Bioinformatics, 33(6), 2017, 854–862
doi: 10.1093/bioinformatics/btw730
Advance Access Publication Date: 21 December 2016
Original Paper

OXFORD

Sequence analysis

RBPPred: predicting RNA-binding proteins from sequence using SVM

Xiaoli Zhang and Shiyong Liu*

School of Physics and Key Laboratory of Molecular Biophysics of the Ministry of Education, Huazhong University of Science and Technology, Wuhan, Hubei 430074, China

*To whom correspondence should be addressed.

Associate Editor: Anna Tramontano

Received on June 25, 2016; revised on October 19, 2016; editorial decision November 12, 2016; accepted on November 16, 2016

Abstract

Motivation: Detection of RNA-binding proteins (RBPs) is essential since the RNA-binding proteins play critical roles in post-transcriptional regulation and have diverse roles in various biological processes. Moreover, identifying RBPs by computational prediction is much more efficient than experimental methods and may have guiding significance on the experiment design.

Results: In this study, we present the RBPPred (an RNA-binding protein predictor), a new method based on the support vector machine, to predict whether a protein binds RNAs, based on a comprehensive feature representation. By integrating the physicochemical properties with the evolutionary information of protein sequences, the new approach RBPPred performed much better than state-of-the-art methods. The results show that RBPPred correctly predicted 83% of 2780 RBPs and 96% out of 7093 non-RBPs with MCC of 0.808 using the 10-fold cross validation. Furthermore, we achieved a sensitivity of 84%, specificity of 97% and MCC of 0.788 on the testing set of human proteome. In addition we tested the capability of RBPPred to identify new RBPs, which further confirmed the practicability and predictability of the method.

Availability and Implementation: RBPPred program can be accessed at: http://rnabinding.com/RBPPred.html.

Contact: liushiyong@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

RNA-binding proteins (RBPs) are proteins that bind to the mRNA or non-coding RNA. In cell, the RBP and RNA can form an RNA–protein complex, which plays an important role in many biological processes, such as posttranscriptional gene regulation, alternative splicing and translation. RBPs interact with tens of thousands of all kinds of RNAs, such as the mRNA (Baltz *et al.*, 2012; Beckmann *et al.*, 2015; Castello *et al.*, 2012; Kwon *et al.*, 2013), long noncoding RNA and tRNA. Recently, high-throughput experimental techniques are developed to identify numerous RBPs, which include 860 RBPs in human HeLa cells (Castello *et al.*, 2012), 797 RBPs in human embryonic kidney cell line (Baltz *et al.*, 2012), 555 mRNA-binding proteins from mouse embryonic stem cells (Kwon *et al.*, 2013) and 120 RBPs from *S. cerevisiae* cells (Mitchell *et al.*, 2013). Despite great efforts to experimentally capture RBPs, we still have an incomplete understanding of how many RBPs exist in all species. Computational prediction approaches are therefore urgent and essential to build an RBP repertoire and RNA-RBP interaction network. As far as we know, there are several computational approaches available predicting RBPs as listed in Table 1. Especially, several SVM-based approaches are developed for RNA-binding proteins prediction (Cai *et al.*, 2003; Cai and Lin, 2003; Han *et al.*, 2004; Kumar *et al.*, 2011; Shao *et al.*, 2009; Shazman and Mandel-Gutfreund, 2008; Spriggs *et al.*, 2009; Yu *et al.*, 2006). Approaches vary in the features employed in those studies, which include the amino acid composition, hydrophobicity, charge, predicted secondary structure and solvent accessible area of residues.

Table 1. Methods for RNA-binding proteins prediction

Method	Means of classification	Level	Properties	Availability
NAbind (Shazman and Mandel-Gutfreund, 2008)	SVM–Gist	Structure-based	Patch size, patch surface accessibly, percent hydrogen bond in patch, protein surface accessibility, dipole, quadrupole moment, the molecular weight, the size of the largest clefts, number of atoms in the negative patch and so on	http://journals.plos.org/ploscompbiol/article/file?id=info%3Adoi/10.1371/journal.pcbi.1000146.s002&type=supplementary
RNApred (Kumar *et al.*, 2011)	SVM–SVM^{light}	Sequence-based	the amino acid, dipeptide, four-part amino acid compositions, predicted binding residues by PPRINT and PSSM	http://www.imtech.res.in/raghava/rnapred/
SPOT-stru (Zhao *et al.*, 2011a,b)	Template-based	Structure-based	The combination of structural alignment and binding affinity	http://sparks-lab.org/pmwiki/download/index.php
SPOT-seq (Zhao *et al.*, 2011a,b)	Template-based	Sequence-based	Sequence-structure match and binding affinity	http://sparks-lab.org/pmwiki/download/index.php
SPalign (Yang *et al.*, 2012)	Template-based	Structure-based	Structure alignment	http://sparks-lab.org/pmwiki/download/index.php
BindUP (Paz *et al.*, 2016)	SVM–Gist	Structure-based	The same as NAbind	http://bindup.technion.ac.il/

*Only the available approaches are listed here, accessed through a web server or program for downloading. Other unavailable RBPs prediction methods include: Cai *et al.* (2003), SVMProt (Han *et al.*, 2004), Yu *et al.* (2006), Ahmad *et al.* (Ahmad and Sarai, 2011), Shao *et al.* (2009), Spriggs *et al.* (2009), Peng *et al.* (Peng *et al.*, 2011), PRBP (Ma *et al.*, 2015a,b) and Ma *et al.* (2015a,b). Gist and SVM^{light} are the chosen tools for SVM classification in corresponding work.

In 2011, Kumar *et al.* (2011) firstly applied evolutional information in the form of position specific scoring matrix (PSSM) to RNA-binding proteins prediction. A SVM-based approach RNApred was developed with a best MCC of 0.62 based on the PSSM-400 on a non-redundant set of 377 RBPs and 377 non-RBPs. Different from the SVM methods, Zhao *et al.* proposed two template-based approaches for predicting RBPs. One is a structure-based method SPOT-stru (Zhao *et al.*, 2011a,b) and the other is a sequence-based method SPOT-seq (Zhao *et al.*, 2011a,b). In SPOT-stru they combined relative structural similarity in the form of Z-score and a statistical energy function DFIRE to predict RBPs. The results show that the combination of Z-score and DFIRE energy function achieved the best performance with MCC of 0.57 on the benchmark of 212 RNA-binding domains and 6761 non-RNA binding domains. Different from the structure-based SPOT-stru, SPOT-seq employed the fold recognition between the target sequence and template structures using the defined sequence-structure matching score. As reported, it achieved a MCC of 0.62 for RBP prediction on a set of 215 RBP chains and 5765 non-binding proteins.

