Amino Acids (2011) 40:239–248
DOI 10.1007/s00726-010-0639-7

ORIGINAL ARTICLE

# Identification of RNA-binding sites in proteins by integrating various sequence information

**Cui-cui Wang · Yaping Fang · Jiamin Xiao · Menglong Li**

Received: 6 October 2009 / Accepted: 22 May 2010 / Published online: 12 June 2010


**Abstract** RNA–protein interactions play a pivotal role in various biological processes, such as mRNA processing, protein synthesis, assembly, and function of ribosome. In this work, we have introduced a computational method for predicting RNA-binding sites in proteins based on support vector machines by using a variety of features from amino acid sequence information including position-specific scoring matrix (PSSM) profiles, physicochemical properties and predicted solvent accessibility. Considering the influence of the surrounding residues of an amino acid and the dependency effect from the neighboring amino acids, a sliding window and a smoothing window are used to encode the PSSM profiles. The outer fivefold cross-validation method is evaluated on the data set of 77 RNA-binding proteins (RBP77). It achieves an overall accuracy of 88.66% with the Matthew's correlation coefficient (MCC) of 0.69. Furthermore, an independent data set of 39 RNA-binding proteins (RBP39) is employed to further evaluate the performance and achieves an overall accuracy of 82.36% with the MCC of 0.44. The result shows that our method has good generalization abilities in predicting RNA-binding sites for novel proteins. Compared with other previous methods, our method performs well on the same data set. The prediction results suggest that the used features are effective in predicting RNA-binding sites in proteins. The code and all data sets used in this article are freely available at http://cic.scu.edu.cn/bioinformatics/Predict_RBP.rar.

**Keywords** RNA–protein interactions · RNA-binding sites prediction · Support vector machines · Position-specific scoring matrix (PSSM)

## Introduction

There is a growing interest in the prediction of RNA-binding sites in proteins because RNA–protein interactions play a key role in various biological processes (Hall 2002; Tian et al. 2004). For example, the ribosome is a protein synthesis complex composed of ribosomal RNAs (rRNAs) and proteins. The identification of rRNA by ribosomal proteins is important for both assembly and function of ribosome (Noller 2005). Transfer RNAs (tRNAs) can be bound to specific proteins for the genetic code translation during protein synthesis (Moras 1992). Some ribonucleoprotein (RNP) particles can take part in the post-transcriptional regulation of gene expression (Varani and Nagai 1998). Thus, identification of the RNA recognition amino acid residues can significantly improve our understanding the mechanisms of various critical biological processes such as mRNA processing, gene expression, protein synthesis, viral replication, cellular defense and developmental regulation (Tuschl 2003), and it can further contribute to advances in drug discovery and drug design, such as providing useful information for the design of RNA drugs for the binding protein (Hermann and Westhof 1998; Ecker and Griffey 1999; Sucheck and Wong 2000).

At present, the structures of known RNA–protein complexes are solved by X-ray crystallography (Berman

Yaping Fang and Cui-cui Wang contributed equally to this work

**Electronic supplementary material** The online version of this article (doi:10.1007/s00726-010-0639-7) contains supplementary material, which is available to authorized users.

C. Wang · Y. Fang · J. Xiao · M. Li (✉)
Key Laboratory of Green Chemistry and Technology, College of Chemistry, Ministry of Education, Sichuan University, Chengdu 610064, China
e-mail: liml@scu.edu.cn

et al. 2000). However, the experimental determination of RNA–protein complex structures remains time consuming and involves expensive experimental technologies. Moreover, the sequence data of proteins are rapidly accumulating from many species, but the structures of most proteins are not available (Wang and Brown 2006a). Therefore, it is necessary to develop effective and reliable computational approaches to identify RNA-binding sites based on the primary amino acid sequence information only. Recently, several computational methods have been developed to identify RNA-binding sites from amino acid sequence information. Jeong et al. (2004) proposed a method based on artificial neural network (ANN) to predict RNA-interacting residues using only the amino acid sequence information and predicted secondary structure information and achieved a MCC of 0.29. Later, the MCC value was improved to 0.41 when the weighted position-specific scoring matrix (PSSM) profiles were integrated into the prediction (Jeong and Miyano 2006). Terribilini et al. (2006) developed the RNABindR method using a Naïve Bayes classifier based on amino acid sequence. Wang and Brown (2006b) provided the BindN method, which was based on support vector machine (SVM) and used the physicochemical properties (PP) including side chain p*K*a value, hydrophobicity index, and molecular mass of an amino acid as input. Kumar et al. (2008) proposed the PPRint method by combining evolutionary information from the PSSM and SVM. Wang et al. (2008) developed the PRINTR method for predicting RNA-binding sites in proteins using PSSM profile and SVM. Recently, Spriggs et al. (2009) presented the method PiRaNha for predicting RNA-binding residues from protein sequence properties information based on SVM. Although there have been many methods for the prediction of RNA-binding sites in proteins, there is also a great scope for improvement in the overall accuracy and the MCC value.

Against this background, we present a method for successfully predicting the RNA-binding sites in proteins based on SVMs by incorporating a variety of features from amino acid sequence information including PSSM profiles, PP and predicted solvent accessibility (PSA). For the PSSM profiles, considering the influence of the surrounding residues of an amino acid and the dependency effect from the neighboring amino acids, a sliding window and a smoothing window are used to encode. Our method is evaluated by the outer fivefold cross-validation on data sets RBP77, RBP107, and RBP109 and achieves good performances in overall accuracy and MCC value. Furthermore, an independent data set RBP39 is also employed to further evaluate the performance of the model on the data set RBP77. The results indicate that this method can perform well in predicting RNA-binding sites in novel proteins. The entire framework of the prediction system is shown in Fig. 1.

