Available online freely at www.isisn.org

Bioscience Research

Print ISSN: 1811-9506 Online ISSN: 2218-3973

Journal by Innovative Scientific Information & Services Network

RESEARCH ARTICLE BIOSCIENCE RESEARCH, 2019 16(3):2699-2709. **OPEN ACCESS**

A novel computational approach for predicting RNA binding proteins using optimum path forest with hybrid features

Zeinab Abd El Haliem[1], Mohammad Nassef[2], Amr Badr[2] and Khaled T. Wassif[2]

[1] Faculty of Computer Science, Modern Science and Arts University, Giza, **Egypt**
[2] Faculty of Computer and Information, Cairo University, Giza, **Egypt**

***Correspondence:** ztaha@msa.eun.eg **Accepted: 27** July 2019 **Published online: 04** Aug 2019

RNA – Protein interactions have vital roles in several cellular processes such as RNA transfer, gene regulation at the transcriptional processes and sequence encoding. RNA-binding prediction is a very important aspect of the analysis that helps in identifying the motifs that bind to DNA and for gene regulations. Predicting and recognizing the proteins that bind to RNA is a major challenging and complex process due to structural biology. Previously, several computational methods have been used and developed for predicting RNA-binding proteins (RBPs) using Support Vector Machine other than many other machine learning techniques. This paper proposes a novel computational approach for predicting RBPs using Optimum Path Forest (OPF) classifier in conjunction with the information of predicted RNA-binding residues. Moreover, the statistical information, mainly the singlet and doublet propensity, have been taken into consideration. For a given protein, its RNA-binding residues are predicted and then checked whether the protein binds to RNA or not through positive and negative samples based on the information from that prediction methodology. The results for the previous step can be classified as "Binding Protein", "Nonbinding Protein", "Binding Protein predicted as Non-Binding Protein" and "Non-binding Protein predicted as Binding Protein", and in this case if the protein cannot be identified then the OPF classifier is used to determine the protein prediction status. The OPF classifier is used incorporated with the amino acid composition feature. The results showed that the statistical information and the binding propensity measures of the predicted RNA-binding residues especially contributed to the prediction process. In addition, the classifier has improved the overall performance of RBPs prediction process.

Keywords: Gene Regulations, Optimum Path Forest Classifier, Prediction, RNA-binding proteins (RBPs), Non- binding proteins, Motifs, Transcriptional Processes, Biomolecules interactions.

INTRODUCTION

RNA interactions and proteins that associate with and bind to RNA have important roles in several cellular processes such as RNA transfer, gene regulation at the transcriptional processes and sequence encoding (Ibba and Soll, 1996, De Guzman et al., 1998).The previous studies showed that about 6% up to 8% of proteins are RNA binding proteins (RBPs) (Cusack, 1997), these RBPs also play an essential role in the gene regulation and gene expression, so the identification and prediction process of RBPs and its motifs is very important and vital in protein function annotation (Jacobs Anderson and Parker, 2000). Due to some limitations, such as experimental methods and X-ray crystallography are very expensive, time- consuming and labor intensive only a few studies concentrated on

proteins that bind to RNA, in spite of several studies related to proteins that bind to DNA.

Consequently, the identification and prediction of RBPs is essential and remains a challenge in the biological and genomics' era (Abdelmohsen, 2008). Numerous computational approaches have been developed and applied to the RBPs identification process (Ma et al., 2015). Due to the challenges in the experimental techniques, computational approaches and tools are required to be developed that could be more reliable, inexpensive and faster identification of RBPs and RNA binding sites. Many computational approaches have been developed in different strategies for RNA- binding site and RBPs, either through protein sequence, or protein structure or by combining different machine learning approaches using a hybrid sequence with structural features which is called protein docking (Si et al., 2015).

The previous studies mainly have focused on prediction based on sequence similarity from amino acid sequence information (Wang and Brown, 2006, Murakami et al., 2010, Wang et al., 2010 and Ma et al., 2011). The Support Vector Machine (SVM) was used by Cai and Lin to predict RBPs from amino acid sequence (Cai and Lin, 2003) and also used to distinguish RBPs from non-RBPs by many different ways (Han et al., 2004, Yu et al., 2006, Shao et al., 2009, and Kumar et al., 2011). Yu et al. developed a new approach by using SVM integrating with physicochemical properties using protein amino acid sequences to predict RPBs (Yu et al., 2006). Shao et al. also developed an approach using "SVM with a conjoint triad feature that extracts information directly from a protein amino acid sequence" to differentiate RBPs from non-RBPS (Shao et al., 2009). (Kumar et al., 2011) developed "RNApred" method that describes a SVM to distinguish RBPs from non-RBPs using a position specific scoring matrix (PSSM) and its input feature. Subsequently, Terribilini et al. developed a classical method to predict RBPs and sites by using Navie Bayes (NB) contributed with RNABindR (Terribilini et al., 2007).

