Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility

Tuo Zhang[1,2,3,*], Hua Zhang[1,2,4], Ke Chen[2], Jishou Ruan[1,5], Shiyi Shen[1,5], and Lukasz Kurgan[2,*]

[1]College of Mathematical Science and LPMC, Nankai University, Tianjin, PRC
[2]Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, CANADA
[3]Indiana University School of Informatics, Indiana University-Purdue University, Indianapolis, IN, USA
[4]School of Computer Science and Information Engineering, Zhejiang Gongshang University, Hangzhou, PRC
[5]Chern Institute of Mathematics, Tianjin, PRC
*Corresponding authors; emails: for TZ freshtuo@gmail.com
for LK lkurgan@ece.ualberta.ca

Abstract

Identification and prediction of RNA-binding residues (RBRs) provides valuable insights into the mechanisms of protein-RNA interactions. We analyzed the contributions of a wide range of factors including amino acid sequence, evolutionary conservation, secondary structure and solvent accessibility, to the prediction/characterization of RBRs. Five feature sets were designed and feature selection was performed to find and investigate relevant features. We demonstrate that (1) interactions with positively charged amino acids Arg and Lys are preferred by the negatively charged nucleotides; (2) Gly provides flexibility for the RNA binding sites; (3) Glu with negatively charged side chain and several hydrophobic residues such as Leu, Val, Ala and Phe are disfavored in the RNA-binding sites; (4) coil residues, especially in long segments, are more flexible (than other secondary structures) and more likely to interact with RNA; (5) helical residues are more rigid and consequently they are less likely to bind RNA; and (6) residues partially exposed to the solvent are more likely to form RNA-binding sites. We introduce a novel sequence-based predictor of RBRs, RBRpred, which utilizes the selected features. RBRpred is comprehensively tested on three datasets with varied atom distance cutoffs by performing both five-fold cross validation and jackknife tests and achieves Matthew's correlation coefficient (MCC) of 0.51, 0.48 and 0.42, respectively. The quality is comparable to or better than that for state-of-the-art predictors that apply the distance-based cutoff definition. We show that the most important factor for RBRs prediction is evolutionary conservation, followed by the amino acid sequence, predicted secondary structure and predicted solvent accessibility. We also investigate

the impact of using native vs. predicted secondary structure and solvent accessibility. The predictions are sufficient for the RBR prediction and the knowledge of the actual solvent accessibility helps in predictions for lower distance cutoffs.

Introduction

The interactions between protein and nucleotides (DNA and RNA) control numerous cellular processes including DNA packaging, replication, transcription regulation, protein synthesis, formation of ribosomes, and catalytic activities. For example, transcription factors bind to DNA sequences and activate or inhibit the transcription of genes that have these sequences close to their promoters. The ribosome is assembled from various ribosomal RNA (rRNA) and protein molecules. The tRNA binds to specific proteins for the translation of the genetic code [1]. Some viruses have an RNA genome surrounded by capsid proteins and require the involvement of host proteins for replication [2]. Many works have investigated the mechanisms of protein-DNA interactions. They can be categorized into two groups. The first group focuses on the binding of DNA sequences in the genome and analyzes the influence of those sequences on the specificity of protein-DNA complexes [3, 4]. The other group concerns the protein sequences and it includes methods that screen proteins that potentially target specific DNA sequences [5, 6] and that locate the binding sites on those proteins [7, 8].

In contrast to the progress in the analysis of protein-DNA interactions, the interactions between proteins and RNA are less well understood. This is primarily because the RNA structures vary more than the DNA structures, resulting in a wider range of mechanisms that implement the protein-RNA interactions, and also due to better availability of the structural information concerning the DNA-protein information [9, 10].

The 3D structures of protein-RNA complexes provide valuable insights into the interaction between proteins and RNA. However, only 684 protein-RNA complexes [11] could be found in the Protein Data Bank (PDB) [12] as of June 2008. This is due to the costly and time consuming experimental determination of the structure of the complex. The above and the widening protein structure-sequence gap motivate the development of computational methods for prediction of RNA-binding residues (RBRs) from the amino acid (AA) sequence. Such methods not only provide the means to annotate protein sequences with unknown structure, but they also help with understanding the mechanisms of the protein-RNA interaction.

Although the investigations into the physical and chemical properties of the protein-RNA interactions have a relatively long history [10, 13-16], the approaches that address prediction of RBRs have surfaced relatively recently. In 2004, Jeong *et al.* built the first RNA-binding predictor using neural network with a single sequence and

predicted secondary structure as the input [17]. This method was improved by Jeong and Miyano by adding weighted profiles [18]. Terribilini *et al.* developed a Naïve Bayes based tool [19] that was used later to develop a web server for identifying binding residues for known protein-RNA complexes and for predicting RNA-binding residues from the sequence for which RNA-bound structure is not available in the PDB [20]. Wang and Brown designed a Support Vector Machine (SVM)-based web tool, BindN, for prediction of the DNA and RNA binding residues using three simple sequence-derived features including the side chain pKa value, hydrophobicity index and molecular mass of an AA [7]. Kim *et al.* introduced residue singlet/doublet interface propensities and used them together with position-specific multiple sequence profiles (PSSM) to propose a structure-based prediction method [21]. In 2008, two methods for the sequence-based prediction of the RNA-protein interacting residues based on the SVM classifier and PSSM were built [22, 23]. Another similar method, RNAProB, which combines a new "smoothed" PSSM with the SVM classifier was proposed in the same year [11]. Around the same time Chen and Lim designed a structure-based prediction method [9]. Most recently, Spriggs *et al.* introduced another sequence-based SVM-based predictor called PiRaNhA that uses PSSM and three amino acid properties as its input [24]. Table 1 summarizes the sequence-based RBRs prediction methods. We note that different methods use definitions of RBRs that can be divided into three categories. In atom distance-based definitions the residues are identified as interacting with RNA if the closest distance between atoms of that residue and the partner RNA is smaller than a cutoff value. In the second group, which includes the structure-based method by Kim *et al.* [21], residues are defined as RNA binding based on a comparison of the solvent accessible surface area of the protein structure with and without RNA. The third approach defines the interacting residues using hydrogen bonding, stacking, electrostatic, hydrophobic and van der Waals interactions, which are found with HBPLUS [25] or ENTANGLE [26]. We concentrate on the atom distance-based definition since most of the existing methods use this definition and since it is also commonly used to define protein-DNA [27] and protein-protein [28] interactions. We note that although Terribilini *et al.* first used the ENTANGLE to define RBRs [19], they later adopted the atom distance-based definition when building their online server [20].

Although several methods have been developed, the sequence-based prediction of RBRs is still a challenging and open problem. For instance, the most recent PiRaNhA method achieves the Matthews Correlation Coefficient (MCC) between 0.4 and 0.5, depending on the dataset used [24]. Prior works utilize different sequence derived information including amino acid sequence, evolutionary conservation, and predicted secondary structure and relative solvent accessibility. Besides the amino acid sequence used by all methods, the first predictor by Jeong *et al.* [17] considered the predicted secondary structure. The works by Jeong and Miyano [18], Kumar *et al.* [23], and Cheng *et al.* [11] used evolutionary conservation, while Wang *et al.* [22] used evolutionary conservation coupled with the predicted secondary structure. Terribilini's work [19, 20] as well as the method by Wang and Brown [7] were based

solely on the sequence. The most recent work by Sprigg *et al.* [24] applied the evolutionary conservation and the predicted solvent accessibility. We emphasize that each of the previous methods focused only on a subset of the above information and none of the studies investigated whether fusing all of these sources could be beneficial.

Table 1. Existing methods for the sequence-based prediction of RNA-binding residues (RBRs). The methods are sorted by the publication date. The abbreviations used in the "classifier" and "inputs" columns include Neural Network (NN), Naïve Bayes (NB), Support Vector Machine (SVM), predicted secondary structure (PSS), evolutionary conservation (EC), and predicted relative solvent accessibility (PRSA).

Reference	Definition of RBRs	Classifier	Inputs
Jeong *et al.* (2004) [17]	Atom distance (6.0Å)	NN	Sequence, PSS
Jeong & Miyano (2006) [18]	Atom distance (6.0Å)	NN	Sequence, EC
Terribilini *et al.* (2006) [19]	ENTANGLE[1]	NB	Sequence
Wang & Brown (2006) [7]	Atom distance (3.5Å)	SVM	Sequence
Terribilini *et al.* (2007) [20]	Atom distance (5.0Å)	NB	Sequence
Wang *et al*. (2008) [22]	ENTANGLE[1]	SVM	Sequence, EC, PSS
Kumar *et al.* (2008) [23]	Atom distance (3.5Å & 6.0Å)	SVM	Sequence, EC
Cheng *et al.* (2008) [11]	Atom distance (3.5Å & 6.0Å)	SVM	Sequence, EC
Sprigg *et al.* (2009) [24]	HBPLUS[2]	SVM	Sequence, EC, PRSA

[1] RBRs are defined by ENTANGLE program [26], which searches for hydrogen bonding, stacking, electrostatic, hydrophobic and van der Waals interactions.

[2] A residue is defined as RNA-binding if any of its non-hydrogen atoms are within vdW contact or hydrogen bonding distance, which are computed by HBPLUS program [25], to any RNA non-hydrogen atom directly or indirectly via a bridging water molecule.

We propose a sequence-based model for the prediction of RBRs (RBRpred) that aims at providing high prediction quality and we also investigate factors associated with the prediction of RNA binding residues. We implemented five feature sets based on the sequence, evolutionary conservation, predicted secondary structure, predicted relative solvent accessibility, and combination of the predicted secondary structure and solvent accessibility. These features were processed by using feature selection and fed into SVM classifier to build RBRpred. The proposed predictor was compared against existing sequence-based methods that consider atom distance-based definition and showed comparable or better performance. Our analysis of the contribution of different features and feature sets, as well as the impact of native secondary structure and solvent accessibility, reveals new and confirms existing factors related to prediction/characterization of the RNA binding residues.

Materials and Methods

Datasets

We prepared four datasets to design and test the proposed method. Three of them, including RB86, RB147 and RB106, were derived from previous studies and were used to perform comparative analysis. The remaining dataset, named by RB48, was constructed to perform an additional, independent test of the proposed method. Table 2 summarizes the four datasets.

Table 2. Summary of four benchmark datasets.

Dataset	Refs	Number of RBRs	Number of non-RBRs	Ratio = RBRs / non-RBRs
RB86	[11, 18, 23]	4568	15503	1:4
RB147	[20]	6157	26167	1:4
RB106	[11, 7, 23]	2555	19496	1:7
RB48	this paper	2262	3926	1:2

The RB86 dataset consists of 86 RNA-interacting protein chains extracted from protein-RNA complexes. Sequence redundancy in this dataset was reduced such that no two chains have identity of above 70%. A cutoff of 6Å was used to define RBRs, i.e., a residue was defined as the RBR if the closest distance between atoms of the interacting RNA molecule and that residue was below 6Å. The number of RBRs and non-RBRs were 4568 and 15503, respectively. The RB86 dataset had been used in Jeong *et al.* [18], Kumar *et al.* [23] and Cheng *et al.* [11]. We selected the RB86 dataset to perform feature selection and parameterization, and to build our predictor. This was motivated by Kumar [23] which also chose this dataset. The RB147 dataset was extracted using PISCES by Terribilini [20]. This dataset contains 147 protein chains with pairwise sequence identity below 30%, and a total of 32324 residues (6157 RBRs and 26167 non-RBRs). The RBRs were defined based on the cutoff distance of 5.0 Å. The RB106 dataset includes 106 protein chains with the pairwise sequence identity below 25%. The RBRs were identified based on the threshold of 3.5 Å, which results in 2555 RBRs and 19496 non-RBRs. This set was used by Wang and Brown [7], Kumar *et al.* [23], and Cheng *et al.* [11].

We designed a new dataset, RB48, to perform an independent test of the proposed predictor. In contrast to the above three sets for which cross-validation tests were performed, we trained our model on the RB86 dataset and tested it on the RB48 set. First, we extracted all 565 protein-RNA complexes from PDB [12]. Second, we retained 41 complexes that were released after 2008. This was because some of the features were based on the secondary structure predicted with PSIPRED [29] and the relative solvent accessibility predicted with Real-SPINE [30], and these two predictors were published before 2008. This should remove a potential bias in

predictions from PSIPRED and Real-SPINE, since the above protein chains would not be included in the training sets of these methods. Next, we run blastclust [31] on the protein chains from the 41 protein-RNA complexes to reduce the sequence identity within this set and between this set and our training dataset RB86. The protein in each complex may have multiple chains and we consider all of these chains that interact with RNA. We found total of 48 chains that had pairwise sequence identity lower than 25% within the set and when compared with chains from the RB86 dataset. These chains constitute the RB48 dataset. We use the cutoff distance of 6 Å (the same as in the RB86 dataset) to identify RBRs, which results in 2262 RBRs and 3926 non-RBRs, respectively.