## Materials and methods

### Data sets

#### *RBP77 data set*

The non-redundant RNA-binding proteins data set RBP77 is derived from the 86 RNA-binding proteins (Cheng et al. 2008; Kumar et al. 2008), which contains 86 protein chains extracted from the structures of RNA–protein complexes determined by X-ray crystallography with a resolution better than 3 Å in protein data bank (PDB) (Berman et al. 2000; Jeong and Miyano 2006). In the RNA–protein complexes, a residue is considered to be an RNA-interacting residue if any atom of the residue in the protein falls within a 6 Å distance from any atom of the RNA molecule. Otherwise, the residue is a non-binding residue. Sequence redundancy in the data set 86 RNA-binding proteins is 70%, so the blast-clust program (Altschul et al. 1997) is applied to remove the redundant sequences so that the mutual sequence identity in the new data set RBP77 is less than 25%. At last, the resultant data set RBP77 contains 77 protein chains with a total of 18,449 residues. The number

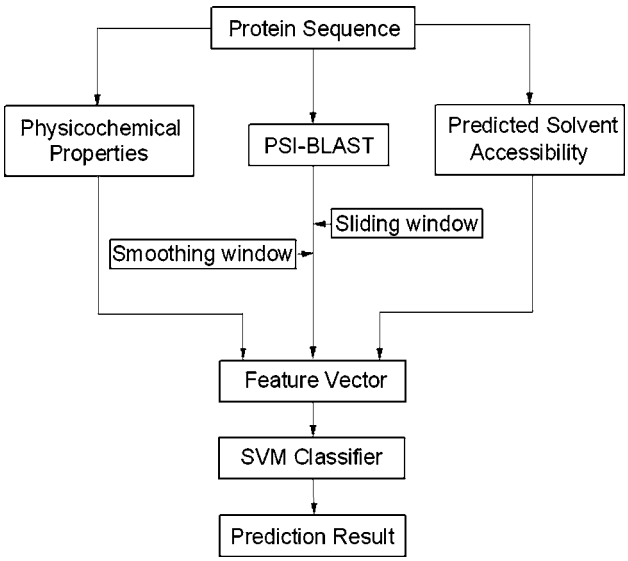


**Fig. 1** The SVM system used for the prediction of RNA-binding sites in proteins. The protein sequence is converted into a feature vector that includes information from PSSM original from PSI-BLAST, physicochemical properties and predicted solvent accessibility. A sliding window and smoothing window were used on the PSSM profiles. The feature vectors were fed in the SVM classifier and obtained the prediction results

of RNA-interacting residues and non-interacting residues are 4,254 and 14,195, respectively.

*RBP107 data set*

The data set RBP107 derived from 61 RNA–protein complexes in PDB (Berman et al. 2000), contains 107 protein chains determined by X-ray crystallography with a resolution better than 3.5 Å. For any two protein chains, the sequence identity is no more than 25%. Used the cut-off of distance 3.5 Å, the resultant data set contains 2,555 RNA-interacting residues and 19,496 non-interacting residues (Wang and Brown 2006b; Cheng et al. 2008; Kumar et al. 2008).

*RBP109 data set*

The data set RBP109 is comprised of 109 non-redundant protein sequences, which are extracted from 56 RNA–protein complexes solved by X-ray crystallography in PDB (Berman et al. 2000) with a resolution better than 3.5 Å. The mutual sequence identity of the non-redundant data set is <30%. Using the ENTANGLE (Allers and Shamoo 2001) default parameters, 3,581 RNA-interacting residues and 21,526 non-interacting residues are obtained, respectively (Terribilini et al. 2006; Wang et al. 2008).

*Independent data set*

To further evaluate the performance of our method in predicting novel proteins, an independent data set RBP39 is employed to test in this work. Data set RBP39 is comprised of 39 non-redundant protein chains derived from PRNA147 (Tong et al. 2008). Using blast-clust program (Altschul et al. 1997), the mutual sequence identity in the data set RBP39 is <25%. Meanwhile, the sequence identity is also <25% between the data set RBP39 and the data set RBP77. At last, the data set RBP39 contains 39 protein chains with 1,451 RNA-interacting residues and 5,930 non-interacting residues.

Support vector machines

The SVM is a kind of machine learning approach based on statistical learning theory proposed by Vapnik (1998). A brief and clear description of how to use the SVM was given by Chou and Cai (2002) and Cai et al. (2002, 2003). In this work the SVM is used to identify RNA-binding residues. LIBSVM (version 2.84; http://www.csie.ntu.edu.tw/~cjlin/libsvm) (Chang and Lin 2001) is employed to carry out the SVM work. This learning machine algorithm has been adopted in our later works and achieved satisfactory performances (Tan et al. 2007; Fang et al. 2008; Guo et al. 2008; Yang et al. 2010). The radial basis function (RBF) is chosen as the kernel function and is defined as follows:

$$K(\vec{x}, \vec{y}) = \exp(-\gamma||\vec{x} - \vec{y}||^2). \quad (1)$$

where $\vec{x}$ and $\vec{y}$ are two data vectors and $\gamma$ is a training parameter. Different values for the regularization parameter $C$ and the kernel width parameter $\gamma$ are tested to obtain the best optimization performance.

The outer fivefold cross-validation method is used to evaluate the performance of all the models in this work. Firstly, the complete data set is randomly divided into five equally sized sets. Four of the sets are used for training on SVM. The performance at each parameter point is evaluated by $n$-fold cross-validation on the training data set using the grid search approach. Then, the remaining fifth set is testing on the resulted model using the optimized parameter $C$ and $\gamma$. This process is repeated five times with different combinations of training (four sets) and testing (the remaining fifth sets) data sets in an outer fivefold cross-validation loop called 'outer-fivefold-cv' method (Batuwita and Palade 2009), and the final result is the average results of these five processes.

