Amino Acids (2008) 35: 295–302
DOI 10.1007/s00726-007-0634-9
Printed in The Netherlands

Amino Acids

PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles

Y. Wang[1], **Z. Xue**[2], **G. Shen**[2], and **J. Xu**[3]

[1] Institute of Biophysics and Biochemistry, School of Life Science, Huazhong University of Science and Technology, Wuhan City, China
[2] Software College, Huazhong University of Science and Technology, Wuhan City, China
[3] Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan City, China

Received September 29, 2007
Accepted November 5, 2007
Published online January 31, 2008;

Summary. Protein–RNA interactions play a key role in a number of biological processes such as protein synthesis, mRNA processing, assembly and function of ribosomes and eukaryotic spliceosomes. A reliable identification of RNA-binding sites in RNA-binding proteins is important for functional annotation and site-directed mutagenesis. We developed a novel method for the prediction of protein residues that interact with RNA using support vector machine (SVM) and position-specific scoring matrices (PSSMs). Two cases have been considered in the prediction of protein residues at RNA-binding surfaces. One is given the sequence information of a protein chain that is known to interact with RNA; the other is given the structural information. Thus, five different inputs have been tested. Coupled with PSI-BLAST profiles and predicted secondary structure, the present approach yields a Matthews correlation coefficient (MCC) of 0.432 by a 7-fold cross-validation, which is the best among all previous reported RNA-binding sites prediction methods. When given the structural information, we have obtained the MCC value of 0.457, with PSSMs, observed secondary structure and solvent accessibility information assigned by DSSP as input. A web server implementing the prediction method is available at the following URL: http://210.42.106.80/printr/.

Keywords: Protein–RNA interactions – RNA-binding sites – Support vector machine – Multiple sequence alignment

1. Introduction

Biological macromolecules' recognizing each other is the basis of many cellular biological processes such as signals transduction, gene expression, gene regulation, and protein synthesis (Chou, 2005a, b). Identifying the recognition sites can help to understand these biological processes. Numerous methods (Jones and Thornton, 1997; Zhou and Shan, 2001; Ofran and Rost, 2003; Koike and Takagi, 2004) have been developed to predict protein–protein interaction site, with considerable prediction accuracies. Some approaches (Ahmad et al., 2004, 2005) have been proposed to identify DNA binding sites. But only very few methods have been designed to discriminate RNA binding sites, owing to the smaller number of experimentally determined structures of protein–RNA complexes. Because of the importance of protein–RNA interactions in protein synthesis, mRNA processing, assembly and function of ribosomes and eukaryotic spliceosomes, it is necessary to develop a reliable method to predict protein–RNA interacting sites.

Although the analysis of physical and chemical properties of the interfaces of protein–RNA complexes has a long history (Draper, 1994, 1999; Allers and Shamoo, 2001; Jones et al., 2001; Treger and Westhof, 2001; Kim et al., 2004), the identification of RNA-binding sites has only begun recently. Jeong et al. (2004) proposed a neural network method to identify RNA-interacting residues with single sequence and predicted secondary structure as input. Later, Jeong and Miyano (2006) improved the prediction performance of the neural network classifier by using weighted profiles. Recently, Terribilini et al. (2006) developed a Naive Bayes based computational tool for predicting which amino acids of an RNA binding protein participate in RNA–protein interactions, with only the single protein sequence as input. Very recently, Wang and Brown (2006) developed an SVM-based web tool BindN for prediction of DNA and RNA binding sites, using single sequence plus three simple sequence features including the side chain pKa value, hydrophobicity index and molecular mass of an amino acid.

In the current work, we proposed PRINTR, a robust method, for predicting RNA–protein interacting residues based on SVM and PSSMs. As discussed above, the methods developed for the prediction of RNA-binding residues were all based on sequence or structure information predicted from sequence. Therefore, we considered two cases in predicting RNA binding sites. One is given the sequence information of a protein chain that is known to interact with RNA. The other is given the structural information. Five different encoding schemes have been discussed. Results show that a great improvement in the prediction performance has been achieved by combining the PSSMs profiles and SVM analysis. The accuracy levels we have achieved are comparable or better than previous methods.