Evaluation of prediction performance

The performance of the proposed method is evaluated based on *n* fold cross validation (FCV) performed on the RB86, RB147 and RB106 datasets. The protein chains were randomly divided into *n* subsets, at each time using *n*-1 subsets to train the model then testing on the remaining subset. This process was repeated *n* times so that each fold was used once as the test set. The 5 FCV and jackknife test (*n* equals the number of sequences in dataset) were performed. Furthermore, an independent test was also performed by training the prediction model on the RB86 dataset and testing on the RB48 dataset. The above three evaluation tests are often used to examine effectiveness of predictors [32]. One of the desirable aspects of the jackknife test, in contrast to cross-validation tests with a lower number of folds, is that it always yields a unique result for a given benchmark dataset [33, 34], although at the same time it requires more computations. This test type has been increasingly used and widely recognized by investigators to examine the predictive quality (see, e.g. [35-47]), which motivates its application in this contribution.

We adopted five quality indices to validate the proposed method:
Sensitivity = TP/ (TP+FN)
Specificity = TN/ (TN+FP)
Precision = TP/ (TP+FP)
Accuracy = (TP+TN)/ (TP+TN+FP+FN)

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

where TP, FP, TN and FN denote true positives (correctly predicted RBRs), false positives (non-RBRs that are incorrectly predicted as RBRs), true negatives (correctly predicted non-RBRs) and false negatives (RBRs that are incorrectly predicted as non-RBRs), respectively. MCC ranges between -1 and 1 where 1 represents a perfect prediction, 0 a random prediction, and -1 a case where all predictions are incorrect. The remaining quality indices range between 0 and 1, where higher value of the index corresponds to better prediction. We selected MCC as the main measure to perform design of the proposed method (including feature selection and classifier parameterization) as well as to compare with the existing methods. This is motivated

by the fact that virtually all modern sequence-based RBR predictors are evaluated using MCC [7, 11, 19, 20, 22-24]. In addition to the above measures, we also use the receiver operating characteristic (ROC) curve [48] and area under the ROC curve (AUC) [49] to evaluate the performance of the proposed method. The output from the SVM classifier was thresholded to plot the ROC curve.

Feature vector

Twelve types of features were utilized to design the proposed method. These features were derived from four sources including the AA sequence (sequence-based features), the evolutionary conservation represented by PSI-BLAST profile (PSSM-based features), the predicted secondary structure (SS-based features), and the predicted relative solvent accessibility (RSA-based features). We also combined the latter two sources to derive SS&RSA-based features.

Sequence-based features

Weiss and Narayana have shown that Arginine-rich motifs are abundant in RNA binding sites [50]. Other strong biases for different types of AAs present at the RNA-protein interfaces have also been reported in prior studies [16, 17, 19, 23, 51, 52]. The above motivates the inclusion of the sequence-based features as inputs for prediction of RBRs. We used binary encoding, i.e., 20-dimensional binary vector, to represent the AA type of a given residue (*RT* features). Considering the tendency of protein-RNA interface residues to be clustered along the primary protein sequences [19, 20], a sliding window of size 15, which includes 7 neighboring residues on both sides of the predicted residue, was used. The selection of window size was motivated by Kumar *et al.* [23] and Wang *et al.* [22] who used the same size. Zero vectors were used to fill in blanks for residues at the sequence termini. A total of 300 features, which corresponds to 15 20-dimentional vectors, were computed for each input residue.

PSSM-based features

Functional residues are usually more conserved when compared with non-functional residues, and evolutionary information is often used to locate the functional sites [53]. Previous studies demonstrate that evolutionary information provides an effective source of information for the prediction of RBRs [11, 18, 21, 23]. The evolutionary information quantified via PSSM has also been used to predict numerous other protein features, such as membrane protein types [54], enzyme functional classes [55], to functionally discriminate membrane proteins [56], and to predict protease types [57], protein fold types [58], protein quaternary structural attribute [59], as well as protein subcellular localizations [60], human protein subcellular localizations [61] and Gram-positive bacterial protein subcellular localizations [62]. Similarly as in the prior works, we used PSI-BLAST [31] to perform multiple alignment of the input sequence with the E-value equal 10^{-3} and three-iterations against the NCBI's non-redundant protein sequence database (NR database). The PSI-BLAST's output includes a 20-dimensional PSSM (position-specific scoring matrix) and a 20-dimensional WOP

(weighted observed percentage) vector for each residue of the input sequence. We note that the WOP vector was not used before in the prediction of the RBRs. Three types of PSSM-based features were extracted, namely *PSSM*, *EntWOP* and *CNCC* (close neighbor correlation coefficient). A window size of 15 was used, as motivated above.

Similarly as in [63], *PSSM* and *EntWOP* features were obtained from the PSSM and WOP vectors and were computed using logistic function $f_{ij}(a_{ij})=1/(1+\exp(-a_{ij}))$ and entropy estimate $EntWOP_i = \sum_j - p_{ij} \log_{20} p_{ij}$, respectively, where $p_{ij} = n_{ij} / \sum_j n_{ij}$, a_{ij} and n_{ii} correspond to the j^{th} values of the PSSM and WOP vector for the i^{th} residue in the sequence.

There are 15×20 = 300 *PSSM* and 15 *EntWOP* features for each predicted residue. In Kim's work [21], structure-derived information concerning adjacent residues was used to predict RBRs, which resulted in an improved prediction quality. However, since the proposed method is based solely on the sequence, we designed *CNCC* features to approximate the features used by Kim *et al*. Given that the PSSM vector for i^{th} residue is denoted as (a_{i1}, a_{i2},…, a_{i20}), the *CNCC* features form a 14-dimensional vector ($C_{i\,i-7}$, $C_{i\,i-6}$,…, $C_{i\,i-1}$, $C_{i\,i+1}$, $C_{i\,i+2}$, $C_{i\,i+7}$), where

$$C_{ij} = \frac{\sum_{k=1}^{20}(a_{ik} - \bar{a}_i)(a_{jk} - \bar{a}_j)}{\sqrt{\sum_{k=1}^{20}(a_{ik} - \bar{a}_i)^2 \sum_{k=1}^{20}(a_{jk} - \bar{a}_j)^2}}$$, and $\bar{a}_i$ and $\bar{a}_j$ are the average values.

The C_{ij} values correspond to the Pearson correlation coefficients between the PSSM vector of the predicted residue and that of the adjacent residues in the sliding window. These features were originally proposed by Cheng and Baldi for the prediction of protein contact maps [64].

SS-based features

Knowledge of the secondary structure has been shown to be helpful in understanding of protein folding [65, 66], and in prediction of protein structure [67, 68], function [69], protein-protein interactions [70], and the RNA-binding interactions [17]. In one of the earlier studies, Draper found two canonical protein-RNA contact types at the secondary structure level, (1) binding between α-helix or loop and a groove of the RNA pockets; and (2) binding between β-sheet surface and unpaired RNA bases [14]. Several previous studies also used predicted secondary structure in prediction of RBRs [17][22]. We estimate the RNA binding propensity of the three secondary structures by

$$propensity_{ss} = \log_2\left(\frac{percentage\ of\ RBRs\ whose\ secondary\ structure\ is\ ss}{percentage\ of\ all\ residues\ whose\ secondary\ is\ ss}\right)$$

where ss = {H (helix), E (strand), C (coil)}. The propensity quantifies the degree to which a certain type of the secondary structure is preferred in the RNA binding sites. The values greater than zero show that the occurrence of a given secondary structure

in the RNA binding sites is higher than that in whole sequence, otherwise the occurrence in the RNA binding sites is either lower or the same, if equal to zero. We computed the propensities on the RB86, RB147 and RB106 datasets, see Figure 1, where DSSP program [71] was used to assign native (actual) secondary structures. We observe that residues in coil conformation are more likely to bind RNA, while helix and strand residues are less likely to form RNA binding sites. The propensities vary with the datasets, which is likely influenced by the different cutoff thresholds. More specifically, the propensity of the strand residues decreases as the cutoff values decrease, i.e., fewer residues which are closer to RNA are considered as RBRs.

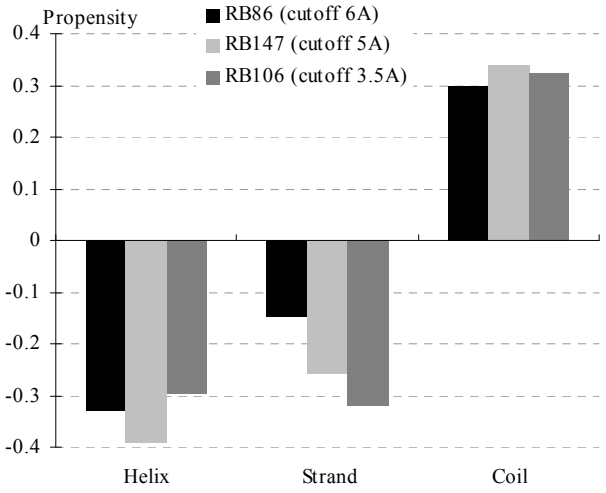


Figure 1. RNA-binding propensity of the three secondary structures on the RB86, RB147, and RB106 datasets, where the RBRs were defined based on the 6Å, 5Å, and 3.5Å thresholds, respectively. The propensity is defined as the percentage of RNA-binding residues with certain secondary structure type among all binding residues divided by the percentage of all residues with the same secondary structure type among all residues.

Our analysis demonstrates that the secondary structure should be useful in the context of the RBRs prediction. We use the secondary structure predictions provided by PSIPRED [29], since this method (1) is well-known, widely used, and has been shown to provide superior accuracy when compared with other modern secondary structure prediction methods [72]; (2) was frequently used in other related prediction methods that concern solvent accessibility [73], protein folds [74], residue depth [75], beta-turns [76], and alpha-turns [77], to name a few; and (3) was used in sequence-based prediction of RBRs [22]. The PSIPRED's output includes the secondary structure state for each residue together with the corresponding probability. The 3-state accuracy (Q_3) of PSIPRED predictions on the RB86, RB147 and RB106 datasets equal 79.9%, 78.2% and 79.2%, respectively. This suggests that PSIPRED did not overfit these datasets since its originally reported accuracy is about 78% [78].

The existing sequence-based predictors of RBRs use only the three-state prediction of

the secondary structure of the predicted residue [17, 22]. In contrast, we designed five types of features based on the outputs from the PSIPRED, including *SSProb*, *SSCont*, *TriSS*, *SegLen* and *SegDB*. These features are based on the sliding window of size 15 and they reflect local secondary structure information. *SSProb* features are composed of 15×3 = 45 features that correspond to probabilities of 15 neighboring residues, where the secondary structure of each residue is represented by a 3-dimensional probability vector, i.e., probability of coil, strand and helix prediction. *SSCont* features encode the three secondary structure contents in the sliding window, i.e., the fraction of residues in a given secondary structure among the residues in the window. We also use a secondary structure triplet to record the secondary structure of target residue and one adjacent residue on both sides. This triplet has 3^3=27 combinations, thus it was encoded by a 27-dimensional binary vector (*TriSS* features). The remaining two types of features, *SegLen* and *SegDB*, concern a secondary structure segment that includes the predicted residue. The segment was defined as a consecutive sequence of residues that were predicted in the same secondary structure state. *SegLen* features are encoded with a 3-dimensional vector which corresponds to the three secondary structure states. If the segment was not composed by a given secondary structure type, then the corresponding dimension is set to 0, otherwise it records the length of the segment which is normalized by the window size of 15. *SegDB* features depict the relative position of predicted residue in the segment, i.e., the distance between the target residue and the termini of the segment. Similarly as *SegLen*, we used two 3-dimensional vectors (6 features) to denote the minimum/maximum distance, respectively. The distance was normalized by half size of the sliding window, which equals 7.

RSA-based features

Solvent accessible surface area (ASA) delineates the surface area of a residue that is accessible to a solvent, and has been widely studied due to the fact that surface residues are directly involved in the interaction with other biological molecules [79-81]. The ASA was widely used in the context of protein structure [67, 68], function [69], stability [82], folding [83, 84], flexibility [85] and fold recognition [86, 87]. Ahmad *et al.* [88] demonstrated the importance of the role of the solvent accessibility of AAs in determining the protein-DNA binding. We investigate whether ASA helps in prediction of the protein-RNA interaction. Given the bias in the ASA values of different AA types, i.e., AAs with larger size potentially have larger ASA, relative solvent accessibility (RSA), which is defined by the ASA of a residue in the protein divided by ASA observed in an extended conformation (Ala-X-Ala) [88], was used.

Similarly as in the case of the secondary structures, we analyze the propensity of residues with different RSA values to form RNA-binding sites. DSSP [71] was used to compute the native ASA values, which were normalized by the area of the extended conformation to find the RSA values. The RNA binding propensity for residues with RSA values binned into 11 intervals is defined as follow:

$$propensity_r = \log_2\left(\frac{percentage\ of\ RBRs\ whose\ RSA\ is\ in\ range\ r}{percentage\ of\ all\ residues\ whose\ RSA\ is\ in\ range\ r}\right)$$

where $r \in \{[0, 0], (0, 0.1], (0.1, 0.2], \ldots, (0.9, 1.0]\}$, i.e. residues were divided into ten equal-sized bins of non-zero RSA values and one bin that includes residues with RSA values of zero. Figure 2 shows the RNA binding propensity on the RB86, RB147 and RB106 datasets. As expected, the residues with large RSA (exposed residues) appear more frequently in the RNA binding regions, while those with small RSA (buried residues) have lower chance to bind RNA. The magnitude of the propensity is larger for datasets with a smaller cutoff. This means that RSA would provide better discrimination between the RNA binding and other residues for smaller distance cutoffs.