A grid search approach was employed to optimize $C$ and $\gamma$ using cross-validation. For each process of the outer fivefold cross-validation, the training data set is separated to $n$ folds. Sequentially a fold is considered as the validation set and the rest are for training. The average of accuracy on predicting the validation sets is the cross-validation accuracy. Different pairs of ($C$, $\gamma$) are tried and the one with the best cross-validation accuracy is selected. The cross-validation procedure can prevent the overfitting problem. It is found that trying exponentially growing sequences of $C$ and $\gamma$ is a practical method to identify the best parameters (e.g., $\log C = -5, -3, \ldots, 15$; $\log \gamma = -15, -13, \ldots, 3$). The 'grid.py' in 'python' subdirectory in libsvm archives automatically executing the procedure above, tries all parameters by calling svm-train within the region specified the model. Grid.py also plots the result by gnuplot (see Fig S1 in the Support Information). In our research, in order to save time for computing, the range for grids is selected as [log C, log $\gamma$] = [−5, 10]*[−10, 3] with the step 2 and the default fold $n = 5$. A more detailed description of the grid search approach was given by Hsu and Lin (2010).

From the above process, it is seen that it should be five best values for ($C$, $\gamma$) for each data set because of outer fivefold cross-validation processes. However, we obtained the same best value for ($C$, $\gamma$) during the optimizing process. For the data sets RBP77, RBP107, and RBP109, the final results were obtained by the average of accuracy on the outer fivefold cross-validation procedures accuracy, respectively.

## Feature extraction and representation

To build a predictor that can distinguish between RNA-interacting residues and non-interacting residues, we have extracted a number of features based on PSSM, PP and PSA for training the SVM.

### *Position-specific score matrix*

The evolutionary information has been used in predicting DNA-binding sites in proteins (Ahmad and Sarai 2005). It also shows that it is effective for the RNA-binding site prediction (Kumar et al. 2008). In this work, we use three iterations of PSI-BLAST to search against the Swiss-Prot database (version 54.4, released on 25 Oct 2007) for multiple sequence alignment against each protein and to generate a PSSM based on BLOSUM62 substitution matrix with $E$ value as 0.001. Then, we obtain a PSSM profile, which contains the evolutionary information vectors comprised of log-likelihoods for 20 different amino acids for the residue at each position in a protein sequence. The size of PSSM matrix of a protein is $20 \times n$ (the length of a protein).

In the original PSSM profiles, each amino acid $a_i$ in a protein sequence can be encoded as an evolutionary information vector of dimension 20 using the $i$th row of PSSM. Furthermore, considering the influence of the neighbor residues of an amino acid, a sliding window is used to incorporate the evolutionary information from upstream and downstream neighbors. Thus, for an amino acid residue $a_i$ in sequence position $i$, we construct a feature vector represented by a $V_i$ of dimension $20 \times w$. $w$ is an odd number which stands for the size of sliding window.

$$V_i = v[a_{i-(w-1)/2}], \ldots, v[a_i], \ldots, v[a_{i+(w-1)/2}] \quad (2)$$

where $v[a_i]$ is the $i$th PSSM row relative to the residue $a_i$. If the window extends beyond the sequence of N-terminal and C-terminal then $(w-1)/2$ zero vectors of dimension 20 are appended on empty positions before the first and after the last residue of a PSSM profile. The profile adding a sliding window is defined as the standard PSSM profile (Cheng et al. 2008).

In the standard PSSM profile, the value at each position is calculated based on the assumption that each position is independent of others. Terribilini et al. (2006) analyzed RNA-binding residues and demonstrated that RNA-binding residues tend to occur in clusters. Thus, considering the effects of the neighbor residues, a smoothed encoding scheme (Cheng et al. 2008) is applied to incorporate the dependency and correlation of the neighbor residues of a central residue. The smoothed PSSM profiles are encoded by calculating the evolutionary information of a central position based on the sum of the surrounding residues' evolutionary information. This process is similar to the spatial domain method used in the research field of image processing (Gonzalez and Woods 2002). For encoding a smoothed PSSM profile, a smoothing window (ws, an odd number) is used. Each row vector of a residue $\alpha_i$ is represented and smoothed by the summation of ws surrounding row vectors:

$$V_{\text{smoothed\_}i} = v[a_{i-(\text{ws}-1)/2}] + \cdots + v[a_i] + \cdots + v[a_{i+(\text{ws}-1)/2}] \quad (3)$$

For the N-terminal and the C-terminal of a protein the $(ws-1)/2$ zero vectors are appended on the empty positions before the first and after the last of a smoothed PSSM profile. After using the sliding and smoothing window encoding scheme, the feature vector of the residue $\alpha_i$ is represented by a vector $A_i$ of dimension $20 \times w$ as follows:

$$A_i = V_{\text{smoothed\_}i-(w-1)/2}, \ldots, V_{\text{smoothed\_}i}, \ldots, V_{\text{smoothed\_}i+(w-1)/2} \quad (4)$$

Figure 2 illustrates an example of an original PSSM profile, a standard PSSM profile and a smoothed PSSM profile. For the residue $a_7$, the feature vector is represented by [$v_6$, $v_7$, $v_8$] when a sliding window size 3 is applied on the original vector in profile a. The vector of dimension $20 \times 3$ is obtained in profile b. Then, for the residue $a_7$ in profile b, the corresponding value of amino acid '$A$' is represented by the sum of $[(-2) + 0 + (-2) + (-3) + (-5)]$ when a smoothing window size 5 is applied to encode as seen in profile c.

### *Physicochemical properties*

We extract eight physicochemical properties including hydrophobicity (Tanford 1962), hydrophobic moments, net charge index of side chains of amino acids (NCI) (Zhou et al. 2006), NCI moments, propensity, propensity moment, mass, and side chain p$K$a value. The original values of the physicochemical properties for each amino acid are listed in Table S1 in the Support Information.