2. Materials and methods

2.1 Data sets

The method has been developed on the data set clustered by Terribilini et al. (2006). It is comprised of 109 nonredundant protein chains with a sequence homology value <30%. They belong to 56 RNA–protein complexes that were determined by X-ray crystallography with resolution better than 3.5 Å. Atomic coordinates for each complex were downloaded from the Protein Data Bank (Berman et al., 2000). We consider protein–RNA interactions that include hydrogen bonding, stacking, electrostatic, van der Waals and hydrophobic interactions. Residues in protein–RNA interface were extracted using ENTANGLE (Allers and Shamoo, 2001). The ENTANGLE program uses structural models in their PDB format and searches for appropriate hydrogen bonding (donor to acceptor $\leq$3.9 Å; geometric cutoff $\leq$90°), stacking (dihedral angle $\leq$30°; center-to-center distance $\leq$5 Å), electrostatic (only considering the interaction of lysine and arginine with O-1 and O-2 of the phosphodiester backbone), hydrophobic (non-polar atoms that are $\leq$5 Å apart) and van der Waals interactions (calculated as the sum of the van der Waals radii of the two atoms plus a maximum distance (default $\leq$0.8 Å). The more detailed default cutoffs are refered to Allers and Shamoo (2001). The total number of amino acid residues consisting of RNA-binding site in the data set is 14%. The list of the 109 protein chains is available in the supplementary material.

2.2 PSI-BLAST profiles

In this work, PSI-BLAST program (Altschul et al., 1997) was employed to generate multiple sequence alignment profiles. Firstly, we collected the updated nonredundant (NR) protein sequence database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/). Secondly, we applied a program named PFILT (Jones, 1999) for masking out low-complexity regions, coiled coil segments, and transmembrane spans. Finally, position-specific score matrices were generated using PSI-BLAST with three rounds against the filtered NR database with a cutoff E-value of 0.001.

2.3 Support vector machine

SVM is a supervised learning algorithm proposed by Vapnik (1995), which has been successfully applied to many pattern recognition problems in biology (Chou and Cai, 2002; Cai et al., 2003; Wang et al., 2004; Yang and Chou, 2004; Ding et al., 2007; Shen et al., 2007b). Distinguishing RNA-binding residues from non-RNA-binding residues is essentially a binary classification problem. To solve the binary classification problem, the SVM maps the input vectors into a higher dimensional feature space by a mapping function $\phi(x)$, and does a linear separation here. It tries to find the separating hyperplane with the largest distance between the two classes within the feature space by solving the following optimization problem:

$$\min_{\omega,\xi_i} \quad \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{l}\xi_i \tag{1}$$

$$\text{s.t.} \quad y_i(\omega \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \;\; i = 1, \ldots, l$$

where x_i represents an input vector, $y_i = \pm 1$ according to whether x_i belongs to the interacting class or non-interacting class, l is the number of training data, ω is a weight vector, b is a bias, ξ_i denotes a slack variable and C is a regularization parameter that controls the trade off between margin and classification error. $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$ denotes the kernel function. Then the corresponding dual quadratic programming problem can be written as:

$$\min_{\alpha_i} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(x_i \cdot x_j) - \sum_{j=1}^{l}\alpha_j \tag{2}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0, \quad 0 \leq \alpha_i \leq C, \;\; i = 1, \ldots, l$$

where α_i are the solutions of the dual formulation. An unlabelled residue x_j can subsequently be classified by the following discriminant function:

$$f(x_j) = \sum_{i=1}^{l} \alpha_i y_i K(x_i, x_j) + b \tag{3}$$

The *SVMlight* package created by Joachims (1999) was used to construct the SVM classifiers. The default output threshold used by SVMlight is 0. If $f(x_j)$ is positive, the residue x_j is classified as the binding class; otherwise, it is classified as the non-binding class. The threshold can be adjusted to get the best prediction performance.

