Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

Prediction of RNA-Binding Residues in Protein Sequences Using Support Vector Machines

Liangjiang Wang, Susan J. Brown

Abstract— **Understanding the molecular recognition between RNA and proteins is central to elucidation of many biological processes in the cell. Although structural data are available for some protein-RNA complexes, the interaction patterns are still mostly unclear. In this study, support vector machines as well as artificial neural networks have been trained to predict RNA-binding residues from five sequence-derived features, including the solvent accessible surface area, BLAST-based conservation score, hydrophobicity index, side chain pK_a value and molecular mass of an amino acid. It is found that support vector machines outperform neural networks for prediction of RNA-binding residues. The best support vector machine achieves 70.74% of prediction strength (average of sensitivity and specificity), whereas the performance measure reaches 67.79% for the neural networks. The results suggest that RNA-binding residues can be predicted directly from amino acid sequence information. Online prediction of RNA-binding residues is available at http://bioinformatics.ksu.edu/bindn/.**

I. Introduction

Knowledge of protein-RNA recognition is critical for understanding many biological processes, including RNA splicing, turnover and translation. For example, the cellular machinery for protein synthesis, or ribosome, is assembled from various ribosomal RNA (rRNA) and protein molecules. The recognition of rRNA by ribosomal proteins is important for both assembly and function of ribosomes. Furthermore, since some viruses have a RNA genome surrounded by capsid proteins and require the involvement of host proteins for replication, identification of the amino acid residues that bind to viral RNA may provide useful information for antiviral drug design [1].

Structural data of protein-RNA complexes provide valuable information for understanding the molecular mechanisms of protein-RNA recognition. Analysis of the available structures at the atomic level suggests that protein-RNA recognition involves a complex combination of hydrogen bonds, van der Waals contacts and electrostatic interactions between amino acid residues and RNA bases [2]. As for residue-wise patterns, the basic amino acids, arginine and lysine, occur more frequently at protein-RNA interfaces than non-binding sites, whereas the acidic amino acids, aspartic acid and glutamic acid, are rarely found as RNA-binding residues due to the negative charge of the RNA backbone [3].

However, it is still challenging to predict RNA-binding residues directly from amino acid sequence information. The sequence-based approach is needed because sequence data are rapidly accumulating from many species but the structures of most proteins are not available. The problem for machine learning can be specified as follows: given the amino acid sequence of a protein that is supposed to bind RNA, the task is to predict which amino acid residues may be located at the interaction interface. Since both the structure of the protein and the sequence of the target RNA are assumed to be unknown, potential RNA-binding residues need to be predicted from the amino acid properties and local sequence patterns in the protein.

Despite the importance of protein-nucleic acid interactions and the need of predictive methods for protein sequence analysis, only a few studies have been reported for prediction of binding residues from amino acid sequence information. For DNA-binding residues, artificial neural networks were trained with sequence and residue solvent accessibility information, and the predictor achieved 40.3% sensitivity and 81.8% specificity [4]. Since the dataset was imbalanced with more negative data instances than positive ones, the prediction strength was measured by the average of sensitivity and specificity, which was 61.1% for the above predictor. Evolutionary information in terms of position-specific scoring matrices (PSSMs) was found to enhance the prediction strength to 67.1% with 68.2% sensitivity and 66.0% specificity [5]. In a related study, machine learning approaches were developed to predict RNA-binding proteins based on primary sequence information [6]. However, RNA-binding residues were not predicted in that study.

In the present study, support vector machines (SVMs) are used to predict RNA-binding residues from amino acid sequence information. SVM is a relatively new machine learning algorithm [7], and has recently been applied to a variety of biological problems for pattern classification [8]. In biological applications, support vector machines often outperform the other machine learning algorithms such as neural networks due to SVM's superior generalization power and its ability to avoid overfitting. The SVM learning algorithm is particularly appealing for prediction of RNA-binding residues. Since the sequence data space of RNA-binding sites appears to be very large but the observations of protein-RNA interactions based on structural data are scarce, model overfitting is the major concern in this case.