The first two features are hydrophobicity and hydrophobic moment. The hydrophobicity plays an important role in amino acid side chain packing and protein folding (Gallet et al. 2000). It has been used in predicting the protein–protein interaction, DNA–protein interaction and RNA–protein interaction (Wang and Brown 2006b; Fang et al. 2008; Chen and Jeong 2009). The hydrophobicity and hydrophobic moment were initially used to distinguish membrane α-helix proteins from soluble proteins (Eisenberg et al. 1984). Recently, they are also used to predict the protein interaction site from sequences (Gallet et al. 2000; Chen and Jeong 2009). The Eisenberg's algorithm is used to calculate the

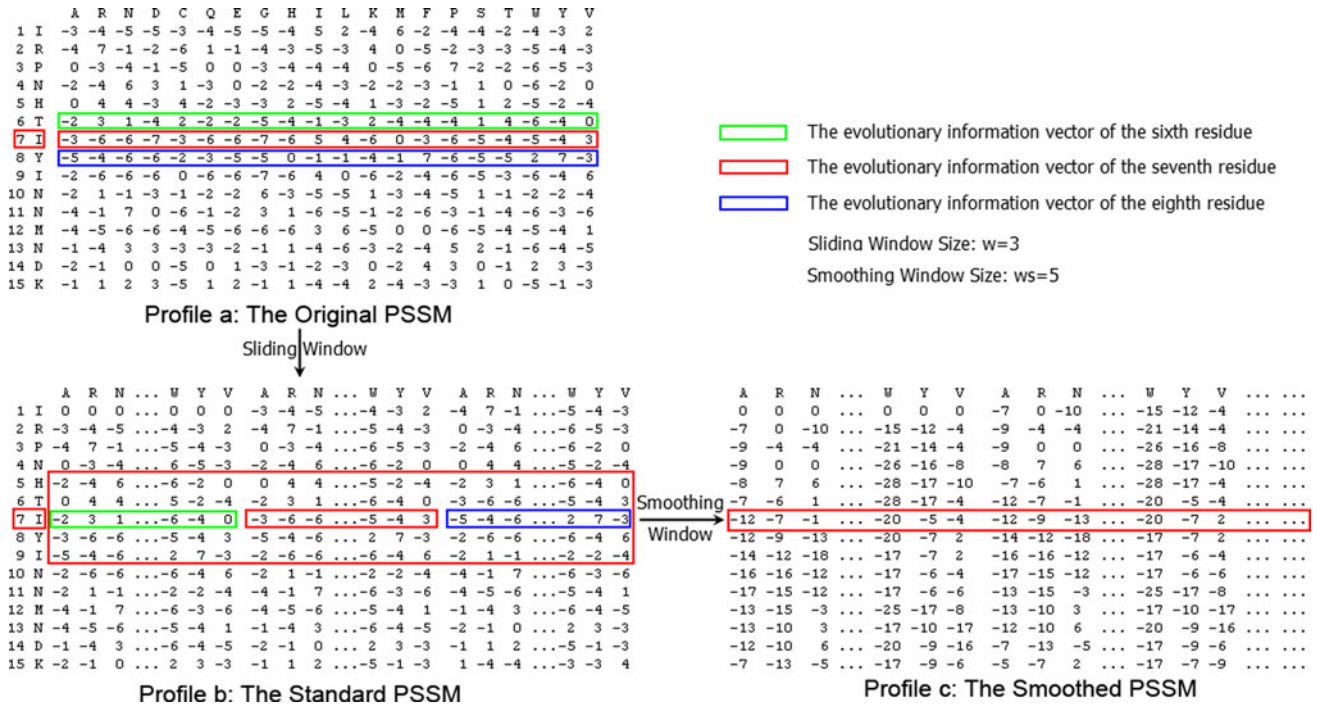


**Fig. 2** Examples of the encoding scheme for the PSSM profiles using sliding window and smoothing window

mean hydrophobicity and mean hydrophobic moment (Eisenberg et al. 1982) as follows:

$$\langle H_i \rangle = \frac{1}{2N+1} \sum_{n=-N}^{N} h_n^{(i)} \tag{5}$$

$$\langle \mu H_i \rangle = \frac{1}{2N+1} \left[ \left( \sum_{n=-N}^{N} h_n^{(i)} \sin(\delta n) \right)^2 + \left( \sum_{n=-N}^{N} h_n^{(i)} \cos(\delta n) \right)^2 \right]^{1/2} \tag{6}$$

An odd sliding window is moved along the protein sequence. The mean hydrophobicity and mean hydrophobic moment are assigned to the center amino acid $i$. $N$ is the size of the sliding window centered around amino acid $i$. $h_n^{(i)}$ is the hydrophobicity of the amino acid that is the $n$th away from the amino acid $i$, and $\delta n$ is the gyration angle between two consecutive residues in the sequence. Gallet et al. (2000) found the method to achieve the most successfully when they used $N = 5$ and $\delta n = 100°$.

Similar to the mean hydrophobicity and mean hydrophobic moment, other feature moments are also calculated, such as mean NCI, mean NCI moment, mean propensity, mean propensity moment, mean mass, and mean p$K$a value.

The NCI describes the integral electric property of the side chains of amino acids (Zhou et al. 2006). The interface propensities of the residues describe whether an amino acid is possibly exposed to solvent or buried in an interface (Jones and Thornton 1997; Chen and Jeong 2009). The side chain p$K$a value determines the ionization state of a residue in protein sequence. The ionization state of amino acid side chains may play an important role in RNA–protein interactions because the RNA molecules are negatively charged. The side chain p$K$a values used in this work are from the book (David and Cox 2000) and the p$K$a value is set to 7 for the amino acid without a side chain p$K$a value. The mass is the molecular mass of an amino acid. Since each amino acid has a unique value of mass, which may be relative to the space volume of a residue taken up in structures. The mass has been used in predicting RNA-bind sites in proteins and performs well (Wang and Brown 2006b).

*Predicted solvent accessibility*

The predicted solvent accessibility is used to represent the solvent exposure of the residue. We use SABLE (version 2.0) to predict the relative solvent accessibility of each residue in each protein sequence (Wagner et al. 2005; Spriggs et al. 2009).