2.4 Encoding schemes

In this study, we employed five different encoding schemes, i.e. single sequence, multiple sequence alignment, single sequence plus predicted secondary structure by PSIPRED (Jones, 1999), multiple sequence alignment plus predicted secondary structure by PSIPRED, multiple sequence alignment plus secondary structure and solvent accessibility information assigned by DSSP (Kabsch and Sander, 1983).

SVMs were trained and tested with residue-wise data instances. The feature vector representing a residue is extracted by the sliding window technique. Whether a residue belongs to the RNA-binding class or not is determined by its neighbor residues. In the case of single sequence, each residue is encoded as an orthogonal binary vector (1,0,0...) or (0,1,0...), etc. The vector is 20-dimensional. Among the twenty units of the vector, each unit stands for one type of amino acids. In order to allow a window to extend over the N-terminus and the C-terminus, a "null" residue represented by a 20-zero vector is used. In the case of multiple sequence alignment, the PSI-BLAST-generated position-specific scoring matrix has been used as input to the SVM classifier. The profile matrix elements are scaled to 0–1 range by the standard logistic function:

$$f(x) = \frac{1}{1 + \exp(-x)} \tag{4}$$

where x is the raw profile matrix value.

Solvent accessibility of a residue i was calculated as the ratio of the solvent-exposed surface area of the residue observed in a given structure, denoted as SA_i, and the maximum possible solvent-exposed surface area

for the residue in an extended tripeptide (Ala-X-Ala) conformation, denoted as MSA_i:

$$RSA_i = \frac{SA_i}{MSA_i} \quad (5)$$

We used the DSSP program to compute the residue solvent-accessible surface area, SA_i. The extended state MSA_i values were taken from Ahmad et al. (2003).

Secondary structure information is encoded as follows: helix → (1,0,0), strand → (0,1,0), and coil → (0,0,1).

2.5 Performance measures

The jackknife test and *n*-fold cross validation are widely used to examine the quality of a method on a data set. The former is more objective and rigorous (Chou and Zhang, 1995) as illustrated by a penetrating analysis in a recent review (Chou and Shen, 2007d), and is used by more and more investigators (Gao et al., 2005; Wang et al., 2005; Xiao et al., 2005, 2006; Chou and Shen, 2006a, b; 2007a, b; Guo et al., 2006; Kedarisetti et al., 2006; Mondal et al., 2006; Niu et al., 2006; Sun and Huang, 2006; Wen et al., 2006; Zhang et al., 2006; Chen et al., 2007; Diao et al., 2007a, b; Ding et al., 2007; Jahandideh et al., 2007; Liu et al., 2007; Mundra et al., 2007; Shen and Chou, 2007b, c, f, g; Shen et al., 2007a; Shi et al., 2007; Tan et al., 2007; Xiao and Chou, 2007; Zhang and Ding, 2007). Due to the limited computational power, we used a 7-fold cross-validation to evaluate the prediction performance. The whole dataset was randomly partitioned into 7 groups with approximately equal size. A classifier was then trained on 6 groups and tested on the remaining group. This process was repeated for 7 iterations, each time setting aside a different test group.

The performance of a particular prediction is assessed by the following five measures: (1) total accuracy is the percentage of correctly predicted residues, which tends to give a highly misleading impression of the prediction quality when the dataset is unbalanced, (2) accuracy is the percentage of correctly predicted interaction sites, (3) coverage, also called sensitivity, is the ratio of correctly predicted RNA binding residues and observed RNA-binding residues, (4) specificity is the ratio of correctly predicted to be non-interacting residues and observed non-interacting residues, and (5) MCC is a more robust measure of the prediction quality, which takes into account both over- and underpredictions. They can be calculated by the following equations:

$$\text{Total accuracy} = \frac{p+n}{t} \quad (6)$$

$$\text{Accuracy} = \frac{p}{p+o} \quad (7)$$

$$\text{Coverage} = \frac{p}{p+u} \quad (8)$$

$$\text{Specificity} = \frac{n}{n+o} \quad (9)$$

$$\text{MCC} = \frac{pn - ou}{\sqrt{(p+o)(p+u)(n+o)(n+u)}} \quad (10)$$

Here, p is the number of correctly classified RNA-binding residues, n is the number of correctly classified non-RNA-binding residues, o is the number of non-RNA-binding residues incorrectly classified as RNA-binding residues, u is the number of RNA-binding residues incorrectly classified as non-RNA-binding residues, and t is the total number of residues.