Manuscript received April 3, 2006. This work is supported by the K-INBRE Bioinformatics Core (NIH grant number P20 RR016475).

L. Wang and S. J. Brown are with the Bioinformatics Center, Division of Biology, Kansas State University, Manhattan, KS 66506, USA (corresponding author: L. Wang, phone: 1-785-532-6347; fax: 1-785-532-6653; e-mail: ljwang@ksu.edu).

Furthermore, a new method has been developed to encode each residue with sequence-derived features. These features provide biological information, which may not be learned directly from the sequences data. Feature extraction also reduces the dimensionality of the sequence data, and thus may overcome the problem of the small training dataset. The results suggest that our approach is effective for accurate prediction of RNA-binding residues from amino acid sequence information.

II. Methods

A. Data Preprocessing

Structural data of protein-RNA complexes were retrieved from the Protein Data Bank (http://www.rcsb.org/pdb/). Structures that had been determined by X-ray crystallography with resolution better than 3.5 Å were selected for this study. The structure dataset had 174 protein-RNA complexes.

The structures were then analyzed for identification of RNA-binding residues. An amino acid residue was designated as a binding site if the side chain or backbone atoms of the residue fell within a cutoff distance of 3.5 Å from any atoms of the RNA molecule in the complex. All the other residues were regarded as non-binding sites. The same criterion was used in the previous studies for identification of DNA-binding residues [4], [5]. A Perl program was developed to take a set of structure files as the input and create an output file of amino acid sequences with each residue labeled as a RNA-binding or non-binding site.

To remove redundancy among the amino acid sequences, clustering analysis was performed using the *blastclust* program (http://www.ncbi.nlm.nih.gov/BLAST/) with the sequence identity threshold set to 25%. From each cluster, the longest sequence was selected. The non-redundant dataset, named PRINR25, had 107 sequences with 3,239 RNA-binding residues and 18,519 non-binding residues.

B. Feature Extraction

Five different sequence features (*A*, *B*, *H*, *K* and *M*) have been selected to encode an amino acid residue. The *A* feature is the relative solvent accessible surface area (ASA) of a residue, which was previously used for prediction of DNA-binding residues [4]. In this study, relative ASA was predicted from sequence data using the PHDacc program (http://cubic.bioc.columbia.edu/pp/).

The *B* feature indicates how well a sequence position is conserved in a BLAST search against a reference database. Let $H_p = \{h_1, h_2, \ldots, h_n\}$ be the set of n hits ($n > 0$) in the BLAST search for a given protein sequence p. Each hit may include one or more pair-wise sequence alignments, in which the BLAST program indicates whether two aligned residues are identical or show similarity based on the BLOSUM62 scoring matrix [9]. The feature value for the residue a_i at position i in p is computed as follows:

$$B_{a_i}^{p} = \frac{\sum_{h_j \in H_p} f(a_i, h_j)}{n + \frac{c}{n}}$$

where $f(a_i, h_j)$ is set to 1 if a_i is aligned to an identical or similar residue in h_j, or 0 otherwise, and c is a pseudo-count, which was set to 10 in this work. The term (c/n) is used to scale the feature value, and it becomes smaller (close to 0) when n gets larger. If p has no BLAST hit in the reference database ($n = 0$), the feature value is set to 0. The protein sequence dataset UniProtKB (http://www.pir.uniprot.org/) was used as the reference database, and the E-value threshold for the BLAST search was set to 1e-5 in this study.