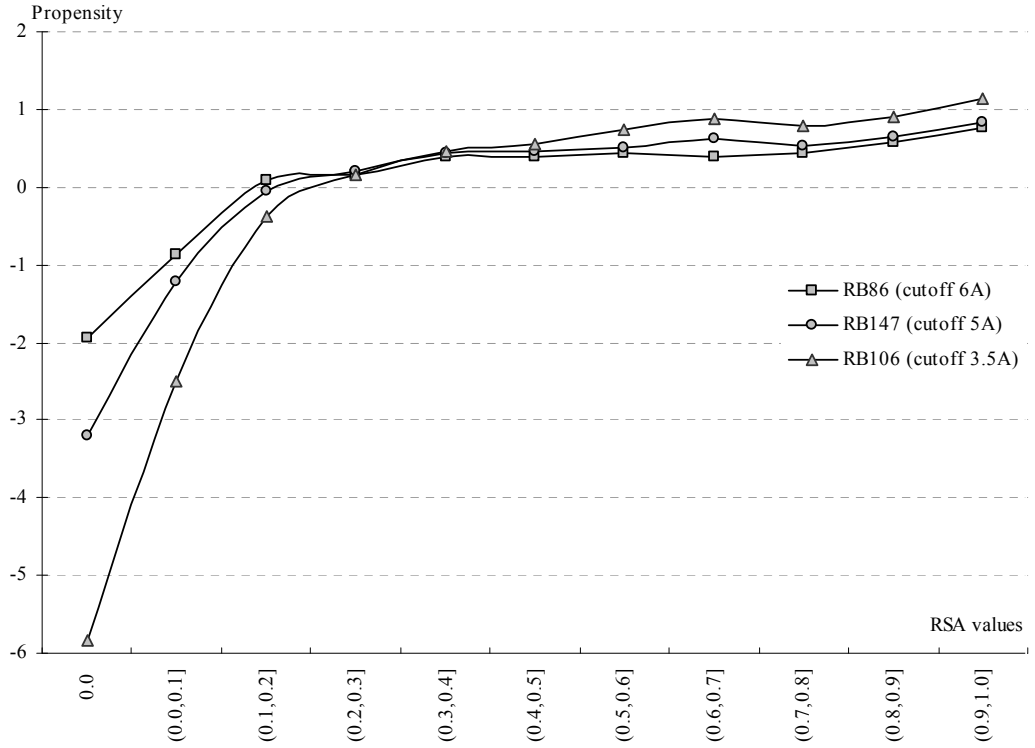


Figure 2. RNA-binding propensity for residues with different RSA values on the RB86, RB147, and RB106 datasets, where the RBRs were defined based on the 6Å, 5Å, and 3.5Å cutoffs, respectively. RNA-binding propensity is defined as the percentage of RNA-binding residues with RSA in a given range among all binding residues divided by the percentage of all residues with RSA in the same range among all residues.

Several methods for the prediction of RSA are available [30, 73, 88-92]. We performed predictions using one of the most recent methods, Real-SPINE [30]. This is motivated by the high quality of its predictions, i.e., a correlation of 0.74 was reported [30]. The predictions with Real-SPINE on the RB86, RB147 and RB106 datasets yield correlations of 0.71, 0.69 and 0.70, respectively. Since these results are

consistent with the originally reported values, we assume that this method did not overfit the three datasets. Only one existing sequence-based predictor of RBRs uses the predicted solvent accessibility [24]. Two types of features, *RSA* and *AveRSA*, were designed based on the predicted RSA. Similarly to [24] where a window is use, a 15-dimensional vector of *RSA* features records the RSA values of the residues in a sliding window of size 15, i.e., each dimension corresponded to one position in the window. We also introduced 7 *AveRSA* features by computing the average of the RSA values in the sliding window by varying window sizes between 3 and 15. The latter features reflect the solvent exposure in a local environment.

Table 3. Summary of the features, divided into five sets and twelve types, before and after feature selection. The last column identifies feature types that were not used before in the sequence-based prediction of the RBRs.

Feature set	Feature type	# features before feature selection	# selected features	Features never used before in the sequence-based RBRs prediction
Sequence-based	*RT*	300	33	
PSSM-based	*PSSM*	300	260	
	EntWOP	15	15	√
	CNCC	14	13	√
SS-based	*SSProb*	45	45	√
	SSCont	3	2	√
	TriSS	27	2	√
	SegLen	3	2	√
	SegDB	6	4	√
RSA-based	*RSA*	15	15	
	AveRSA	7	7	√
SS&RSA-based	*SR*	54	22	√
Total number		789	420	

SS&RSA-based features

These features combine the information concerning predicted secondary structure and the predicted RSA for the predicted residues. We binarized the RSA values to categorize the residue as either exposed (RSA higher than a cutoff value) or buried (RSA lower than a cutoff value). We used 9 cutoff values, 0.1, 0.2, 0.3, …, 0.9. As a result, we defined a 54-dimensional (9×2×3=54) binary vector, called *SR*, in which we encode the possible combinations of 3 secondary structures and 2 exposure states with 9 cutoffs. The vector includes all zeros except 9 positions that correspond to the predicted secondary structure and exposure states for the nine thresholds.

Overall, we produced 789 features, see Table 3. Those features were divided into five feature sets: sequence-based feature set (*RT*), PSSM-based feature set (*PSSM*, *EntWOP*, and *CNCC*), SS-based feature set (*SSProb*, *SSCont*, *TriSS*, *SegLen*, and *SegDB*), RSA-based feature set (*RSA* and *AveRSA*) and SS&RSA-based feature set

(*SR*).

Prediction method

Support vector machine (SVM)

The motivation behind the choice of the SVM comes from wide-spread applications of SVM in various bioinformatics problems, such as prediction of secondary structure [93, 94], catalytic residues [63], subcellular localization [95, 96], protein-protein interaction site [97], and the successful application in the existing method for RBR predictions [7, 11, 22-24]. We note that four most recent sequence-based predictors of the RNA-binding residues are based on the SVM classifier, see Table 1. SVM is a linear large-margin classifier which can be extended to non-linear classification with the use of a kernel function [98]. We used SVMlight [99] to develop and test the proposed method. The SVMlight is a well-known SVM package that has been used in previous RBR predictions [7, 22, 23]. Radial basis function (RBF) is chosen as the kernel function due to its competitive performance for solving nonlinear problems when compared with other kernel functions [63, 75] and since this kernel was also selected in the most recent sequence-based RBRs predictors [11, 22, 24].

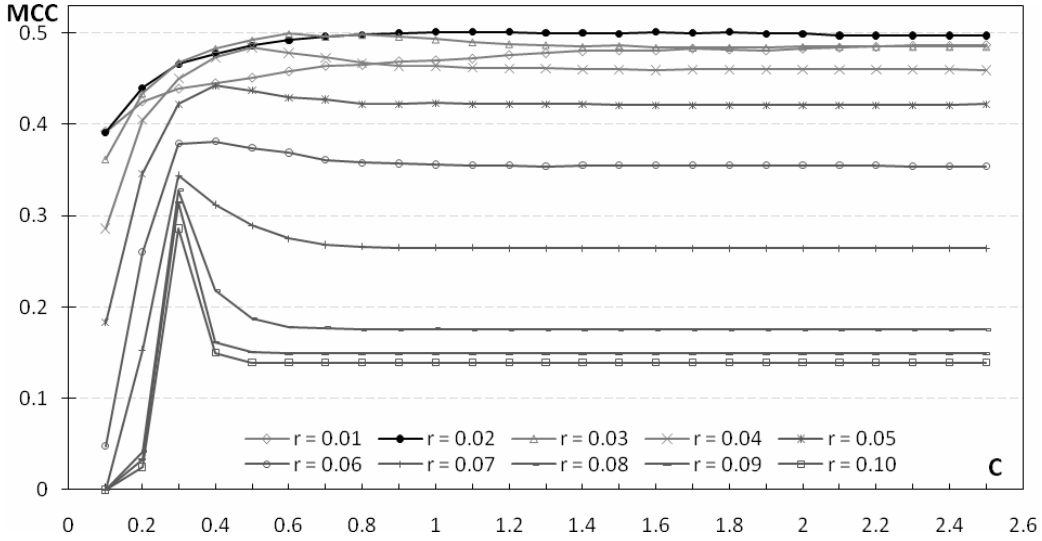


Figure 3. First-stage parameterization. The MCC values (y-axis) of RBR predictors built by choosing soft-margin constant *C* (x-axis) and RBF kernel width γ (each curve corresponds to a different γ value).

The considered SVM classifier has two parameters, soft-margin constant *C* and RBF kernel width γ. A two-stage parameterization was performed. First, we used the entire set of 789 features to perform a grid search over *C* and γ values based on 5 FCV on the RB86 dataset. After several trials, the *C* and γ values were constrained to the Cartesian product of {0.1, 0.2, 0.3, …, 2.5}×{0.01, 0.02, 0.03, …, 0.10}. Figure 3 shows the MCC values of the predictors using different pairs of parameters. Choosing

a smaller γ yields better MCC, irrespectively of the value of *C*. The best MCC = 0.501 was obtained for *C* = 1.0 and γ = 0.02. The same MCC value was also observed for γ = 0.02 and *C* = 1.1/1.2/1.6/1.8. Here we chose the smallest *C* to save computational time. These parameters were used to perform feature selection (see the "Feature selection" section), after which the parameterization was repeated using the selected subset of features. We again performed 5 FCV on the RB86 dataset considering the grid search over the Cartesian product of *C* = {0.8, 0.9, 1.0, 1.1, 1.2} and γ = {0.005, 0.010, 0.015, …, 0.040}. We narrowed the range of the parameter values and used a finer grid step that was centered on the optimal parameters obtained in the first round of the parameterization, see Figure 4. The selection of γ in range 0.025 to 0.040 results in larger MCC values. The best MCC = 0.513 was obtained for *C* = 1.1 and γ = 0.025, and these parameters were used through all empirical tests.

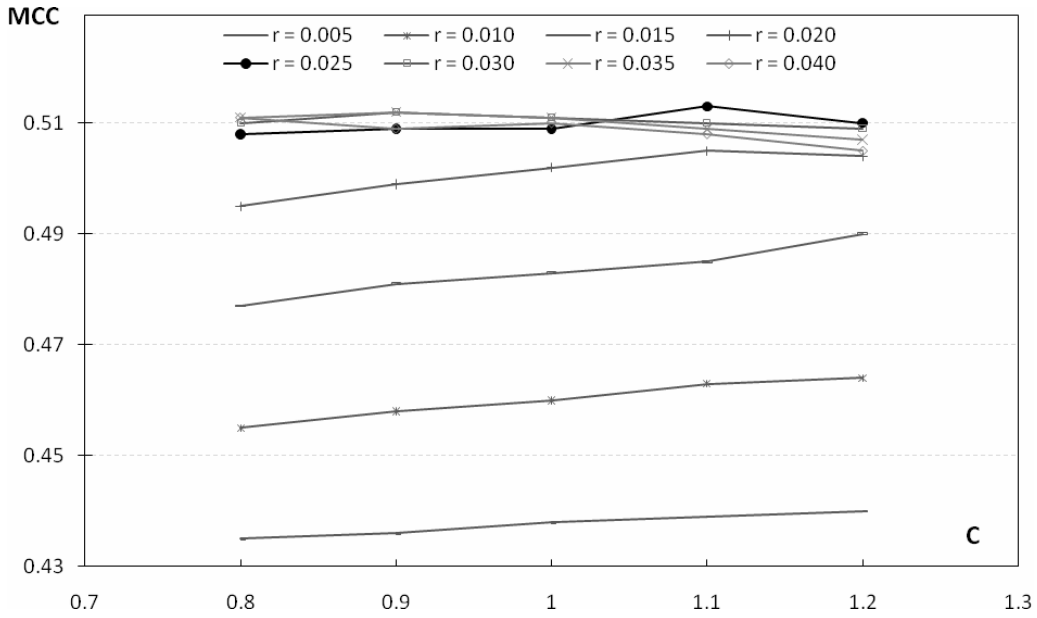


Figure 4. Second-stage parameterization. The MCC values (y-axis) of RBR predictors built by choosing soft-margin constant *C* (x-axis) and RBF kernel width γ (each curve corresponds to a different γ value).

