Research Article

Journal of Molecular Recognition

Received: 17 October 2009, Revised: 3 May 2010, Accepted: 13 May 2010, Published online in Wiley Online Library: 31 July 2010

(wileyonlinelibrary.com) DOI:10.1002/jmr.1061

SVM based prediction of RNA-binding proteins using binding residues and evolutionary information

Manish Kumar[a], M. Michael Gromiha[b] and Gajendra P. S. Raghava[c]*

RNA-binding proteins (RBPs) play crucial role in transcription and gene-regulation. This paper describes a support vector machine (SVM) based method for discriminating and classifying RNA-binding and non-binding proteins using sequence features. With the threshold of 30% interacting residues, RNA-binding amino acid prediction method PPRINT achieved the Matthews correlation coefficient (MCC) of 0.32. BLAST and PSI-BLAST identified RBPs with the coverage of 32.63 and 33.16%, respectively, at the e-value of 1e-4. The SVM models developed with amino acid, dipeptide and four-part amino acid compositions showed the MCC of 0.60, 0.46, and 0.53, respectively. This is the first study in which evolutionary information in form of position specific scoring matrix (PSSM) profile has been successfully used for predicting RBPs. We achieved the maximum MCC of 0.62 using SVM model based on PSSM called PSSM-400. Finally, we developed different hybrid approaches and achieved maximum MCC of 0.66. We also developed a method for predicting three subclasses of RNA binding proteins (e.g., rRNA, tRNA, mRNA binding proteins). The performance of the method was also evaluated on an independent dataset of 69 RBPs and 100 non-RBPs (NBPs). An additional benchmarking was also performed using gene ontology (GO) based annotation. Based on the hybrid approach a web-server RNApred has been developed for predicting RNA binding proteins from amino acid sequences (http://www.imtech.res.in/raghava/rnapred/).

Supporting information can be found in the online version of this paper.

Keywords: evolutionary information; hybrid prediction method; RNA interacting residues; RNA binding proteins; SVM

INTRODUCTION

RNA participates in several essential and diverse functions of cell. It is a constituent part of ribosome (Moore, 1998), spliceosome (Luhrmann *et al*., 1990) and also reported to have catalytic activity (Scott and Klug, 1996). In most cases, interaction of RNA with protein is the common phenomena of RNA functions. Proteins tend to interact with RNA mainly at the secondary structure elements, such as stem-loops and bulges to form the complex (Nagai, 1996). Besides, non-Watson-Crick base pairing also plays a critical role in interaction (Steitz, 1999). RNA recognition by proteins is primarily mediated by certain classes of RNA-binding domains and motifs that are divided into two main classes on the basis of RNA recognition mode: (i) groove binding and (ii) β-sheet binding. In the former, interaction occurs through binding of α-helix or loop with groove of RNA helix whereas in later, residues in β-sheets interact with unpaired RNA bases (Draper, 1999). In spite of the apparent functional importance, protein–RNA interaction has attracted lesser attention than protein–DNA interaction. This has been evidenced with the keyword search, "protein–DNA complex" and "protein–RNA complex" in PDB (Berman *et al*., 2000). Out of approximately 54 000 structures deposited, about 966 and 384 structures have been retrieved for DNA and RNA complexes, respectively.

The most common method to identify the class of unknown proteins is searching of structurally or sequentially similar proteins. Similarity-searching based methods are limited by the absence of experimentally annotated homologous proteins in database. Hence it is important to develop computational tools to identify RNA-binding proteins using amino acid sequence alone. In past, attempts have been made to predict RNA-binding proteins using amino acid composition coupled with charge, hydrophobicity, and accessible surface area of residues (Cai and Lin, 2003). Han *et al* (2004) developed a SVM based method using three types of descriptors to describe the global composition of each protein, (i) composition: percent compositions of amino acids of a particular property (such as hydrophobicity), (ii) transition: percent frequency with which amino acids of a particular property followed by amino acids of a different property, and (iii) distribution: amino acids of a particular property within the first 25, 50, 75, and 100% of the protein (Han *et al*., 2004). Another sequence-based method is developed by Yu *et al.* (2006) that

* *Correspondence to: G. P. S. Raghava, Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh-160036, India. E-mail: raghava@imtech.res.in*

a *M. Kumar*
Department of Biology, McGill University, Montreal, QC, H3A 1B1, Canada

b *M. M. Gromiha*
Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology, AIST Tokyo Waterfront Bio-IT Research Building, 2-4-7 Aomi, Koto-ku, Tokyo 135-0064, Japan

c *G. P. S. Raghava*
Bioinformatics Centre, Institute of Microbial Technology, Sector 39A, Chandigarh-160036, India

predicts rRNA, RNA-, and DNA-binding proteins. They used the three types of global composition descriptors (composition, transition, and distribution), earlier used by Han *et al* (2004), along with the physicochemical properties of proteins (hydrophobicity, predicted secondary structure, predicted solvent accessibility, normalized Van der Waals volume, polarity, and polarizability) as input for SVM. Using the jack-knife test they achieved 83.98%, 77.51%, and 71.64% accuracy of prediction for rRNA-, RNA-, and DNA-binding proteins, respectively. In order to show the consistency of prediction, they also carried out the self-consistency test (SVM was trained and tested with the same data set) of the SVM models and observed an average accuracy of 92.84%, 83.21%, and 74.37% for rRNA-, RNA-, and DNA-binding proteins, respectively. Recently, Shazman and Mandel-Gutfreund developed a method based on structural properties of RBPs (Shazman and Mandel-Gutfreund, 2008). Despite the availability of several methods, identification of RNA-binding proteins using sequence information with high accuracy is still a major challenge.

In the present study, a systematic attempt has been made to develop models for predicting RBPs with high accuracy using only sequence-based information. We have evaluated the performance of similarity search methods BLAST and PSI-BLAST as well as the method developed for predicting RNA-binding residues, PPRINT (Kumar *et al*., 2007a), to identify the RBPs with specific thresholds. We have observed that although PPRINT achieved very high specificity, the sensitivity was low. Hence, SVM modules were developed using different composition forms (single amino acid, dipeptide, and four part composition) and PSSM profile. We also evaluated the performance of hybrid methods developed using different combinations. Among different types of RNAs, mRNA, rRNA, and tRNA constitute a major fraction. In all the three types of RBPs, no exclusive RNA-binding motif is present. In several cases same motif (such as the KH motif or the zinc finger motif) was found in proteins that binds to different class of RNA (Lunde *et al*., 2007). Hence, we also developed a sub-classification method to differentiate between different classes of RBPs using multi-class SVM approach.