Recently, another three approaches are developed for RBPs prediction, (Ma *et al.*, 2015a,b; Paz *et al.*, 2016). The two methods proposed by Ma *et al.* differ in the features used to train the random forest model for predicting. One (Ma *et al.*, 2015a,b) encodes the EIPP and amino acid composition for the protein sequence, while the other (Ma *et al.*, 2015a,b) employs features of a different EIPP property, conjoint triad, binding and non-binding propensity. In Ma *et al.*'s work, the evolutionary information in the form of PSSM was combined with 6 physicochemical properties to form a new vector named EIPP with 120 dimension. BindUP (Paz *et al.*, 2016) is a structure-based approach, available through a web server for predicting RBP for a given protein structure or structural model using the SVM classifier. Based on the electrostatic features of protein surface and other properties, sensitivity of 0.71, specificity of 0.96 are achieved on an independent testing set of 323 structures of DNA and RNA binding proteins and a control set of an equal number extracted from PDB.

As pointed out in SPOT-stru (Zhao *et al.*, 2011a,b), earlier studies did not eliminate the homologous proteins in the training or testing set (Cai *et al.*, 2003; Cai and Lin, 2003; Han *et al.*, 2004) and the performance was not evaluated using the unbiased measurement of the area under the receiver operating characteristic curve (AUC) or Matthews correlation coefficient (MCC) (Cai *et al.*, 2003; Cai and Lin, 2003; Han *et al.*, 2004; Yu *et al.*, 2006). Moreover, SVM model were trained or tested on more or less equal number of RNA-binding and non-binding proteins (Kumar *et al.*, 2011; Peng *et al.*, 2011; Shao *et al.*, 2009; Spriggs *et al.*, 2009; Yu *et al.*, 2006), which was inconsistent with the real-word simulation where the proportion of discovered RBPs was just a very small fraction of all the proteins (UniProt, 2008). However, the templated-based approaches often performed worse in identifying novel RBPs with low sensitivity since they were based on the homology or similarity between the target protein and the template. Moreover, analysis of the numerous experimentally identified RBPs showed that, 402 of 860 HeLa mRNA-binding proteins (47%) (Castello *et al.*, 2012) lacked known RNA-binding motifs and 216 of the 555 mRNA-binding proteins (39%) (Kwon *et al.*, 2013) lacked known RNA-binding domains. About 39–47% RBPs without known RNA-binding domains imply that they could be missed by sequence homology search tools (Gerstberger *et al.*, 2014; Ghosh and Sowdhamini, 2016) or an RNA-binding domain-based RBP prediction algorithm (Zhao *et al.*, 2014).

In the present work, we proposed a computational approach named RBPPred (an RNA-binding protein predictor) motivated by the previous studies (Kumar *et al.*, 2011; Wang *et al.*, 2013; Yu *et al.*, 2006). By combing important features used in the three work, we employed more significant features for RBPs prediction. The important features include hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure,

predicted solvent accessibility, side chain's charge and polarity in protein-RNA interaction and the PSSM profile of the protein sequence. The SVM classifier was chosen to distinguish the RBPs from non-RBPs based on the trained model and tested on independent testing sets.

2 Methods

2.1 Datasets

2.1.1 Training set

To develop and evaluate the PBPPred method, we constructed a non-redundant training set of RBPs and non-RBPs. The flow chart of building our training set was described in Figure 1. By using the GO term 'RNA binding' to search the UniProt database (Apweiler *et al.*, 2004), we obtained 68084 reviewed RBP chains. For non-RBPs, we adopted the construction method from the approach SPOT-stru (Zhao *et al.*, 2011a,b), by using PISCES (Wang and Dunbrack, 2003) with sequence identity of 25%, sequence length between 50 and 10 000 amino acids and resolution of X-ray better than 3.0 Å. As a result, 14 389 chains were picked out. Protein chains with the PDB records of 'ribosomal', 'RNA', 'Nucleoprotein', 'unknown function', 'uncharacterized', or 'hypothetical' in the title were removed. This generated 12790 protein sequences, which were regarded as non-RBPs.

Afterwards, the RBPs and non-RBPs were mixed to remove the redundant sequences with sequence identity cutoff $\geq$25% using the psi-cd-hit program in the CD-HIT package (Li and Godzik, 2006). Hence, we obtained 2878 RBPs and 7098 non-RBPs. To keep consistent with the protein length in the non-RBPs dataset, the proteins with length less than 50 or more than 10 000 amino acids in the RBPs set were discarded. Meanwhile, proteins with the title 'Fragment' were also abandoned from the RBPs dataset. After that, 2782 RBPs remained.

Figure 1 shows the flow chart of the construction of the training set which finally includes 2780 RBPs and 7093 non-RBPs. 2 RBPs and 5 non-RBPs were removed since no secondary structure or evolutionary information results were generated from the programs for secondary structure prediction or evolutionary information searching.