Consequently, for each residue in proteins, the total numbers of the input features are $20 \times w$, 8 and 1 for group (a), group (b) and group (c), respectively.

Performance measures

To assess the performance of various models developed in this work, the following measures including sensitivity

(Se), specificity (Sp), Matthew's correlation coefficient (MCC) and overall accuracy (Acc) are used to evaluate. These measures were also used in previous works (Guo et al. 2006; Wang and Brown 2006a; Wang and Brown 2006b; Wen et al. 2007; Guo et al. 2008; Kumar et al. 2008). Sensitivity is the ratio of correctly predicted RNA-binding residues. Specificity is the ratio of correctly predicted non-interacting residues. Overall accuracy is the percentage of correctly predicted RNA-binding residues and non-interacting residues, which tends to give a highly misleading impression of the prediction quality when the data set is unbalanced. MCC is a more robust measure of the prediction quality, which takes into account both over- and under-predictions and gives a complementary measure of the prediction performance (Matthews 1975; Baldi et al. 2000). MCC = 1, MCC = 0, MCC = −1 denote a perfect prediction, a completely random assignment, and a perfectly reverse correlation, respectively. The sensitivity, specificity, MCC, and overall accuracy (Acc) are calculated in the following equations, respectively.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{7}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{8}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \times (\text{TP} + \text{FN}) \times (\text{TN} + \text{FP}) \times (\text{TN} + \text{FN})}} \tag{9}$$

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100 \tag{10}$$

where TP, TN, FP, and FN denote the numbers of true positives (residues predicted to be RNA-binding residues that are in fact RNA-binding residues), true negatives (residues predicted to be non-RNA-binding residues that are in fact not RNA-binding residues), false positives (residues predicted to be RNA-binding residues that are in fact not RNA-binding residues), and false negatives (residues predicted to be non-RNA-binding residues that are in fact RNA-binding residues), respectively.

To further evaluate the performance of our method, we also use the receiver operating characteristic (ROC) curve (Swets 1988), which is probably one of the most robust approaches for classifier evaluation. The ROC curve is obtained by plotting true positive rate (sensitivity) on the *y*-axis against the false positive rate (1-specificity) on the *x*-axis. The area under the ROC curve (AUC) (Bradley 1997) can be used as a reliable measure for the prediction performance. The maximum value of AUC (1) denotes a perfect prediction. A random guess receives an AUC value close to 0.5.

## Results and discussion

### Window size optimization and selection

In order to obtain the best performance for predicting RNA–protein binding sites, the sliding window size and smoothing window size are optimized. In this work, the best window size is optimized with the respect to the overall accuracy and MCC value. The optimization sliding window size is obtained by testing the performance of different sliding window sizes $w$ from 3 to 25 with the smoothing window size of 11 and the default parameters $C$ and $\gamma$ in SVM. Figure 3a shows the results of different sliding window sizes on the data set RBP77. One can see that setting the sliding window size to 15 achieves the best predicting performance. Therefore, the optimization sliding window size is set to 15 in the following study. Similar to the optimization of sliding window size, we also test the performance of different smoothing window sizes ws from 1 to 21 with the sliding window size of 15 and the default parameters C and $\gamma$ in SVM. The results on the data set RBP77 are shown in Fig. 3b. The results indicate that the smoothing window encoding scheme can improve the performance significantly. Especially, when ws is set to 9, we obtain the highest overall accuracy and MCC value. Therefore, 9 is selected as the smoothing window size in the following study.

### Prediction performance of SVM method using various features on the data set RBP77

For the data set RBP77, the outer fivefold cross-validation method is used to evaluate the prediction performance. We train the SVM by all possible combination of the three groups of features using the optimization window sizes and the optimized parameters in SVM ($C = 8$; $\gamma = 0.03125$; weight parameter = 4). The prediction results of the SVM models using various features are listed in Table 1. It shows that the prediction performance increases significantly while using the encoded PSSM profiles as input. When the PSSM profile with combination of one or more additional features are used as input, the performance is improved slightly. The best performance is achieved as the MCC of 0.688 and overall accuracy of 88.66% (with sensitivity 78.51% and specificity 91.71%) by combination of all the features as input. The results indicate that the combination of all the features is capable of capturing more information for predicting RNA-binding sites and non-binding sites in proteins. So, the model by combination of all the features is selected as the final model. In addition, we also calculate the area under the ROC curve. The ROC curve for the prediction performance on the data set RBP77 is

