PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Prediction of RNA binding sites in a protein using SVM and PSSM profile

Manish Kumar,[1] M. Michael Gromiha,[2] and G. P. S. Raghava[1]*

[1] Bioinformatics Centre, Institute of Microbial Technology, Chandigarh-160036, India

[2] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, Tokyo 135-0064, Japan

ABSTRACT

RNA-binding proteins (RBPs) play key roles in post-transcriptional control of gene expression, which, along with transcriptional regulation, is a major way to regulate patterns of gene expression during development. Thus, the identification and prediction of RNA binding sites is an important step in comprehensive understanding of how RBPs control organism development. Combining evolutionary information and support vector machine (SVM), we have developed an improved method for predicting RNA binding sites or RNA interacting residues in a protein sequence. The prediction models developed in this study have been trained and tested on 86 RNA binding protein chains and evaluated using fivefold cross validation technique. First, a SVM model was developed that achieved a maximum Matthew's correlation coefficient (MCC) of 0.31. The performance of this SVM model further improved the MCC from 0.31 to 0.45, when multiple sequence alignment in the form of PSSM profiles was used as input to the SVM, which is far better than the maximum MCC achieved by previous methods (0.41) on the same dataset. In addition, SVM models were also developed on an alternative dataset that contained 107 RBP chains. Utilizing PSSM as input information to the SVM, the training/testing on this alternate dataset achieved a maximum MCC of 0.32. Conclusively, the prediction performance of SVM models developed in this study is better than the existing methods on the same datasets. A web server 'Pprint' was also developed for predicting RNA binding residues in a protein sequence which is freely available at http://www.imtech.res.in/raghava/pprint/.

Proteins 2008; 71:189–194.

Key words: evolutionary information; interacting residue; protein; RNA; SVM.

INTRODUCTION

The RNA-binding proteins (RBPs) play a vital role in living organisms. One of the important questions to understand the function of RBPs in an organism development is how RBPs distinguish their targets from nontargets *in vivo*. In other words, how RBPs specifically recognize their RNA targets. RBPs may recognize specific sequences, structures, or both, which are present in their RNA targets. Understanding of RNA binding specificity of an RBP can be another way to identify unknown targets that contain similar features, if they have enough information to distinguish targets from nontargets computationally. Thus, the identification of RBPs and their binding sites is a major challenge in the field of molecular recognition. During the past one-decade, the number of RBPs has increased exponentially because of the large scale sequencing projects in progress. Despite tremendous progress in the annotation of RBPs, identification of RNA interacting residues in proteins is still a major challenge. Although, it is not difficult to identify the RNA interacting residues in a protein from the structures of RNA-protein complexes, the experimental determination of a complex structure is costly and time consuming. Thus, the development of methods for predicting RNA-binding site in a protein from its amino acid sequence is important for understanding the function of these RBPs. In 2004, Joeng *et al.*,[1] developed an artificial neural network (ANN)-based method for predicting RNA interacting residues using amino acid sequence and secondary structure information and achieved a maximum *MCC* of 0.29. Using evolutionary information extracted from PSI-BLAST profiles and CLUSTALW alignment, Jeong and Miyano[2] improved the *MCC* to 0.41. Wang and Brown[3] developed a SVM-based method using side chain pK_a, hydrophobicity index and molecular mass of amino acids and achieved a maximum accuracy of 69.32% with 66.28% sensitivity. Recently, Terribilini *et al.*[4] developed a method for predicting RNA interacting residues using Naïve Bayes Classifier, and achieved maximum *MCC* of 0.35. In the present study, a systematic attempt has been made to

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat.

Grant sponsors: Council of Scientific and Industrial Research (CSIR), Department of Biotechnology (DBT).

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

Received 18 April 2007; Accepted 26 May 2007
Published online 11 October 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21677

improve the prediction accuracy of RNA interacting residues using SVM and evolutionary information.

MATERIALS AND METHODS

Main dataset

The main dataset contained 86 RNA interacting protein chains extracted from the structures of RNA-protein complexes.[2] These structures (solved at 3 Å or better resolution) were obtained from protein data bank (PDB).[5] Using PSI-BLAST[6] only the nonredundant protein chains, where no two chains had sequence similarity more than 70% were included in this main dataset. Each protein chain in the dataset had at least four RNA interacting residues. Terribilini *et al.*[4] evaluated their Naive Bayes Classifier on the same dataset. In the present study, we used a cutoff of 6 Å to define the RNA interacting residues in consideration with the experimental noise. Hence, a residue was considered to be RNA-interacting if the closest distance between atoms of the protein and the partner RNA was within the cutoff (6 Å). The protein chains in main dataset consisted of a total 20,071 residues out of which 4568 were RNA-interacting residues.