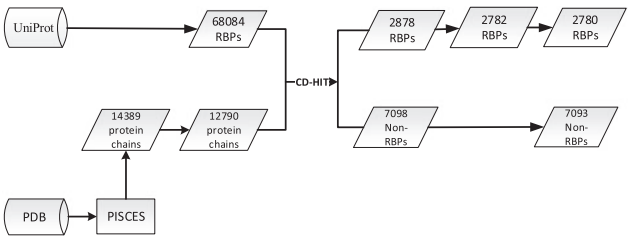

Fig. 1. The flowchart of building training set. The initial RBPs collection of 68084 reviewed RBPs was obtained by using 'GO: 0003723 (RNA binding)' to search the UniProt database. The initial collection of 12790 non-RBPs was retrieved through using the PISCES tool to cut protein sequences from PDB and removing those proteins with unknown function or related to RNA binding. The redundancy between the proteins of the initial 68084 RBPs and 12790 non-RBPs was removed by using the CD-HIT tool with sequence identity cutoff of 25%, which generated the non-redundant collections of 2878 RBPs and 7098 non-RBPs. To keep the length consistency with the proteins in the collection of non-RBPs, some proteins were deleted from the non-redundant set of RBPs, which resulted in a set of 2782 RBPs. When using SSPro to predict the secondary structure, ACCPro to predict solvent accessibility (Magnan and Baldi, 2014) or PSI-BLAST to generate PSSM (Altschul *et al.*, 1997) for the non-redundant 2782 RBPs and 7098 non-RBPs, we found that there were no results output for some proteins. So we discarded these proteins and the remaining 2780 RBPs and 7093 non-RBPs constituted the final non-redundant training set

2.1.2 Independent testing set

The RBPPred was tested on 3 species, human and other two model organisms, *Saccharomyces cerevisiae* (*S. cerevisiae*) and *Arabidopsis thaliana* (*A. thaliana*).

Retrieval with the GO term 'RNA binding' to search UniProt, we collected 1551, 560 and 603 RBPs for human, *S. cerevisiae* and *A. thaliana* respectively. For the negative samples, we extracted the non-RBPs belonging to the three species from PISCES (Wang and Dunbrack, 2003) by searching a different version of PDB, which formed three non-redundant negative sets, including1350 non-RBPs of human, 395 non-RBPs of *S. cerevisiae* and 102 of *A. thaliana* proteomes, respectively.

Some proteins were removed due to the same deletion reason with the training set, that is, no secondary structure or evolutionary information results were generated for these proteins. Moreover, in order to test objectively, the same sequences between each of the three testing sets and training set were deleted. Three independent testing sets were eventually constructed, which contained 967 RBPs and 597 non-RBPs for human, 354 RBPs and 135 non-RBPs for *S. cerevisiae*, 456 RBPs and 37 non-RBPs for *A. thaliana*, respectively.

2.2 Sequence feature and vector encoding

The construction of the feature vector for each protein sequence was based on the amino acids composition and evolutionary information of the primary sequence. We encoded eight properties with a vector of 576 dimension to represent a protein sequence, as shown in Figure 2. The dimension of each feature group was also listed in Supplementary Table S1 (See Supplemental Material). Five vectors with 21 dimension representing the properties of hydrophobicity, predicted secondary structure, normalized van der Waals volume, polarity and polarizability, respectively, a 7 dimensional vector indicating solvent accessibility, a 64 dimensional vector indicating

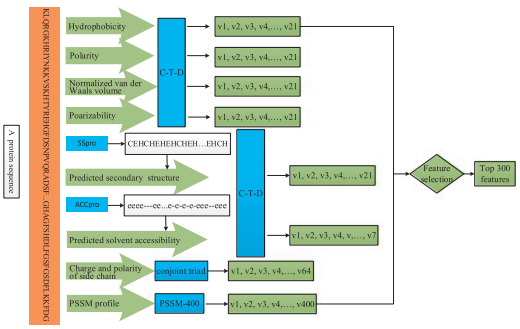

Fig. 2. Procedure of encoding the protein sequences into feature vectors. For secondary structure, SSPro was used to predict the secondary structure in the form of 'C', 'E' and 'H'. For relative solvent accessibility, ACCPro was used to present the solvent accessibility in the form of 'e' and '-'. For the five properties of hydrophobicity, polarity, normalized van der Waals volume, polarizability and predicted secondary structure, the global protein sequence descriptors (C-T-D) was employed to encode each feature vector with 21 dimension (v1, v2, v3, v4,..., v21). For predicted solvent accessibility, C-T-D was applied to encode a feature vector of 7 dimension (v1, v2, v3, v4,..., v7). According to charge and polarity of side chain, the protein sequence was encoded to a vector of 64 dimension (v1, v2, v3, v4,..., v64) through the conjoint triad encoding method. The sequence's PSSM profile obtained by PSI-BLAST was encoded to a 400 dimensional vector (v1, v2, v3, v4,..., v400) by PSSM-400 approach

charge and polarity of side chain and evolutionary information with a vector of 400 dimension are encoded for each protein sequence.

2.2.1 Physicochemical properties

For an example protein sequence displayed in Figure 2, we used the global composition feature encoding method (composition, transition and distribution, i.e. C-T-D) to encode the properties of hydrophobicity, polarity, normalized van der Waals volume, polarizability, predicted secondary structure and solvent accessibility, which we adopted from (Han *et al.*, 2004). The global protein sequence descriptors (C-T-D) was firstly proposed to describe global composition of amino acid sequence in protein folding class prediction (Dubchak *et al.*, 1995).

In the method, the first index C describes the percent composition of each group in the protein sequence. Meanwhile, the second descriptor T represents the transition probability between two contiguous amino acids belonging to two different groups. Last but not the least, the distribution of amino acid (the position of the first, 25%, 50%, 75% and the last amino acid) of each group along the sequence is expressed by the third descriptor D. 20 amino acids were classified into 3 groups (Dubchak *et al.*, 1999) according to their hydrophobicity, normalized van der Waals volume, polarity, polarizability. For the predicted secondary structure and solvent accessibility, we employed the SSpro and ACCpro program (Magnan and Baldi, 2014) to predict because of its declaratively remarkable prediction performance and reasonable speed. As reported, it achieved accuracies of 92.9% for secondary structure prediction and 90% for relative solvent accessibility prediction by combining sequence similarity and sequence-based structural similarity. As described in Figure 2, SSpro output the secondary structure in the form of H (representing helix), E (representing strand) and C (representing the rest) for each amino acid along the sequence. In a similar way, ACCpro exported 'e' for the exposed residues and '-' for the buried representing predictive solvent accessibility. Using the encoding method and classification of amino acids described above, we finally constructed a 21 dimensional vector representing the physicochemical properties of hydrophobicity, normalized van der Waals volume, polarity, polarizability, predicted secondary structure and 7 for solvent accessibility, respectively.