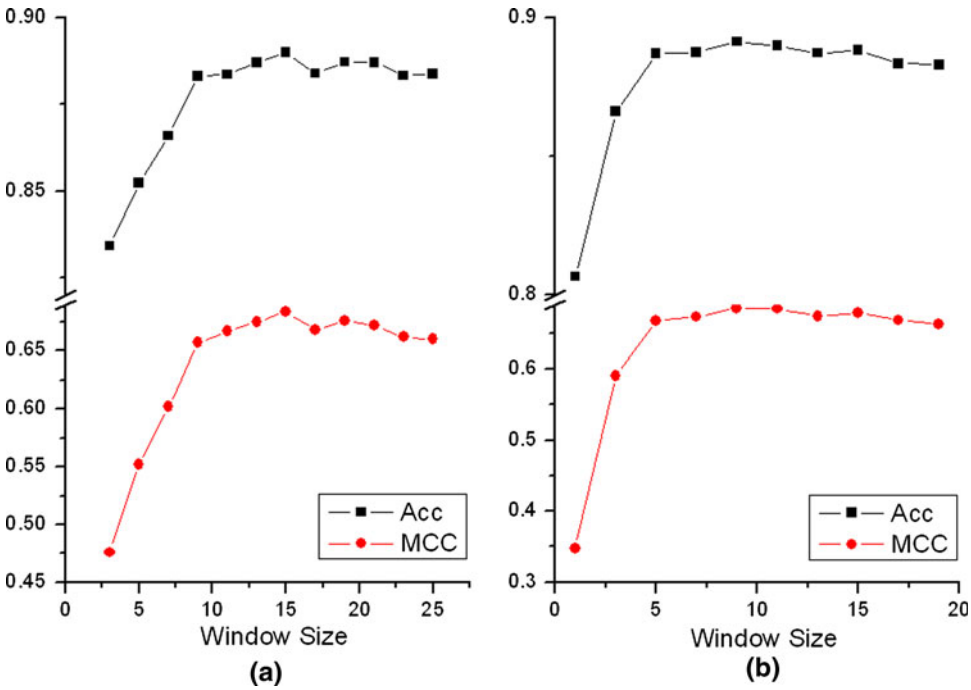


**Fig. 3** **a** The performance on the data set RBP77 with different sliding window sizes. **b** The performance on the data set RBP77 with different smoothing window sizes

**Table 1** The prediction performance of the SVM models based on various features by outer fivefold cross-validation

| Feature vector | Acc (%) | Se (%) | Sp (%) | MCC |
|---|---|---|---|---|
| PP | 69.03 | 54.55 | 73.31 | 0.252 |
| PSA + PP | 69.15 | 55.89 | 73.12 | 0.261 |
| PSSM | 88.56 | 78.97 | 91.26 | 0.687 |
| PSSM + PP | 88.45 | 78.42 | 91.47 | 0.683 |
| PSSM + PSA | 88.56 | 78.74 | 91.50 | 0.686 |
| PSSM + PSA + PP | 88.66 | 78.51 | 91.71 | 0.688 |

*PSA* predicted solvent accessibility, *PP* physicochemical properties

shown in Fig. 4 in red. The AUC value is 0.93, which is higher than other methods on the same data set.

## Prediction performance on the data set RBP107 and RBP109

To further evaluate the performance of our method and to compare with other methods, another two data sets RBP107 and RBP109 are employed to construct models for prediction. For the data set RBP107, it achieves the MCC value of 0.44 and the overall accuracy of 83.94% (with sensitivity 69.57% and specificity 85.83%) by the outer fivefold cross-validation based on the optimized parameters ($C = 0.35$; $\gamma = 0.07$) and the weight parameter 8 in SVM. We can obtain high sensitivity with a slight decrease in specificity when the weight parameter is added because the ratio of the number of non-interacting residues to that of interacting residues in the data set RBP107 is 7.6:1. The ROC curve for the prediction performance on the data set RBP107 is shown in Fig. 4 in green. The AUC value is 0.82, which is significantly higher than a random guess 0.5. Similar to the data set RBP107, the outer fivefold cross-validation method is also used to evaluate the performance on the data set RBP109. It achieves the MCC value of 0.58 and overall accuracy of 89.35% (with sensitivity 67.56% and specificity 92.97%) with the optimized parameters ($C = 8$; $\gamma = 0.03125$) and the weight parameter 6 in SVM. The ROC curve for the prediction performance on the data set RBP109 is shown in Fig. 4 in blue and the AUC value is 0.88.

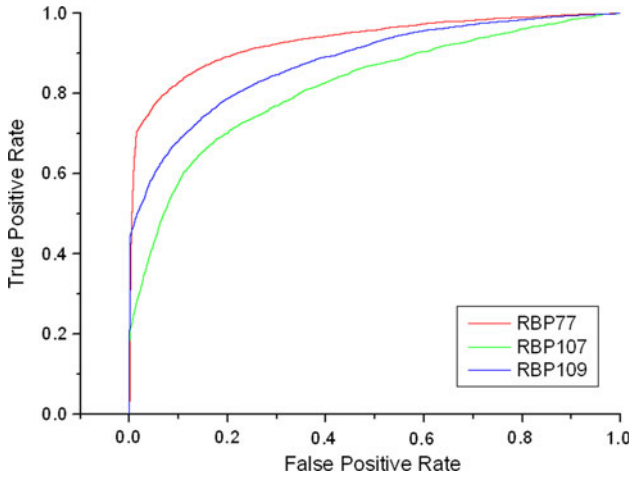


**Fig. 4** The ROC curves for the prediction of RNA-binding sites in proteins by outer fivefold cross-validation on the data sets RBP77, RBP107 and RBP109

### Prediction performance of SVM on the independent data set

In order to demonstrate and evaluate the performance of the proposed method, an independent data set RBP39 is also used for testing. The prediction is carried out on the best model of the data set RBP77. It achieves the MCC value of 0.44 and overall accuracy of 82.36% (with sensitivity 54.90% and specificity 89.07%). For comparison purposes, the prediction is also carried out on the PiPaNhA (Spriggs et al. 2009), PPRInt (Kumar et al. 2008) and BindN (Wang and Brown 2006b) servers using default parameters. The results are listed in Table 2. The results of all the methods show a difference bias in sensitivity and specificity. PPRInt achieves a higher sensitivity but with a lower specificity, and our method achieves better than PiPaNhA's in sensitivity but a bit lower in specificity. Based on the performance measures MCC and overall accuracy, our method outperforms the other two methods BindN and PPRInt. Moreover, our method achieves higher sensitivity than PiPaNhA. Overall, the result shows that our method has good generalization abilities in predicting RNA-binding sites in novel proteins.