Another measure is also considered to assess the prediction accuracy as a function of the residue type, which defined as:

$$\text{Accuracy propensity } (i) = \frac{N_{pi}/N_p}{N_{oi}/N_o} \quad (11)$$

where i is the different residues type, N_{pi} is the number of RNA-binding predicted residue i, N_p is the total number of RNA-binding predicted residues, N_{oi} is the number of RNA-binding residue i ($i = 1$–20), and N_o is the total number of RNA-binding residues.

Sensitivity and specificity are considered for the receiver operating characteristic (ROC) as a threshold independent measure. A ROC curve is obtained by plotting all sensitivity values on the *y*-axis against their equivalent (1-specificity) for all available thresholds on the *x*-axis. The area under the ROC curve (AUC) is a robust measure of prediction performance. A perfect classifier has AUC of 1, and a random classifier receives a score of 0.5.

3. Results and discussion

3.1 Parameter optimization of the prediction system

For training an SVM, we only need to select the kernel function and the regularization parameters. We selected the kernel function based on previous studies. The radial basis function (RBF) kernel performed better than other kernels in many studies such as protein secondary structure prediction (Hua and Sun, 2001), protein subcellular location prediction (Hua and Sun, 2001), and protein relative solvent accessibility prediction. Accordingly, in this work, we used the RBF kernel:

$$k(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2) \quad (12)$$

Then the parameters to be ascertained are the kernel parameter γ and the regularization parameter C. The disparity of binding residues and non-binding residues will lead to the SVM overpredict the larger class of non-binding residues. Thus, another parameter cost factor j is adopted to ensure that the total potential cost of false positives equals the total potential cost of false negatives. Then the penalty term $C\sum_i \xi_i$ in Eq. (1) becomes: $C_+ \sum_{i:y_i=1} \xi_i + C_- \sum_{i:y_i=-1} \xi_i$. The parameters C_+ and C_- are chosen to obey (Morik et al., 1999):

$$\frac{C_+}{C_-} = \frac{N_-}{N_+} \quad (13)$$

In this study, N_+ is the number of interacting residues and N_- is the number of non-interacting residues. The optimized parameters are: $\gamma = 0.0625$, $C = 5$ ($C_- = C$; $C_+ = j^* C$), and $j = 6$.

3.2 Window size optimization

A proper window size can lead to a good prediction performance. A too short window size can lose some important classification information. But there is an up limit for improving prediction accuracy by increasing the window length because a too long window size will decrease the signal-noise ratio. The optimal window size is obtained by

Table 1. Dependence of testing accuracy on the window length

	$L=11$	$L=13$	$L=15$	$L=17$	$L=19$
Total accuracy (%)	86.8	87.0	87.1	87.0	86.8
Accuracy (%)	54.3	55.1	55.9	55.4	54.7
Coverage (%)	44.7	45.2	45.6	45.4	44.8
MCC	0.419	0.425	0.432	0.428	0.421

Results were generated with multiple sequence alignment profiles plus predicted secondary structure as input

testing different window length of 11–19. Results based on different window lengths are shown in Table 1. As can be seen from Table 1, the best prediction performance is achieved when the window length = 15. We adopted the optimum window size of 15 in the following analysis of this study. It is instructive to point out that recently a flexible window size approach was developed for predicting signal peptides of proteins (Chou and Shen, 2007e; Shen and Chou, 2007e) that might be very promising to deal with this kind of problems.

3.3 Influence of PSSMs and secondary structures on prediction performance

As pointed out in the data set section, the ratio of the number of RNA-binding residues to that of non-binding residues is about 1:6. This presents particular difficulties for training a binary classifier. We solved this problem by a two-stage method. The first step is to interpolate between the over- and underprediction of binding residues at the training stage by setting different regularization parameters C_+ and C_-. The second step is realized by use of an optimal output threshold at the decision stage. After testing different thresholds at the decision stage, we obtained the optimal output threshold of -0.3.