For charge and polarity of side chain in protein-RNA interaction, we followed the conjoint triad encoding strategy, used for protein-RNA interaction prediction (Wang *et al.*, 2013), which was firstly proposed for protein-protein interaction prediction (Shen *et al.*, 2007). Twenty amino acids were grouped into 4 classes, which are acidic [DE], basic [HRK], polar [CGNQSTY] and non-polar [AFILMPVW] (Cheng *et al.*, 2008; Wang *et al.*, 2013; Yu *et al.*, 2006). Three successive amino acids were treated as a unit and the different amino acids belonging to the same type were treated as the same. Using conjoint triad method to encode the four types of amino acids, we obtained a 4 * 4 * 4= 64 dimensional vector representing each protein sequence. Each value in the 64 dimensional vector was the normalized probability of a specific triad along the sequence.

2.2.2 Evolutionary information

In RNApred method, evolutionary information in the form of position specific scoring matrix (PSSM) was firstly used for predicting RNA-binding proteins with a maximum MCC of 0.62 (Kumar *et al.*, 2011). The PSSM profile was generated for each protein sequence using PSI-BLAST to search the NR (non-redundant) protein database using three iterations with e-value threshold of 0.001 for inclusion of sequences during constructing profiles. The probability of each of 20 amino acids at each position of a query protein sequence is an essential part of the PSSM profile.

In our study, we performed PSSM constructing by using PSI-BLAST (Altschul *et al.*, 1997) (BLAST 2.2.30+ released) to search the NCBI-NR90 database with three iterations and e-value threshold of 0.001 for saving hits. The other parameters for PSI-BLAST searching were as default. NR90 database is a representative subset of the NR database, which was derived by using CD-HIT (Li and Godzik, 2006) with sequence identity of 90% to remove the homologous protein sequences from NR database. Because it took much less computational time and achieved only slightly poorer performance than the NR database, NR90 was used to execute the PSI-BLAST search (Ahmad and Sarai, 2005; Carson *et al.*, 2010; Si *et al.*, 2009).

Then, we obtained the normalized PSSM by using formula $1/(1+\hat{e}(-x))$, where x is the value in the PSSM. In order to convert the L * 20 PSSM profile (L is the number of amino acids in the query protein sequence) into a fixed dimension, we adopted the strategy used by Kumar *et al.* for DNA- and RNA-binding proteins prediction. Firstly, for each column, the values belonging to the same amino acid in all rows were summed to form a vector of 20 dimension. Secondly, 20 vectors were combined together to form a 20 * 20 = 400 dimensional vector (Kumar *et al.*, 2007, 2011).

2.3 SVM classifier

The support vector machine (SVM) method was used for processing classification and regression problems. In the study, LIBSVM-3.17 package (Chang and Lin, 2011) was used as a stand-alone program to train the model and perform RBP prediction using the radial basis function (RBF) kernel. The optimal values of tunable parameter C and γ were determined by the grid search method and C = 185363.800047, $\gamma = 0.000690533966002$ were obtained for the training set with the selecting top 300 features of the vectors using 20 CPUs in about 13 hours.

2.4 Performance evaluation

The performance of RBPPred was measured by the ten-fold cross-validation approach. To perform this cross-validation, the training set was randomly divided into ten parts of equal size. For each cross-validation, the nine parts were combined as the sub-training set while the remaining one part was used as the sub-testing set for testing. This process was repeated ten times to ensure each part was once used as the sub-testing set. We evaluated the average performance of all the ten sub-testing sets by using sensitivity (SN), specificity (SP), precision (PRE), accuracy (ACC), F-measure and Matthews correlation coefficient (MCC), which are defined as:

$$\text{Sensitivity (SN)} = \text{TP}/(\text{TP}+\text{FN})$$

$$\text{Specificity (SP)} = \text{TN}/(\text{TN}+\text{FP})$$

$$\text{Precision (PRE)} = TP/(TP+FP)$$

$$\text{Accuracy (ACC)} = (\text{TP}+\text{TN})/(\text{TP}+\text{FN}+\text{TN}+\text{FP})$$

$$\text{F}-\text{measure} = (2*\text{PRE}*\text{SN})/(\text{PRE}+\text{SN})$$

Matthews Correlation Coefficient (MCC)

$$= (\text{TP}*\text{TN} - \text{FP}*\text{FN})/\sqrt{(\text{TP}+\text{FN})*(\text{TP}+\text{FP})*(\text{TN}+\text{FP})*(\text{TN}+\text{FN})}$$

where, TP refers to true positive, FN, TN, FP represents false negative, true negative and false positive respectively. MCC gives an

Table 2. Comparison between predictive abilities of individual feature group and the performance of RBPPred (all features combined) on 2780 RBPs and 7093 non-RBPs

Feature group	SN (%)	SP (%)	PRE (%)	ACC (%)	F-measure	AUC	MCC
Evolutionary information (PSSM)	74.67 ± 2.93	96.45 ± 1.19	89.20 ± 3.56	90.32 ± 1.16	0.812 ± 0.023	0.946 ± 0.007	0.754 ± 0.029
Polarizability (Pz)	65.08 ± 2.42	92.68 ± 0.79	77.56 ± 3.46	84.93 ± 0.75	0.708 ± 0.026	0.904 ± 0.005	0.611 ± 0.029
Hydrophobicity (Hb)	57.96 ± 2.71	95.51 ± 0.79	83.52 ± 2.24	84.93 ± 1.48	0.684 ± 0.022	0.917 ± 0.006	0.606 ± 0.027
Van der Waals volume (VDWV)	59.82 ± 3.05	93.99 ± 0.90	79.47 ± 3.70	84.37 ± 1.12	0.682 ± 0.028	0.903 ± 0.008	0.592 ± 0.033
Polarity (Pl)	55.17 ± 3.78	95.66 ± 0.67	83.23 ± 2.55	84.26 ± 1.22	0.663 ± 0.029	0.912 ± 0.009	0.586 ± 0.028
Relative solvent accessibility (RSA)	39.96 ± 3.79	98.60 ± 0.71	91.86 ± 3.70	82.08 ± 1.35	0.556 ± 0.035	0.796 ± 0.008	0.528 ± 0.026
Secondary structure (SS)	44.60 ± 3.89	96.92 ± 0.87	85.01 ± 3.84	82.17 ± 1.46	0.584 ± 0.033	0.825 ± 0.023	0.526 ± 0.032
Charge and polarity of side chain (CPSA)	42.14 ± 4.60	95.10 ± 0.69	77.08 ± 2.09	80.23 ± 1.20	0.544 ± 0.039	0.861 ± 0.011	0.464 ± 0.031
All features (RBPPred)	82.77 ± 1.65	96.50 ± 0.80	90.18 ± 2.50	92.64 ± 0.61	0.863 ± 0.016	0.975 ± 0.003	0.814 ± 0.019

*The results were calculated by using 10-fold cross-validation and the values were listed in the form of the average ± the standard deviation.

overall measurement of the performance while SN (or SP) assesses the correct prediction rate in the positive (or negative) set. Another objective evaluation index AUC used here is the area under the receiver operation characteristic (ROC) curve.