Alternate dataset

In addition to main dataset, we also used an alternate dataset that consisted of 107 RNA interacting chains obtained from 61 RNA-interacting proteins. This nonredundant dataset, where no two chains had sequence similarity more than 25%, was obtained from BindN server (http://bioinformatics.ksu.edu/bindn/) and was used by Wang and Brown[3] for developing method 'BindN' to predict RNA binding residues. In this alternate dataset, we also used a cutoff of 3.5 Å to define the RNA interacting residues similar to criteria used by BindN researchers. Among the main and alternate datasets, 38 protein chains were found to be common. The protein chains in alternate dataset consisted of a total 22,051 residues out of which 2555 were RNA-interacting residues.

Five-fold cross-validation

The fivefold cross-validation technique was used to evaluate the performance of all the methods attempted by us. In this technique, proteins are randomly divided into five sets of which four sets are used for training and the remaining fifth set for testing. This process is repeated five times in such a way that each set is used once for testing. The final performance is obtained by averaging the performance of all the five sets.

Pattern or window size

For each sequence, we created overlapping patterns (segments) of different size (or window size) 11, 13, 15 etc. If the central residue of pattern was RNA interacting residue, then we classified the pattern as positive or RNA interacting pattern and otherwise it was termed as noninteracting or negative pattern. To create a pattern corresponding to the terminal residues in a protein chain, we added $(L - 1)/2$ dummy residue "X" at both terminals of protein (where L is the length of pattern). It means for window size 11, we added 5 "X" before N-terminal and 5 "X" after C-terminal, in order to create L patterns from sequence of length L. There is a pattern corresponding to each residue in a protein sequence. It is similar to the approach adopted by Singh and Raghava for prediction of MHC class II binding peptide prediction.[7]

Support vector machine (SVM)

In this study, SVM technique was implemented using SVM_light package.[8,9] This package is very powerful and user-friendly where one can adjust the parameters and kernel functions like Polynomial, RBF, Linear, and Sigmoid. In the past also, SVM technique has been used successfully for developing a wide range of bioinformatics tools.[10,11]

Evolutionary information

This was obtained from position-specific scoring matrix (PSSM) generated during PSI-BLAST[6] search against nonredundant (nr) database of protein sequences at NCBI. The PSSM matrix was generated by three iterations of searching at cutoff *e*-value of 0.001 for inclusion of sequences in next iteration. The PSSM thus generated contained the probability of occurrence of each type of amino acid residues at each position along with insertion/deletion. Hence, PSSM is considered as a measure of residue conservation in a given location. This means that evolutionary information for each amino acid is encapsulated in a vector of 21 dimensions where the size of PSSM matrix of a protein with N residues is $21 \times N$.

Performance measures

To assess the performance of various modules developed in this study, we computed following threshold dependent parameters: sensitivity (Sn) or percent coverage of RNA interacting residues; specificity (Sp) or percent coverage of noninteracting residues; overall accuracy (Ac); percent probability of correct prediction of RNA interacting residues (PPV), also called as accuracy of interacting residues and Matthew's correlation coefficient (MCC) using following equations:

$$Sn = \frac{tp}{tp + fn} \times 100 \quad (1)$$

$$Sp = \frac{tn}{tn + fp} \times 100 \quad (2)$$

$$Ac = \frac{tp + tn}{tp + fn + tn + fp} \times 100 \quad (3)$$

$$MCC = \frac{(tp)(tn) - (fp)(fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \times 100 \quad (4)$$

$$PPV = \frac{tp}{tp + fp} \times 100 \quad (5)$$

where *tp* and *tn* are correctly predicted positive and negative examples, respectively. Similarly, *fp* and *fn* are wrongly predicted as positive and negative patterns, respectively.