The second strategy for predicting RBPs is through using structure based methods, and the protein structure is available. Therefore the prediction process will be simple and reliable. There are some structured based methods that used for predicting RBPs and sites. (Chen and lim, 2008) developed a prediction method depending on a protein structure information like evolution and geometry for predicting RNA binding sites. (Zhao et al., 2011) furthermore developed a predictor depending on a protein structure information by combining RNA binding affinity with a structural similarity for RNA binding domains. Yang et al. proposed a structural pattern in prediction package named SPOT-Seq-RNA for "Predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction" (Yang et al., 2014). Another strategy is docking which aimed at modelling interaction of macromolecular complexes (Moreira et al., 2010). Although, research on protein 3D structure modeling is very hard and complex, demonstrating protein structure of RNA complex is very essential and helpful to understand the tools of communication. Many docking methods are developed and used to predict RNA protein complexes and protein structure depending on known RNA (Katchalski – Katzir et al., 1992; Ritchie and Kemp, 2000; Schneidman-Duhovny et al., 2005; and Gabb et al., 1997).

From the above results of prediction methods for RNA binding sites, RNA binding residues and RBPs, the accuracy is approximately 60% up to 80% and the specificity and sensitivity range of these methods is extensively extended (Nagarajan, and Gromiha, 2014). Every method in different strategy has its own perspective because of algorithms and techniques that used, the various datasets, the input features, and the predictors. In addition to the web servers that the researcher developed to show the proposed method and its results.

Taking into consideration the limitations from the previous studies, our present study proposes a novel computational approach for predicting RBPs using optimum path forest classifier in conjunction with the binding propensity information extracted from predicted RNA binding residues. Moreover, we consider the statistical information mainly the singlet and doublet propensity into our methodology. For a given protein, we predict its RNA binding residues and then check whether the protein binds to RNA or not through positive and negative samples based on the information from that prediction. For judging the query protein we tried to construct two or three binding measures using statistical propensity (singlet and doublet), and although these measures cannot judge all proteins, therefore optimum path forest model was developed in conjunction with a hybrid feature containing binding measures combined with statistical propensity information and amino acid

composition.

The rest of this paper is structured as follows: Section 2 describes the theoretical background including dataset, statistical propensity measures, binding measures, feature vector models, and the optimum path forest classifier. Section 3 describes the proposed approach, showing its effectiveness in solving that problem. Section 4 shows the evaluation and results using evaluation criteria. Section 5 discusses the results of our model with comparable results Finally, Section 6 concludes the proposed model and highlights some points for future work.

MATERIALS AND METHODS

Theoretical background

This section firstly describes used dataset in details. Secondly, the statistical singlet and doublet propensity, then, binding measures, subsequently the feature vector models and recently the optimum path forest classifier.

Data set

The UniProt database (http://www.uniport.org/) is used in this study to extract RBPs and Non-RBPs (Ma. et al., 2005). Data selected manually which is annotated and reviewed protein sequences (Consortium et al., 2012). Selected data divided into a positive dataset and a negative dataset as follow:

Positive Dataset

The Positive dataset means extracting RNA Binding protein sequences from a UniProt database when searching the database we retrieved about 54,550 RNA binding protein sequences.

The retrieved data designated using “Rough Positive” dataset that is used in previous researches (Yu et al., 2006; Shao et al., 2009; Kumar et al., 2011 and Huang et al., 2010). The dataset was processed and filtered by removing protein sequences with more than 6000 amino acids that may be protein complexes, sequences of protein with amino acids less than 50 because it might be fragmented and also sequences that contain irregular symbol like “X” and “Z” were removed.

As a result of the previous issues, only 3,712 positive protein sequences were obtained and used in our work.

Negative Dataset

The Negative dataset also obtained and retrieved from the UniProt database through a list of keywords such as Non-binding, and DNA/RNA binding using logic operator or. The Negative dataset contains about 140,387 protein sequences. As similar as in positive dataset, a rough negative dataset was used to filter our retrieved data to be 94,770 protein sequences in our negative dataset (Cai and Lin, 2003).

From the extracted positive and negative dataset we noticed that there is imbalance problem between numbers of protein sequences, and to deal with this problem, 3,712 non-RBPs were selected randomly from the negative dataset to make our dataset stable with the same size from the positive and negative samples.

The Testing dataset contains a composition of equal sizes of proteins from the positive and negative datasets, and this combination named “RNAt_7424” as used in previous work (Huang et al., 2010). As well as the testing dataset is used to evaluate the performance of our methodology.

For predicting RNA binding residues we used Protein Data Bank (PDB) to extract RNA protein complexes as used in previous studies (Ma. et al 2015 and Huang et al., 2010).

Statistical Propensity Measures

The residue interface propensity is needed to measure the importance of different amino acid types that is exists in RNA binding sequence interface.

There is singlet and doublet interface propensity, the singlet interface propensity (P_i) is calculated for each amino acid type (i=1, 2 …20) by the following equation:

$$P_i = \frac{\bar{f}_i}{f_i} \text{ Where } f_i = \frac{n_i}{\sum_{i=1}^{20} n_i} \text{ and } \bar{f}_i = \frac{\overline{n_i}}{\sum_{i=1}^{20} n_i} \quad \textbf{(1)}$$

Where n_i is the number of amino acid type i on the protein surface, and $\overline{n_i}$ is also the number of amino acid type i but in the RNA interface. The P_i is more than one as amino acid of type i may occurs more frequently in the RNA interface than on the protein surface (Liu and Gong, 2012).

The doublet propensity interface gives a measure of pairing preferences of amino acid types in RNA protein interfaces. In amino acid sequences, the doublet propensity considered from amino acid type i and the neighboring amino acid j if the distance between their atoms is less than or equal to a certain threshold.