MATERIALS AND METHODS

Datasets

Main dataset

Yu *et al* (2006) have created a dataset of sequences obtained from Swiss-Prot. It has 88 rRNA, 377 RNA, and 1153 DNA-binding protein sequences in addition to 17 779 sequences that do not belong to either class (negative dataset). Yu *et al*. generated this non-redundant dataset with the sequence identity of less than 25% (using CD-HIT (Li *et al*., 2001) and PISCES (Wang and Dunbrack, 2003)) and removed ambiguous proteins, such as proteins having more than 6000 or less than 50 amino acid residues, or having irregular amino acids (e.g., 'X' and 'Z') (Yu *et al*., 2006). We used this dataset for developing different models described in this paper, which is called as *main dataset* henceforth. In other words the main dataset was a subset of Yu's dataset, having 377 RBPs, and equal number of randomly selected non-binding proteins (NBPs) from negative dataset. Among the 377 RBPs 64 binds to mRNA, 82 to rRNA, 23 to tRNA, and 208 either binds to more than one class of RNAs or specific class of binding information is not available (for details see Table S1).

RNA-binding proteins sub-classification dataset

RNA-binding proteins can be sub-classified into different categories such as, rRNA-, tRNA-, mRNA- snRNA, snoRNA-, miRNA-, and siRNA-binding proteins. Hence, it is pertinent to develop a method that can further sub-classify the predicted RNA-binding proteins into a potential sub-class. In the proposed work, we have confined ourselves with only three classes of RBPs namely rRNA-, tRNA-, and mRNA-binding proteins. In their work, Yu *et al* (2006) have compiled three classes of proteins (rRNA-, RNA-, and DNA-binding proteins and we used the 88 rRNA-binding proteins for sub-classification. The remaining two classes of proteins (tRNA- and mRNA-binding proteins) were retrieved from the Swiss-Prot database. We got a total of 413 and 235 mRNA- and tRNA-binding proteins, respectively from Swiss-Prot. After removing proteins having less than 50 residues and more than 6000 amino acids, we created the final dataset with less than 25% sequence identity using PISCES web-server (Wang and Dunbrack, 2003). Finally, we got 83 tRNA- and 64 mRNA-binding proteins.

Independent dataset

We used two different independent datasets to evaluate the real life efficiency of our method.

RNAiset1: We have randomly selected 100 Swiss-Prot sequences that do not interact with DNA or RNA. In addition, 107 RNA-binding protein chains were added from protein-RNA complexes of PDB. Originally both sets of sequences were compiled by Wang and Brown to evaluate their DNA/RNA binding residue prediction web-server BindN (Wang and Brown, 2006). We noticed that out of 107 RNA-binding protein chains 38 were present in the dataset used to develop PPRINT. We removed these 38 common sequences. Hence the final *RNAiset1* dataset has 69 RNA-binding protein chains.

RNAiset2: The main purpose for developing the present method is to use it for prediction of RBPs in different available proteomes. In order to show the real life efficiency we used GO annotation (Ashburner *et al*., 2000) as benchmark. We created a new independent dataset having 100 RBPs and 1000 NBPs. The pfam2go section of GO database (http://www.geneontology.org/external2go/pfam2go) contains mapping of each Pfam domain to the GO terms. It means pfam2go can give information about the function of Pfam domains. Firstly, we collected all Pfam domains that contain GO term having RNA-binding property. Then from UniProt database (http://www.uniprot.org/) we collected protein sequences that have any one of the RNA- binding Pfam domains (collected in the previous step). We randomly selected 100 proteins having RNA-binding Pfam domains, taking care that no domain was present more than once. For making the non-RNA-binding protein dataset we collected 1000 sequences from UniProt that do not have Pfam domains found in RBPs. Both these classes of sequences are given as supplementary data.

Performance measures

In the present work we used standard parameters like sensitivity, specificity, accuracy, and MCC (Matthews, 1975) for evaluating the models. The detailed descriptions of these parameters have been described in our earlier publication (Kumar *et al*., 2006). In addition threshold independent measure receiver operating characteristic (ROC) and area under curve (AUC) were also used to evaluate the performance of models.

Journal of Molecular Recognition

RNA-binding amino acid prediction

Prediction of RNA-binding amino acids was done using our earlier developed method PPRINT (Kumar *et al*., 2007a). Each protein was submitted to PPRINT web-server (http://www.imtech.res.in/raghava/pprint) and fraction of RNA-binding amino acids was calculated by dividing the number of binding residues with total amino acid residues in protein. In order to avoid the false positive RNA-binding residues prediction, a higher threshold (0.5) was used, instead of the default threshold −0.2.

Five-fold cross validation

All models except RNA-binding residue prediction based modules developed during this study were evaluated using five-fold cross validation, in which binding as well as non-binding proteins were randomly divided into five sets. Each set contain one fifth of RNA-binding and one fifth of non-RNA-binding proteins. All the models developed in this study were trained on four sets and tested on remaining fifth set. This process is repeated five times so that each set is used once for testing. The final performance of a model is average performance of five sets.

During the development of sub-classification SVM models we considered the corresponding proteins as positive example and other class of RBPs as negative example. Besides this, equal number of NBPs (Negative data of main dataset) was also added. It means SVM models for mRNA-, tRNA-, and rRNA-binding protein predictions were developed on unbalanced dataset. For example SVM model for rRNA-binding protein was created using a dataset of 88 rRNA-binding protein, 64 mRNA-, 83 tRNA-binding, and 88 NBPs. Similarly SVM model for tRNA-binding protein prediction was developed on 83 tRNA-binding proteins and 235 non-tRNA-binding proteins (83 NBP, 64 mRNA-binding, 88 rRNA-binding proteins). Further, mRNA-binding protein prediction model was trained on 235 negative (64 NBP, 88 rRNA-, and 83 tRNA-binding proteins) and 64 mRNA-binding proteins.

BLAST and PSI-BLAST

In the present study, BLAST (Altschul *et al*., 1990) and PSI-BLAST (Altschul *et al*., 1997) based searching was done against a database of RBPs and NBPs. The PSI-BLAST was used in addition to BLAST, as it is capable to search remotely homologues proteins. In this study PSI-BLAST search was done for three iterations at a cut-off e-value of 0.001. In order to distinguish this searching from general BLAST and PSI-BLAST searching, the database contains only the sequences of main dataset. Proteins of test set were searched against the corresponding training set and depending on the class of the top-most hit, the search performance was calculated.

Sequence feature and vector encoding

Residue compositions

In this study, SVM based models have been developed using (i) percent amino acid composition in which each protein was represented by a vector of dimension 20; (ii) percent dipeptide composition that represented each sequence by a vector of dimension 400; (iii) four-parts composition where sequence was divided into four equal non-overlapping parts and composition of each part was calculated separately. Thus, a protein was represented by a vector of dimension 80 (20×4) in case of four-part composition. The detailed descriptions of amino acid and dipeptide compositions have been described in our earlier work (Kumar *et al*., 2006).

