RPI2241). The dataset contains 952 protein sequences and 443 RNA sequences. This dataset gives the positive samples in the model but no such dataset exist for negative samples; but to train the model efficiently without any biasing, non-interacting RNA-protein pairs were generated by randomly pairing the 943 RNA-protein complexes. In the random pairing of protein and RNA chains, those chains were removed which were already available in positive sample dataset. Thus a balanced dataset comprising of both positive and negative samples were obtained. The dataset used for testing the model was obtained from RNA-Protein interaction database (PRD-http://pri.hgc.jp/). From the database of 340 interactions, 131 interacting pairs were obtained in which protein and RNA chains have same origin of species. The test dataset includes pairs from 4 different species: Homo sapiens (50 pairs), Drosophila melanogaster

(22 pairs), Mus musculus (37 pairs) and Saccharomyces cerevisiae (22 pairs.

3. FEATURE EXTRACTION TECHNIQUES

In this study, sequence-based features of proteins and RNA chains were obtained, six different types of feature extraction techniques were implemented and models corresponding to each technique were developed.

3.1 Amino Acid Composition (AAC)

Each protein and RNA is made of 20 amino acids and 4 nucleotides respectively. In this technique, protein chain is expressed in 20 features and RNA chain in 4 features. Thus, a RNA-protein pair is represented by a 24-dimension vector [7] [8]. The feature represents the normalized frequency of each amino acid or nucleotide in protein or RNA chain respectively. If protein chain is p and RNA chain is r, also the frequencies of occurrence of its constituent amino acids and nucleotides are f_1 (p_i) and f_2 (r_j) respectively, the composition of amino acids and nucleotides is given by,

$$X(p_i) = \frac{f_1(p_i)}{\sum_{i=1}^{20} f_1(p_i)} \quad i = 1,2,3 \dots \dots 20 \qquad (1)$$

$$Y(r_j) = \frac{f_2(r_j)}{\sum_{j=1}^{4} f_2(r_j)} \quad j = 1,2,3,4 \qquad (2)$$

Combining the features obtained from above 2 equations a 24 dimension vector is obtained.

code.

3.2. Dipeptide Composition (DIPEP)

In this technique, the protein chain is represented by normalized frequency of each dipeptide possible from 20 amino acids. Thus, a 400 (20x20) feature vector is obtained for each protein chain. Similarly, in RNA chain 4 nucleotides are present, therefore 16 (4x4) pairs are possible. So, for RNA sequence a 16 dimension vector is obtained. The dipeptide composition is given by:

$$X(p_i, p_j) = \frac{f_1(p_i, p_j)}{\sum_{i=1}^{20}\sum_{j=1}^{20} f_1(p_i, p_j)} \quad i, j = 1,2 \dots 20 \qquad (3)$$

$$Y(r_i, r_j) = \frac{f_2(r_i, r_j)}{\sum_{i=1}^{4}\sum_{j=1}^{4} f_2(r_i, r_j)} \quad i, j = 1,2,3,4 \qquad (4)$$

Combining the feature obtained from above 2 equations a 416-dimension vector is obtained.

3.3. Pseudo Amino Acid Composition (PseAAC)

The composition based methods suffer from loss of sequence information. In order to obtain information from sequence, Chou [9] proposed Pseudo Amino Acid Composition (PseAAC). In this technique, the correlation factor between amino acids is found using its physiochemical properties. In this study Electron Ion Interaction Potential (EIIP) of each amino acid [10] and nucleotide [11] is used to calculate the PseAAC values. The PseAAC feature extraction process has been followed from [12].