3 Results

In this work, we proposed a sequence-based approach named RBPPred for RNA-binding proteins prediction. To perform the RBPs prediction, we extracted important features from each protein sequence, including 5 physicochemical properties, predicted secondary structure information, predicted relative solvent accessibility and the evolutionary information. SVM was trained using the encoded features to present the model for prediction. We also applied our method to proteomes, and compared it with previous approaches.

3.1 Performance of RBPPred on the training set using 10-fold cross-validation

The performance of RBPPred was measured using 10-fold cross-validation, by a number of indicators (SN, SP, PRE, ACC, F-measure, AUC and MCC). Our RBPPred successfully predicted RBPs using 10-fold cross-validation on 2780 RNA-binding, 7093 non-binding proteins, and the best performance was achieved with the average SN of 82.77%, SP of 96.50%, F-measure of 0.863, AUC of 0.975 and MCC of 0.814 by all the eight properties combined, as listed in Table 2. We also compared the contribution of individual property to the prediction and observed that MCC ranged from 0.464 to 0.754, suggesting the difference in the prediction ability of each feature group. In all of the single properties, the prediction ability of evolutionary information was the best, with the highest SN, AUC and MCC values of 74.67%, 0.946 and 0.754, respectively.

However, not all the single property we have chosen performed well for RBP prediction, some of which gave poor prediction results. In addition, the generated features may be redundant with each other. Therefore, we employed feature selection method to remove the redundancy among the eight properties and picked out the top ranked features according to their predictive contribution. mRMR algorithm was proposed by Peng *et al.* for selecting good features in pattern classification system based on mutual information with the minimal redundancy, maximal relevance criteria (Peng *et al.*, 2005). The program ranks each feature with the corresponding mRMR score and the higher score represents the stronger prediction ability of the feature.

Table 3. The performance of prediction models by employing mRMR-based feature selection method on 2780 RBPs and 7093 non-RBPs

Model	Top 100	Top 200	Top 300
SN (%)	77.97 ± 1.97	80.97 ± 2.19	83.07 ± 2.13
SP (%)	96.39 ± 0.40	96.17 ± 0.69	96.00 ± 0.80
PRE (%)	89.38 ± 1.43	89.15 ± 2.41	89.00 ± 2.61
ACC (%)	91.23 ± 0.59	91.90 ± 0.81	92.36 ± 0.75
F-measure	0.833 ± 0.016	0.848 ± 0.019	0.859 ± 0.018
AUC	0.965 ± 0.003	0.970 ± 0.004	0.975 ± 0.003
MCC	0.777 ± 0.018	0.795 ± 0.023	0.808 ± 0.022

We chose mRMR to filter the top 100, 200 and 300 important features from the total 576 features encoded from the 2780 RBPs and 7093 non-RBPs in the whole training set. The selected features and the number of them in each feature group were listed in Supplementary Table S1 and S2 (See Supplemental Material). It can be seen that the selected features covered the eight feature groups which demonstrates that the encoded features are helpful to the prediction of RNA-binding proteins. Particularly, the evolutionary information generated by PSI-BLAST occupied the largest number in the selected top 100, 200 and 300 features, which proved its significant role in RNA-binding proteins prediction. The second largest proportion is charge and polarity of side chain, with selected 34, 22 and 10 features among the top 300, 200 and 100 features, respectively. The number of features picked out from polarity and hydrophobicity properties was more than from the other three properties with 21 dimensionality. For relative solvent accessibility, 5 features were retained in the three types of feature selection. Among the selected features, the firstly ranked was the feature from evolutionary information, followed by the features from secondary structure and from polarity. Seventeen features were selected from the evolutionary information among the top 30 features listed in Supplementary Table S2.

We applied the selected top 100, 200, 300 features from the entire training set to each subset. Only the selected features were encoded for each subset of the training set to conduct the ten-fold cross-validation. The prediction performance by taking different number of features were summarized in Table 3. As is shown, the top 300 features gave the best results with the average SN of 83.07%, SP of 96%, AUC of 0.975 and MCC of 0.808, which were chosen to create the final model.

3.2 Performance of RBPPred on independent datasets

RBPPred was tested on three independent datasets from human, *S. cerevisiae* and *A. thaliana* species respectively. The whole training set, which containing 2780 RBPs and 7093 non-RBPs was used to construct the model. As shown in Table 4, RBPPred achieved excellent performance with a best MCC of 0.788, sensitivity of 84.28% and specificity of 96.65% for 967 RBPs and 597 non-RBPs for human. Moreover, for the whole 1551 RBPs in human proteome with GO annotation of RNA binding from UniProt, RBPPred successfully predicted 84.28%, which performed much better than SPOT-seq (42.6% reported) (Zhao *et al.*, 2014). Likely, prediction results were obtained for the other two testing sets, with sensitivities of 86.16% and 86.40%, specificities of 91.85% and 94.59%, MCCs of 0.729 and 0.537 for *S. cerevisiae* (354 RBPs and 135 non-redundant non-RBPs) and *A. thaliana* species (456 RBPs and 37 non-redundant non-RBPs).