Evolutionary information

In past, evolutionary information in the form of PSSM was used in secondary structure prediction (Jones, 1999), turns (Kaur and Raghava, 2003a; Kaur and Raghava, 2003b; Kaur and Raghava, 2004a; Kaur and Raghava, 2004b), and β-hairpin prediction (Kumar *et al*., 2005). PSSM based prediction methods perform better than single sequence-based methods because it also provides the patterns in sequence variability and the location of insertions and deletions along with information of amino acid sequence. In this study, first time we have used evolutionary information in the form of PSSM for predicting RBPs. We performed three iterations of PSI-BLAST search against NCBI 'nr' (non-redundant) protein database, which is a non-redundant collection of GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF. The e-value threshold for inclusion of sequences during profile construction was 0.001. All other parameters of PSI-BLAST search were default parameters. The PSSM contains occurrence probability of all 20 amino acids at each residue position of protein sequence. This means evolutionary information in PSSM is presented by a matrix of dimension L^*21 (L rows and 21 columns) for a protein of length L. Here 20 columns represent occurrence/substitution of 20 naturally occurring amino acids and remaining one column for insertion/deletion. One of the major limitations of machine learning algorithms is the requirement of fixed length input patterns. In proteins, the number of amino acids (L) is not the same; hence we can not use PSI-BLAST PSSM directly to train SVM. In order to convert variable size $L \times 21$ dimension PSSM into fixed size 400 dimension input vector (PSSM-400) strategy used in DNA-binding proteins prediction method DNA binder was adopted (Kumar *et al*., 2007b). In short it was done in following steps: (a) all values of PSSM were normalized in range of 0–1 using formula (Value-minimum)/(maximum-minimum); (b) All rows belonging to the same amino acid were pooled together to form 20 matrices of size $N_{AA} \times 20$, where N_{AA} is the number of amino acid of type AA; (c) the summation of each column in new matrices (each daughter matrix) will produce a 20 dimensional vector. Since there were 20 matrices we get $20 \times 20 = 400$ dimension vector.

Support vector machine

SVM is a machine-learning method based on the structural risk minimization principle from statistical learning theory. It takes a set of feature vectors, along with their real output as input and use them for training of model. After training, learned model can be used for prediction of unknown examples. A detailed description of SVM is available in Vapnik (Vapnik, 1995). In this work, the SVM training has been carried out by optimization of various kernel function parameters and the value of the regularization parameter C. We have used a freely downloadable package SVMlight available at http://svmlight.joachims.org/to implement SVM in this work.

Hybrid approach

We also developed methods which combines two or more than two approaches called hybrid methods. We combined residue based prediction method PPRINT and composition based SVM model called as hybrid modules.

Hybrid1 Module: In this module, first RNA binding residues are predicted using PPRINT, if the percentage of predicted residues in a protein is more than the specific threshold then the protein is assigned as a RNA-binding protein. In case percentage of predicted residues in the protein is less than threshold then PSSM-400 based SVM model is used to predict whether protein is RNA-binding or not.

Hybrid2 Module: In this module, if a protein have percent of binding residue (predicted using PPRINT) more than upper threshold then protein is assigned as RBP; if a protein have percent of binding residue less than lower threshold then protein is assigned as NBP; otherwise SVM model is used to predict whether protein is RBP or NBP.

Hybrid3 Module: In this module PSSM-400 and similarity based approaches (e.g., BLAST, PSI-BLAST) are combined. The proteins in the test sets were searched as query against a database containing the proteins of corresponding training sets. BLAST/PSI-BLAST hits were given the priority over PSSM-400; SVM predictions were used only for proteins where BLAST/PSI-BLAST showed insignificant hits.

Hybrid4 Module: In this module, similarity search methods (BLAST and PSI-BLAST) are combined with Hybrid2 module. Firstly RNA-binding residue prediction method PPRINT was used to predict the RNA-binding residues in the query sequence. On the basis of percentage of RNA-binding residue in the query protein, it was predicted as RBP or NBP. In case the fraction of predicted residues lie between the lower and upper threshold limits then BLAST/PSI-BLAST search was used in five-fold cross validation mode. On the basis of class of top hit, the class of query protein was decided. In case BLAST/PSI-BLAST search also did not find any hit then SVM was used for prediction. In short first priority was given to PPRINT prediction, second to similarity search methods, and third or lowest to SVM. In Hybrid4 method, lower and upper threshold limits were kept at 3 and 30%, respectively, while BLAST/PSI-BLAST we used an e-value of 1e-4.

RESULTS

RNA-binding residue prediction based approach

Recently, we developed a method PPRINT for predicting RNA-binding residues (Kumar *et al*., 2007a). Using PPRINT, RNA-binding residues were predicted for all the proteins in the main dataset and RBPs were assigned with different thresholds. In this approach if fraction of predicted RNA-interacting residues in a protein is greater than a certain threshold then it was predicted as a RNA-binding protein. As shown in Table 1, with the threshold limit of 30% we predicted all RBPs with 100% specificity. However, the sensitivity was very low (18.83%). This result showed that the interacting residue prediction approach could discriminate RNA binding and non-binding proteins with very high specificity but low sensitivity.

Prediction by similarity search methods

We evaluated the performance of similarity search approaches using BLAST and PSI-BLAST on main dataset using five-fold cross validation approach. Proteins of each test set were searched against the corresponding training set sequences. As shown in Table 2, at e-value 1e-4, BLAST got 129 hits out of total 377 RNA-binding proteins, among which 123 were correct (coverage 32.63%). On the other hand, 371 hits were obtained at e-value 10, among which only 288 were correct. PSI-BLAST should be able to identify even remotely homologous proteins because it searches the sequence database iteratively and builds the scoring matrix on the basis of sequences found in previous searches. But in this study minor difference was observed in performance of BLAST and PSI-BLAST. At e-value of 1e-4 and 10, the coverage of PSI-BLAST search was 33.16 and 77.45%, respectively (Table 2).

Table 1. Performance of PPRINT to identify RNA-binding proteins at different thresholds. The % cut-off implies the minimum fraction of residues predicted as binding

% Cut-off	Sensitivity	Specificity	Accuracy	MCC
5	77.72	71.09	74.40	0.49
10	53.85	91.25	72.55	0.49
15	36.87	96.82	66.84	0.42
20	28.12	99.47	63.79	0.39
25	24.93	99.73	62.33	0.37
30	18.83	100.00	59.42	0.32
35	14.32	100.00	57.16	0.28
40	10.61	100.00	55.31	0.24
45	6.90	100.00	53.45	0.19

Amino acid composition analysis

It has been shown that the amino acid composition can be used to classify proteins of different folding types and develop prediction methods. Hence we analyzed the amino acid compositions of RNA-binding and non-binding proteins of main dataset. We observed that the charged residues (Glu, Lys and Arg) were more abundant in RNA-binding proteins (Figure 1). It was expected because of the importance of electrostatic interactions in protein–RNA interactions. The higher level of Glu can also be explained due to the negative charged nature of RNA. Interestingly, no significant difference was found in composition of other negative charged amino acid Asp. In non-binding proteins Cys, Phe, Ile, Leu, Trp, and Tyr were found to be over-represented. We also carried out the calculation for statistical significance of amino acid composition of Glu, Lys, Arg, Cys, Phe, Ile, Leu, Trp, and Tyr at significance value $p < 0.0001$. We found that the differences in composition of all amino acid residues are statistically significant.