In this work, the distance (threshold) is set to be 7.0 A^o, and this threshold value is chosen for

the neighboring residues. The doublet propensity is calculated by the following equation:

$P_{ij}=\frac{\bar{f}_{ij} \ * f_i \ * f_j}{f_{ij} \ * \bar{f}_i \ * \bar{f}_j}$ where $\overline{f_{ij}}=\frac{\bar{n}_{ij}}{\sum_{i=1}^{20}\sum_{j=1}^{20}\bar{n}_{ij}}$

And $f_{ij}=\frac{n_{ij}}{\sum_{i=1}^{20}\sum_{j=1}^{20}n_{ij}}$ **(2)**

Where $n_{ij}$is number of doublet of amino acid type ij on the protein surface, and $\bar{n}_{ij}$ is also number of doublet of amino acid type ij but on the RNA interface.

Binding Propensity Measures

A method for predicting RBPs can be built after predicting RNA residues from the RNA interface and also after predicting amino acid preferences. So to find the binding propensity measures we have to consider that RNA binding residues should located in the RBPs, and should also appear on the surface of RBPs. And non-RBPs should contain less binding residues comparable with RBPs. To complete the prediction of RBPs we applied two binding measures depending on RNA binding residues, as used in (Huang et al., 2010).

BPM (1) = $\frac{\sum_{i=1}^{n}PR(i)}{10N}$ **(3)**

Where BPM is Binding Propensity Measure, n is the number of RNA binding residues, PR is the predictive reliability of RNA binding residues i that predicted using optimum path forest classifier, and N is the number of amino acids.

BPM1 describes the information in the amino acids depending on the appearance of RNA binding residues, and the reliability of RNA binding residues.

BPM (2) = $\frac{\sum_{i=1}^{N-1}2^{-i+1} \quad \sum_{k=1}^{n(i)}\overline{PR(k)}}{10\ (N-1)}$ **(4)**

Where N is the number of amino acids, i is the distance between RNA residue and its neighbor, n (i) is the number of two RNA binding residues of i amino acid within the distance, $\overline{PR(k)}$ represents the mean predictive reliability of RNA binding residue k and binding residue k+i, so the total part of $\sum_{k=1}^{n(i)}\overline{PR(k)}$ measures the mean predictive reliability of each pair in the RNA binding residues.

BPM 2 describes the relation between RNA binding residues with different distances from 1 to N-1 amino acids, so we can say that this measure represents the association between RNA binding residues in the amino acid sequences (Ma et al., 2015).

Moreover, this binding feature is similar to the doublet statistical propensity in measuring the distances between RBPs Paris.

Feature Vector

In our work there is four classes of feature vector representations.

Model (1) Simulated using amino acid sequence vector

M_1= (X_{-n}, X_{-n+1}… X_{t-1}, X_t, X_{t+1}… X_{n-1}, X_n)

Where X is the amino acid representation and X_n, X_{-n} is a segment of amino acid sequence. X_t represents the interface that may contain 0 or 1 which means interface or non-interface (Liu and Gong, 2012).

Model (2) simulated using amino acid composition (AAC) or amino acid alignment

The Amino Acid alignment means a vector that consists of the rate of recurrence for the 20 amino acids types. In a protein query, the AAC is calculated as follow:

$p_i=\frac{l_i}{L}$ Where (i=1, 2, 3… 20) **(5)**

P_i represents the rate of recurrence of the i_{th} type of amino acid, l_i is the number of i_{th} type of amino acid that is exists in the protein sequence, and L is the total number of amino acids in the protein sequence.

Model (3) Simulated using the singlet Propensity of amino acids

M_3= (PS_{-n}, PS_{-n+1}… PS_{t-1}, PS_t, PS_{t+1}… PS_{n-1}, PS_n)

Where PS_i represents the singlet propensity of amino acid i.

Model (4) Simulated using the doublet Propensity of neighboring amino acids

M_4= ($PD_{-n,-n+1}$ $PS_{-n+1,-n+2}$… $PS_{t-1,\ t}$, $PS_{t,\ t+1}$, $PS_{n-2,\ n-1}$, $PS_{n-1,\ n}$)

Where PD_{ij} id the doublet propensity of amino acid i and its neighbor j in the protein sequence.

Optimum Path Forest Classifier(OPF)

Papa et al., presented OPF as a simple, fast, efficient, and parameter independent classifier (Papa et al., 2009, 2012). OPF is a supervised classification method that has shown good results in many classification problems (Papa et al., 2008, 2010., 2012), and the training dataset can be represented as a complete graph. It represents the samples as graph nodes whose arcs are weighted by using any distance function. In the graph, each node is represented as a feature vector, and each edge connects a pair of nodes, constituting a fully connected graph (Sayed et al., 2016).

OPF classification process consists of two steps, fit and predict. In the fit step, OPF classifier chooses the training samples to be prototypes, but in the predict step the OPF classifier assigning

the label of prototype that offers the lower path cost to the test samples using a cost function.