3.4. Conjoint Triad

In this technique, the interacting chains of protein and RNA are represented by 599-dimensional vector features [13]. The protein chain is represented by 343 feature vectors and the RNA chain is represented by 256 feature vectors. In this representation, the 20 amino acids are divided into 7 groups according to their dipole moments and volume of their side chains: A, G, V, I, L, F, P, Y, M, T, S, H, N, Q, W, R, K, D, E, C. Amino acids belonging to the same group are represented by same group number. Hence, the protein chain is now displayed in numbers from one to seven. The protein is represented by conjoint triad feature (CTF), where each feature represents normalized frequency of 3-mer in the 7-number representation of the protein sequence. Thus, 3-mer of 7 numbers makes 343 (7x7x7) features. Similarly, in RNA chain instead of 3-mer, normalized frequency of 4-mer is obtained. Thus, RNA sequence is represented by 256 (4x4x4x4) features (see [13] for details). Both these features are merged together to form a 599 dimension vectors.

3.5 Conjoint Ternion

This technique is pretty much similar to Conjoint Triad, with a slight difference. In this representation of protein and RNA chains, the features constitute normalized frequency of 3-mer protein sequence as well as RNA sequence; different than conjoint triad where 4-mer of RNA chain was chosen.

Thus, the protein sequence is represented by 343 dimensional vectors and RNA sequence is represented by 64 (4x4x4) feature vectors. The features of interacting pairs of protein and RNA chain are merged together to form 409-dimensional vector features.

3.6. Conjoint Dyad

In this technique, the protein and RNA chains are represented by the features constituted from normalizing the frequency of 2-mer protein as well as RNA sequence. Thus, the protein sequence is represented by 49 feature vectors (7x7) and the RNA sequence is represented by 16 feature vectors (4x4). Therefore, total number of features representing the pair of RNA-protein pair is 65-dimensional vectors.

4. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is a class of learning machines, from statistical learning theory, based on optimization principle. SVM is an effective technique for classification and has been widely used in many applications with great performance It separates the input data with a maximum possible margin, while maintaining a reasonable computing efficiency. In this algorithm, each data item is

plotted as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. An optimal separating hyperplane is constructed in the feature space, which maximizes the margin between the two classes and thereby separates the data into different classes. The hyperplane is defined as:

$$g(x) = w^{\mathrm{T}}x + \mathrm{c} = 0 \qquad (5)$$

In equation (5), ω is an adjustable weight vector and c is a bias. The hyperplane can be found by minimizing the following cost function:

$$K(w) = \frac{1}{2}w^{\mathrm{T}}w = \frac{1}{2}||w||^2 \qquad (6)$$

This is subjected to the constraints:

$$d_i[w^{\mathrm{T}}x_i + \mathrm{c}] \geq 1, \ \mathrm{i} = 1,2,..\mathrm{N} \qquad (7)$$

SVMs employ kernel functions to map input feature vectors from a lower dimension into a higher dimensional space and construct an optimal separating hyperplane in this higher dimensional space. Some common kernel functions are: the linear kernel, the polynomial kernel, the radial basis function (RBF).

5. RESULTS & DISCUSSION

5.1. Performance Evaluation

Given a protein-RNA sequence as input we have tried to predict whether a given pair of protein-RNA pair interacts or not. The training of the machine have been done by(SV M)light[14] algorithm. $(SVM)^{light}$ is an implementation of Vapnik's Support Vector Machine [15] for the problem of pattern recognition, for the problem of regression, and for the problem of learning a ranking function. The advantage of using (SV M)light is that it has fast optimization algorithm and can handle many thousands of support vectors. RNA-Protein interaction predictions have been done only using sequence information. The performance of the machine is evaluated by using the true positive, true negative, false positive and false negative rate and the following parameters are calculated:

$$Accuracy\ (A) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Precision\ (P) = \frac{TP}{TP + FP}$$

$$Sensitivity\ (Se) = \frac{TP}{TP + FN} \quad Specificity\ (Sp) = \frac{TN}{TN + FP}$$

$$MCC = \frac{(TP * TN - FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

True Positive (TP) the number of correct predictions that the instance is positive. False Positive (FP) is the number of incorrect predictions that an instance is positive. True Negative (TN) is the number of correct predictions that an instance is negative. False Negative (FN) is the number of incorrect predictions that an instance if negative.