Composition based SVM models

The results obtained with similarity search methods and binding site prediction approach clearly showed that they are not capable of discriminating RNA-binding proteins at high accuracy. Hence, we developed SVM models that require only amino acid sequence for prediction. Using amino acid composition we achieved the accuracy of 79.97% and MCC of 0.60. The MCC decreased to 0.46 (accuracy = 73.09%) with dipeptide composition. We obtained the MCC of 0.53 (accuracy = 76.53%) using four-part amino acid composition, which is lower than the performance of simple amino acid composition based SVM model. We also calculated the AUC statistics of different SVM models. The AUC values of amino acid, dipeptide, and 4-parts split amino acid compositions were 0.85, 0.80 and 0.83, respectively. The detailed performance of all composition based SVM modules are shown in Table 3 and S2.

Journal of Molecular Recognition

Table 2. Performance of similarity search methods BLAST and PSI-BLAST on RNA-binding proteins in main dataset at different e-value thresholds

	RNA binding proteins				Non RNA binding proteins			
E-value	TH	CH	Prob	Cov	TH	CH	Prob	Cov
				Summary of BLAST search				
1e-4	129	123	95.35	32.63	16	11	68.75	2.92
1e-3	137	129	94.16	34.22	19	11	57.89	2.92
1e-2	163	151	92.64	40.05	27	16	59.26	4.24
0.1	200	182	91.00	48.28	49	25	51.02	6.63
1	281	238	84.70	63.13	161	81	50.31	21.49
10	371	288	77.63	76.39	358	168	46.93	44.56
				Summary of PSI-BLAST search				
1e-4	131	125	95.42	33.16	16	11	68.75	2.92
1e-3	146	138	94.52	36.60	20	11	55.00	2.92
1e-2	165	153	92.72	40.58	28	16	57.14	4.24
0.1	203	184	90.64	48.81	55	28	50.91	7.43
1	284	242	85.21	64.19	197	94	47.72	24.93
10	375	292	77.87	77.45	371	176	47.44	46.68

TH = Total number of hits obtained during similarity searching; CH = Number of proteins whose top hit is same as class of query protein; Prob = Probability of correct prediction; Cov = Coverage of searching [(CH/total number of searches) × 100]. Total number of proteins searched is 377 RBPs and 377 NBPs.

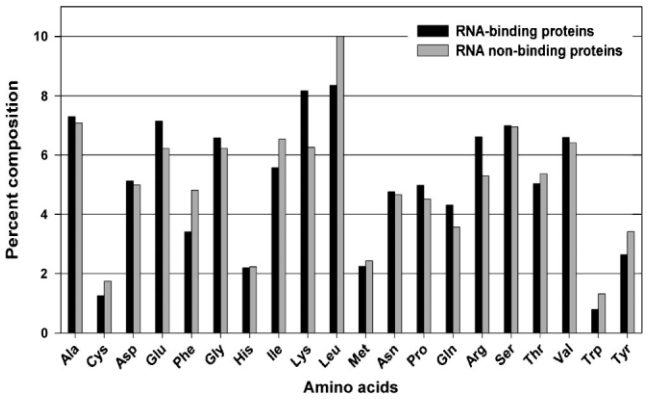

Figure 1. Amino acid composition of RNA-binding and non-binding proteins in main dataset.

Evolutionary information based SVM model

To further enhance the prediction performance, evolutionary information encoded in PSSM was used as input to SVM. PSSM was generated by PSI-BLAST search against "NCBI nr protein database" and normalized into 400 dimension input vector. We achieved the maximum MCC of 0.62 with an accuracy of 80.90% using PSSM (Table 3 and S2). The ROC plot (Figure 2) and AUC (Table 3) also show that PSSM based SVM model performs better than composition based SVM models.

Hybrid approach

As shown in Table 1, PPRINT was very effective in discriminating NBPs. In order to take advantage of high specificity obtained with PPRINT; we combined it with PSSM-400 SVM model. In the hybrid method if 30% residues of a protein were predicted as RNA interacting residues then protein was predicted as RBP. For remaining proteins PSSM-400 model was used to predict type of protein. Using this approach (Hybrid1) the MCC marginally improved to 0.63 (Table 4). Since RNA-binding residue prediction approach was very efficient in discriminating the NBPs we also imposed a lower threshold limit henceforth referred as Hybrid2 method. In the Hybrid2 if the % of predicted RNA binding residues in a protein is less than lower threshold limit (e.g., 1, 2, 3, 4, 5 in Table 5) it was predicted as non-binding and if it is above the upper limit (30%) it is predicted as binding. For the proteins having percent of predicted residue in between lower and upper limits we used PSSM-400 SVM model for predicting type of

Table 3. SVM results on main dataset

Input	Thr.	Sens. (%)	Spec. (%)	Accu. (%)	MCC	SVM_light learning parameters	AUC
AAC	0.4	80.63	79.31	79.97	0.60	$J=2$; $t=2$; $g=0.001$; $c=75$	0.85
DP	0.0	74.03	72.15	73.09	0.46	$J=1$; $t=1$; $d=2$	0.80
4-AAC	0.0	76.67	76.38	76.53	0.53	$J=1$; $t=1$; $d=4$	0.83
PSSM-400	−0.2	81.95	79.84	80.90	0.62	$J=1$; $t=1$; $d=1$; $c=0.01$	0.87

4-ACC = 4 part amino acid composition; DP = dipeptide composition; PSSM-400 = PSSM normalized in the form of 400 input vectors; Thr = Threshold; Sens = Sensitivity; Spec = Specificity; Accu = Accuracy; MCC = Matthews correlation coefficient; AUC = Area under curve.

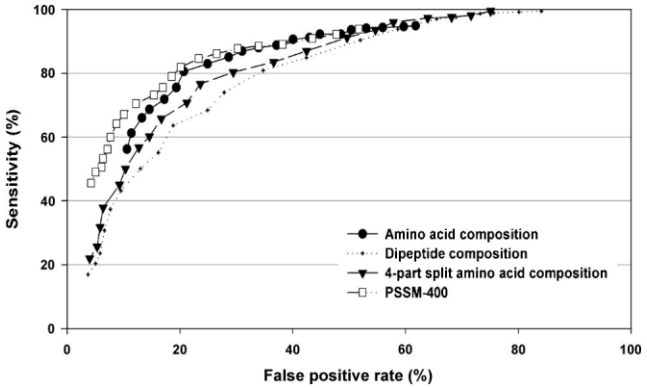

Figure 2. ROC plot analysis of SVM prediction with different inputs on main dataset.

proteins. As shown in Table 5 we achieved the maximum MCC of 0.66 with lower limit 3%. With this approach the specificity increased considerably to 88%.

As shown in Table 2 when RBPs are searched with five fold cross validation mode at e-value of 1e-4, BLAST has picked up 123 RBPs correctly out of 129 retrieved hits. On the other hand PSI-BLAST identified 125 RBPs correctly out of 131 total hits. These results showed that the coverage of similarity search methods is poor although the probability of finding a correct homologue is very high (~95%). Hence we combined the similarity search methods BLAST and PSI-BLAST with PSSM-400 SVM model (henceforth referred as Hybrid3). In contrary to our expectation we did not find improvement in MCC with Hybrid3 (Table S3).