In our study, the dataset Z is divided into two parts Z_1 and Z_2 where Z_1 is the training set and Z_2 is the testing set, and Z is a fully labeled dataset. Let (Z_1, A) is a complete graph whose nodes are the samples in this set and any pair of samples represents an arc in A=Z_1 X Z_1. Let π_S be a path in the graph in sample S $\in$ Z_1 (training set), and (π_S. (S, t)) Is the concatenation between π_S and the arc (s, t) where t$\in$ Z_1, and S $\subset$ Z_1 as a set of key prototypes of all classes (samples). The past cost can be computed by using the following equation:

$$f_{max} = \begin{cases} 0 & if s \in S, \\ +\infty & \text{otherwise}, \end{cases} \quad \textbf{(6)}$$

$$f(\pi_s . (s,t)) = \max\{f(\pi_s), d(s,t)\}, \quad \textbf{(7)}$$

Where *d(s, t)* is the distance between node *s* and node *t*

OPF is a practical classifier as it is sensible to any outliers, since the prototypes choosing based on the Minimum Spanning Tree (MST) may choose noisy samples to become prototypes and these samples have great impact on OPF's classification decision. A group of prototypes which can be represented as S^* (represents an optimal set of prototypes) that can be found using the representation of MST in the complete graph (Z_1, A). A MST can be described as optimum when the sum of its arc weights is the lowest amount compared to any other spanning tree in the complete graph. A MST contains just one optimum path tree for any selected root node, and to get it, the closest elements of this tree have to be selected with different labels in Z_1 (Papa et al., 2009). Every pair of samples in the MST is connected by a single path that can be checked and evaluated as minimum or not by equation (6).

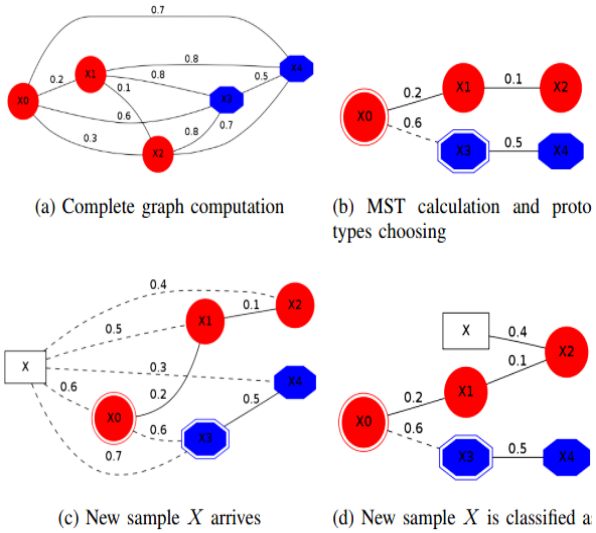
Consequently, in the graph, nodes represent all the samples of Z_1, and the arcs are weighted by the distance *d* between any neighboring or contiguous samples.

The training phase of this classifier starts with nodes (prototypes) to minimize the cost between each pair or sample in the training set samples. After that, it gets an optimum path forest which can be described as a collection of optimum path trees rooted at each node or prototype. Conversely, in the testing/classification phase all the arcs are taken into consideration especially those connecting a *t* sample in the testing data Z_2 with samples s $\in$ Z_1 (training set), so the sample *t* was a part of the training graph. The optimum path $P^*(t)$ can be found by evaluating all possible paths from S^* to the sample *t*, and label *t* with the most strongly connected prototype in all paths S^* by $\lambda(R(t)) \in S^*$, where λ (t) is the function that assigns the correct class label, and R (t) is the function that gets the root of *t* and this root is one of the prototypes R (t) $\in$S (Papa et al., 2012). We can identify this path by calculating the optimum cost equation (8) as follows:

$$C(t) = \min\{\max\{C(s), d(s,t)\}\}, \forall s \in Z_1 \quad \textbf{(8)}$$

According to Eq. (6), it can be assumed that the node *P (t)* is the predecessor in the optimum path $P^*(t)$, and $S^* \in Z_1$ is the node that satisfies the equation too. Given that $L(S^*) = \lambda(R(t))$ as the class *t*.

An example of optimum path forest classifier example is shown in figure1.(Souza et al., 2012) that represents a complete undirected graph where the training set Z1 is divided into 5 folds named X0, X1, X2, X3, and X4 where X0, X1, and X2 from class "RED" and X3, and X4 from class "Blue". Figure 1(a) represents a complete graph computation where X0 and X3 are chosen as prototypes. Figure 1(b) represents the Minimum Spanning Tree calculation with chosen prototypes. Figure 1(c) shows a new sample X that arrives and compute the distance between it and to every sample in the training set. Figure 1(d) shows that X is assigned to tree rooted in X0, so X is classified as a new member of class "RED". We can say that the first 2 steps showed in (a) and (b) represents the fitting phase and the other steps (c) and (d) represents the prediction phase.