RESULTS

Compositional analysis

We computed the composition of interacting and non-interacting residues and found that Gly, His, Lys, Asn, Gln, and Arg were more abundant in RNA interacting residues than in noninteracting residues (Fig. 1). The dominance of these amino acid residues indicated their major contribution in interactions. Among these residues, the higher occurrence of Lys and Arg was expected since these are positively charged and can easily interact with negatively charged RNA. This is similar to DNA-interacting residues as shown by Ahmad *et al.*[12] Among these two positively charged amino acids, Arg was found to be more prevalent which can be due to its unique capacity to form multiple and bifurcated hydrogen bonds and to simultaneously bind multiple nucleotides. The over-representation of His can be attributed to its peculiar pK_a value because depending on the pH value, His can also be positively charged. Further, it can participate in stacking interactions with RNA bases through its imidazole ring. For Asn and Gln, their H-bonding capability can be the reason of over-representation in binding state. Similarly, small size and flexibility of Gly residues is probably making it suitable for the structural adjustments required during the interactions. Among the under-represented residues, the most prominent were Ala, Glu, Phe, Ile, Leu, and Val. The less occurrence of Glu can be due to its negative charged side chain. But interestingly, other negatively charged amino acid, Asp did not show any difference. The question arises that whether the amino acid residues surrounding the interacting residue also shows preference for some particular residues. To examine this, we created 20 web logos for each type of amino acid. At first, the interacting pattern of length 25 was generated for each type of interacting residue at center position of pattern. These patterns were submitted to http://weblogo.berkeley.edu/ to create web-logos. The web logos for Cys and Trp are presented in Figure 2, where Cys and Trp were found at the center position in interacting patterns. As shown in Figure 2, there are preferences for amino acids at few neighboring locations around the interacting residue.

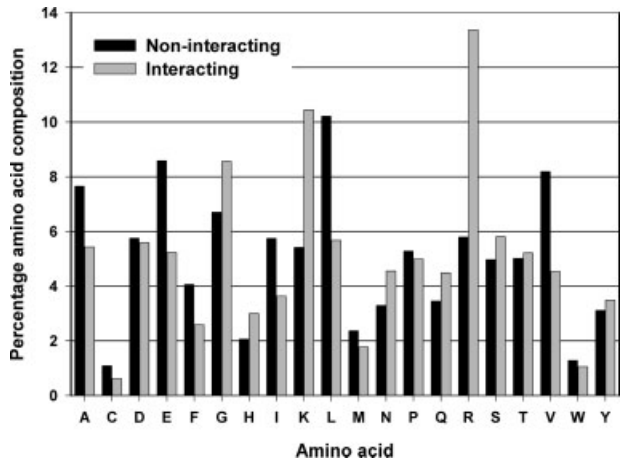

Figure 1

Percent composition of interacting and non-interacting residues in 86 protein chains (main dataset) used for development of Pprint.

SVM model using amino acid sequence

Fixed length patterns were generated from RNA interacting chains, where a pattern was assigned positive, if the centre residue was found to be interacting residue, otherwise negative pattern. These sequence patterns were converted to binary patterns, where a pattern of length *N* was represented by a vector of dimension $N \times 21$. Each amino acid was represented by a vector of 21 (e.g. Ala by 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0) which contained 20 amino acids and one dummy amino acid "X." As shown in Table I, this SVM-based model was able to achieve a maximum *MCC* of 0.31 with 76.05% accuracy.

SVM model using evolutionary information

In the past, it has been shown in some studies that evolutionary information obtained from multiple sequence alignment provides more comprehensive information about the protein than a single sequence.[13,14] In the present study, the evolutionary information obtained from PSSM generated from PSI-BLAST profiles was also used for predicting RNA interacting residues. As shown in Table II, performance increased significantly when PSSM was used as input instead of single sequence. A maximum *MCC* of 0.45 was achieved with 81.16% accuracy.

Performance of SVM on alternate dataset

Recently, BindN server[3] was developed on 107 non-redundant RNA-interacting chains that had achieved the