We have taken into the advantage of other methods, (i) highly specific prediction of RNA-binding residue based approach, (ii) high probability of correct prediction of similarity search methods and (iii) generalized SVM method, and developed a hybrid method by combining all these three methods (henceforth referred as Hybrid4). With Hybrid4 we achieved the maximum MCC of 0.66 (Table S4), which was equal to the performance of Hybrid2 method.

Table 4. The performance of combined approach (Hybrid1), prediction of RNA binding residues and PSSM-400 SVM model

PSSM-400 SVM threshold	Sensitivity	Specificity	Accuracy	MCC
−1.0	93.90	48.01	70.95	0.47
−0.9	92.31	52.25	72.28	0.49
−0.8	90.98	56.50	73.74	0.51
−0.7	89.12	61.80	75.46	0.53
−0.6	88.59	66.05	77.32	0.56
−0.5	87.80	69.76	78.78	0.59
−0.4	86.47	73.47	79.97	0.60
−0.3	84.88	76.66	80.77	0.62
−0.2	*83.29*	*79.84*	*81.56*	*0.63*
−0.1	80.64	81.43	81.03	0.62
0.0	77.45	83.02	80.24	0.61
0.1	75.07	84.62	79.84	0.60
0.2	72.68	87.80	80.24	0.61
0.3	69.23	89.92	79.58	0.60
0.4	66.58	91.25	78.91	0.60
0.5	62.86	92.31	77.59	0.58
0.6	60.48	92.84	76.66	0.56
0.7	58.62	93.63	76.13	0.56
0.8	56.50	93.90	75.20	0.54
0.9	55.44	94.96	75.20	0.55
1.0	53.05	95.76	74.40	0.54

In the hybrid method, if predicted RNA-interacting residues in a protein was more than or equal to 30% of total amino acids, then it was predicted as RBP. Otherwise PSSM-400 based SVM model was used for prediction.

Table 5. The performance of hybrid method (Hybrid2) with different minimum % threshold for non-RNA binding proteins prediction. The threshold for RBPs prediction was 30% of total amino acids was predicted as RNA-interacting by PPRINT

% Minimum cut-off	Sensitivity	Specificity	Accuracy	MCC
1	82.23	82.49	82.36	0.65
2	79.84	84.88	82.36	0.65
3	77.72	88.06	82.89	0.66
4	73.21	89.92	81.56	0.64
5	70.56	91.78	81.17	0.64

Classification of RNA-binding proteins into different classes

In addition, we have developed a method to distinguish different classes of RNA (tRNA, rRNA, or mRNA)-binding proteins using multiclass SVM approach. We built three new SVM classifiers, which were used to discriminate rRNA-binding proteins, mRNA-binding proteins, and tRNA-binding proteins. As shown above, PSSM-400 model perform better than other composition-based SVM models and hence we developed PSSM-400 based model for sub-classification. Using tRNA-binding proteins as positive example we achieved the accuracy of 76.70% and MCC of 0.49 (Table 6). The accuracy and MCC increased to 79.28% and 0.50, when mRNA-binding proteins were considered as positive example. The accuracy and MCC for classifying rRNA-binding proteins are 90.14% and 0.77, respectively. The AUC values of tRNA, mRNA, and rRNA-binding protein prediction SVM model were 0.82, 0.83, and 0.96, respectively. The ROC plots of these also show high performance (Figure S1).

All the three SVM models for predicting different class of RNA-binding proteins were developed using the same dataset. The only difference between them is that they were developed for predicting only a subset of RBPs. Hence it is pertinent to analyze the rate of false positive predictions with different classes of RBPs. As shown in Table 7, when SVM model developed on tRNA-binding proteins was used for predicting mRNA- and rRNA-binding proteins our method correctly predicted 47 and 83 proteins, respectively, as non-tRNA-binding proteins. Similarly, the SVM model trained for predicting mRNA-binding proteins correctly excluded 80.72% and 94.32% of tRNA- and rRNA-binding proteins, respectively. SVM model developed for predicting rRNA-binding proteins correctly identified 91.57 and

Journal of Molecular Recognition

Table 6. Sub-classification of RNA-binding proteins using PSSM-400 SVM approach

	Thr.	Sens. (%)	Spec. (%)	Accu. (%)	MCC	SVM_light learning parameters	AUC
tRNA	−0.30	75.88	76.98	76.70	0.49	$J=2$; $t=2$; $g=0.0001$	0.82
mRNA	0.00	77.95	79.61	79.28	0.50	$j=5$; $t=2$; $g=0.0001$	0.83
rRNA	0.40	89.93	90.22	90.14	0.77	$j=4$; $t=2$; $g=0.001$	0.96

Thr = Threshold; Sens = Sensitivity; Spec = Specificity; Accu = Accuracy; MCC = Matthews correlation coefficient; AUC = Area Under Curve.

90.63% non rRNA-binding proteins (tRNA-binding and mRNA-binding proteins, respectively).

Evaluation on independent dataset

N-fold (e.g., five-fold) cross validation is a standard practice used for evaluating prediction methods, where set of proteins used for training and testing are mutually exclusive. But it has been observed in past that performance of *N*-fold cross validation technique is biased with optimization (Bhasin and Raghava, 2004b). Thus, it is important to evaluate a newly developed method on an independent dataset (proteins not used for developing the method). We evaluated the performance of our best SVM model as well as hybrid approaches on an independent dataset (*RNAiset1*). Among different hybrid methods we selected Hybrid2 for independent testing since it showed the maximum accuracy along with Hybrid4. As shown in Table 8, at default threshold PSSM-400 SVM model predicted with MCC and accuracy of 0.45 and 73.37%, respectively. When Hybrid1 method was applied on independent dataset the sensitivity increased to 73.91% and MCC to 0.49. With Hyrid2 method there was minor decrease in sensitivity but specificity increased by 3% with MCC of 0.51. These results clearly show that our prediction model performs well in real life situation and the method is reliable. We named the Hybrid2 method as RNApred, which was the final prediction method for RBPs.

Independent evaluation of RNApred on GO annotation

Recent developments in high-throughput techniques resulted in accumulation of large amount of functional data. Gene ontology consortium (www.geneontology.org) is an attempt to systematically organize different forms of data in a hierarchical manner (Ashburner *et al*., 2000). GO represents the function of each gene using three main ontologies—molecular function, biological process, and cellular component. In this work, we evaluated the performance of RNApred on independent dataset (*RNAiset2*) that had 100 RNA-binding and 1000 non-binding proteins. At default thresholds of Hybrid2 module of RNApred, it correctly predicted 73 proteins as RNA-binding and 833 proteins as non-binding (see supplementary file 2 for detail).