(a) Complete graph computation (b) MST calculation and proto types choosing

(c) New sample X arrives (d) New sample X is classified a

Figure 1: Example of OPF phases workflow

Proposed solution

In this paper, we developed a novel computational approach for predicting RBPs using statistical information and binding information of predicted RNA binding residues in conjunction with optimum path forest supervised classifier. We established the following plan for predicting RBPs and RNA non-binding proteins as follow: (1) identifying a training and testing dataset to be used with our predictor; (2) We formulate the statistical and binding propensity based on the prediction of RBPs and residues in the query protein; (3) We select the important features to institute the predictor; (4) Select OPF classifier to do the prediction; (5) from the above 4 steps we developed an effective method using our binding propensity models and OPF predictor for predicting RBPs. Figure (2) shows the flowchart of the proposed methodology. The workflow of our methodology starts with predicting RNA binding residues and calculating the statistical and binding propensity measures from the prediction results. After that we use the binding propensity measure threshold to check whether the query protein binds to RNA or not. In case one of the two binding propensity is greater than the upper threshold assigned for each one then the protein assigned to be RBP. As well as in case one of them is less than the required lower threshold of it then the protein assigned to be Non-RBP. Otherwise, if the protein sequence cannot be classified to be RBP or non-RBP, OPF classifier is used to predict and check whether the query protein is RNA binding protein or Not.

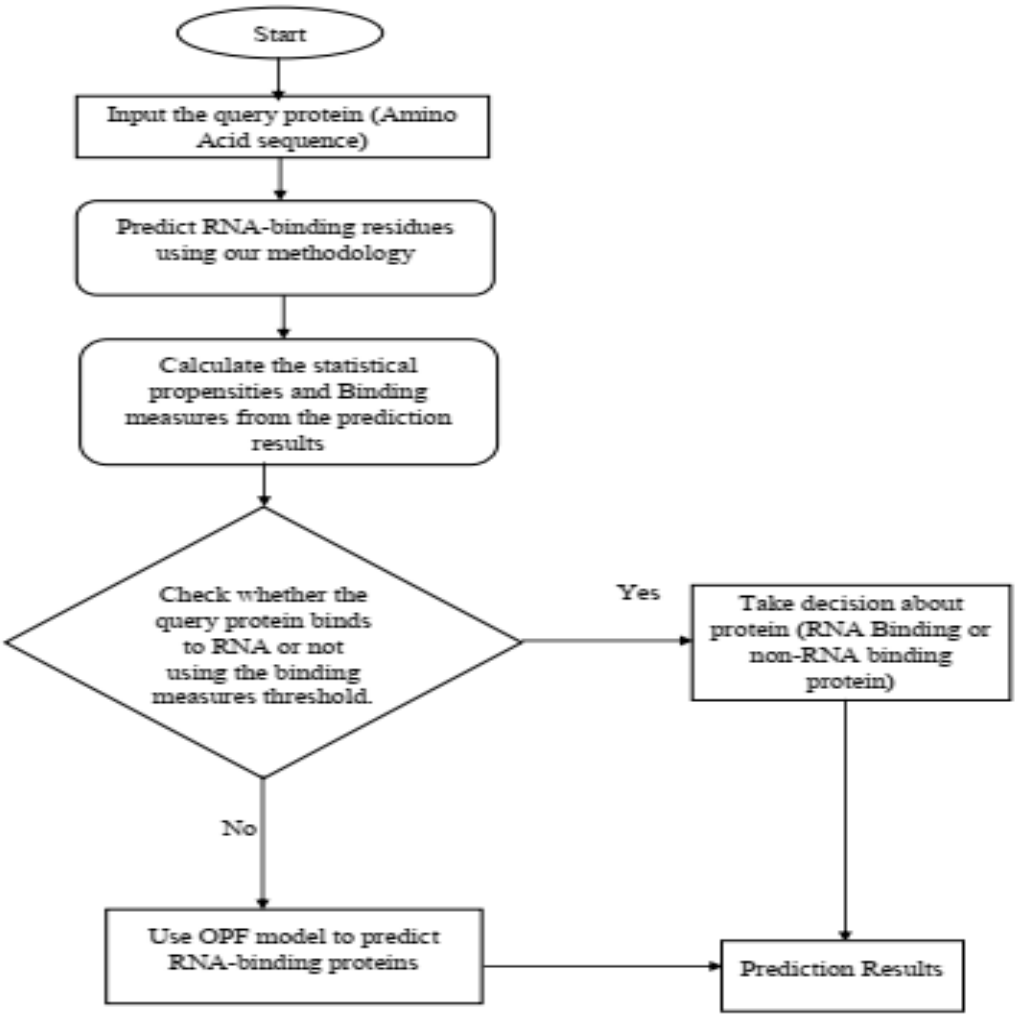


Figure 2. Flowchart of the proposed methodology

RESULTS

EXPERIMENTAL RESULTS

This section starts by illustrating the evaluation criteria, and after that shows the results for our methodology depending on feature vector models.

Evaluation Criteria

To evaluate the results of our prediction methodology, we used five measures: (1) Sensitivity (SE); (2) Specificity (SP); (3) Accuracy (ACC); (4) Matthew correlation coefficient (MCC); and Area under the Curve (AUC) to evaluate our OPF classifier and to differentiate between the results for each model in the feature vector.

Sensitivity is defined by the following equation:

$$SE = \frac{TP}{TP + FN} \quad \textbf{(9)}$$

Specificity is defined by the following equation:

$$SP = \frac{TN}{TN + FP} \quad \textbf{(10)}$$

Accuracy is defined by the following equation:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \quad \textbf{(11)}$$

Matthew correlation coefficient is defined by the following equation:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TN+FN)(TP+FN)(TN+FP)}} \quad \textbf{(12)}$$

Where TP, TN, FP, FN is the number of true positive, true negative, false positive and false negative results.