Supplementary Table S3 was presented to analyze which classes of RNA binding proteins were not predicted well by RBPPred for the independent datasets by cross-referencing the predictions against the PFAM database (Finn *et al.*, 2016). The RNA-binding proteins from three independent datasets were analyzed to find the Pfam families they belonging to. The number of proteins in the families from the above RBPs set and the proportion of families which were incorrectly predicted were presented in Supplementary Table S3. The results showed that for the larger families, such as PF00076.20, PF00271.29 and PF00270.27, 97% of 488 PF00076.20, 95% of 143 PF00271.29 and 95% of 141 PF00270.27 families were correctly predicted by RBPPred. For the small families which have number of proteins less than or equal to 10 from independent sets, our method also achieved high success rate with 79% of 1482 small families being completely predicted successfully and only 13% of these small families were completely wrongly predicted. Especially, PF04857.18, PF13041.4 and PF00191.18 families were poorly predicted with the error rates of 96%, 67% and 67%, respectively.

3.3 Application of RBPPred to human RBPs census

Recently, a census of 1542 RBPs in human proteome (Gerstberger *et al.*, 2014) are extracted from Pfam database (Finn *et al.*, 2010) starting from protein domains, which were tested to verify our method. The whole RBPs were classified to two groups, experimentally validated and computationally recognized RBPs. Our method achieved better performance on the experimentally validated RBPs which have experiment evidence of binding RNAs.

Firstly, we classified these proteins identified by Baltz *et al.* (2012), Castello *et al.* (2012) or included in RBPDB (Cook *et al.*, 2011) as experimentally validated RBPs and the others in the list of 1542 RBPs as computationally recognized RBPs. From the analysis of 1542 proteins ids extracted form Gerstberger *et al.* we find that there are 8 different protein ids only corresponding to 4 unique protein sequences. In this situation, only one protein id was kept in the set. This generated the whole set of 1538 RBPs. Finally, 916 RBPs with experimentally evidence of interacting with RNAs and 622 computationally recognized RBPs were included in the testing set.

Table 5 listed the prediction results of RBPPred on the census of 1538 human RBPs (Gerstberger-1538) and 1284 RBPs where the identical sequences with the training set were removed (Gerstberger-1284). As shown, 81% RBPs were successfully predicted for 916 experimentally validated RBPs, and the sensitivity declined much with the success rate of 58% for 622 computationally recognized RBPs. In addition, the success rate was consistently slightly dropped when sequences included in the training set were removed (for example, from 72% to 68%).

Furthermore, we have compared the data from our independent testing set of human species with the 1538 RBPs from a census of human RBPs by Gerstberger *et al.* in Figure 3. There are 705 RBPs present both in the positive samples of 1551 RBPs reviewed from UniProt and the collection of 1538 putative RBPs from the census of human RBPs. Among the 705 common RBPs identified, 81.42% were correctly identified by our method. However, 213 proteins identified as RBPs by Gerstberger *et al.* belongs to the set of 9647 unlabeled human proteins which have no direct evidence of binding an RNA and still 66.20% of them were predicted as RBPs by RBPPred. Last, the rest of 620 proteins only appearing in the census

Table 4. Performance on the RBPs and negative samples from PDB for three proteomes

Dataset	Human	S. cerevisiae	A. thaliana
SN (%)	84.28	86.16	86.40
SP (%)	96.65	91.85	94.59
PRE (%)	97.60	96.52	94.59
ACC (%)	89.00	87.73	87.02
F-measure	0.905	0.910	0.925
MCC	0.788	0.729	0.537

*Human dataset is the combination of 597 non-RBPs from PISCES with 25% sequence identity and the 967 RBPs without elimination of redundancy from the human proteome, as described in the section 'Datasets'; *S. cerevisiae* dataset is the combination of 135 non-RBPs with 25% sequence identity from PISCES and the 354 RBPs without elimination of redundancy from the *S. cerevisiae* proteome, as described in the section 'Datasets'; *A. thaliana* dataset is the combination of 37 non-RBPs from PISCES with 25% sequence identity and the 456 RBPs without elimination of redundancy from the *A. thaliana* proteome, as described in the section 'Datasets'.

Table 5. Prediction results of 1538 human RBPs from work of Gerstberger *et al.*

Dataset		SN
Gerstberger-1538	Total (1538)	0.72
	Experimental (916)	0.81
	Computational (622)	0.58
Gerstberger-1284	Total (1284)	0.68
	Experimental (718)	0.78
	Computational (566)	0.56

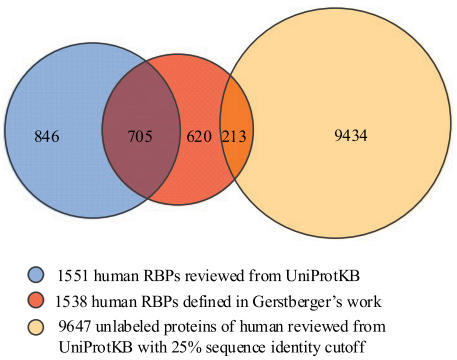

Fig. 3. Comparison of 1538 RBPs from a census of human RBPs by Gerstberger *et al.* with the 1551 RBPs and 9647 unlabeled proteins with sequence identity of 25% reviewed from UniProt

as RBPs but absent in our set were predicted as binding with percentage of 62.58.

3.4 Capacity of forecasting new RBPs and comparison with existing approaches

To further evaluate our method, we tested its ability of predicting novel RBPs and made comparisons with other methods on different sets. The results show that our method RBPPred performed much better than SPOT-seq and RNApred.

As we summarized in Table 1, there are several methods developed for RNA-binding proteins prediction but only two sequence-based methods can be accessed (Kumar *et al.*, 2011; Zhao *et al.*, 2011a,b), either through a web server or code was public for downloading. In this research, RNApred was compared in the module of amino acid composition, which was the only supported prediction module on its web server now. The other two prediction modules of PSSM profile and hybrid prediction could not produce results in tolerable time on the web.