Comparison with existing methods

A direct comparison with all existing methods is not appropriate because of the differences in the data sets, descriptors, and classification methods. But it is important to compare a newly developed method with existing methods in order to examine the novelty of new method. To the best of our knowledge four methods were reported in the literature for predicting RNA-binding proteins. Cai and Lin developed a rRNA-, RNA-, and DNA-binding proteins prediction method using 10 fold cross-validation (Cai and Lin, 2003). They achieved 96.84 and 85.74% accuracy in discriminating rRNA- and RNA-binding proteins, respectively. In a first glance it appears that the accuracy is higher than our method, but it does not definitely means that their method is superior to ours. There were two main reasons behind the high accuracy. Firstly, they considered Swiss-Prot annotation regardless of the fact that the function had been determined experimentally or predicted and secondly, they have not removed the redundancy in training dataset. Han *et al* also used SVM for predicting rRNA-, mRNA-, tRNA-, and all RNA-binding proteins (Han *et al*., 2004). The predicted sensitivity reported for each class of proteins was 94.1, 79.3, 94.1, and 97.8%, respectively. On the other hand the specificity achieved for each group of non-RNA-binding proteins, is 98.7, 96.5, 99.9 and 96.0%, respectively. Similar to performance reported by Cai and Lin (2003) the performance reported by Han *et al* is higher than our method.

Table 7. Specificity of different SVM sub-classification models on remaining two classes of RNA-binding proteins

Test protein binding class	True negative	Specificity (%)
tRNA-binding protein prediction SVM model		
mRNA	47	73.44
rRNA	83	94.32
mRNA-binding protein prediction SVM model		
tRNA	67	80.72
rRNA	83	94.32
rRNA-binding protein prediction SVM model		
tRNA	76	91.57
mRNA	58	90.63

Table 8. Performance of different RBPs prediction methods developed in this study using an independent dataset (*RNAiset1*)

	Sensitivity	Specificity	Accuracy	MCC
PSSM-400	69.57	76.00	73.37	0.45
Hybrid1	73.91	76.00	75.15	0.49
Hybrid2	72.46	79.00	76.33	0.51

Hyrbid1 corresponds to method in which threshold limit of 30% was used for RNA-binding protein prediction. Hybrid2 refers to the method in which both lower (3%) and upper (30%) threshold limits were used for prediction.

However, Han *et al* also did not follow the established way (making the dataset non-redundant, removing sequences whose existence or function (here RNA-binding property) is not experimentally proven) of making dataset for a prediction method. They have not specified the degree of similarity among training dataset sequences. Rather they mentioned that all distinct members in each group were used to construct positive samples for training, testing, and independent evaluation of the SVM classification system (Han *et al*., 2004). The structural and electrostatic feature based method of Shazman and Gutfreund (2008) was able to predict RBPs from NBP with very high accuracy. The SVM model developed to classify between RBPs to NNBP (non-nucleic acid binding protein chains) achieved 80% sensitivity and 90% specificity (MCC, 0.67). They also developed a SVM model to differentiate RBPs from NNBP having large electrostatic patch with 80% sensitivity and 91% specificity (MCC, 0.72). Though the performance of Shazman's method was very high it is of limited utility due to the requirement of 3D structure, which is not directly available for most of the sequences. As described earlier, our method has achieved the maximum accuracy of 81% using PSSM-400 SVM model, which is better than that obtained (77.5%) by Yu *et al.* on the same dataset (Yu *et al*., 2006). The main difference is the development of Hybrid1 and Hybrid2 methods. Incorporating the RNA-binding residue prediction of PPRINT further increased the accuracy to ~83%.

Benchmarking of existing RBP prediction methods on independent dataset

There is only one publicly available web-server for predicting RBPs, which is based on the method developed by Han *et al*. (2004). So, it would be interesting to benchmark the RNAPred vis-à-vis SVMProt using an independent data. For this we have submitted the *RNAiset1* proteins to SVMProt. Out of 69 RBPs, SVMProt correctly predicted 42 as RBPs with the sensitivity of 61%. Out of 100 NBPs, 94 were predicted correctly as NBP. This shows that SVMProt has very high specificity of prediction.

Description of RNApred Web-server

Based on amino acid composition and PSSM-400 models and hybrid method, we developed a web-server RNApred for identifying RNA-binding proteins from amino acid sequences. RNApred is freely accessible from http://www.imtech.res.in/raghava/rnapred/. The common gateway interface (CGI) script of RNApred is written using PERL version 5.03. The web-server is hosted on a Sun Server (420E) under UNIX (Solaris 7) environment. In order to do prediction, the users have to provide protein sequence in FASTA format. The server generates PSI-BLAST PSSM and then converts it to PSSM-400 input vector for prediction. At RNApred web-server there are three modes for RNA-binding protein prediction. First mode takes amino acid composition of submitted protein as input; second mode utilizes PSSM generated during PSI-BLAST search (PSSM-400) of query protein as input. The third option uses hybrid of PSSM-400 and PPRINT (referred as Hybrid2 method in this manuscript).

DISCUSSION

RNAs in cells are associated with different RNA-binding proteins in order to perform their functions. The RBPs influence the structure and interactions of the RNAs and play critical roles in their biogenesis, stability, function, transport, and cellular localization. Despite a very important role in cellular metabolism and regulation, RBPs got lesser attention than DNA-binding proteins. In the present work, we have developed SVM based methods for identifying RNA-binding proteins using a systematic approach. All modules developed during this work were based on a dataset of 754 proteins containing equal number of RBPs and NBPs. It was originally compiled by Yu *et al*. (2006) and contains proteins from all branches of life (Table S1).

Each RBP contains some amino acids that bind with the RNA. The RNA-binding residues should be more abundant in RBPs in comparison to NBPs. Hence RBP can also be predicted indirectly by RNA-binding residue prediction. Hence firstly we evaluated our method PPRINT developed for predicting RNA-binding amino acids in a protein. It was observed that this approach worked well as we expected. Binding residue based prediction approach can successfully discriminate between RBP and NBP; however, the sensitivity was poor and specificity was very high (Table 1). It was also observed that NBPs contain fewer binding amino acids.

Similarity search method is one of the most common approaches for predicting the function of a protein. In this method a query protein is searched against a database of annotated proteins and assigns the function of most similar target protein to query protein. Similarity based approach is highly accurate if an experimentally annotated homologous protein is found. We evaluated the performance of commonly used similarity searching tools, BLAST, and PSI-BLAST on our dataset using five-fold cross validation. The performance of both BLAST and PSI-BLAT was poor on our dataset, which indicates that the proteins in our dataset have low similarity to each other (Table 2).

To annotate a sequence with unknown molecular function, a biologist first search it against a sequence database such as NCBI 'nr' or UniProt for homologous sequences using sequence alignment methods BLAST or PSI-BLAST. On the basis of alignment score and e-value, homology between sequences is inferred. If query sequence is found homologous to any experimentally annotated database sequence then function of both sequences is considered same. The real problem arises when no homologous sequence is found during database searching. In this case predictive methods (for example machine learning) remains the only option. In other words, sequence similarity and machine learning methods are useful in different phases of sequence functional annotation. In the present work we used sequence similarity methods on a small set of equal number of binding and non-RNA binding proteins. The reason of using a small dataset is the simplicity in comparison among different methods of sequence functional annotation. Since performance of all approaches was calculated on same dataset, we can benchmark their strength and weakness without any biasing. Further the aforementioned dataset represents all RNA-binding proteins available in the sequence database. If BLAST can find a significant hit by searching on the small dataset, then it would also get when searched against full UniProt or NCBI nr databases. Further absence of a significant hit with a RBP shows that BLAST/PSI-BLAST will fail invariably even on a full-length sequence database. We also considered our previously developed RNA-binding amino acid prediction method PPRINT as an existing method of RNA-binding protein prediction because it can also be used for RBP prediction on the basis of RNA-binding amino acids. In order to reduce the false positive predictions of PPRINT we used higher threshold during prediction. As shown in Table 1, the

Journal of Molecular Recognition

PPRINT predicted very small number of RNA-binding residues in RBPs, which is, even less in non-binding proteins. But as shown in Table 1, we were able to predict nearly 78% RBPs. If we further try to increase the sensitivity, the proportion of false positive prediction will increase. Hence it was not logical to lower the percentage of PPRINT predicted RNA-binding residues. So, in further studies PPRINT prediction was only used as an initial filter to discriminate between binding and non-binding protein.