Feature selection

First, the 789 features were ranked based on the χ^2-statistic [100] between their values and the class labels (annotation of RBRs). The motivation to use the χ^2-statistic comes from the observation that such ranking results in selection of features that provide improved precision in the subsequent classification [101]. We note that precision, which is defined as the success rate among all predicted RBRs, is one of the key indicators of the quality of the prediction of RBRs. We also note that the same feature selection was successfully used in designing a relevant, recent, to date top-performing, sequence-based method for the prediction of catalytic residues [63]. The χ^2-scores were computed using 5 FCV on the RB86 dataset, i.e., average over the 5 training folds was computed to avoid overfitting. Features were ranked according to the

descending values of the average χ^2-score. We started with the top 10 ranked feature and we added 10 features at a time according to their rank. We used these features to generate the SVM predictor (with C = 1.0 and RBF kernel with γ = 0.02) based on 5 FCV on the RB86 dataset. Figure 5 shows the MCC values of these SVMs for different numbers of the input features n. The MCC values initially increase as n grows larger, and at n = 420 the MCC values saturate and we observe only small fluctuations. As a result, we selected the top 420 features to build the proposed prediction model. The breakdown of the selected features is shown in Table 3. We note that a substantial computational cost of building SVM classifiers (given the large size of the problem that includes 5 folds with 16,000 samples and n attributes in each fold to run for each of the 79 feature set evaluations) forced us to use the above relatively simple feature selection procedure.

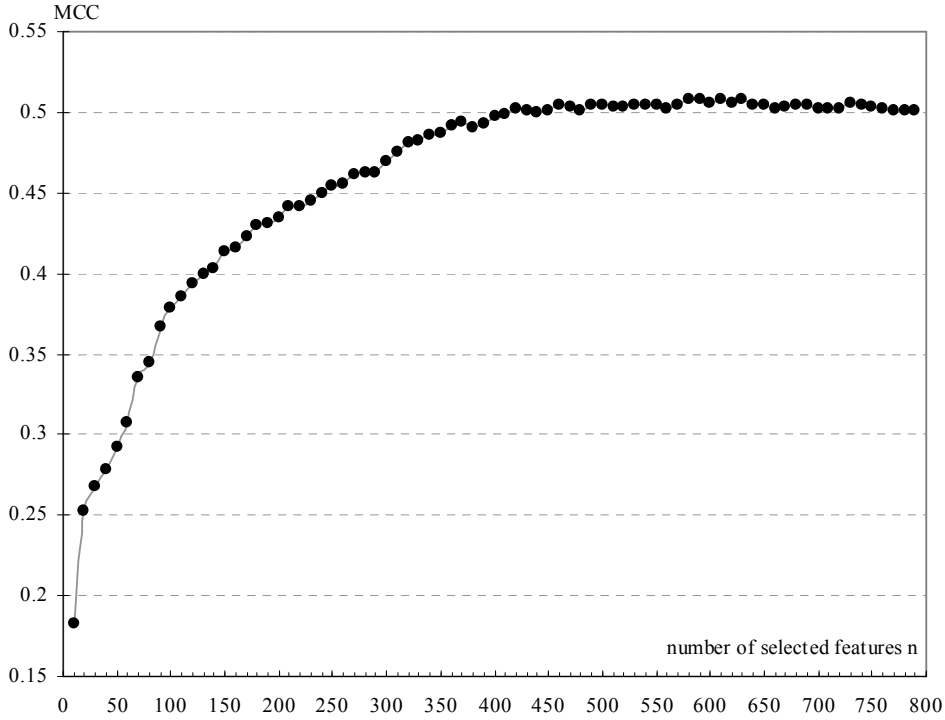


Figure 5. The MCC values (*y*-axis) of the SVM-based RBRs predictors built using *n* (*x*-axis) top ranked features. The features were ranked according to their χ^2 -score.

Overview of the prediction system

Figure 6 shows the architecture of the proposed system. Twelve sets of features are extracted based on sequence, PSI-BLAST profiles from the PSI-BLAST [31], predicted secondary structure from the PSIPRED [29] and predicted relative solvent accessibility from the Real-SPINE [30]. A total of 420 features, which were selected among 789 considered features, are fed into the SVM classifier to predict RBRs.

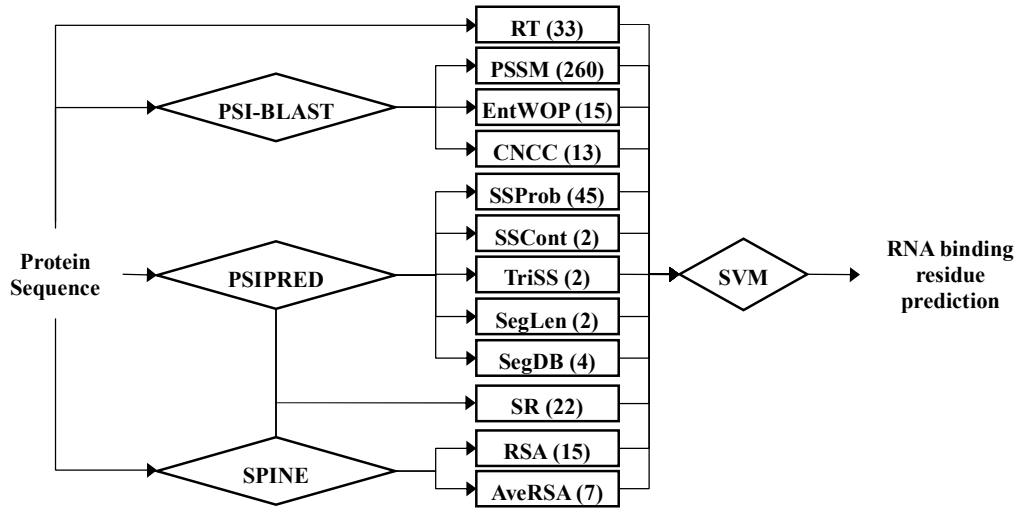


Figure 6. Block diagram of the proposed prediction system. The numbers of selected features are shown in parentheses.

Results and Discussion

Comparison with existing sequence-based RBR prediction methods

The existing sequence-based predictors use different definitions of RBRs. As discussed above, the proposed method, RBRpred, applies the atom distance-based definition and is compared against all other modern predictors that use this definition. The competing methods include the neural network based method (ANN) and its improved version (ANN_WP) by Jeong *et al.* [17, 18], three SVM based methods including BindN by Wang and Brown [7], Pprint by Kumar *et al.* [23], and RNAProB by Cheng *et al.* [11], and the Naïve Bayes based method, RNABindR, by Terribilini *et al.* [20]. The six existing methods were evaluated using different test procedures. ANN and ANN_WP were tested with the 10 FCV; BindN, Pprint and RNAProB were evaluated based on the 5 FCV; and authors of RNABindR applied the jackknife test. We performed both 5 FCV and jackknife tests on the RB86, RB147 and RB106 datasets, see Table 4. We did not run the 10 FCV for two reasons. First, the 10 FCV was only used to evaluate the ANN and ANN_WP methods and Kumar *et al.* demonstrated that the Pprint method did better than the two neural network-based methods [23]. Thus, comparison with Pprint under the 5 FCV setting should also be indicative of the comparison with the ANN and ANN_WP methods. Second, since the results of 5 FCV and jackknife tests are quite similar on all three datasets, i.e., the MCC values vary by between 0 and 0.02 and the accuracies differ by between 0.09 and 0.53 when comparing the 5 FCV and jackknife tests for the RBRpred, we anticipate that the 10 FCV would likely yield similar results.

Table 4. Comparison between the proposed RBRpred and the six competing sequence-based RBR prediction methods on three datasets.

Dataset	Method	Ref.	Test type	Sensitivity	Specificity	Precision	Accuracy	MCC
RB86	ANN	[17]	10 FCV	43.40	91.04	58.80	80.20	0.39
	ANN_WP	[18]	10 FCV	NR[3]	NR[3]	NR[3]	NR[3]	0.41
	Pprint	[23]	5 FCV	53.05	89.55	59.93	81.16	0.45
	RNAProB[1]	[11]	5 FCV	79.95	90.36	70.96	87.99	0.68
	RNAProB_std[2]	[11]	5 FCV	NR[3]	NR[3]	NR[3]	83.39	0.50
	RBRpred	this paper	5 FCV	60.88	89.67	63.46	83.12	0.51
	RBRpred	this paper	Jackknife	61.14	89.27	62.68	82.87	0.51
RB147	RNABindR	[20]	Jackknife	33.00	95.00	61.00	83.19	0.36
	RBRpred	this paper	5 FCV	52.90	91.02	58.09	83.76	0.46
	RBRpred	this paper	Jackknife	54.88	91.21	59.50	84.29	0.48
RB106	BindN	[7]	5 FCV	66.28	69.84	27.76	69.32	0.27
	Pprint	[23]	5 FCV	70.09	75.54	27.30	75.43	0.32
	RNAProB[1]	[11]	5 FCV	77.14	80.87	54.30	80.44	0.42
	RNAProB_std[2]	[11]	5 FCV	NR[3]	NR[3]	NR[3]	77.80	0.36
	RBRpred	this paper	5 FCV	39.92	95.68	54.78	89.22	0.41
	RBRpred	this paper	Jackknife	41.41	95.58	55.13	89.31	0.42

[1]SVM-based method using smoothed PSSM

[2]SVM-based method using standard PSSM

[3]The result was not reported and cannot be duplicated

RBRpred provides a relatively high MCC, i.e., 0.51, 0.48 and 0.42 on the RB86, RB147 and RB106 datasets, respectively. When compared with ANN and ANN_WP, the MCC is higher by 0.12 and 0.10. Similar improvements of 0.15 and 0.12 are observed when comparing against BindN and RNABindR, respectively. A bit smaller, although still relatively substantial improvements of 0.06 and 0.10 are obtained against Pprint on the RB86 and RB106 datasets, respectively. Although the RNAProB outperforms RBRpred on the RB86 dataset, the two methods provide similar quality on the RB106 dataset. This advantage of RNAProB on the RB86 dataset is likely due to the use of the smoothed PSSM encoding scheme since a comparable result, i.e., 0.50 for RNAProB_std and 0.51 for RBRpred, is obtained when using the standard PSSM encoding. At the same time, for the RB106 dataset the RNAProB_std is outperformed by RBRpred by a margin of 0.05. We observe that the RNAProB method was extensively parameterized by the authors. They considered not only the values of C and γ, but also tuned values of smoothing window size and two weight parameters. Additionally, the authors re-tuned their prediction model for each dataset, effectively creating dataset-tuned models, while we propose a single model that was tuned on the RB86 dataset and applied (without re-tuning) on the other datasets.

Table 5. Comparison between the proposed RBRpred and the six competing sequence-based RBR prediction methods on three datasets. Each competing method is compared with RBRpred at equal sensitivity and at equal precision. The RBRpred results are based on the 5 FCV and the matching sensitivity and precision values are underlined.

Dataset	Method	Sensitivity	Specificity	Precision	Accuracy	MCC
RB86	ANN	43.40	91.04	58.80	80.20	0.39
	RBRpred[1]	43.41	95.65	74.60	83.76	0.48
	RBRpred[2]	66.44	86.28	58.80	81.76	0.51
	Pprint	53.05	89.55	59.93	81.16	0.45
	RBRpred[1]	53.06	92.76	68.36	83.73	0.50
	RBRpred[2]	64.60	87.28	59.94	82.12	0.51
	RNAProB	79.95	90.36	70.96	87.99	0.68
	RBRpred[1]	79.95	71.62	45.36	73.51	0.44
	RBRpred[2]	49.65	94.01	70.96	83.92	0.50
RB147	RNABindR	33.00	95.00	61.00	83.19	0.36
	RBRpred[1]	33.00	97.48	75.51	85.20	0.43
	RBRpred[2]	49.49	92.56	61.00	84.35	0.46
RB106	BindN	66.28	69.84	27.76	69.32	0.27
	RBRpred[1]	66.30	84.12	35.37	82.06	0.39
	RBRpred[2]	77.57	73.55	27.76	74.01	0.35
	Pprint	70.09	75.54	27.30	75.43	0.32
	RBRpred[1]	70.10	80.77	32.33	79.53	0.38
	RBRpred[2]	78.20	72.71	27.30	73.34	0.35
	RNAProB	77.14	80.87	34.57	80.44	0.42
	RBRpred[1]	77.14	74.05	28.04	74.41	0.35
	RBRpred[2]	67.12	83.36	34.58	81.47	0.39

[1]results at equal sensitivity

[2]results at equal precision

The RBRpred is characterized by consistently, across all 3 datasets, high accuracy (>82%). At the same time, sensitivity, specificity and precision record fluctuations and cannot be reliably compared using Table 4. To facilitate comparison, we concentrate on two indices that quantify predictions of RBRs (as opposed to the non-RBRs), sensitivity and precision. We report the four remaining indices obtained by RBRpred at sensitivity equal to that of a given competing method, and similarly we report the prediction quality at equal precisions, see Table 5. These values for RBRpred are computed by thresholding the outputs of the SVM classifier. The conclusions are similar to those for Table 4. The proposed method outperforms, in terms of providing higher MCC, the ANN, BindN, RNABindR and Pprint methods. When compared with the ANN and RNABindR, RBRpred obtains comparable specificity and accuracy and higher precision (by 15.8% and 14.5%, respectively) at equal sensitivity, and

higher sensitivity (by 23.0% and 16.5%, respectively) at the equal precision. RBRpred's precision at the equal sensitivity when contrasted against the Pprint on the RB86 and the RB106 datasets is 8.4% and 5.0% higher, respectively, and similarly sensitivity at the equal precision is 11.6% and 8.1% higher, respectively. The BindN method is shown to provide 14.3%, 7.6% and 12.7% lower specificity, precision and accuracy at the equal sensitivity, respectively, and 11.3%, 3.7% and 4.7% lower sensitivity, specificity and accuracy at the equal precision, respectively, when compared with the RBRpred. Finally, the RNAProB is confirmed to improve over RBRpred on the RB86 dataset. More specifically, at the equal sensitivity, the RNAProB improves specificity by 18.7%, precision by 25.6% and accuracy by 14.4%, and at the equal precision, it offers sensitivity that is higher by 30.3% and comparable accuracy and specificity. At the same time, results on the RB106 dataset reveal that the RNAProB obtains a slightly higher (by about 6%) specificity, precision and accuracy at the equal sensitivity, and a better sensitivity (by 10.0%), similar accuracy and a slightly lower specificity (by about 2.5%) at the equal precision.