Table 2. Results with different encoding schemes

Method	Total accuracy (%)	Accuracy (%)	Coverage (%)	MCC
(1)	82.1	37.2	37.4	0.268
(2)	82.5	38.3	37.9	0.279
(3)	87.0	55.5	45.4	0.429
(4)	87.1	55.9	45.6	0.432
(5)	87.4	58.0	48.2	0.457

Results were based on the following encoding schemes: (1) single sequence; (2) single sequence plus predicted secondary structure; (3) multiple sequence alignment profiles; (4) multiple sequence alignment profiles plus predicted secondary structure; and (5) multiple sequence alignment profiles plus secondary structure and solvent accessibility assigned by DSSP

In Table 2, the effect of different encoding schemes on prediction accuracy is presented. It is clear that the use of multiple sequence alignment information greatly improves the prediction accuracy as the MCC improved from 0.268 to 0.429 and the total accuracy from 82.1 to 87.0%. The positive effect of PSSMs has been shown in many problems that predict functional or structural properties of a given residue in a protein sequence, such as protein secondary structure prediction (Jones, 1999), protein-protein interaction sites prediction, and DNA binding sites prediction (Ahmad et al., 2005). This is due to the fact that PSSMs contain the information whether a residue is conserved or occurs by chance. The PSSM approach, or more accurately, the PsePSSM approach parallel to the PseAA composition approach (Chou, 2001, 2005c), was successfully used to predict membrane protein type (Chou and Shen, 2007c), enzyme functional class (Shen and Chou, 2007a), and protein subnuclear location (Shen and Chou, 2007d), respectively.

Predicted secondary structure information contributes little to the prediction accuracy, especially when multiple sequence alignment information is used. But given the structural information of a protein chain that is known to interact with RNA, we got the highest MCC value of 0.457 with multiple sequence alignment, solvent accessibility and secondary structure information assigned by DSSP as input.

We also assessed the prediction performance of the SVM model with different encoding schemes by calculating the area under the receiver operating characteristic curve. Figure 1 shows the ROC curves for five different

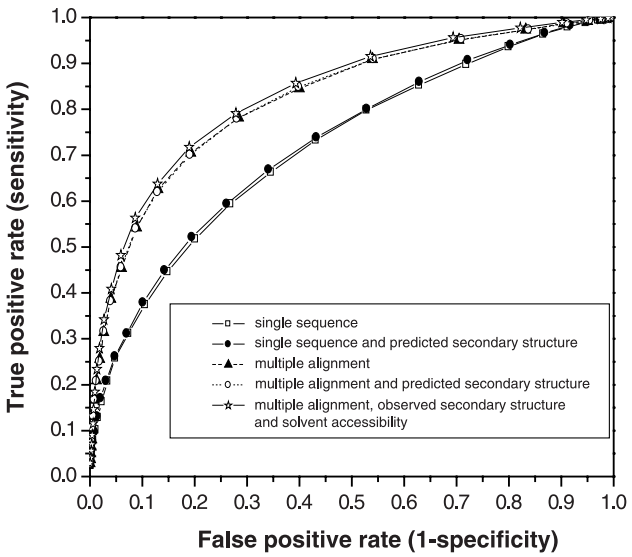

Fig. 1. The ROC curves of the five SVM predictions

encoding schemes. The corresponding area under the curves is as follows: single sequence, 0.7188; single sequence with predicted secondary structure, 0.7251; multiple alignment, 0.8248; multiple alignment with predicted secondary structure, 0.8251 and multiple alignment with observed secondary structure and real solvent accessible area, 0.8357.