Results

The experimental results of our methodology were discussed regarding the evaluation criteria discussed before.

RNA binding residues were predicted before in previous studies as stated before, but also here we predicted RNA binding residues to assist our prediction of RBPs throughout the binding propensity measures discussed before. At first RBPs and RNA non-binding proteins are assigned based on statistical doublet propensity threshold with amino acid composition from the feature vector models. Table 1 shows the best results of feature vector models over our measurements and shows the comparison of our models related to the statistical propensity (singlet, doublet) on the feature vector using 10- fold cross validation. The results shows that each model has a better performance at some point and may be increased at one of the measures but decreased at the other, but we can say that the doublet propensity and AAC vector models have a good results than the other ones by means of we can noticed that the sensitivity increases without much more decreases in the specificity and accuracy. Moreover, we can see that the results of the singlet propensity model is much more similar to amino acid sequence model.

The second part in this work is related to RNA binding residues that predicted for all proteins in the combined dataset RNAt_7424. Then RBPs is defined depending on the binding propensity measure thresholds.

In this stage to determine that the query protein binds to RNA or not, we have to test RBPs and RNA non-binding proteins on different threshold of the binding measures. We have two cases, case 1 depending on the upper threshold of BPM1, and BPM2, and case 2 depending on the lower threshold of them. The protein is assigned to be RBP in case of BPM1 of that protein is more than a certain upper threshold of BPM1, or BPM2 of that protein is more than a certain upper threshold of BPM2. Otherwise, the protein is assigned to be non-binding protein in case of BPM1 or BPM2 is less than a certain threshold assigned for each one. Table 2 shows the results of predicting RBPs at different Upper thresholds using BPM1.Table 3 shows the same as table 2 but at different lower thresholds of BPM1. Table 4 shows the results of predicting RNA binding protein at different upper threshold of BPM2, and Table 5 is similar as table 4 but at different lower thresholds of BPM2. In this stage to determine that the query protein binds to RNA or not, we have to test RBPs and RNA non-binding proteins on different threshold of the binding measures. We have two cases, case 1 depending on the upper threshold of BPM1, and BPM2, and case 2 depending on the lower threshold of them. The protein is assigned to be RBP in case of BPM1 of that protein is more than a certain upper threshold of BPM1, or BPM2 of that protein is more than a certain upper threshold of BPM2. Otherwise, the protein is assigned to be non-binding protein in case of BPM1 or BPM2 is less than a certain threshold assigned for each one. Table 2 shows the results of predicting RBPs at different Upper thresholds using BPM1.Table 3 shows the same as table 2 but at different lower thresholds of BPM1. Table 4 shows the results of predicting RNA binding protein at different upper threshold of BPM2, and Table 5 is similar as table 4 but at different lower thresholds of BPM2.

Table 1: Comparison of the best results of feature vector models using 10-fold cross validation

Evaluation Criteria		Feature Vector Models		
	M1	M2	M3	M4
SE	0.347	0.413	0.312	0.462
SP	0.921	0.872	0.881	0.927
ACC	0.802	0.785	0.782	0.804
MCC	0.0553	0.0524	0.0123	0.0516
AUC	0.753	0.742	0.743	0.752

Table 2: The Results of predicting RBPs at different Upper thresholds of BPM1

Upper Threshold	Measurements			
	SE	**SP**	**ACC**	**MCC**
1.0	0.29	0.943	0.596	0.1123
2.0	**0.27**	**0.962**	**0.542**	**0.1114**
3.0	0.25	0.921	0.451	0.0598
4.0	0.226	0.886	0.582	0.1824
5.0	0.351	0.778	0.604	0.1916

Table 3: The Results of predicting RBPs at different Lower thresholds of BPM1

Lower Threshold	Measurements			
	SE	**SP**	**ACC**	**MCC**
0.01	0.743	0.16	0.696	0.0523
0.02	0.862	0.23	0.742	0.0114
0.03	0.981	0.225	0.751	0.0098
0.04	0.996	0.226	0.802	0.0124
0.05	**0.998**	**0.03**	**0.804**	**0.0116**

Table 4: The Results of predicting RBPs proteins at different Upper thresholds of BPM2

Upper Threshold	Measurements			
	SE	**SP**	**ACC**	**MCC**
1.1	0.29	0.943	0.546	0.0323
1.2	0.27	0.962	0.542	0.0214
1.3	**0.05**	**0.992**	**0.551**	**0.0198**
1.4	0.022	0.986	0.582	0.0182
1.5	0.035	0.998	0.604	0.0171

Table 5: The Results of predicting RBPs at different Lower thresholds of BPM2

Lower Threshold	Measurements			
	SE	**SP**	**ACC**	**MCC**
0.01	0.743	0.06	0.696	0.0183
0.02	0.862	0.08	0.742	0.0114
0.03	0.981	0.041	0.761	0.0148
0.04	0.996	0.032	0.825	0.0082
0.05	**0.997**	**0.025**	**0.83**	**0.0086**

From Table 2, and 4 we can classify RBPs and non-binding proteins with a high value of sensitivity but not specificity as seen in table 2 at 2.0 sensitivity is 0.27 but specificity is 0.962. Also in table 3 at 1.3 sensitivity is 0.05 but specificity is 0.992.