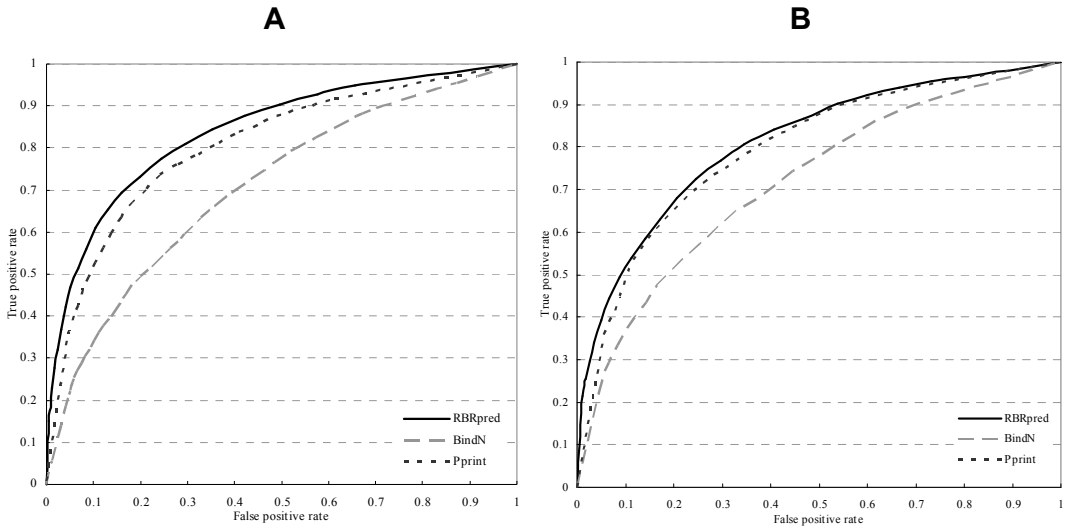


Figure 7. The ROC curves where TP rate is on the *y*-axis and FP rate on the *x*-axis for the RBRpred, BindN and Pprint methods on two datasets, (A) for the RB86 dataset; and (B) for the RB48 dataset.

The ROC curve for the proposed predictor and two existing methods, BindN and Pprint, for the RB86 dataset is given in Figure 7A. The BindN and Pprint predictions were obtained from the BindN webserver [7] and Table II in [23], respectively. We could not retrieve predictions for individual proteins which are necessary to draw the curve for the ANN, ANN_WP, RNABindR and RNAProB methods. The curve demonstrates that RBRpred improves over the other two predictors for the entire range of the true positive and the false positive rates.

We observe differences in MCC achieved by the RBRpred for the three datasets. The

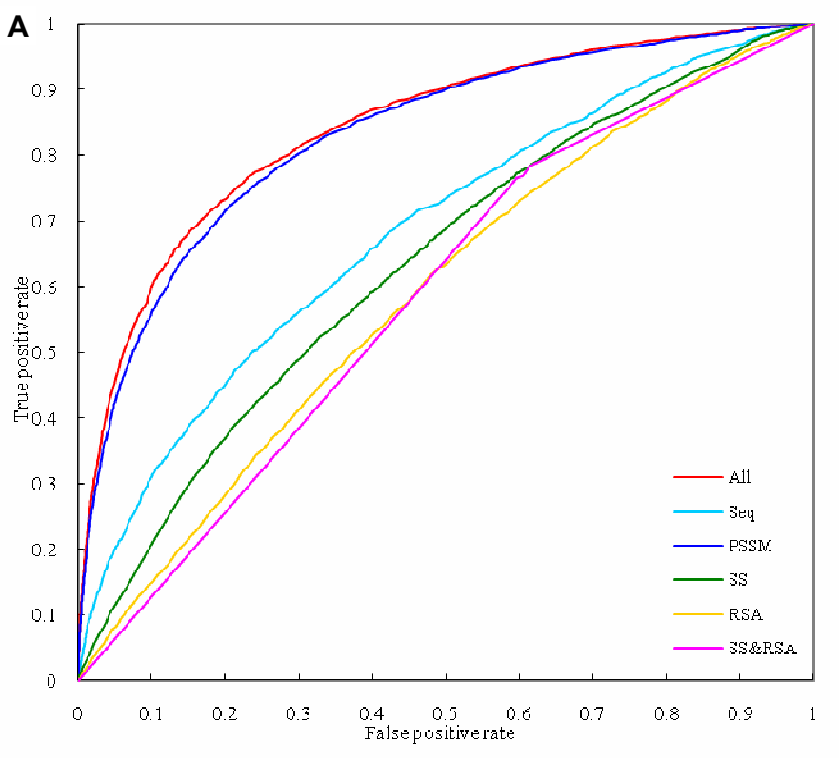
highest MCC was obtained on the RB86 dataset, second highest on the RB147, and the lowest on the RB106 dataset. This could be due to the intrinsic difficulty in predicting RBRs defined by a smaller distance cutoff, i.e., the three datasets use progressively smaller cutoffs, and since as a result the datasets are more imbalanced, see Table 2.

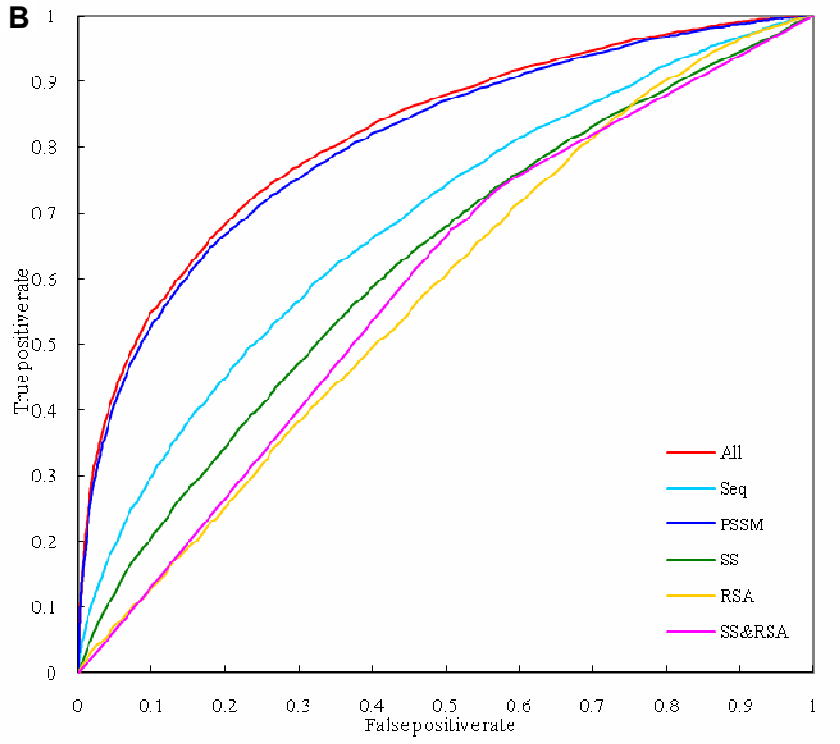
Test on the RB48 dataset

In this test the prediction model, which is generated on the RB86 dataset using the selected 420 features and parameterized SVM (with C = 1.1 and γ = 0.025), is tested on the RB48 dataset. The test dataset is characterized by low pairwise sequence identity (<25%) with respect to the RB86 dataset. The corresponding ROC curves are shown in Figure 7B. The results demonstrate that RBRpred outperforms BindN and Pprint, and that these improvements are consistent with the cross validation results on the RB86 dataset, see Figure 7A.

Analysis of selected features

We considered total of 789 features, among which 420 were selected to build the proposed method. Table 3 shows the number of features before/after the feature selection for each feature set and type. About two thirds of the selected features are PSSM-based, indicating the importance of the evolutionary conservation in predicting RBRs. In order to evaluate the contributions of features derived from different sources, the selected features were grouped into the five sets which were used separately to build the prediction model. This was performed using 5 FCV tests with the parameterized SVM (with C = 1.1 and γ = 0.025) on the RB86, RB147 and RB106 datasets. We show ROC curves for each of the five feature sets, together with the curve for entire set of 420 features, for the three datasets in Figure 8. Since some curves intersect with each other, we also give the AUC values in Table 6.


A
True positive rate
False positive rate
All
Seq
PSSM
3S
RSA
3S&RSA


B
True positive rate
False positive rate
All
Seq
PSSM
SS
RSA
SS&RSA

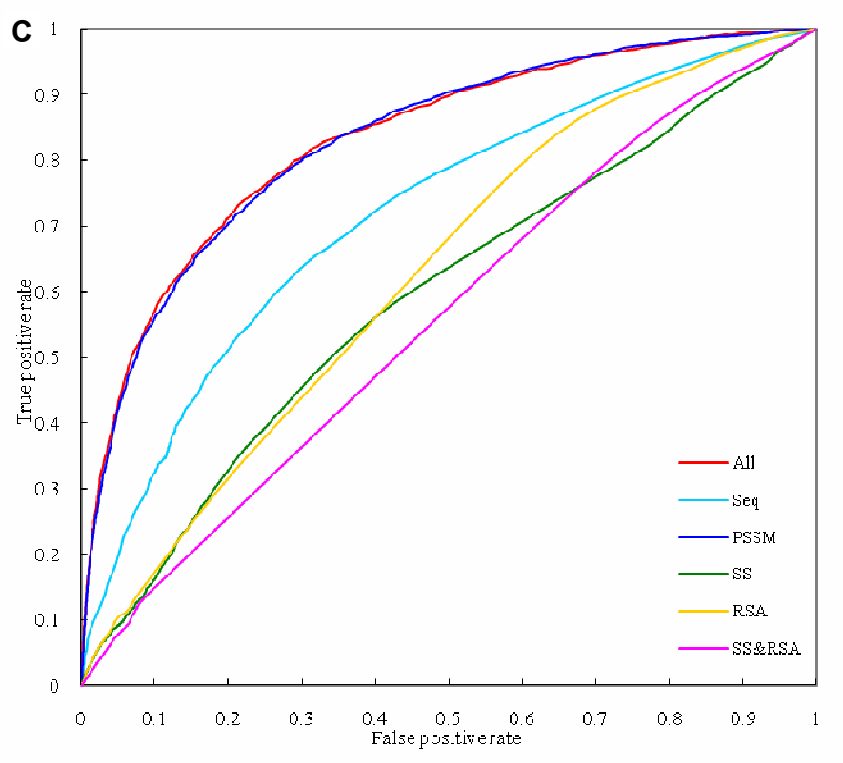


Figure 8. The ROC curves where TP rate is on the *y*-axis and FP rate on the *x*-axis for the SVM-based predictors that use the 33 selected sequence-based features (Seq), 288 selected PSSM-based features (PSSM), 55 selected SS-based features (SS), 22 selected RSA-based features (RSA), and 22 selected SS&RSA-based features (SS&RSA). The tests were performed using 5 FCV on the (A) RB86 dataset, (B) RB147 dataset and (C) RB106 dataset.

Table 6. The AUC values for the SVM-based predictors that use the 33 selected sequence-based features (Seq), 288 selected PSSM-based features (PSSM), 55 selected SS-based features (SS), 22 selected RSA-based features (RSA), and 22 selected SS&RSA-based features (SS&RSA), which were measured based on the 5 FCV on the RB86, RB147, and RB106 datasets. The feature sets are sorted by the AUC value on the RB86 dataset.