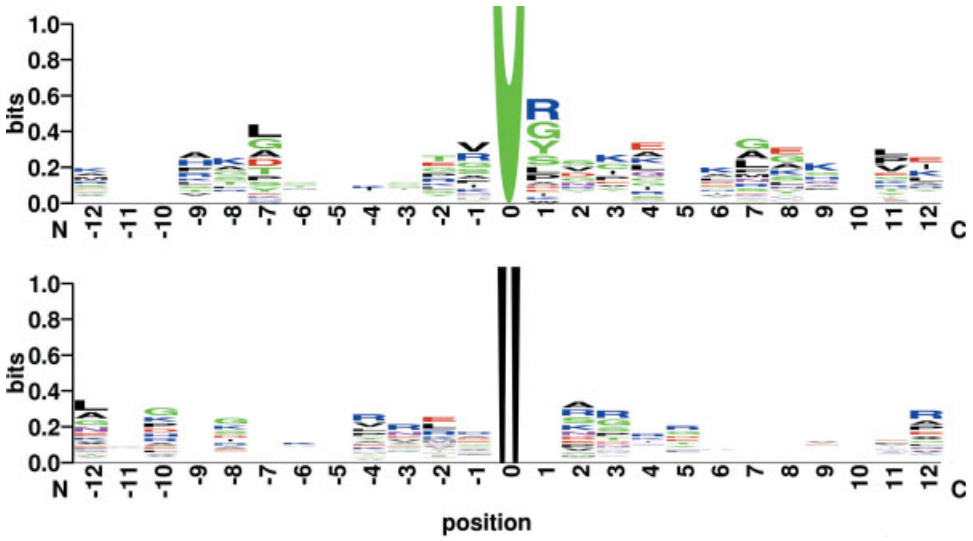

Figure 2

Sequence logo of Cys and Trp along with upstream and downstream 12 amino acid residues. Position 0 is the central residue. Negative and positive represents upstream and downstream positions, respectively.

maximum *MCC* of 0.27 with 69.32% accuracy (sensitivity, 66.28%; specificity, 69.84%). To compare our method with BindN, we also developed SVM models on 107 RNA binding chains obtained from BindN server; called alternate dataset in this study. We achieved the *MCC* of 0.28 with 72.76% accuracy (sensitivity, 66.09%; specificity, 73.49%) from SVM model (learning parameter: $-z\ c\ -j\ 7\ -t\ 2\ -g\ 0.001$) using amino acid sequence on alternate dataset.[3] In addition, SVM model ($-z\ c\ -j\ 6\ -t\ 0$) using evolutionary information was also developed that

Table I

The Performance of SVM Model (Learning Parameter: $-z\ c\ -j\ 4\ -t\ 2\ -g\ 0.01$) on Main Dataset Using Amino Acid Sequence

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
−1.00	93.46	20.59	37.82	0.16
−0.90	92.04	24.51	40.47	0.17
−0.80	90.28	28.91	43.43	0.19
−0.70	88.40	33.56	46.54	0.21
−0.60	86.33	38.34	49.72	0.22
−0.50	83.89	43.14	52.84	0.24
−0.40	81.00	48.06	55.93	0.25
−0.30	77.80	53.00	58.97	0.26
−0.20	74.56	58.01	62.05	0.28
−0.10	70.22	62.71	64.62	0.28
0.00	**66.62**	**67.03**	**67.09**	**0.29**
0.10	62.37	71.28	69.36	0.30
0.20	57.99	75.32	71.43	0.30
0.30	53.71	79.15	73.37	0.31
0.40	49.45	82.33	74.80	0.31
0.50	44.98	85.32	76.04	0.31
0.60	***40.66***	***87.72***	***76.87***	***0.31***
0.70	36.63	90.06	77.67	0.30
0.80	33.32	92.34	78.62	0.31
0.90	29.36	93.80	78.82	0.30
1.00	25.70	94.98	78.85	0.29

Values in bold indicate the point where sensitivity and specificity is roughly equal. Bold and italics are the point of maximum MCC.

Table II

The Performance of SVM model (Learning Parameter: $-z\ c\ -j\ 4\ -t\ 1\ -d\ 2$) on Main Dataset Using PSI-BLAST Profile

Threshold	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
−1.00	91.48	39.65	51.94	0.28
−0.90	89.90	44.91	55.62	0.30
−0.80	87.79	50.23	59.19	0.33
−0.70	85.40	55.24	62.49	0.34
−0.60	83.22	60.16	65.74	0.37
−0.50	80.63	64.50	68.46	0.38
−0.40	78.32	68.51	70.98	0.40
−0.30	76.08	72.51	73.52	0.42
−0.20	**73.58**	**75.91**	**75.53**	**0.44**
−0.10	70.17	79.04	77.12	0.44
0.00	67.16	81.79	78.54	0.45
0.10	63.64	84.02	79.42	0.45
0.20	60.08	86.09	80.16	0.45
0.30	56.56	87.97	80.76	0.45
0.40	***53.05***	***89.55***	***81.16***	***0.45***
0.50	49.46	90.93	81.38	0.44
0.60	46.16	92.20	81.56	0.44
0.70	42.79	93.24	81.55	0.42
0.80	39.93	94.28	81.67	0.42
0.90	36.81	95.08	81.56	0.41
1.00	34.06	95.70	81.39	0.40

The number in bold shows performance of model at threshold, where sensitivity and specificity are nearly equal. The number in bold and italics shows performance of method at the threshold, where MCC is maximum.

achieved the *MCC* of 0.32 with 75.43% accuracy (sensitivity, 70.09%; specificity, 75.54%). This indicates that models developed using evolutionary information performs better than the methods solely based on the single sequence information as well as it performs better than previous methods on same dataset.