3.4 Evaluating RPINTR predictions in the context of three-dimensional structures

To demonstrate that our method can provide useful information for understanding protein–RNA interactions, we have examined the predicted binding residues in the context of three-dimensional structures. Figure 2 depicts the prediction results of protein 1urn:A (U1A). As can be seen from Fig. 2, 15 out of 25 RNA-binding residues (60%) are predicted correctly. These true positives are highlighted in red. The residues in pink are the ten false negatives. For the 72 non-binding residues, 70 or 97.2% are predicted correctly. In addition, two of the non-binding residues are predicted incorrectly. The MCC value obtained for this protein is 0.658. Therefore, the predictions made by our method can provide valuable information for understanding protein–RNA interactions.

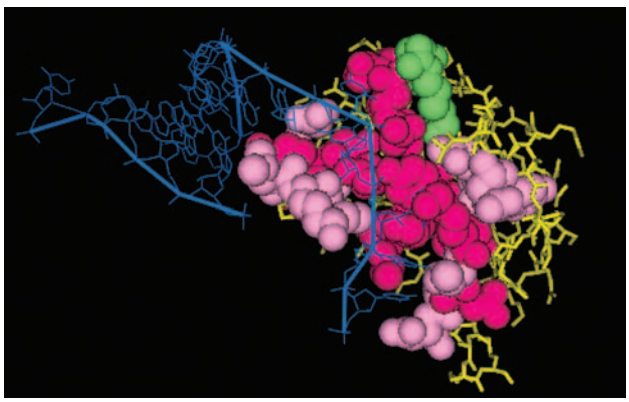

Fig. 2. Prediction results shown in the context of three-dimensional structures for 1urn: A. The prediction was performed using PSSMs and predicted secondary structure as input. The correctly predicted binding residues (*true positives*) are in red; the correctly predicted non-binding residues (*true negatives*) are in yellow; the binding residues but predicted as negatives (*false negatives*) are in pink; the non-binding residues but predicted as positives (*false positives*) are in green. The true positives are T10, Y12, N14, N15, L43, R46, R51, G52, Q53, F55, K79, A86, K87, T88, and D89. The false positives are E4, E18, K21, L48, K49, M50, Q84, Y85, S90, and D91. The figure was generated using PyMol (http://www.pymol.org) (for an interpretation of the reference to colour in this figure, the reader is referred to the online version of this paper under www.springerlink.com)

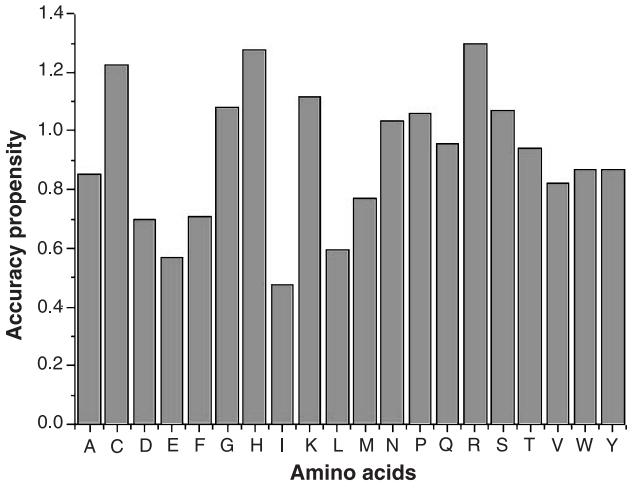

Fig. 3. Accuracy propensity for 20 amino acids. The results were generated with PSSMs and predicted secondary structure as input

3.5 Evaluating RPINTR predictions based on the residue type

We also assessed the prediction accuracy as a function of the residue type with Eq. (11). The results are shown in Fig. 3. Figure 3 shows that arginine is the best-predicted amino acid, followed by histidine, cysteine, lysine, and glycine. These positively charged amino acids arginine, lysine and histidine are also highly prefered in protein–RNA interfaces (Terribilini et al., 2006). The worst predicted amino acid is isoleucine, followed by glutamic acid, leucine, aspartic acid, and phenylalanine. These negatively charged and hydrophobic amino acids are all significantly underrepresented in interfaces (Terribilini et al., 2006).