Feature sets	RB86	RB147	RB106
All features	0.842	0.818	0.833
PSSM-based	0.832	0.807	0.831
Sequence-based	0.683	0.684	0.717
SS-based	0.635	0.624	0.591
RSA-based	0.590	0.577	0.623
SS&RSA-based	0.586	0.589	0.555

The PSSM-based features contribute the most among the five feature sets. This is

expected as residues from the RNA binding sites are usually conserved. The selected sequence-based features rank second, and the remaining three feature sets provide lower and comparable amount of information for the prediction of RBRs. The selected SS-based features (derived from the predicted secondary structure) are more beneficial on the RB86 and RB147 datasets, while the selected RSA-based features are better than the other two feature sets on the RB106 dataset. The improved value of the RSA-based features for the RB106 dataset could be explained by the smaller cutoff used to define RBR, i.e., the RBR residues are on average more exposed to the solvent in this dataset when compared with the other two datasets. As noted above, comparisons with existing methods show variations of MCC on the different datasets, which is likely due to the different distance cutoffs that were used. This is also observed in Figure 8 and Table 6. The cutoff used in the RB147 dataset is closer to that used in the RB86 dataset, and these two datasets share relatively similar ROC curves (Figures 8-A and 8-B) and AUC values for the corresponding feature sets. On the contrary, the ROC curves (Figure 8C) and AUC values for the RB106 dataset, which uses a more stringent (lower) cutoff, are different. For instance, usage of the PSSM-based features results in predictions that are of similar quality when compared with the prediction using the entire set of 420 features. This suggests that the proposed prediction model that is designed (including feature selection and parameterization) on the RB86 dataset for which a distance cutoff of 6Å is used, may not work as well on the other two datasets which apply different cutoffs (5Å for the RB147 and 3.5Å for the RB106). Although RBRpred shows better results on the RB147 and RB106 datasets when compared to several existing methods, there may still be space to improve the predictions on these two datasets.

Table 7. Comparison between the proposed RBRpred and the predictions based on the features extracted from the native secondary structure (SS) and/or native relative solvent accessibility (RSA).

Dataset	SS	RSA	Sensitivity	Specificity	Precision	Accuracy	MCC
RB86	Predicted	Predicted	60.88	89.67	63.46	83.12	0.51
	Native	Native	58.82	91.09	66.05	83.75	0.52
	Native	Predicted	58.30	90.50	64.39	83.17	0.51
	Predicted	Native	61.43	90.00	64.40	83.49	0.52
RB147	Predicted	Predicted	52.90	91.02	58.09	83.76	0.46
	Native	Native	52.53	92.75	63.04	85.09	0.49
	Native	Predicted	50.30	91.95	59.51	84.01	0.45
	Predicted	Native	54.86	91.72	60.93	84.70	0.49
RB106	Predicted	Predicted	39.92	95.68	54.78	89.22	0.41
	Native	Native	42.23	96.47	61.03	90.18	0.46
	Native	Predicted	37.22	96.45	57.85	89.58	0.41
	Predicted	Native	45.13	96.01	59.71	90.11	0.47

Prediction using native secondary structure and relative solvent accessibility

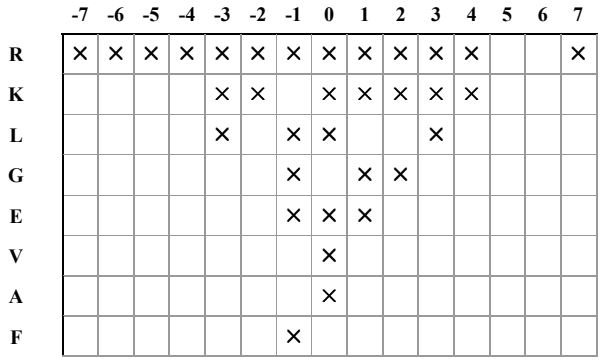
The proposed method integrates information concerning the secondary structure predicted with PSIPRED [29] and relative solvent accessibility predicted with Real-SPINE [30]. Although both PSIPRED and Real-SPINE provide high quality predictions, the predicted values differ from the native values. We investigate whether the usage of the native secondary structure and/or solvent accessibility would further increase the quality of the prediction of RBRs. We performed 5 FCV tests on the RB86, RB147 and RB106 datasets using the four combinations of native/predicted values of the secondary structure (SS) and the relative solvent accessibility (RSA). The features are computed using native SS and native RSA, native SS and predicted RSA, and predicted SS and native RSA. Predictions using the above features are compared with original predictions (using predicted SS and predicted RSA), see Table 7. The native SS and RSA were extracted with the DSSP program [71] and the remaining features (sequence- and PSSM-based) were used in all four cases. The inclusion of the native SS and RSA does not lead to improvements in prediction quality for the RB86 dataset. At the same time, it helps with the predictions on the RB147 and RB106 datasets. We observe improvement of MCC by 0.03 and 0.05, respectively, and precision by 5% and 6%, respectively. The improvement is due to the use of the native RSA, since the usage of the predicted secondary structure does not seem to lower the prediction quality. This is likely since these two datasets apply lower cutoffs to define RBRs, and thus the knowledge of the actual RSA would be more helpful when compared with the RB86 dataset. Overall, in our view the results demonstrate that the predicted SS and RSA are sufficient for high quality predictions of the RBRs.

Relationship between residue types and RBRs

We use a sliding window to encode AA types of the target residue and its neighbors. Among the 300 *RT* features, 33 are selected and they include eight AA types, namely Arg (R), Lys (K), Leu (L), Gly (G), Val (V), Ala (A), Glu (E) and Phe (F). The positively charged AAs Arg and Lys show higher propensity to form RNA-binding sites, likely due to their ability to participate in interactions with the negatively charged phosphate backbone of RNA [9, 19]. Gly is small and provides flexibility for the protein-RNA interactions [19, 23]. The Glu, Leu, Val, Ala and Phe are disfavored in the RNA-binding sites. This is likely since Glu has a negatively charged side chain while the other four residue types are hydrophobic [19, 23]. Table 8 lists the selected *RT* features along the sliding window. The selected residues are symmetrical against the central position that denotes the predicted residue. Arg is selected at virtually all window positions, which is likely due to the abundance of Arginine-rich motifs in the RNA binding sites [50].

Table 8. The selected *RT* features along the sliding window. Rows correspond to AA types (only the selected AA types are listed) and columns to positions in the sliding

window where 0 represents the predicted residue, +*i*/-*i* denote the i^{th} neighboring residue towards C-/N-terminus, and crosses denote selected features.



	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7
R	×	×	×	×	×	×	×	×	×	×	×	×			×
K					×	×		×	×	×	×	×			
L					×		×	×			×				
G							×		×	×					
E							×	×	×						
V								×							
A								×							
F							×								

Table 9. Summary of the selected PSSM features along the sliding window. Rows represent positions in the sliding window where 0 represents the predicted residue and +*i*/-*i* denote the i^{th} neighboring residue towards C-/N-terminus. Columns represent AA types. The cells in the table represent PSSM features where shading denotes the results of the feature selection. Darker shading corresponds to higher ranked features (according to the χ^2-scores) and white shading to features that were not selected. The last row/column shows the average χ^2-score of the selected PSSM features for each window position/ AA type. The χ^2-scores for each feature are obtained by using χ^2 feature selection method where a higher χ^2-score corresponds to a higher rank in the feature selection.

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	Avg. score
-7																					82.48
-6																					74.76
-5																					92.03
-4																					104.66
-3																					114.40
-2																					111.04
-1																					152.92
0																					252.58
+1																					124.54
+2																					117.09
+3																					127.90
+4																					110.92
+5																					88.56
+6																					78.07
+7																					92.63
Avg. score	82.94	218.19	62.85	234.86	91.40	59.38	181.43	63.96	63.49	111.54	113.05	173.20	114.76	96.76	60.73	57.75	67.56	79.64	135.56	130.96	

Relationship between sequence conservation and RBRs

The feature selection demonstrates the fundamental role of the sequence conservation in predicting RBRs. A total of 260 out of the 300 *PSSM* features are selected, see Table 9. The average χ^2-scores of the selected features at each window position are computed and shown in the figure. As expected, the distribution of the average scores is relatively symmetric and the values diminish with the linear distance with respect to the central position in the window. We found an exception at the +2 /-2 positions, in which case the average χ^2-score is lower than that for positions +1/-1 and +3/-3. This could be due to weaker interactions between the central residue and the two residues at +2/-2 positions when compared to the residues at +3/-3 or +4/-4 positions which may form hydrogen bonds with the central residue [75]. We also compute the average χ^2-scores for the 20 AA types. The six AA types with the highest scores are Asp (D), Arg (R), Glu (E), Lys (K), Tyr (Y) and Val (V). The Arg, Lys and Tyr have positively charged side chains and thus they have a higher chance to interact with the negatively charged RNA. The Asp and Glu that have negatively charged side chains and the hydrophobic Val are disfavored in the RNA binding sites [19].

The selected features also include 15 *EntWOP* features. The χ^2-scores of those features are not symmetrical with respect to the central position in the window. The residues with higher scores are located towards the N-terminus side of the window. Currently, we have no explanation for this skewed distribution. Among the 14 *CNCC* features, only one feature (corresponding to the correlation between the central residue and the residue at the 1st position towards the N-terminus) was removed. The top scoring *CNCC* features correspond to the residues at +2/-2 position in the window.

Relationship between secondary structure and RBRs

During the feature selection, none of the *SSProb* features were removed, indicating that the predicted secondary structure information of the target and the neighboring residues is helpful in distinguishing binding/non-binding residues. Two out of 27 *TriSS* features were selected and they correspond to secondary structure triplets "CCC" and "HHH". Among the *SSCont*, *SegLen* and *SegDB* features, only the coil and helix related features are selected. The reason could be that coil residues provide flexibility for RNA binding sites while helix residues are disfavored due to their rigidity. Treger and Westhof [16] found that RNA interface residues in helices that interacted with the RNA molecules through main-chain contacts were less numerous than expected. Ellis *et al.*'s work shows that helices are disfavored in protein-RNA interface while non-helical structures may occur more frequently due to their potential flexibility, which complements the flexible nature of the bound RNA structures [102].

Among the three *SSCont* features, we analyze two features that represent the predicted coil/helix content in the sliding window. We contrast values of these two features on the RB86 dataset between the windows in which the central residue binds RNA (RBR windows) and the windows where the central residue does not bind RNA (non-RBR

windows). The average coil content for the RBR and non-RBR windows is 49.3% and 39.6%, respectively, while the helix content equals 29.7% and 40.4%, respectively. This confirms that coil conformation, in contrast to the helical conformation, is more frequent in a sequence window centered on the RBRs.

Two *SegLen* features (segment length of predicted coils/helixes) and four *SegDB* features (minimum/maximum distance between the central position and the boundaries of the predicted coil/helix segments) were picked by the feature selection. Figures 9 and 10 present the RNA binding propensities for the residues with varied *SegLen* and *SegDB* values that were computed on the RB86 dataset. Propensity larger than zero indicates that more residues with certain *Seglen*/*SegDB* value are observed in the RNA-binding sites than that observed in other sequence positions, i.e., RBRs prefer sequence patterns with certain *Seglen*/*SegDB* value. On the contrary, propensity smaller than zero means residues with certain *Seglen*/*SegDB* value are less likely to be RBRs. Propensity equal to zero means that no preference is found. Figure 9 demonstrates that residues in long coil segments are more likely to bind RNA, while residues in long helix segments are disfavored in the RNA binding sites. Additional insights are provided in Figure 10 which reveals that the positive/negative propensity is higher for residues that are located farther from the termini of the predicted coil/helix segments, i.e., inside of longer segments. The underlying reason behind these observations could be that coil structures are more flexible and can provide flexibility needed for the RNA-binding process, while helix structures are relatively more rigid.

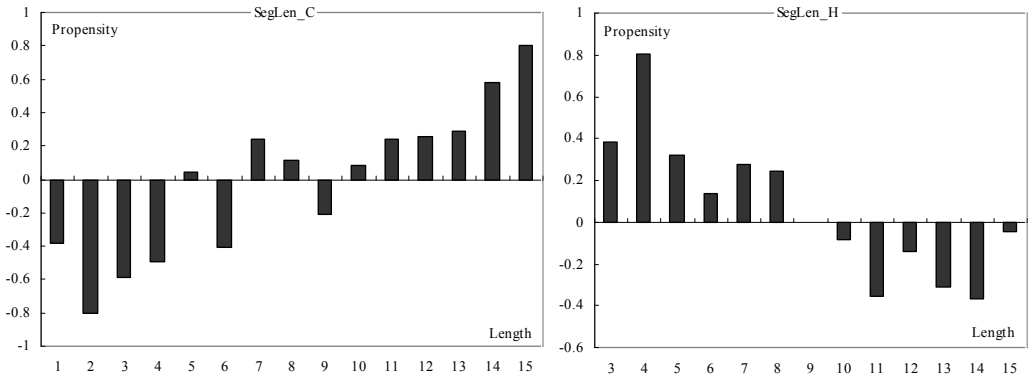


Figure 9. RNA binding propensity for residues with varied *SegLen* values, which are defined as the length of the predicted secondary structure segment which includes the central residue. The *SegLen*_C and *SegLen*_H correspond to the coil and helix segments, respectively. The *x*-axis represents the length of corresponding segment and the *y*-axis denotes the RNA binding propensity defined as percentage of RBRs with certain *SegLen* value among all RBRs divided by the percentage of all residues with the same *SegLen* value among all residues in the RB86 dataset. The length of helix segments starts from 3 since at least three consecutive helical residues are necessary to form a helix segment. The propensities were computed using the RB86 dataset.