### Comparison with other methods

Jeong and Miyano (2006) proposed the method based on ANN using multiple sequence profiles and weighted profiles, respectively. They achieved the best MCC of 0.41 when used weighted PSSM profiles. Wang and Brown (2006b) proposed the method BindN on the data set 107 RNA-binding proteins and achieved the accuracy of 69.32% with the MCC of 0.27. Terribilini et al. (2006) developed the RNABindR method on the data set 109 RNA-binding proteins using a Naïve Bayes classifier on amino acid sequence and achieved the MCC of 0.35. Kumar et al. (2008) developed an improved approach combining evolutionary information and SVM and achieved the MCC of 0.45 on the data set 86 RNA-binding proteins and 0.32 on the data set 107 RNA-binding proteins. Cheng et al. (2008) proposed the method RNAProB on three benchmark data sets and achieved good results. Wang et al. (2008) developed a PRINTR for predicting RNA-binding sites in proteins using PSSM profile and SVM and achieved the accuracy of 87.10 with the MCC of 0.43. Recently, Spriggs et al. (2009) presented the method PiRaNha based on SVM and achieved the accuracy of 87.2% with the MCC of 0.50.

In this work, a SVM classifier is developed on the data set RBP77 using PSSM profile, physicochemical properties and predicted solvent accessibility as input. It achieves the comparatively level of the MCC 0.69 and the overall accuracy of 88.66% (with sensitivity 78.51% and specificity 91.71%). The results show that our method improves the sensitivity significantly compared with the PiRaNhA and improved the specificity significantly compared with the RNAProB (Table 3). Especially, our data set has a less sequence identify than other methods'. Additionally, SVM models are also developed on data sets RBP107 and RBP109 for a convenient comparison with other methods. The compared results are listed in Table 3. On the data set RBP107, our method achieves the best overall accuracy and MCC value. On the data set RBP109, the results show that our method reaches the comparatively level of MCC 0.58 and overall accuracy 89.35%. In addition, our method improves the sensitivity significantly to 67.56% compared with the best previous methods on the data set RBP109. On the whole, the results demonstrate that our method improves the sensitivity significantly and achieves well performance for predicting RNA-binding sites in proteins.

**Table 2** Prediction for the independent data set RBP39 using our method and using BindN, PPRInt and PiRaNhA servers

| Methods (reference) | Data set | Acc (%) | Se (%) | Sp (%) | MCC | AUC |
|---|---|---|---|---|---|---|
| Jeong 2006 (Jeong and Miyano 2006) | RBP86 | 80.20 | 43.40 | 91.04 | 0.39 | – |
| Pprint (Kumar et al. 2008) | | 81.20 | 53.05 | 89.55 | 0.45 | – |
| RNAProB (Cheng et al. 2008) | | 87.99 | 79.95 | 90.36 | 0.68 | – |
| PiRaNhA (Spriggs et al. 2009) | RNAset81 | 87.2 | 56.3 | 92.8 | 0.50 | 0.86 |
| Our method | RBP77 | 88.66 | 78.51 | 91.71 | 0.69 | 0.93 |
| BindN-PCP (Wang and Brown 2006b) | RBP107 | 69.32 | 66.28 | 69.84 | 0.27 | 0.73 |
| RNAProB (Cheng et al. 2008) | | 80.44 | 77.14 | 80.87 | 0.42 | – |
| Pprint (Kumar et al. 2008) | | 75.43 | 70.09 | 75.54 | 0.32 | – |
| Our method | | 83.94 | 69.57 | 85.83 | 0.44 | 0.82 |
| RNABindR (Terribilini et al. 2006) | RBP109 | 84.80 | 38.00 | 93.00 | 0.35 | – |
| RNAProB (Cheng et al. 2008) | | 89.70 | 64.62 | 93.88 | 0.58 | – |
| PRINTR (Wang et al. 2008) | | 87.10 | 55.90 | – | 0.43 | – |
| Our method | | 89.35 | 67.56 | 92.97 | 0.58 | 0.88 |

**Table 3** Performance comparison between our method and other methods on the benchmark data sets

| Method (reference) | Acc (%) | Se (%) | Sp (%) | MCC |
|---|---|---|---|---|
| BindN (Wang and Brown 2006b) | 73.43 | 44.83 | 80.42 | 0.24 |
| Pprint (Kumar et al. 2008) | 75.79 | 68.21 | 77.64 | 0.39 |
| PiRaNhA (Spriggs et al. 2009) | 84.03 | 54.21 | 91.32 | 0.47 |
| Our method | 82.36 | 54.90 | 89.07 | 0.44 |

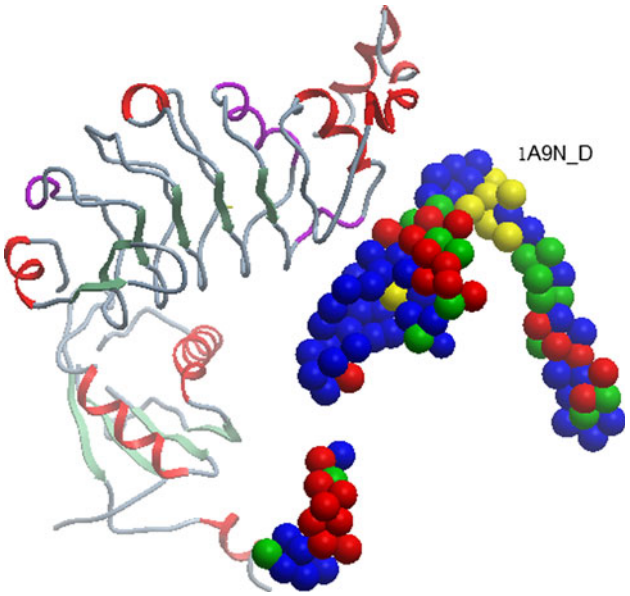


**Fig. 5** Representative prediction results shown in the context of three-dimensional structures for 1A9N_D in spacefill. The correctly predicted RNA-binding residues (true positives) are in *red* (*black*); the correctly predicted non-binding residues (true negatives) are in *blue* (*dark grey*); the binding residues but predicted as negatives (false negatives) are in *green* (*deep grey*); the non-binding residues but predicted as positives (false positives) are in *yellow* (*light grey*)

### Evaluation the prediction in the context of three-dimensional structures

In Fig. 5, we evaluate an example of the predicted RNA-binding residues of the protein 1A9N_D in the context of three-dimensional structures using the best model trained on the data set RBP77. The protein 1A9N_D is not included in the data set RBP77. The true positives are in red; the true negatives are in blue; the false negatives are in green; the false positives are in yellow. The overall accuracy of the prediction is 74.19%. For 54 non-binding residues, 47 residues in blue are predicted correctly (87.04%). The results demonstrate that the predictions of our method can provide useful information for understanding RNA–proteins interactions, which may be guide experimental studies such as the design of RNA drugs for the binding protein.