MCC [16] is used as a measure of the quality of binary classifications. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes.

Table 1 shows the performance of the various feature extraction techniques. The features have been extracted by using the protein and RNA sequence in the RPI2241 dataset. RPI2241 dataset consists of 2241 pairs of protein-RNA pairs. It is a benchmark dataset which is has been obtained from PRIDB which a comprehensive database of RNA-protein complexes extracted from the PDB [17]. The features extracted from RPI2241 are fed for training in the SVM^{light} model. The machine was

Table 1. Performance Parameter of SVM Model on the training dataset RPI2241

Method	Std. Deviation of RBF (σ)	Box constraint	A (%)	P (%)	Sp (%)	Se (%)	MCC (%)
Conjoint Ternion	12	50	89.67	86.58	85.45	93.89	79.62
Conjoint Triad	10	50	88.26	84.56	82.91	93.62	76.97
Conjoint Dyad	10	100	86.39	82.32	80.10	92.68	73.36
Dipep	5	100	83.51	81.11	87.37	79.65	67.22
AAC	10	100	78.67	74.42	69.97	87.37	58.23
PseAAC	10	100	59.24	55.76	29.09	89.38	23.15

trained by changing the parameters, rbf kernel and box constraint and the models were obtained for each of the techniques. The training accuracy was between 89.67% in Conjoint Triple to 59.24% in Pseudo Amino Acid Composition (PseAAC). testing, 4 different datasets which belongs to different species have been used. These independent test datasets have been used in the paper published by Muppirala [5]. All these four datasets are tested using the model of the feature extraction techniques which were generated during the training. The accuracy for all the datasets have been given in Table 2. The overall accuracy was between 83.23% in Conjoint Triple method to 58.01% in Pseudo Amino Acid Composition (PseAAC).

For testing, 4 different datasets which belongs to different species have been used. These independent test datasets have been used in the paper published by Muppirala [5]. All these four datasets are tested using the model of the feature extraction techniques which were generated during the

training. The accuracy for all the datasets have been given in Table 2. The overall accuracy was between 83.23% in Conjoint Triple method to 58.01% in Pseudo Amino Acid Composition (PseAAC).

The data currently available for RPI prediction is limited. Hence it is one of the major problem in this field. In this work the classifiers were trained using only RPIs for which experimental structures are available. RPI2241 is a non-redundant training dataset and it consists of 2241 interacting RNA-protein pairs, and it also includes a wide variety of different functional classes of proteins and RNA like ribosomal RNA (rRNA), messenger RNA (mRNA), micro RNA (miRNA), transfer RNA (tRNA) [5] [18]. rRNA ribosomal protein pairs consists approximately 40% of the total pairs which shows that the predominance of the dataset.

Table 2. Performance of the SVM Model on the Independent Dataset

Datasets	No. of Samples	Conjoint Ternion (%)	Conjoint Triad (%)	Conjoint Dyad (%)	Dipep (%)	AAC (%)	PseAAC (%)
H. sapiens	50	88	86	82	84	86	68
D. melanogaster	22	90.91	95.45	81.82	86.36	86.36	45.45
M. musculous	37	78.38	70.27	91.89	94.59	89.19	64.86
S. cerevisiae	22	72.73	81.82	77.27	95.48	50	36.36
Overall Accuracy	131	83.23	82.44	83.96	89.31	80.91	58.01

In this study, the feature extraction techniques which were used have accurately predicted RPI predictions using the RPI2241 dataset which is our benchmark training dataset and the independent datasets which have been used for testing. The data used in this study represents a small fraction of RNA-protein complexes so with more and more reliable data it is expected that the accuracy will eventually increase and the results will become more reliable.