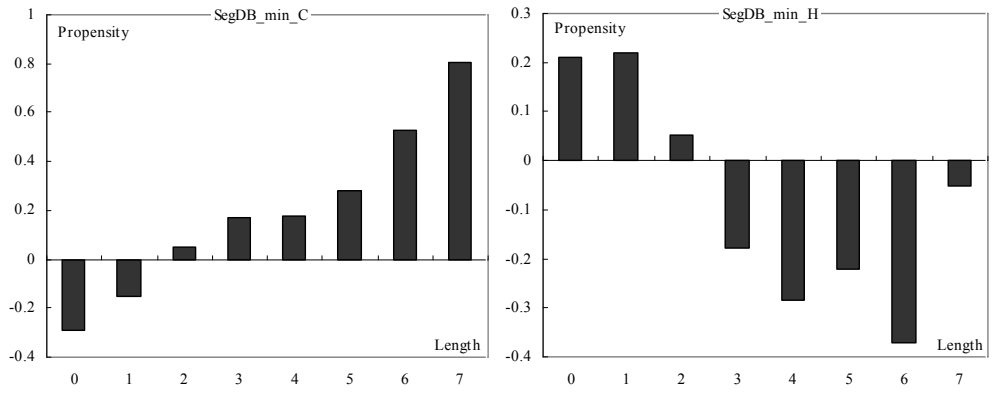


Figure 10. RNA binding propensity for residues with varied *SegDB_min* values, which are defined as the minimal distance between the central position and the boundaries of the predicted secondary structure segments. The *SegDB_min_C* and *SegDB_min_H* correspond to the distance to the nearest coil and helix segments, respectively. The *x*-axis represents the minimum distance and the *y*-axis denotes the RNA binding propensity defined as percentage of RBRs with certain *SegDB_min* value among all RBRs divided by the percentage of all residues with the same *SegDB_min* value among all residues in RB86 dataset. The propensity distribution for the maximal distance is not shown since it is not as consistent as the distribution for the minimal distance. The propensities were computed using the RB86 dataset.

Relationship between solvent accessibility and RBRs

All 15 *RSA* features and 7 *AveRSA* features were picked in the performed feature selection. We study the relation between these features and the RBRs using the *RSA-0* feature, which represents the relative solvent accessibility of the target residue. This is since this feature obtains the highest χ^2-score among all 15 *RSA* features. Figure 11 shows the distribution of *RSA-0* values for the RBRs and non-RBRs. We observe that RBRs tend to have higher *RSA-0* values, i.e., they are more solvent exposed, as the corresponding distribution peaks around the [0.30, 0.35) interval. On the contrary, the distribution of the RSA values for the non-RBRs is skewed towards lower values and it peaks around the [0, 0.1) interval. This indicates that residues that are partially solvent exposed are more likely to form the RNA binding sites.

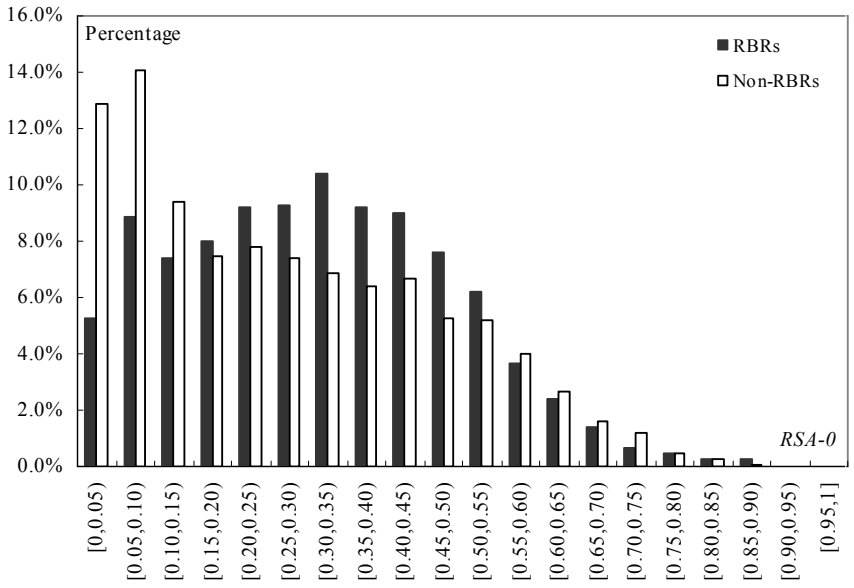


Figure 11. Distribution of predicted RSA values for the RBRs/non-RBRs in central position in the sliding window (the *RSA-0* feature). Values of the predicted RSA that vary between 0 (fully buried) and 1 (fully exposed) are divided into 20 equally-sized bins (*x*-axis). The *y*-axis denotes the percentage of RBRs/non-RBRs with RSA values in a certain bin.

A total of 22 out of 54 *SR* features, which incorporate both the predicted secondary structure and solvent accessibility, are selected. All of them but two, which have low χ^2-scores, concern the coil and helix structures. This is consistent with the selected SS-based features, and demonstrates the importance of the coil and helix structures in the RBR prediction. The top three *SR* features according to the χ^2-scores are *SR_[0.1,1.0]_C*, *SR_[0.2,1.0]_C* and *SR_[0,0.2)_H*. The first two features correspond to the exposed coil residues which are more likely to form RNA binding sites and the last feature denotes the buried helical residues which have lower chance to bind RNA.

Conclusions

We developed a novel sequence-based method, called RBRpred, for the prediction of RNA binding residues (RBRs). RBRpred utilizes a wide range of information derived from the amino acid (AA) sequence including PSI-BLAST profiles, predicted secondary structure (SS) and predicted relative solvent accessibility (RSA). This information is converted into five custom-designed sets of features (sequence-based, PSSM-based, SS-based, RSA-based and SS&RSA-based features) which are fed into SVM classifier to generate the predictions.

We applied feature selection to reduce the dimensionality of the input vector and to investigate the relations between the input features and residues that interact with

RNA. Analysis of the selected features reveals that: (1) sequence conservation plays a fundamental role in predicting RNA-binding residues; (2) the positively charged AAs Arg and Lys show higher propensity to form RNA-binding sites due to their ability to interact with the negatively charged phosphate backbone of the RNA; (3) Gly also has higher propensity since it provides flexibility for the protein-RNA interactions; (4) Glu that has negatively charged side chain and a few hydrophobic residues such as Leu, Val, Ala and Phe are disfavored in the RNA-binding sites; (5) residues in the coil conformation, especially those in long coil segments, are more flexible and are more likely to interact with RNA; (6) residues in the helix conformation are more stable (rigid) and consequently they are less likely to bind RNA; and (7) residues that are partially exposed to the solvent are more likely to be in the RNA-binding sites.

We evaluated contributions of each of the five feature sets to the prediction of the RNA binding residues. PSSM-based features that express evolutionary conservation account for over 60% of the selected features and they contribute the most to the prediction. The sequence-based features rank second, and the remaining three feature sets have the lowest and comparable impact on the prediction. We also investigated the impact of the usage of the native SS and RSA when compared with the predicted values. We conclude that predicted SS and RSA are sufficient for the prediction of RNA binding residues, and that knowledge of the native RSA values helps in predictions of RBRs defined using lower distance cutoff.

The RBRpred method was compared with state-of-the-art sequence-based prediction methods on three benchmark datasets using both 5 fold cross validation and jackknife test. We also performed a blind test of the proposed method on an independent dataset. The results demonstrate that RBRpred is characterized by quality comparable to or better than the existing methods.

Abbreviations

RBR	RNA-Binding Residue
SVM	Support Vector Machine
AA	Amino Acid
PSSM	Position Specific Scoring Matrix
MCC	Matthews Correlation Coefficient
FCV	Fold Cross Validation
ROC	Receiver Operating Characteristic
AUC	Area Under the ROC Curve
SS	Secondary Structure
ASA	solvent Accessible Surface Area
RSA	Relative Solvent Accessibility
RT	Residue Type (amino acid type of a given residue)
WOP	Weighted Observed Percentage (generated by PSI-BLAST)

EntWOP	Entropy computed based on WOP vector
CNCC	Close Neighbor Correlation Coefficient (correlation coefficient of PSSM vectors of neighboring residues)
SSprob	Secondary Structure Probabilities (probability of a given residue to be predicted as helix, strand or coil by PSIPRED)
SScont	Secondary Structure Content
TriSS	Triplet Secondary Structure
SegLen	secondary structure Segment Length e.g. SegLen_C and SegLen_H correspond to the length of coil and helix segments, respectively.
SegDB	in one secondary structure Segment, Distance between the given residue and the Boundaries of the segment e.g. SegDB_min_C and SegDB_min_H represent the minimum distance from the given residue to the boundaries of coil and helix segment, respectively.
AveRSA	Average RSA value in a sliding window
SR	SS and RSA e.g. SR_[0.1,1.0]_C indicates the given residue is a coil residue and the RSA value of that residue is within range [0.1,1.0].

Acknowledgments

The authors thank Manish Kumar and Michael Terribilini for providing their datasets, which were supplemented with helpful explanations. The authors are also grateful to Dr. Zhou for providing and explaining the Real-SPINE program.

References

[1] Moras, D. Aminoacyl-tRNA synthetases. *Curr. Opin. Struct. Biol.*, **1992**, *2*, 138-142.

[2] Freed, E.O.; Mouland, A.J. The cell biology of HIV-1 and other retroviruses. *Retrovirology*, **2006**, *3*, 77.

[3] Gromiha, M.M. Influence of DNA stiffness in protein-DNA recognition. *J. Biotechnol.*, **2005**, *117*, 137-45.

[4] Gromiha, M.M.; Selvaraj, J.G.S.; Kono, H.; Sarai, A. Intermolecular and intramolecular readout mechanisms in protein-DNA recognition. *J. Mol. Biol.*, **2004**, *337*, 285-94.

[5] Kumar, M.; Gromiha, M.M.; Raghava, G.P. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics*, **2007**, *8*, 463.

[6] Bhardwaj, N.; Langlois, R.E.; Zhao, G.; Lu, H. Kernel-based machine learning protocol for predicting DNA-binding proteins. *Nucleic Acids Res.,* **2005**, ***33***(20), 6486-93.

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

[8] Bhardwaj, N.; Lu, H. Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions. *FEBS Lett.*, **2007**, *581*(5), 1058-66.

[9] Chen, Y.C.; Lim, C. Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res.*, **2008**, *36*(5), e29.

[10] Jones, S.; Daley, D.T.; Luscombe, N.M.; Berman, H.M.; Thornton, J.M. Protein-RNA interaction: a structural analysis. *Nucleic Acids Res.*, **2001**, *29*, 943–954.

[11] Cheng, C.W.; Su, E.C.; Hwang, J.K.; Sung, T.Y.; Hsu, W.L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, **2008**, *9*(Suppl 12), S6.

[12] Berman, H.M.; Battistuz, T.; Bhat, T.N.; Bluhm, W.F.; Bourne, P.E.; Burkhardt, K.; Feng, Z.; Gilliland, G.L.; Iype, L.; Jain, S.; Fagan, P.; Marvin, J.; Padilla, D.; Ravichandran, V.; Schneider, B.; Thanki, N.; Weissig, H.; Westbrook, J.D.; Zardecki, C. The Protein Data Bank. *Acta. Crystallogr. D. Biol. Crystallogr.*, **2002**, *58*(Pt 6 No 1), 899-907.

[13] Draper, D.E. Protein-RNA recognition. *Annu. Rev. Biochem.*, **1994**, *64*, 593–620.

[14] Draper, D.E. Themes in RNA-protein recognition. *J. Mol. Biol.*, **1999**, *293*, 255-270.

[15] Allers, J.; Shamoo, Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **2001**, *311*, 75–86.

[16] Treger, M.; Westhof, E. Statistical analysis of atomic contacts at RNA-protein interfaces. *J. Mol. Recogn.*, **2001**, *14*, 199–214.

[17] Jeong, E.; Chung, I.F.; Miyano, S. A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.*, **2004**, *15*, 105-116.

[18] Jeong, E.; Miyano, S. A weighted profile based method for protein-RNA interacting residue prediction. In: *Lecture notes in computer science*; Corrado, P.; Luca, C.; Stephen, E., Ed.; Berlin/Heidelberg: Springer, **2006**; Vol. *3939*, pp 123–139.

[19] Terribilini, M.; Lee, J.H.; Yan, C.; Jernigan, R.L.; Honavar, V.; Dobbs, D. Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **2006**, *12*, 1450-1462.

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

[21] Kim, O.T.P.; Yura, K.; Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **2006**, *34*(22), 6450-60.

[22] Wang, Y.; Xue, Z.; Shen, G.; Xu, J. PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, **2008**, *35*(2), 295-302.

[23] Kumar, M.; Gromiha, M.M.; Raghava, G.P.S. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **2008**, *71*, 189-194.

[24] Sprigg, R.V.; Murakami, Y.; Nakamura, H.; Jones, S. Protein function annotation from sequence prediction of residues interacting with RNA. *Bioinformatics*, **2009**, *25*(12), 1492-1497.