As time goes on, some new RBPs were identified or annotated. We collected the proteins annotated with the function of binding RNAs between June 9, 2015 and April 13, 2016 as new RBPs. Using the term 'GO:0003723' to search the UniProt database, we reviewed 31 new RBPs of human, 49 new RBPs of *S. cerevisiae* and 65 of *A. thaliana* proteomes. Apply our method RBPPred and the other two methods to the newly annotated RBPs, we compared the number of RBPs which are correctly predicted as binding. From Table 6 we can see that our method RBPPred achieved much better performance than SPOT-seq. RBPPred correctly identified 25 out of 31 novel human RBPs. Furthermore, 43 in 49 newly recognized RBPs from *S. cerevisiae* and 57 amongst 65 recently annotated RBPs from *A. thaliana* proteomes are also predicted successfully. But for the new RBPs of human and *A. thaliana*, our method RBPPred performed little worse than RNApred. However, as can be seen from Table 7, RNApred achieved prediction results with much high false positive rate. The detailed information of the newly annotated RBPs of the three proteomes, including RNA binding type, evidence for RNA binding were described in Supplementary Table S4.

In order to compare effectively, the three methods were tested on the independent testing sets from human, S. cerevisiae and A. thaliana proteome, including both the positive and negative data. The prediction results of human testing set, which consisting 967 positive and 597 negative samples, were listed in Table 7. RBPPred performed much better than SPOT-seq and RNApred, with accuracy of 89% versus 57% and 71%, MCC of 0.788 versus 0.330 and 0.366 on the human set. The prediction results of the other two testing sets (See Supplemental Material) further strengthened the better performance of RBPPred than the other two methods with MCC of 0.729 versus 0.312 and 0.446 for *S. cerevisiae* (Supplementary Table S5), and MCC of 0.537 versus 0.312 and 0.162 for *A. thaliana* (Supplementary Table S6), respectively.

From the above comparison of results, it seems that RNApred and SPOT-seq tend to achieve high performance on one data type (positive or negative). RNApred achieved high sensitivity for the positive data but low specificity for the negative data while SPOT-seq achieved low sensitivity for the positive data but high specificity for the negative data. In contract, our method RBPPred can strike a balance between the two kinds of data with both high sensitivity and specificity.

Table 6. The number of new RBPs that was correctly predicted by RNApred, SPOT-seq and RBPPred for human, *S. cerevisiae* and *A. thaliana* species

Organism	Number of new RBPs	RNApred	SPOT-seq	RBPPred
Human	31	28	10	25
S. cerevisiae	49	42	19	43
A. thaliana	65	65	23	57

Table 7. Method comparison for RNA-binding proteins prediction on human testing set

Method	SN (%)	SP (%)	PRE (%)	ACC (%)	F-measure	MCC
RNApred	88.52	43.55	71.75	71.36	0.793	0.366
SPOT-seq	34.54	94.30	90.76	57.35	0.500	0.330
RBPPred	84.28	96.65	97.60	89.00	0.905	0.788

3.5 Discussion

RNA-binding proteins play important and various roles in cells and biological processes. Some experimental technologies and calculation methods were developed and used to detect or predict the interactions of a protein and RNA. Overall, the computational methods can be divided into three levels: the prediction of RBPs, the prediction of binding sites in the protein or RNA sequence (Carson *et al.*, 2010; Choi and Han, 2013; El-Manzalawy *et al.*, 2016; Kumar *et al.*, 2008; Liu *et al.*, 2010; Livi *et al.*, 2016; Miao and Westhof, 2015; Muppirala *et al.*, 2016; Sun *et al.*, 2016; Walia *et al.*, 2012, 2014; Wu and Zhou, 2013; Yang *et al.*, 2014), the prediction of protein-RNA pairs (Agostini *et al.*, 2013; Akbaripour-Elahabad *et al.*, 2016; Bellucci *et al.*, 2011; Cheng *et al.*, 2015; Lu *et al.*, 2013; Muppirala *et al.*, 2011; Suresh *et al.*, 2015). Relative to the other two kinds of prediction, the available approaches of RBPs prediction are much less. In this work, we proposed a new SVM-based predictor RBPPred for RBPs prediction by integrating the features used in previous works, which have some improvement in prediction performance.

From the performance we can see, the evolutionary features have a major impact on the SVM performance. To illustrate the universality of our training set and the possible bias of the data between the training set and the three testing species, we turned back to analyze the species composition of the proteins in the training set (See Supplemental Material). The analysis showed that the 2780 RBPs and 7093 non-RBPs came from hundreds of organisms and the top nine organisms were listed in Supplementary Table S7 and S8. The proteins of human, *S. cerevisiae* and *A. thaliana* only occupied 20%, 7%, 5% for the 2780 RBPs and 16%, 4% and 1% for 7093 non-RBPs of the training set, respectively. Besides from the proteins of the three testing species, the great majority of proteins in the training set were from other species. However, since the limited number of available RBPs in the three testing sets, only the identical proteins (rather than redundancy proteins) with training set were removed from testing sets, which may lead to some bias of the data between the training and testing sets.

In the application to human proteome, 9615 non-redundant proteins with sequence identity cutoff of 25% were reviewed after removing those RBPs from human proteome. RBPPred only predicted 31% of them as non-RBPs and forecasted the rest may have the potential of binding RNAs with a probability score from 0.5 to 1. The protein list contains 6657 possible RBPs with the probability score cutoff of 0.5 (Supplementary Table S9). For each of the proteins, we specified its known functions and prediction probability score. Predicted probability score is the predicted probability value

between 0 and 1 of a protein potentially to be a RBP; the higher the value is, the greater the probability to be a new RBP. In the table, we further marked those proteins simultaneously predicted as RBPs by SPOT-seq, a total of 1134 proteins. The results show that there are many possibly potential RBPs waiting to be discovered.

Acknowledgements

We thank the National Supercomputing Center in Shenzhen and the National Supercomputer Center in Guangzhou for support of computing resources.

Funding

This work has been supported by the National Natural Science Foundation of China [31100522]; the National High Technology Research and Development Program of China [2012AA020402]; Special Program for Applied Research on Super Computation of the NSFC-Guangdong Joint Fund (the second phase) and the Fundamental Research Funds for the Central Universities [2016YXMS017].

Conflict of Interest: none declared.

References

Agostini,F. *et al*. (2013) catRAPID omics: a web server for large-scale prediction of protein–RNA interactions. *Bioinformatics*, **29**, 2928–2930.

Ahmad,S. and Sarai,A. (2005) PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinform*., **6**, 33.

Ahmad,S. and Sarai,A. (2011) Analysis of electric moments of RNA-binding proteins: implications for mechanism and prediction. *BMC Struct. Biol*., **11**, 8.