3.6 Comparison with other studies

Jeong and Miyano (2006) compared three kinds of weighted profiles generated from the PSI-BLAST alignment to predict residues that interact with RNA. When using weighted PSSMs as input, they obtained the best MCC value of 0.41 after a 10-fold cross-validation. The MCC value obtained by our SVM model with PSSMs as input was 0.429 by a 7-fold cross-validation. We also tested our method with PSSMs by a 10-fold cross-validation, attaining an MCC value of 0.426. Because the criterion used to choose non-homologoues dataset we used is more stringent than theirs (30% for ours, 70% for theirs), it can be concluded that the present SVM model outperforms the artificial neural network model. Terribilini et al. used a Naive Bayes classifier to predict RNA binding sites

and used a leave-one-out cross-validation to evaluate the performance. The optimal MCC value they obtained was 0.35 with single sequence as input. They also reported that inclusion of sequence conservation information obtained from the HSSP database did not improve the classification performance and that using PSSMs as inputs resulted in improved prediction performance comparable with that of Jenog and Miyano (2006). Wang and Brown (2006) used SVM to predict DNA and RNA binding sites with combined sequence features as input. They used a dataset consisted of 107 protein chains with the sequence identity <25% for predicting RNA-binding residues. The optimal AUC value they obtained for RNA binding residues prediction was 0.7308 in a 5-fold cross validation, which is similar to that of our model with single sequence and predicted secondary structure but significantly lower than that of our model with PSSMs as input. In sum, we achieved the best performance when only given the sequence information, with an MCC of 0.43. Furthermore, the MCC value was improved to 0.46, when given the structure information of the unbound protein interacting with RNA.

The good performance of the present method can be ascribed to the following three aspects of our approach. Firstly, SVM has been adopted. The superior features of SVM are that it can not only effectively avoid overfitting with the use of structural risk minimization but also condense the information of training sets by identifying support vectors. Secondly, we have used a two-step method to handle the unbalanced dataset. Thirdly, a combination of the powerful binary-classification tool SVM and PSSMs greatly improves the prediction performance.

4. Conclusions

We have developed a robust approach for predicting RNA-binding residues based on SVMs and PSSMs. When only given the sequence information, the best performance was achieved by using multiple alignment and predicted secondary structure as input, with the MCC value of 0.432. The prediction performance is better when the structural information of the unbound protein interacting with RNA is known. The multiple alignment information contributes much to the prediction quality. This approach can be a complementary one to other RNA binding sites prediction methods. In addition, we have developed a web server called RPINTR to predict RNA–protein interacting residues. It is available at http://210.42.106.80/printr.

Acknowledgements

The authors would like to thank M. Terribilini et al. for providing the data set. This work is supported by a National Natural Science Grant (China) (No. 60274026).

References

Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins Struct Funct Genet 50: 629–635

Ahmad S, Gromiha MM, Sarai A (2004) Analysis and prediction of DNA-binding proteins and their binding residues based on sequence and structural information. Bioinformatics 20: 477–486

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

Allers J, Shamoo Y (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol 311: 75–86

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 databases and search programs. Nucleic Acids Res 25: 3389–3402

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

Cai YD, Zhou GP, Chou KC (2003) Support vector machines for predicting membrane protein types by using functional domain composition. Biophys J 84: 3257–3263

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet (Erratum: ibid., 2001, Vol. 44 (60)) 43: 246–255

Chou KC (2005a) Coupling interaction between thromboxane A2 receptor and alpha-13 subunit of guanine nucleotide-binding protein. J Proteome Res 4: 1681–1686

Chou KC (2005b) Insights from modeling the 3D structure of DNA-CBF3b complex. J Proteome Res 4: 1657–1660

Chou KC (2005c) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21: 10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277: 45765–45769

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007c) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360: 339–345

Chou KC, Shen HB (2007d) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370: 1–16

Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357: 633–640

Chou KC, Zhang CT (1995) Prediction of protein structural classes. Crit Rev Biochem Mol Biol 30: 275–349

Diao Y, Li M, Feng Z, Yin J, Pan Y (2007a) The community structure of human cellular signaling network. J Theor Biol 247: 608–615

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2007b) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids, DOI: 10.1007/s00726-007-0550-z.