On the other side from table 3, and 5 we can classify RBPs and non- binding proteins with a high value of sensitivity but low specificity, at table 3 when 0.05 was selected as the lower bound threshold, the prediction of RBPs were predicted with 0.998 sensitivity but low specificity 0.03, same as point 0.05 in table 5 with 0.997 sensitivity but 0.025 specificity.

From the above results we can conclude that the protein is assigned to be RNA binding protein, if the BPM1 of that protein is greater than 2.0, or BPM2 is greater than 1.3. And the protein is assigned to be non- binding protein, if the BPM1 or BPM2 is less than 0.05. In the RNAt_7424 testing dataset 75 proteins are predicted as RBPs, 140 are predicted as non-binding proteins, and the rest 7209 of proteins are assigned to the OPF classifier to be predicted using our criteria. The combination of these features has great impact of the results which achieved 97.8% sensitivity with 82.05% specificity, with 83.6% accuracy and 0.0523 Matthew's correlation coefficient value.

DISCUSSION

This section discusses the experimental results of the proposed methodology. Although some techniques already achieve good results, their deterministic characteristic might prevent the feasibility of the algorithm in complex situations. Our method shows a very good performance for predicting RNA binding and non-binding proteins for some reasons. At first we combine four models in our system as we combine amino acid sequences with statistical propensity measures (singlet, doublet) and amino acid composition (ACC). Moreover, we use machine learning method by using OPF classifier rather than scoring based method, OPF used in our methodology to assist our prediction results. Furthermore, we tried to put an interval of threshold to the binding propensity measures by predicting RNA binding residues for assisting the prediction of RBPs and non-binding proteins depending on the lower and upper threshold. The OPF classifier performance was evaluated using all features in RNAt-7424 dataset using 10-fold cross validation and achieved high values in sensitivity, specificity, accuracy, and Matthew correlation coefficient respectively.

After that OPF is combined with our model to predict proteins in RNAt-7424 dataset and calculate propensity measures for each sequence depending on upper and lower threshold to assign that sequence to RNA binding or non-binding results. The results showed that 75 proteins in the RNAt-7424 dataset as predicted to be RBPs, 140 are assigned to be non-binding proteins, and OPF classifier is used to predict the other 7209 proteins.

CONCLUSION

In this paper, we proposed and developed an effective method for predicting RBPs from amino acid sequences using optimum path forest classifier in conjunction with statistical and binding propensity measures and amino acid composition models. The two statistical and binding propensity measures are evaluated and analyzed in our method. Then the query protein is made based on the results. After that if the protein not determined to be RNA binding protein or non-RNA binding protein, then the OPF is used to identify whether the protein sequence status depending on the prediction of RNA binding residues. The combination of these features has great impact of the results which achieved 97.8% sensitivity with 82.05% specificity, with 83.6% accuracy and 0.0523 Matthew's correlation coefficient value. As well as we can demonstrate by adding the statistical information and the binding propensity measures of the predicted RNA-binding residues especially contributed to the prediction process. In addition, the OPF classifier has improved the overall performance of RBPs prediction process.
For future work we need to construct a web server for our prediction model to be used by many researches, and to facilitate research's efficient prediction of RBPs.

CONFLICT OF INTEREST

The authors declared that present study was performed in absence of any conflict of interest.

ACKNOWLEGEMENT

I would like to thank my brilliant and truly outstanding directors for expert advice and encouragement through this work.

AUTHOR CONTRIBUTIONS

All authors contributed in collecting and analyzing data. All authors participated in writing every part of this study. All authors read and approved the final version.

Copyrights: © 2019 @ author (s).
This is an open access article distributed under the terms of the **Creative Commons Attribution License (CC BY 4.0)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author(s) and source are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

REFERENCES

Abdelmohsen, K.; Kuwano, Y.; Kim, H.H.; Gorospe, M. (2008) Posttranscriptional gene regulation by RNA-binding proteins during oxidative stress: Implications for cellular senescence. *Biol. Chem 389*, 243–255.

Cai, YD.; Lin, SL. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochim Biophys Acta, vol. 1648, no. 1-2, pp. 127-33, May 30.

Chen,Y.C.; Lim, C. (2008). Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res. 36.*

Consortium, T.U. (2012).Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res,* vol. 40, no. Database issue, pp. D71-D75.

Cusack, S. (1997). "Amino acyl-tRNA synthetizes." Curr Opin Struct Biol 7(6): 881-9.

De Guzman RN., Turner RB, Summers MF. (1998). Protein-RNA recognition. Biopolymers. 48:181–95. Biopolymers, vol. 48, no. 2-3, pp. 181-95, 1998.

Gabb, H.A; Jackson, R.M; Sternberg, M.J. (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. *J. Mol. Biol. 272*, 106–120.

Han, LY; Cai, CZ; Lo, SL; Chung, MC; Chen,YZ. (2004). Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA. 10: 355–368.

Huang, Y.F; Chiu, L.Y; Huang, C.C; Huang, C. K. (2010). Predicting RNA-binding residues from evolutionary information and sequence conservation," *BMC Genomics,* vol. 11 Suppl 4, pp. S2.

Ibba M., Soll D. (1996). Protein-RNA molecular recognition. *Nature*. 381(6584):p. 656. Doi: 10.1038/381656a0.