Since accuracy achieved by existing methods (PPRINT, BLAST, and PSI-BLAST) was poor we used the machine-learning technique, SVM for improving the accuracy. First, we developed SVM modules using various compositions such as amino acid composition, dipeptide composition, and four-part amino acid composition. Among different forms of compositions maximum MCC of 0.60 was achieved with simple amino acid composition (Table 3). It was contrary to previously reported fact that dipeptide and four-part amino acid composition achieve better accuracy due to greater information content (Bhasin and Raghava, 2004a; Bhasin and Raghava, 2004b; Garg *et al*., 2005; Kumar *et al*., 2006). Similar trend in MCC was also reported with DNA-binding protein prediction method developed by our group (Kumar *et al*., 2007b).

It has been reported in past that using evolutionary information as input vector of SVM can drastically increase the prediction accuracy (Kaur and Raghava, 2003b; Bhasin and Raghava, 2004a; Kaur and Raghava, 2004a; Kaur and Raghava, 2004b; Garg *et al*., 2005; Xie *et al*., 2005; Kumar *et al*., 2007b;). We extracted evolutionary information from the PSSM obtained from PSI-BLAST search against NCBI 'nr protein database' and normalized it into fixed length of 400 dimensional input vectors. Evolutionary information in the form of PSSM was first used for predicting the secondary structure of proteins (Jones, 1999). This brought considerable increment in the prediction accuracy. It is believed that PSSM contains more information than the single sequence input because it also contains the information of other residues at a position rather than giving the information of only a single amino acid at a particular position. For the protein secondary structure prediction no processing of PSSM was required due to the fixed pattern size. In the present work, prediction was done to the whole protein and not for a pattern. Since, SVM works only on fixed length input, we processed the $L \times 20$ dimension PSSM (L = number of amino acids in protein) into fixed length of 400 patterns by summing columns of identical amino acids. Hence the PSSM-400 input should have more information content than the amino acid composition. This is the first study that used evolutionary information for predicting RBPs. The MCC increased to 0.62 using PSSM profile (Table 3). This result agrees with the observations reported in the literature that PSSM provides more information than amino acids. However, in the present work the difference between the performance of amino acid composition and PSSM based SVM model (PSSM-400) is small.

As shown in Table 1 the binding residue based prediction approach achieved very high specificity, indicating the capability of predicting RBPs without any single false prediction. In order to benefit from high specificity we created two hybrid methods combining PSSM-400 and RNA-binding residue prediction methods. First, hybrid method (Hybrid1) was used to predict only RBP at a threshold at which we have not found any false NBP predicted as RBP. Using a threshold limit of 30% we observed that MCC increased from 0.62 to 0.63 (Table 4). During evaluation of PPRINT it was found that NBPs contain small number of binding residues that means that fraction of binding residue can be useful to screen out false positive prediction. Hence we developed a second hybrid method (Hybrid2), which further increased the MCC to 0.66 (Table 5).

The analysis on the results obtained with similarity search methods BLAST and PSI-BLAST showed that the coverage was quite low but the probability of correct prediction (fraction of proteins correctly predicted to the class to which they actually belong) was very high. This implies that if BLAST/PSI-BLAST is used along with generalized SVM based prediction method, accuracy should increase. Hence, we constructed a hybrid prediction method by combining BLAST and PSI-BLAST with PSSM-400 based SVM model (Hybrid3). In contrary to our expectation we did not observe improvement in accuracy (Table S3). Thereafter we also incorporated the binding residue prediction method into Hybrid3 to construct Hybrid4. Again no improvement was observed with Hybrid4 (Table S4). In order to find out the reason behind the failure of Hybrid3 and Hybrid4, we analyzed the performance of PSSM-400 SVM model on proteins, which got hit during BLAST and PSI-BLAST searches at e-value threshold of 1e-4 (129 and 131, respectively). It was found that out of 129 and 131 RBPs identified with BLAST and PSI-BLAST, respectively, 122 and 124 were also correctly predicted by SVM. It should be noteworthy that similarity search methods correctly identified 123 and 125 RBPs. On the other hand, out of 16 NBPs obtained with BLAST and PSI-BLAST, 12 were correctly predicted as NBPs by SVM. It shows that performance of SVM is very good with proteins, which are very similar to each other. Machine-learning methods like SVM are used from long time to develop prediction method that can be used even for proteins examples that do not share any similarity with each other because it models the data without considering the similarity between the samples. It also gave us insights behind a minor increase in sensitivity and slight decrease in specificity when similarity search methods were combined with SVM (Table S3). The findings of Table S5 can also be extrapolated to the reason of not getting any significant advantage of combining similarity search methods with RNA-binding residue prediction method and PSSM-400 based SVM model in Hybrid4 (Table S3 and S4).

Depending upon the category of RNA to which RBP binds, we also developed a multi-class SVM for predicting rRNA-, mRNA-, and tRNA-binding proteins (Table 6). We observed that the performance of predicting tRNA- and mRNA-binding proteins was approximately the same, but accuracy of rRNA-binding protein was about 10% point greater than the remaining two classes. We did not comprehend the reason for this difference as our method is an amino acid sequence based method and there is no exclusive RNA-binding domain for any of the three classes of RBPs. It may have occurred due to presence of some higher dimensional protein sequence feature that is common in tRNA- and mRNA-binding proteins, but different in rRNA-binding proteins. We also evaluated the prediction accuracy of SVM model trained with one group on another group of RNA-binding proteins. From the cross-prediction we observed that all three SVM models showed very high specificity (Table 7).

We benchmarked the performance of Hybrid2 method (called as RNApred) in two independent ways using two different datasets (see "Materials and Methods" for details). We found that on both datasets, performance was nearly equal with five fold cross-validation (Table 8 and supplementary file 2). This showed that RNApred is robust enough to work in real life condition.

On searching the literature we noticed four other existing methods for discriminating RBPs. Three methods were developed to predict RBPs on the basis of amino acid sequence alone. The first method was developed by Cai and Lin (2003). They reported 85.74% accuracy in RBPs prediction, which is probably over-prediction (see "comparison with existing methods"). In their work, Han *et al* (2004) reported accuracy higher than our method. But in their work, they also have not adopted the standard process of dataset creation. They took all the proteins, which optimally represent each family of RNA-binding proteins and used for training and testing. The availability of many redundant proteins should have enhanced the accuracy. The third method was developed by Yu *et al* (2006), which achieved the accuracy of 77%. Using the same dataset our method achieved the accuracy of 83%. Recently, a structure-based prediction method was also developed by Shazman and Gutfreund, which is of limited use due to the requirement of 3D structure (Shazman and Mandel-Gutfreund, 2008).