## Conclusions

In this work, a computational method has been described based on the amino acid sequence information with the SVM models for predicting RNA-binding sites in proteins. A variety of physicochemical properties, PSSM profiles and predicted solvent accessibility are effectively extracted to construct the models. Meanwhile, a sliding window and a smoothing window are used to encode the PSSM profiles. The prediction results show that using the sliding window and smoothing window on PSSM profiles performs better than using the original PSSM profiles. The method has been evaluated by the outer fivefold cross-validation method on the data sets RBP77, RBP107 and RBP109. The results show that the model constructed with combination of all features as input has the best performance. Based on an independent data set RBP39, the method also has good generalization abilities in predicting novel proteins.

With the level of success performance in this work, the method proposed for predicting RNA-binding sites in proteins may be helpful in the process of investigating functions of novel proteins. Furthermore, this method could be applied in the RNA-binding function predictions at the level of the whole proteins in future work.

**Acknowledgments** The work was funded by the National Natural Science Foundation of China (No. 20775052). The authors would like to express their cordial thanks to the unknown reviewers for providing comments on the manuscript.

## References

Ahmad S, Sarai A (2005) PSSM-based prediction of DNA binding sites in proteins. BMC Bioinform 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 database search programs. Nucleic Acids Res 25:3389–3402

Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16:412–424

Batuwita R, Palade V (2009) microPred: effective classification of pre-miRNAs for human miRNA gene prediction. Bioinformatics 25:989–995

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

Bradley AP (1997) The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognit 30:1145–1159

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for predicting HIV protease cleavage sites in protein. J Comput Chem 23:267–274

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

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen X, Jeong JC (2009) Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics 25:585–591

Cheng CW, Su E, Hwang JK, Sung TY, Hsu WL (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinform 9:S6

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

David LN, Cox MM (2000) Lehninger principles of biochemistry. Worth Publishers, New York

Ecker DJ, Griffey RH (1999) RNA as a small-molecule drug target: doubling the value of genomics. Drug Discov Today 4:420–429

Eisenberg D, Weiss RM, Terwilliger TC (1982) The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature 299:371–374

Eisenberg D, Schwarz E, Komaromy M, Wall R (1984) Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J Mol Bio 179:125–142

Fang Y, Guo Y, Feng Y, Li M (2008) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids 34:103–109

Gallet X, Charloteaux B, Thomas A, Brasseur R (2000) A fast method to predict protein interaction sites from sequences. J Mol Biol 302:917–926

Gonzalez RC, Woods RE (2002) Digital image processing. Prentice-Hall, Englewood Cliffs

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

Guo Y, Yu L, Wen Z, Li M (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36:3025–3030

Hall KB (2002) RNA–protein interactions. Curr Opin Struct Biol 12:283–288

Hermann T, Westhof E (1998) RNA as a drug target: chemical, modelling, and evolutionary tools. Curr Opin Biotech 9:66–73

Hsu CW, Lin CJ (2010) A practical guide to support vector classification. http://www.csie.ntu.edu.tw/~cjlin/libsvm

Jeong E, Miyano S (2006) A weighted profile based method for protein–RNA interacting residue prediction. Lect Notes Comput Sci 3939:123–139

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

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

Kumar M, Gromiha MM, Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71:189–194

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

Moras D (1992) Aminoacyl-tRNA synthetases. Curr Opin Struct Biol 2:138–142

Noller HF (2005) RNA structure: reading the ribosome. Science 309:1508–1514

Spriggs RV, Murakami Y, Nakamura H, Jones S (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics 25:1492–1497

Sucheck SJ, Wong CH (2000) RNA as a target for small molecules. Curr Opin Chem Biol 4:678–686

Swets JA (1988) Measuring the accuracy of diagnostic systems. Science 240:1285–1293

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 33:669–675

Tanford C (1962) Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. J Am Chem Soc 84:4240–4247

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 12:1450–1462

Tian B, Bevilacqua PC, Diegelman-Parente A, Mathews MB (2004) The double-stranded-RNA-binding motif: interference and much more. Nat Rev Mol Cell Biol 5:1013–1023

Tong J, Jiang P, Lu Z (2008) RISP: A web-based server for prediction of RNA-binding sites in proteins. Comput Methods Programs Biomed 90:148–153

Tuschl T (2003) Functional genomics: RNA sets the standard. Nature 421:220–221

Vapnik V (1998) Statistical learning theory. Springer, New York

Varani G, Nagai K (1998) RNA recognition by RNP proteins during RNA processing. Annu Rev Biophys Biomol Struct 27:407–445

Wagner M, Adamczak R, Porollo A, Meller J (2005) Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 12:355–369

Wang L, Brown SJ (2006a) Prediction of RNA-binding residues in protein sequences using support vector machines. Conf Proc IEEE Eng Med Biol Soc 1:5830–5833

Wang L, Brown SJ (2006b) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34:W243–W248

Wang Y, Xue Z, Shen G, Xu J (2008) PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids 35:295–302

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) 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

Yang L, Li Y, Xiao R, Zeng Y, Xiao J, Tan F, Li M (2010) Using auto covariance method for functional discrimination of membrane proteins based on evolution information. Amino Acids 38:1497–1503

Zhou P, Tian F, Li B, Wu S, Li Z (2006) Genetic algorithm-based virtual screening of combinative mode for peptide/protein. Acta Chim Sinica 64:691–697