Akbaripour-Elahabad,M. *et al*. (2016) rpiCOOL: A tool for In Silico RNA–protein interaction detection using random forest. *J. Theor. Biol*., **402**, 1–8.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Apweiler,R. *et al*. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res*., **32**, D115–11D119.

Baltz,A.G. *et al*. (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.

Beckmann,B.M. *et al*. (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun*., **6**, 10127.

Bellucci,M. *et al*. (2011) Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**, 444–445.

Cai,C.Z. *et al*. (2003) SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res*., **31**, 3692–3697.

Cai,Y.D. and Lin,S.L. (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta*, **1648**, 127–133.

Carson,M.B. *et al*. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res*., **38**, W431–43W435.

Castello,A. *et al*. (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.

Chang,C.C. and Lin,C.J. (2011) LIBSVM. A Library for Support Vector Machines, *ACM Trans. Intell. Syst. Technol. (TIST)*, **2**, 27.

Cheng,C.W. *et al*. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinform*., **9**, S6.

Cheng,Z. *et al*. (2015) Computationally predicting protein–RNA interactions using only positive and unlabeled examples. *J. Bioinform. Comput. Biol*., **13**, 1541005.

Choi,S. and Han,K. (2013) Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Comput. Biol. Med*., **43**, 1687–1697.

Cook,K.B. *et al*. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res*., **39**, D301–D308.

Dubchak,I. *et al*. (1995) Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. U. S. A*., **92**, 8700–8704.

Dubchak,I. *et al*. (1999) Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification. *Proteins*, **35**, 401–407.

El-Manzalawy,Y. *et al*. (2016) FastRNABindR: fast and accurate prediction of protein–RNA interface residues. *PLoS One*, **11**, e0158445.

Finn,R.D. *et al*. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res*., **44**, D279–D285.

Finn,R.D. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res*., **38**, D211–D222.

Gerstberger,S. *et al*. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet*., **15**, 829–845.

Ghosh,P. and Sowdhamini,R. (2016) Genome-wide survey of putative RNA-binding proteins encoded in the human proteome. *Mol. Biosyst*., **12**, 532–540.

Han,L.Y. *et al*. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355–368.

Kumar,M. *et al*. (2007) Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinform*., **8**, 463.

Kumar,M. *et al*. (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Kumar,M. *et al*. (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit*., **24**, 303–313.

Kwon,S.C. *et al*. (2013) The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol*., **20**, 1122+.

Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.

Liu,Z.P. *et al*. (2010) Prediction of protein–RNA binding sites by a random forest method with combined features. *Bioinformatics*, **26**, 1616–1622.

Livi,C.M. *et al*. (2016) catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. *Bioinformatics*, **32**, 773–775.

Lu,Q. *et al*. (2013) Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, **14**, 651.

Ma,X. *et al*. (2015a) Sequence-based prediction of RNA-binding proteins using random forest with minimum redundancy maximum relevance feature selection. *BioMed. Res. Int*., **2015**, 425810.

Ma,X. *et al*. (2015b) PRBP: prediction of RNA-binding proteins using a random forest algorithm combined with an RNA-binding residue predictor. *IEEE/ACM Trans. Comput. Biol. Bioinform*., **12**, 1385–1393.

Magnan,C.N. and Baldi,P. (2014) SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**, 2592–2597.

Miao,Z. and Westhof,E. (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res*., **43**, 5340–5351.

Mitchell,S.F. *et al*. (2013) Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol*., **20**, 127. U161.

Muppirala,U. *et al*. (2016) A motif-based method for predicting interfacial residues in both the rna and protein components of protein-RNA complexes. *Pac. Symp. Biocomput*., **21**, 445–455.

Muppirala,U.K. *et al*. (2011) Predicting RNA–protein interactions using only sequence information. *BMC Bioinform*., **12**, 489.

Paz,I. *et al*. (2016) BindUP: a web server for non-homology-based prediction of DNA and RNA binding proteins. *Nucleic Acids Res*., **44**, W568–W574.

Peng,C.R. *et al*. (2011) Prediction of RNA-binding proteins by voting systems. *J. Biomed. Biotechnol*., **2011**, 506205.

Peng,H. *et al*. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Machine Intell*., **27**, 1226–1238.

Shao,X. *et al*. (2009) Predicting DNA- and RNA-binding proteins from sequences with kernel methods. *J. Theor. Biol*., **258**, 289–293.

Shazman,S. and Mandel-Gutfreund,Y. (2008) Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol*., **4**, e1000146.

Shen,J. *et al*. (2007) Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A.*, **104**, 4337–4341.

Si,J.N. *et al*. (2009) TIM-Finder: a new method for identifying TIM-barrel proteins. *BMC Struct. Biol.*, **9**, 73.

Spriggs,R.V. *et al*. (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, **25**, 1492–1497.

Sun,M. *et al*. (2016) Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors. *BMC Bioinform.*, **17**, 231.

Suresh,V. *et al*. (2015) RPI-Pred: predicting ncRNA–protein interaction using sequence and structural information. *Nucleic Acids Res.*, **43**, 1370–1379.

UniProt,C. (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Walia,R.R. *et al*. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinform.*, **13**, 89.

Walia,R.R. *et al*. (2014) RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. *PLoS One*, **9**, e97725.

Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,Y. *et al*. (2013) De novo prediction of RNA–protein interactions from sequence information. *Mol. Biosyst.*, **9**, 133–142.

Wu,J.S. and Zhou,Z.H. (2013) Sequence-based prediction of microRNA-binding residues in proteins using cost-sensitive Laplacian support vector machines. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **10**, 752–759.

Yang,X.X. *et al*. (2014) RBRDetector: improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. *Proteins*, **82**, 2455–2471.

Yang,Y. *et al*. (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. *Proteins*, **80**, 2080–2088.

Yu,C.S. *et al*. (2006) Prediction of protein subcellular localization. *Proteins*, **64**, 643–651.

Yu,X. *et al*. (2006) Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.*, **240**, 175–184.

Zhao,H. *et al*. (2011a) Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. *RNA Biol.*, **8**, 988–996.

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

Zhao,H.Y. *et al*. (2014) Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. *Proteins Struct. Funct. Bioinform.*, **82**, 640–647.