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Peptide Lett 14: 811–815

Draper DE (1994) Protein-RNA recognition. Annu Rev Biochem 64: 593–620

Draper DE (1999) Themes in RNA-protein recognition. J Mol Biol 293: 255–270

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Hua SJ, Sun Z (2001) A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. J Mol Biol 308: 397–407

Hua SJ, Sun Z (2001) Support vector machine approach for protein subcellular location prediction. Bioinformatics 17: 721–728

Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128: 87–93

Jeong E, Chung IF, Miyano S (2004) A neural network method for identification of RNA-interacting residues in proteins. Genome Inform Ser Workshop Genome Inform 15: 105–116

Jeong E, Miyano S (2006) A weighted profile based method for Protein-RNA interacting residues prediction. Trans Comput Syst Biol IV: 123–139

Joachims T (1999) Making large-scale SVM learning practical. In: Schőlkopf B, Burges C, Smola A (eds) Advances in kernel methods-support vector learning. MIT Press, Cambridge, MA, USA

Jones S, Thornton JM (1997) Prediction of protein-protein interaction sites using patch analysis. J Mol Biol 272: 133–143

Jones DT (1999) Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292: 195–202

Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM (2001) Protein-RNA interaction: a structural analysis. Nucleic Acids Res 29: 943–954

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22: 2577–2637

Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348: 981–988

Koike A, Takagi T (2004) Prediction of protein–protein interaction sites using support vector machines. Protein Eng 17: 165–173

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Morik K, Brockhausen P, Joachims T (1999) Combining statistical learning with a knowledge-based approach – a case study in intensive care monitoring. In: Proceedings of the 16th International Conference on Machine Learning (ICML-99)

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM source. Pattern Recogn Lett 28: 1610–1615

Niu B, Cai YD, Lu WC, Zheng G.Y, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Ofran Y, Rost B (2003) Predicted protein–protein interaction sites from local sequence information. FEBS Lett 544: 236–239

Shen HB, Chou KC (2005) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2006) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364: 53–59

Shen HB, Chou KC (2007b) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20: 39–46

Shen HB, Chou KC (2007c) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007d) Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng Des Sel (DOI: 1093/protein/hzm057)

Shen HB, Chou KC (2007e) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363: 297–303

Shen HB, Chou KC (2007f) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Chou KC (2007g) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85: 233–240

Shen HB, Yang J, Chou KC (2007a) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33: 57–67

Shen HB, Yang J, Chou KC (2007b) Review: methodology development for predicting subcellular localization and other attributes of proteins. Expert Rev Proteomics 4: 453–463

Shi JY, Zhang SW, Pan Q, Cheng YM, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33: 69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Tan F, Feng,X, Fang Z, Li M, Guo Y, Jiang L (2007) Prediction of mitochondrial proteins based on genetic – algorithm partial least squares and support vector machine. Amino Acids (DOI: 10.1007/s00726-006-0465-0)

Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D (2006) Prediction of RNA binding sites in proteins from amino acid sequence. RNA 12: 1–13

Treger M, Westhof E (2001) Statistical analysis of atomic contacts at RNA-protein interfaces. J Mol Recogn 14: 199–214

Vapnik V (1998) The nature of statistical learning theory. Springer, New York

Wang LJ, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res Web Server Issue: W243–W248

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo-amino acid composition. Protein Eng Des Sel 17: 509–516

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids 28: 395–402 (Erratum, ibid. 2005, 29: 301)

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Peptide Lett 14: 871–875

Xiao X, Shao SH, Ding YS, Huang ZD Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27: 478–482

Yang ZR, Chou KC (2004) Bio-support vector machines for computational proteomics. Bioinformatics 20: 735–741

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids (DOI: 10.1007/s00726-007-0496-1)

Authors' address: Yan Wang, Institute of Biophysics and Biochemistry, School of Life Science, Huazhong University of Science and Technology, Wuhan City 430074, China,
Fax: +86-027-87792024, E-mail: yanw@mail.hust.edu.cn