[25] McDonald, I.K.; Thornton, J.M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **1994**, *238*, 777-793.

[26] Allers, J.; Shamoo, Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **2001**, *311*, 75-86.

[27] Wu, J.; Liu, H.; Duan, X.; Ding, Y.; Wu, H.; Bai, Y.; Sun, X. Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. *Bioinformatics*, **2009**, *25*(1), 30-5.

[28] Ezkurdia, I.; Bartoli, L.; Fariselli, P.; Casadio, R.; Valencia, A.; Tress, M.L. Progress and challenges in predicting protein-protein interaction sites. *Brief Bioinform.*, **2009**, *10*(3), 233-46.

[29] Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **1999**, *292*, 195-202.

[30] Dor, O.; Zhou Y. Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins*, **2007**, *68*, 76-81.

[31] Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **1997**, *25*, 3389-3402.

[32] Chou, K. C.; Zhang, C. T. Review: Prediction of protein structural classes *Crit. Rev. Biochem. Mol. Biol.*, **1995**, *30*, 275-349.

[33] Chou, K. C.; Shen, H. B. Cell-PLoc: A package of web-servers for predicting subcellular localization of proteins in various organisms *Nature Protocols*, **2008**, *3*, 153-162.

[34] Chou, K. C.; Shen, H. B. Review: Recent progresses in protein subcellular location prediction. *Anal. Biochem.*, **2007**, *370*, 1-16.

[35] Zhou, G. P. An intriguing controversy over protein structural class prediction *J. Protein Chem.*, **1998**, *17*, 729-738.

[36] Zhou, G. P.; Assa-Munt, N. Some insights into protein structural class prediction *Proteins: Struct., Funct., Genet.*, **2001**, *44*, 57-59.

[37] Zhou, G. P.; Doctor, K. Subcellular location prediction of apoptosis proteins *Proteins: Struct., Funct., Genet.*, **2003**, *50*, 44-48.

[38] Zhou, X. B.; Chen, C.; Li, Z. C.; Zou, X. Y. Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes *J. Theor. Biol.*, **2007**, *248*, 546–551.

[39] Lin, H. The modified Mahalanobis discriminant for predicting outer membrane proteins by using Chou's pseudo amino acid composition *J. Theor. Biol.*, **2008**, *252*, 350-356.

[40] Jiang, X.; Wei, R.; Zhang, T. L.; Gu, Q. Using the concept of Chou's pseudo amino acid composition to predict apoptosis proteins subcellular location: an approach by approximate entropy. *Prot. Pept. Lett.*, **2008**, *15*, 392-396.

[41] Li, F. M.; Li, Q. Z. Predicting protein subcellular location using Chou's pseudo amino acid composition and improved hybrid approach *Prot. Pept. Lett.*, **2008**, *15*, 612-616.

[42] Lin, H.; Ding, H.; Feng-Biao Guo, F. B.; Zhang, A. Y.; Huang, J. Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid

composition *Prot. Pept. Lett.*, **2008**, *15*, 739-744.

[43] Ding, Y. S.; Zhang, T. L. Using Chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: an approach with immune genetic algorithm-based ensemble classifier *Pattern Recogn. Lett.*, **2008**, *29*, 1887-1892.

[44] Chen, C.; Chen, L.; Zou, X.; Cai, P. Prediction of protein secondary structure content by using the concept of Chou's pseudo amino acid composition and support vector machine *Prot. Pept. Lett.*, **2009**, *16*, 27-31.

[45] Lin, H.; Wang, H.; Ding, H.; Chen, Y. L.; Li, Q. Z. Prediction of Subcellular Localization of Apoptosis Protein Using Chou's Pseudo Amino Acid Composition *Acta Biotheor*, **2009**, *57*, 321-330.

[46] Ding, H.; Luo, L.; Lin, H. Prediction of cell wall lytic enzymes using Chou's amphiphilic pseudo amino acid composition *Prot. Pept. Lett.*, **2009**, *16*, 351-355.

[47] Zeng, Y. H.; Guo, Y. Z.; Xiao, R. Q.; Yang, L.; Yu, L. Z.; Li, M. L. Using the augmented Chou's pseudo amino acid composition for predicting protein submitochondria locations based on auto covariance approach *J. Theor. Biol.*, **2009**, *259*, 366–372.

[48] Swets, J.A. Measuring the accuracy of diagnostic systems. *Science*, **1988**, *240*(4857), 1285-1293.

[49] Bradley, A.P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recogn.*, **1997**, *30*(7), 1145-1159.

[50] Weiss, M.A.; Narayana, N. RNA recognition by arginine-rich peptide motifs. *Biopolymers*, **1998**, *48*, 167-180.

[51] Lustig, B.; Arora, S.; Jernigan, R.L. RNA base-amino acid interaction strengths derived from structures and sequences. *Nucleic Acids Res.*, **1997**, *25*, 2562-2565.

[52] Kim, H.; Jeong, E.; Lee, S.W.; Han K. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett.*, **2003**, *552*, 231-239.

[53] Lichtarge, O.; Sowa, M.E. Evolutionary predictions of binding surfaces and interactions. *Curr. Opin. Struct. Biol.*, **2002**, *12*, 21-27.

[54] Chou, K.C.; Shen, H.B. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM *Biochem. Biophys. Res. Comm.*, **2007**, *360*, 339-345.

[55] Shen, H.B.; Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses *Biochem. Biophys. Res. Comm.*, **2007**, *364*, 53-59.

[56] Gromiha, M.M.; Yabuki, Y. Functional discrimination of membrane proteins using machine learning techniques. *BMC Bioinformatics*, **2008**, *9*, 135.

[57] Chou, K.C.; Shen, H.B. ProtIdent: A web server for identifying proteases and their types by fusing functional domain and sequential evolution information *Biochem. Biophys. Res. Comm.*, **2008**, *376*, 321-325.

[58] Shen, H.B.; Chou, K.C. Predicting protein fold pattern with functional domain and sequential evolution information *J. Theor. Biol.*, **2009**, *256*, 441-446.

[59] Shen, H.B.; Chou, K.C. QuatIdent: A web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information *Journal of Proteome Research*, **2009**, *8*, 1577–1584.

[60] Shen, H.B.; Chou K.C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. *Protein Eng Des Sel*, **2007**, *20*, 561-7.

[61] Shen, H.B.; Chou, K.C. A top-down approach to enhance the power of predicting human protein subcellular localization: Hum-mPLoc 2.0 *Anal. Biochem.*, **2009**, *394*, 269–274.

[62] Shen, H.B.; Chou, K.C. Gpos-mPLoc: A top-down approach to improve the quality of predicting subcellular localization of Gram-positive bacterial proteins *Prot. Pept. Lett.*, **2009**, *16*, 1478-1484.

[63] Zhang, T.; Zhang, H.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Accurate sequence-based prediction of catalytic residues. *Bioinformatics*., **2008**, *24*(20), 2329-38.

[64] Cheng, J.; Baldi, P. Improved residue contact prediction using support vector machines and a large feature set. *BMC Bioinformatics*, **2007**, *8*, 113.

[65] Jiang, Y.; Iglinski, P.; Kurgan, L. Prediction of protein folding rates from primary sequences using hybrid sequence representation. *J. Comput. Chem.* **2009**, *30*(5), 772-83.

[66] Ivankov, D.N.; Finkelstein, A.V. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. *Proc. Natl. Acad. Sci. USA*., **2004**, *101*(24), 8942-4.

[67] Faraggi, E.; Yang, Y.; Zhang, S; Zhou, Y. Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure*, **2009**, *17*, 1515-1527.

[68] Xue, B.; Faraggi, E.; Zhou, Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. *Proteins*, **2009**, *76*, 176-183.

[69] Fischer, J.D.; Mayer, C.E.; Söding J. Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics.*, **2008**, *24*(5), 613-20.

[70] Ofran, Y.; Rost, B. ISIS: interaction sites identified from sequence. *Bioinformatics*, **2007**, *23*(2), e13-6.

[71] Kabsch, W.; Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **1983**, *22*, 2577-2637.

[72] Birzele, F.; Kramer, S. A new representation for protein secondary structure prediction based on frequent patterns. *Bioinformatics*, **2006**, *22*, 2628-34.

[73] Garg, A.; Kaur, H.; Raghava, G.P. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*, **2005**, *61*, 318-324.

[74] Chen, K.; Kurgan, L. PFRES: Protein Fold Classification by Using Evolutionary Information and Predicted Secondary Structure. *Bioinformatics*, **2007**, *23*, 2843-2850.

[75] Zhang, H.; Zhang, T.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. Sequence based residue depth prediction using evolutionary information and predicted secondary structure. *BMC Bioinformatics*, **2008**, *9*, 388.

[76] Zheng, C.; Kurgan, L. Prediction of ß-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. *BMC Bioinformatics*, **2008**, *9*, 430.

[77] Wang, Y.; Xue, Z.; Xu, J. Better prediction of the location of alpha-turns in proteins with support vector machine. *Proteins*, **2006**, *65*, 49-54.

[78] Bryson, K.; McGuffin, L.J.; Marsden, R.L.; Ward, J.J.; Sodhi, J.S.; Jones, D.T. Protein structure prediction servers at University College London. *Nucleic Acids Res.* **2005**, *33*, W36-38.

[79] Lee, B.; Richards, F. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **1971**, *55*, 379-400.

[80] Connoly, M.L. Solvent accessibility surfaces of protein and nucleic acids. *Science*, **1983**, *221*, 709-713.

[81] Wang, J.Y.; Lee, H.M.; Ahmad, S. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins*, **2005**, *61*, 481-491.

[82] Gromiha, M.M.; Oobatake, M.; Kono, H.; Uedaira, H.; Sarai, A. Role of structural and sequence information in the prediction of protein stability changes, comparison between buried and partially buried mutations. *Protein Eng.*, **1999**, *12*, 549-555.

[83] Chan, H.S.; Dill, K.A. Origins of structure in globular proteins. *Proc. Natl. Acad. Sci.*, *USA*, **1990**, *87*, 6388-6392.

[84] Eisenberg, D.; McLachlan, A.D. Solvation energy in protein folding and binding. *Nature*, **1986**, *319*, 199-203.

[85] Zhang, H.; Zhang, T.; Chen, K.; Shen, S.; Ruan, J.; Kurgan, L. On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins*, **2009**, *76*(3), 617-36.

[86] Cheng, J.; Baldi, P. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, **2006**, *22*, 1456-1463.

[87] Liu, S.; Zhang, C.; Liang, S.; Zhou, Y. Fold recognition by concurrent use of solvent accessibility and residue depth. *Proteins*, **2007**, *68*, 636-664.

[88] Ahmad, S.; Gromiha, M.M.; Sarai, A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **2003**, *50*, 629–635.

[89] Rost, B.; Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins*, **1994**, *20*, 216-226.

[90] Yuan, Z.; Huang, B. Prediction of protein accessible surface areas by support vector regression. *Proteins*, **2004**, *57*, 558-564.

[91] Adamczak, R.; Porollo, A.; Meller, J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **2004**, *56*, 753-767.

[92] Xu, Z.; Zhang, C.; Liu, S.; Zhou, Y. QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins*, **2006**, *63*, 961-966.

[93] Ward, J.J.; McGuffin, L.J.; Buxton, B.F.; Jones, D.T. Secondary structure prediction with support vector machines. *Bioinformatics*, **2003**, *19*(13), 1650-1655.

[94] Karypis, G. YASSPP: better kernels and coding schemes lead to improvements in protein secondary structure prediction. *Proteins*. **2006**, *64*(3), 575-586.

[95] Yu, C.S.; Chen, Y.C.; Lu, C.H.; Hwang, J.K. Prediction of protein subcellular localization. *Proteins*. **2006**, *64*(3), 643-51.

[96] Shi, J.Y.; Zhang, S.W.; Pan, Q.; Cheng, Y.M.; Xie, J. Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. *Amino Acids*., **2007**, *33*(1), 69-74.

[97] Bradfor, J.R.; Westhead, D.R. Improved prediction of protein-protein binding sites using a support vector machines approach. *Bioinformatics*, **2005**, *21*(8), 1487-94.

[98] Vapnik, V. *The Nature of Statistical Learning Theory*.; Springer-Verlag: New York, USA, **1999**.

[99] Joachims, T. Making large-Scale SVM Learning Practical. In: *Advances in Kernel Methods - Support Vector Learning*. Schölkopf, B.; Burges, C.; Smola, A. Ed.; Cambridge: MIT-Press, **1999**; pp. 169-184.

[100] Liu, H.; Setiono, R. Chi2: feature selection and discretization of numeric attributes. In: *Proceedings of the 7th International Conference Tools with Artificial Intelligence*; IEEE Computer Society: Washington, DC, USA, **1995**, pp. 388-391.

[101] Forman G. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* **2003**, *3*, 1289–1305.

[102] Ellis, J.J.; Broom, M.; Jones, S. Protein–RNA interactions: structural analysis and functional classes. *Proteins*, **2007**, *66*(4), 903-911.