We have also benchmarked the performance of only publicly available RBP prediction server, SVMProt using *RNAiset1* dataset. We found that although sensitivity was lower, but specificity was very high. When we critically analyzed the prediction results of SVMProt, we found that it actually gave the probability of each protein belonging to a particular class. It means it lists the name of probable class to which the query protein may belong. For example out of the 42 true positives prediction, 8 were very low in the list. We have even found cases in which a protein was predicted to have "Plant defences property" as well as viral coat protein. So, in true sense, we can claim that RNAPred has no lower specificity than SVMProt.

CONCLUSIONS

The RNA-binding proteins play very important role in gene-regulation and expression. Hence prediction of RBPs can be an important step toward understanding the gene regulatory mechanism and their interactions. We developed a highly accurate method, RNApred for identifying RNA-binding proteins. Firstly, we evaluated the performance of similarity search methods and RNA-binding amino acid prediction method. Since both were not very effective in prediction, we developed SVM based method that requires only amino acid sequence as input. Maximum accuracy was achieved with PSSM-400 based SVM model. We also developed different hybrid methods using the similarity search, binding residue approach and PSSM-400 SVM model. It was observed that the hybrid method developed using binding residue prediction approach and PSSM-400 SVM model showed the best performance. A web-server RNApred has also been developed to make the prediction method available to the scientific community. We hope that RNApred would help to speed up the rate of protein function prediction.

Acknowledgements

We thank Mr. Nitish Kumar Mishra for help during manuscript revision. This work was supported by grants from Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Government of India. Manish Kumar was a senior research fellow of CSIR. This research article has IMTech communication number 15/2007.

REFERENCES

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402. gka562 [pii].

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410. 10.1006/jmbi. 1990. 9999 0022283680799990S[pii]

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT. others. 2000. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.* **25**: 25–29. 10.1038/75556.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242 gkd090 [pii]

Bhasin M, Raghava GPS. 2004a. Classification of nuclear receptors based on amino acid composition and dipeptide composition. *J. Biol. Chem.* **279**: 23262–23266 10.1074/jbc.M401932200 401932200M [pii]

Bhasin M, Raghava GPS. 2004b. ESLpred: SVM-based method for subcellular localization of eukaryotic proteins using dipeptide composition and PSI-BLAST. *Nucleic Acids Res.* **32**: W414–419. 10.1093/nar/gkh350 32/suppl_2/W414 [pii]

Cai YD, Lin SL. 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. *Biochim. Biophys. Acta.* **1648**: 127–133 S1570963903001122 [pii]

Draper DE. 1999. Themes in RNA-protein recognition. *J. Mol. Biol.* **293**: 255–270 10.1006/jmbi. 1999. 2991 S0022-2836 (99) 92991-1 [pii]

Garg A, Bhasin M, Raghava GP. 2005. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. *J. Biol. Chem.* **280**: 14427–14432 M411789200 [pii] 10.1074/jbc.M411789200

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

Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**: 195–202 10.1006/jmbi. 1999. 3091 S0022-2836 (99) 93091-7 [pii]

Kaur H, Raghava GPS. 2003a. A neural-network based method for prediction of gamma-turns in proteins from multiple sequence alignment. *Protein Sci.* **12**: 923–929.

Kaur H, Raghava GPS. 2003b. Prediction of beta-turns in proteins from multiple alignment using neural network. *Protein Sci.* **12**: 627–634.

Kaur H, Raghava GPS. 2004a. A neural network method for prediction of beta-turn types in proteins using evolutionary information. *Bioinformatics* **20**: 2751–2758 10.1093/bioinformatics/bth322 bth322 [pii]

Kaur H, Raghava GPS. 2004b. Prediction of alpha-turns in proteins using PSI-BLAST profiles and secondary structure information. *Proteins* **55**: 83–90 10.1002/prot.10569

Kumar M, Gromiha MM, Raghava GPS. 2007a. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins 10.1002/prot.21677*

Kumar M, Verma R, Raghava GPS. 2006. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J. Biol. Chem.* **281**: 5357–5363 M511061200 [pii] 10.1074/jbc. M511061200

Kumar M, Bhasin M, Natt NK, Raghava GPS. 2005. BhairPred: prediction of beta-hairpins in a protein from multiple alignment information using ANN and SVM techniques. *Nucleic Acids Res.* **33**: W154–159.

Kumar M, Gromiha MM, Raghava GPS. 2007b. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. *BMC Bioinformatics* **8**: 463. 1471-2105 - 8-463 [pii] 10.1186/1471-2105 -8-463

Li W, Jaroszewski L, Godzik A. 2001. Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* **17**: 282–283.

Luhrmann R, Kastner B, Bach M. 1990. Structure of spliceosomal snRNPs and their role in pre-mRNA splicing. *Biochim. Biophys. Acta.* **1087**: 265–292.

Lunde BM, Moore C, Varani G. 2007. RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.* **8**: 479–490 nrm2178 [pii] 10.1038/nrm2178

Matthews BW. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* **405**: 442–451.

Moore PB. 1998. The three-dimensional structure of the ribosome and its components. *Annu. Rev. Biophys. Biomol. Struct.* **27**: 35–58 10.1146/annurev.biophys.27.1.35

Nagai K. 1996. RNA-protein complexes. *Curr. Opin. Struct. Biol.* **6**: 53–61 S0959-440X(96) 80095-9 [pii]

Scott WG, Klug A. 1996. Ribozymes: structure and mechanism in RNA catalysis. *Trends Biochem. Sci.* **21**: 220–224 0968-0004 (96) 10026-8 [pii]

Shazman S, Mandel-Gutfreund Y. 2008. Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput. Biol.* **4**: e1000146. 10.1371/journal.pcbi.1000146

Steitz T. 1999. RNA recognitition by proteins, Gestel R, Cech T (eds). Cold Spring Harbor Laboratory Press, Cold Spring Harbor: NY.

Vapnik V. 1995. The nature of statistical learning theory. Springer: New York.

Wang G, Dunbrack RL Jr. 2003. PISCES: a protein sequence culling server. *Bioinformatics* **19**: 1589–1591.

Wang L, Brown SJ. 2006. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.* **34**: W243–248. 34/suppl_2/W243 [pii] 10.1093/nar/gkl298

Xie D, Li A, Wang M, Fan Z, Feng H. 2005. LOCSVMPSI: a web server for subcellular localization of eukaryotic proteins using SVM and profile of PSI-BLAST. *Nucleic Acids Res.* **33**: W105–110. 33/suppl_2/W105 [pii] 10.1093/nar/gki359

Yu X, Cao J, Cai Y, Shi T, Li Y. 2006. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J. Theor. Biol.* **240**: 175–184 S0022-5193 (05) 00421-2 [pii] 10.1016/j.jtbi.2005.09.018