Description of web-server

A user-friendly web-server 'Pprint' was developed for the prediction of RNA-interacting residues in a protein (Supplementary Fig. S1). The user may submit the amino acid sequence in standard 'FASTA' format or as simple amino acids in single letter code. The server automatically generates the evolutionary profile of whole sequence by running PSI-BLAST, generates SVM pattern from this PSSM profile and then, predicts RNA interacting residues using SVM model. The server also allows a user to select threshold value, which is important for setting high sensitivity or specificity for predicting RNA binding residues in a protein sequence. The authors suggest that in order to predict interacting residues with high specificity (i.e. prediction of RNA binding residues with high confidence), user should opt for higher threshold. In this case, one has to compromise with the sensitivity of prediction. However, to predict maximum RNA interacting residues in a protein, user should opt for lower threshold value. In this case, the probability of correct prediction of RNA interacting residues will be low. It means that the percent sensitivity can be increased at the expense of specificity and *vice-versa*. The default threshold value was set at −0.2 as at this threshold, sensitivity and specificity was roughly found to be equal during the training and five-fold cross validation procedure. Hence, at this threshold, one may not lose much, both in terms of sensitivity and specificity. The prediction results are presented in a tabular as well as in graphical format where the predicted RNA-binding residues in a protein are displayed in different color (Supplementary Fig. S2).

Comparison with other existing methods

In 2004, Jeong *et al.*[1] developed an ANN-based method for predicting RNA binding residues using amino acid sequence as input and achieved the maximum *MCC* of 0.29 with an overall accuracy of 77.50% (sensitivity, 40.30; specificity, 87.29). They also studied the effect of shifting and filtering on performance of method and achieve maximum *MCC* 0.59 with ±2 shifting and ±1 filtering (see Table V of Jeong *et al.* 2004). In 2006, they developed ANN models[2] using various alignment profiles like Clustalw, PSSM and weighted profile, thereby improved the *MCC* to 0.41. Recently, Terribilini *et al.*[4] developed a Naïve Bayes classifier (NBC) model on a nonredundant dataset of 109 RBP chains, where no two chains have sequence identity more than 30%. They achieved the maximum *MCC* of 0.35 from their NBC model using amino acid sequence. They also evaluated the performance of their NBC method on Jeong *et al.* 2004 dataset in order to provide direct comparison of their method with ANN method of Jeong *et al.* 2004 (see Table III of Terribilini *et al.*).[4] They achieved the *MCC* of 0.30 with 76.60% accuracy (sensitivity, 43.00%; specificity, 86.53%) on Jeong *et al.* 2004 dataset.[1] In the present work, we used the dataset of Jeong *et al.*[1] for developing a SVM module using amino acid sequence and achieved the maximum *MCC* of 0.31 with 76.87% accuracy (sensitivity 40.66%; specificity 87.72%), which is slightly better than other methods on the same dataset using amino acid sequence (Table III). In addition, we also developed a SVM classifier using PSSM profile that achieved the maximum *MCC* of 0.45 with 81.16% accuracy (sensitivity, 53.05%; specificity, 89.55%), which is significantly better than any other existing methods on the same dataset. One of the major advantages of our method over the Jeong and Miyano (2006) method is that our prediction method is available as a web server (Pprint) to the public. Wang and Brown (2006) developed a method BindN[3] on a different dataset (alternate dataset) for predicting RNA interacting residues and achieved the maximum accuracy of 69.32% with *MCC* 0.27. We had achieved the maximum accuracy of 75.43% with *MCC* of 0.32 on the alternate dataset using our PSSM-based SVM model. These results clearly demonstrate the superiority of our method over the existing methods (Table III).