The developed prediction model is assessed with an independent dataset. The performance results are reported in Table 2. The Conjoint Ternion method provides better results. For Homo sapiens it shows 88% accuracy. Similarly in Drosophila, the accuracy of prediction was 90.91% in Conjoint Ternion method and 90.95% in Conjoint Triad method. For the mouse (Mus musculus) and the yeast (Saccharomyces cerevisiae) the accuracy was around 78% and 73% in Conjoint Ternion method.

6. CONCLUSION

In this paper, a machine learning based method is presented to predict whether a pair of RNA-protein interacts or not. Several sequence based features such as Conjoint Ternion, Conjoint Triad, Conjoint Dyad, DIPEP, PseAAC and AAC are derived from the protein and RNA sequences. The conjoint feature shows better results. The Conjoint Trenion feature is found superior. It provides 89.76 % accuracy in a standard database. Also it provides better accuracy of 83% in test dataset which consist of samples from various organisms.

REFERENCES

[1] Si J, Cui J, Cheng J, Wu R," Computational Prediction of RNA-Binding Proteins and Binding Sites" Karabencheva-Christova T, ed. International Journal of Molecular Sciences. 2015;16(11):26303-26317. doi:10.3390/ijms161125952.

[2] Marchese D, de Groot NS, Lorenzo Gotor N, Livi CM, Tartaglia GG," Advances in the characterization of RNAbinding proteins," Wiley Interdisciplinary Reviews RNA. 2016;7(6):793-810. doi:10.1002/wrna.1378.

[3] Towfic F, Caragea C, Gemperline DC, Dobbs D, Honavar V," Struct-NB: Predicting Protein-RNA Binding Sites Using Structural Features," International journal of data mining and bioinformatics. 2010;4(1):21-43.

[4] Kim OTP, Yura K, "Go N. Amino acid residue doublet propensity in the proteinRNA interface and its application to RNA interface prediction," Nucleic Acids Research. 2006;34(22):6450-6460. doi:10.1093/nar/gkl819.

[5] Usha K. Muppirala, Vasant G. Honavar, Drena Dobbs." Predicting RNA-protein Interactions using Only Sequence Information", BMC Bioinformatics 2011, 12:489

[6] Ren H, Shen Y," RNA-binding residues prediction using structural features," BMC Bioinformatics. 2015;16:249. doi:10.1186/s12859-015-0691-0.

[7] Kaundal R, Saini R, Zhao PX," Combining Machine Learning and Homology-Based Approaches to Accurately Predict Subcellular Localization in Arabidopsis," Plant Physiology 2010, 154:36-54.

[8] Sahu SS, Panda G,"A novel feature representation method based on Chous pseudo amino acid composition for protein structural class prediction," Computational Biology and Chemistry 2010, 34:320-327.

[9] Kuo-Chen Chou," Prediction of protein cellular attributes using pseudo-amino-acidcomposition," PROTEINS: Structure, Function, and Genetics, 2001, 43: 246-255

[10] Kawashima, S. and Kanehisa, M,"(2000) Aaindex: amino acid index database," Nucleic Acid Res., 28, 374

[11] Cosic I, IEEE Trans Biomed Eng. 1994;41:12.

[12] Rakesh Kaundal, Sitanshu S. Sahu, RuchiVerma and Tyler Weirick," Identification and characterization of plastid-type proteins from sequenceattributed features using machine learning, "BMC Bioinformatics 14(S14): S7, 2013

[13] 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, 104:4337-41.

[14] T. Joachims," Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning," B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999

[15] Vladimir N. Vapnik, "The Nature of Statistical Learning Theory,"Springer, 1995

[16] Matthews, "Comparison of the predicted and observed secondary structure of T4 phage lysozyme,"Biochimica et Biophysica Acta (BBA) - Protein Structure. 405 (2): 442451
[17] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, "Bourne PE: The Protein Data Bank," Nucleic Acids Res 2000, 28:235-42
[18] Kishore S, Luber S, Zavolan M," Deciphering the role of RNA-binding proteins in the posttranscriptional control of gene expression,"Brief Funct Genomics 2010, 9:391-404.