J.P. Papa, A.X. Falcao , C.T.N. Suzuki, (2009), Supervised pattern classification based on optimum-path forest″, Int. J. Imaging Syst. Technol. 19 (2) (2009) 120–131 .

Jacobs Anderson, J.S.; Parker, R. (2000) Computational identification of cis-acting elements affecting post-transcriptional control of gene expression in Saccharomyces cerevisiae. *Nucleic Acids Res. 28*, 1604–1617.

Katchalski – Katzir, E.; Shariv, I.; Eisenstein, M.; Friesem, A.A.; Aflalo, C.; Vakser, I.A. (1992). Molecular surface recognition: Determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA 89*, 2195 2199.

Kumar, M.; Gromiha, MM; Raghava, G.P. (2011). SVM based prediction of RNA-binding proteins using binding residues and evolutionary information," J Mol Recognit, vol. 24, no. 2, pp. 303-313.

Liu, Xin-Mi; Gong, Xiu-Jun. (2012). Predicting Protein - RNA- binding Sites Using Sequence Statistical Feature of Amino Acids, Physics Procedia, Volume 33, Pages 334-340.

Ma, X.; Guo, J.; Wu, J.; Liu, H.; Yu , J.; Xie , J.; Sun, X. (2011). Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins 79*, 1230–1239.

Ma, X.; Guo, J.; Xiao, K.; and Sun, X. (2015) PRBP: Prediction of RNA-Binding Proteins Using a Random Forest Algorithm Combined with an RNA-Binding Residue Predictor. In *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 12, no. 06, pp. 1385-1393.

Moreira, I.S.; Fernandes, P.A.; Ramos, M.J. (2010). Protein–protein docking dealing with the unknown. *J. Comput. Chem. 31*, 317–342.

Murakami, Y.; Spriggs, R.V.; Nakamura, H.; Jones, S. (2010). PiRaNhA: A server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res 38*, W412–W416.

Nagarajan, R.; Gromiha, M.M. (2014). Prediction of RNA- binding residues: An extensive analysis based on structure and function to select the best predictor. *PLoS ONE 9*,

e91140.

Papa, J. P; Falcao, A. X; de Freitas, G. M; and de Avila, A. M. H. (2010). Robust Pruning of Training Patterns for Optimum-Path Forest Classification Applied to Satellite-Based Rainfall Occurrence Estimation," IEEE Geoscience and Remote Sensing Letters, vol. 7, pp. 396–400.

Papa, J.; Spadotto, A.; Falcao, A.; and Pereira, J. (2008). Optimum path forest classifier applied to laryngeal pathology detection. In Systems, Signals and Image Processing. IWSSIP 2008. 15th International Conference on, june 2008, pp. 249 –252.

Papa, J.P, Falcao, A.X, Albuquerque; V.H.C; Tavares, J.M.R. (2012), Efficient supervised optimum-path forest classification for large datasets, Pattern Recognit. 45 (1) 512–520.

Ritchie, D.W; Kemp, G.J. (2000). Protein docking using spherical polar Fourier correlations. *Proteins 39*, 178–194.

Sayed, S.A., Nabil, E., Badr, A. (2016). A binary clonal flower pollination algorithm for feature selection, Pattern Recognition Letters: North Holland (pp. 21-27).

Schneidman-Duhovny, D.; Inbar, Y.; Nussinov, R.; Wolfson, H.J. (2005). PatchDock and SymmDock. Servers for rigid and symmetric docking. *Nucleic Acids Res. 33*, W363–W367.

Shao, X.; Tian, Y.; Wu, L.; Wang, Y.; Jing, L.; Deng, N. (2009). Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J Theor Biol, vol. 258, no. 2, pp. 289-293.

Si, Jingna; Cui, Jing; Cheng, Jin; Wu, Rongling. (2015). Computational Prediction of RNA-Binding Proteins and Binding Sites. International journal of molecular sciences. 16. 26303-26317. 10.3390/ijms161125952.

Souza, R.; Lotufo, R.; Rittner, L. (2012). "A Comparison between Optimum-Path Forest and k-Nearest Neighbors Classifiers," *2012 25th SIBGRAPI Conference on Graphics, Patterns and Images*, Ouro Preto, pp. 260-267.

Terribilini, M.; Sander, J.D.; Lee, J.H.; Zaback, P.; Jernigan, R.L.; Honavar, V.; Dobbs, D. (2007). RNABindR: A server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res. 35*, W578–W584.

Wang, L.; Brown, S.J. (2006). BindN: A web-based tool for efficient prediction of DNA and RNA- binding sites in amino acid sequences. *Nucleic Acids Res.34*, W243–W248.

Wang, L.; Huang, C.; Yang, M.Q.; Yang, J.Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst. Biol.*, *4* (Suppl. S1).

Yang, Y.; Zhao, H.; Wang, J.; Zhou, Y. (2014). SPOT-Seq-RNA: Predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. *Methods Mol. Biol. 1137*, 119–130.

Yu, X.; Cao, J.; Cai, Y.; Shi, T.; Li, Y. (2006) Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. JTheor Biol, vol. 240, no. 2, pp. 175-184.

Zhao, H.; Yang, Y.; Zhou, Y. (2011). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res.39*, 3017–3025.