Table III
Comparison of Different Prediction Methods

Input	Sensitivity (%)	Specificity (%)	Accuracy (%)	MCC
ANN_amino[a]	40.30	87.29	77.50	0.29
NBC_amino[a]	43.00	86.53	76.60	0.30
SVM_amino[b]	40.66	87.72	76.87	0.31
ANN_PSSM[b]	43.40	91.04	80.20	0.39
ANN_WP[b]	NR	NR	NR	0.41
SVM_PSSM[b]	53.05	89.55	81.16	0.45
SVM_PCP[c]	66.28	69.84	69.32	0.27

Here, ANN_amino is the performance of ANN on amino acid sequence[1]; NBC_amino is the performance of Naïve Bayes Classifier on amino acid sequence,[4] SVM_amino is performance of SVM model on amino acid sequence[Pprint]; ANN_PSSM is performance of ANN on PSSM profile[2]; ANN_WP is performance of ANN on weighted profile[2]; SVM_PSSM is the performance of SVM on PSSM profile [Pprint] and SVM_PCP is performance of SVM on Physico-chemical properties.[3]
NR: Not reported by authors.
[a]Jeong *et al.* (2004) dataset was used.
[b]Jeong and Miyano (2006) dataset was used.
[c]Wang and Brown (2006) dataset was used.

DISCUSSION

Because of the vital significance of RNA protein interaction in cellular metabolism and difficulties in identification of RNA binding residues by biophysical or *in-vitro*

analysis, there is an urgent need for computational methods to identify RNA binding sites on the basis of amino acid sequence of a protein. In this direction, we have made a systematic attempt to develop highly accurate method for predicting RNA interacting residues in a protein sequence. First, we examined the existing methods and found that these methods have been evaluated on two datasets used by Jeong and Miyano[2] and Wang and Brown.[3] To compare the performance of our newly developed method (Pprint), we developed our models using the same datasets as used by these researchers and termed these as 'main' and 'alternate' dataset in the present study. Then, an analysis was done for searching RNA interacting residues and residues near the interacting residues. During this analysis, we came across a number of observations, which are important for understanding RNA-protein interactions. We found significant bias in the type of interacting amino acids as well as flanking regions.

It has been reported in some of the earlier studies that SVM performs better than other machine learning techniques. Wang and Brown[3] also used SVM in predicting the RNA binding residues but they implemented it using Physico-chemical properties of residues. First, we developed SVM model based on amino acid sequence; this model performed marginally better than the previous methods on a single protein sequence. We observed that all techniques based on amino acid sequence perform equally well: (i) ANN used by Jeong *et al.* 2004[1]; (ii) NBC used by Terribilini *et al.* 2006; and (iii) SVM used by Wong and Brown, 2006. However SVM model based on amino acid sequence used in this study perform slightly better than previous studies but margin was not significant. The major improvement in performance comes from evolutionary information used in form of PSSM profile. Both SVM (in Pprint) and ANN (in Jeong and Miyano, 2006[4]) models based on PSSM profile performs significantly better than other methods. Our SVM model based on PSSM profile out-perform all existing methods including ANN model of Jeong and Miyano, 2006. To provide direct access of prediction method developed by us to the scientific community, we developed a web server Pprint, which allows user to predict RNA-interacting residues in their protein.

ACKNOWLEDGMENTS

The authors thank Rakesh Kaundal for critically reading the manuscript. This manuscript has IMTech communication number 054/2006.

REFERENCES

1. Jeong E, Chung IF, Miyano S. A neural network method for identification of RNA-interacting residues in protein. Genome Inform 2004;15:105–116.
2. Jeong E, Miyano S. A Weighted profile based method for protein-RNA interacting residue prediction. In: Corrado P, Luca C, Stephen E, editors. Lecture notes in computer science, Vol. 3939. Berlin/Heidelberg: Springer; 2006. pp 123–139.
3. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006;34 (Web Server issue):W243–W248.
4. Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D. Prediction of RNA binding sites in proteins from amino acid sequence. RNA 2006;12:1450–1462.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
6. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
7. Singh H, Raghava GPS. ProPred: prediction of HLA-DR binding sites. Bioinformatics 2001;17:1236–1237.
8. Joachims T. Making large scale SVM learning practical. In: Scholkopf B, Burges C, Smola A, editors. Advances in kernel methods: Support Vector Learning. Cambridge: MIT Press; 1999, pp 169–184.
9. Vapnik V. The nature of statistical learning theory. New York: Springer; 1995.
10. Kumar M, Verma R, Raghava GPS. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. J Biol Chem 2006;281:5357–5363.
11. Garg A, Bhasin M, Raghava GPS. Support vector machine-based method for subcellular localization of human proteins using amino acid compositions, their order, and similarity search. J Biol Chem 2005;280:14427–14432.
12. Ahmad S, Gromiha MM, Sarai A. Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 2004;20:477–486.
13. Kaur H, Raghava GPS. Prediction of β-turns in proteins from multiple alignment using neural network. Protein Sci 2003;12:627–634.
14. Garg A, Kaur H, Raghava GPS. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins 2005;61:318–324.