The other three features represent biochemical properties of an amino acid. The *H* feature is the hydrophobicity index of an amino acid [10]. Hydrophobicity is a key factor in amino acid side chain packing and protein folding. Hydrophobic amino acids are often located inside globular proteins but rarely found at protein-RNA interfaces. The *K* feature takes the amino acid side chain pK_a value, which determines the ionization state of a residue in protein sequences. Since the phosphate groups of RNA are negatively charged, the ionization state of amino acid side chains may play an important role in protein-RNA interactions. In this study, the side chain pK_a values from [11] were used. The *K* feature value was set to 7 for the amino acids without a side chain pK_a value. The *M* feature is simply the molecular mass of an amino acid. Each amino acid has a unique value of mass, which is related to the volume of space that a residue occupies in structures.

C. Training Strategies

The training and test datasets contained residue-wise data instances extracted from the sequence dataset. Each instance was a subsequence of length w, where w was the sliding window size set to eleven in this study. Other window sizes were also tested, but the classifiers constructed with $w = 11$ gave the best performance. From a protein sequence with n residues, a total of $(n - w + 1)$ data instances were extracted. The target residue was positioned in the middle of the subsequence, and the neighboring residues provided context information for the target residue. A data instance was labeled with 1 (positive) if the target residue was RNA-binding or -1 (negative) if the target residue was non-binding. When the data instances were extracted, each residue was replaced with one or more feature values.

A fivefold cross-validation approach was used to train and test support vector machines as well as neural networks. The positive and negative instances were distributed randomly into five folds. Each fold contained the same number of positive as well as negative instances. In each of the five iterative steps, four of the five folds were used to build a classifier (training), and then the classifier was evaluated using the remaining one fold (testing). The predictions made for the test instances in all the five

iterations were combined and used to compute the results presented in this paper.

D. Support Vector Machines

The *SVMlight* package (http://svmlight.joachims.org/) was used to construct the support vector machine (SVM) classifiers. For a given set of binary-labeled training examples, SVM maps the input space into a higher-dimensional feature space and seeks a hyperplane in the feature space to separate the positive data instances from the negative ones [7]. The optimal hyperplane maximizes the separation margin between the two classes of training data, and is defined by a small fraction of the input data instances close to the hyperplane (the so-called support vectors). The distance measurement between the data points in the high-dimensional feature space is defined by the kernel function. In this study, we used the radial basis function (RBF) kernel, $K(\vec{x},\vec{y}) = \exp(-\gamma \| \vec{x} - \vec{y} \|^2)$, where $\vec{x}$ and $\vec{y}$ are two data vectors, and γ is a training parameter. A smaller γ value makes the decision boundary smoother. Another parameter is the regularization factor C, which controls the tradeoff between low training error and large margin [12]. Different values for the γ and C parameters were tested to optimize the prediction of RNA-binding residues. Since the training dataset was imbalanced, the cost factor was set to 5.7 for giving more weight to training errors on positive examples than errors on negative ones. All the other parameters were set to their default values as specified in *SVMlight*.

E. Artificial Neural Networks

Feed-forward neural networks with the back-propagation learning algorithm were implemented using the *NevProp* package version 3 (http://brain.cs.unr.edu/publications/). Each neural network had three layers of neural units or neurons, including an input layer, a hidden layer and an output layer. The input layer had $k \times w$ units, where k was the number of the features used to encode each residue, and w was the sequence window size. Various numbers of hidden units were tested to optimize the network settings. The output layer had a single neuron, which predicted whether a target residue was RNA-binding or not. The neural networks were fully connected, and the output and hidden neurons used the logistic sigmoid activation function.

F. Classifier Performance Measures

Predictions made for the test data instances are compared with the class labels (RNA-binding or non-binding) to evaluate classifier performance. The overall accuracy equals $(TP + TN)/(TP + TN + FN + FP)$, where TP is the number of true positives (RNA-binding residues with positive predictions); TN is the number of true negatives; FN is the number of false negatives; and FP is the number of false positives. However, the overall accuracy alone could be misleading in this case. Since the dataset is imbalanced, a classifier can achieve over 85% accuracy by simply predicting all the residues as negatives. Thus, sensitivity = $TP/(TP + FN)$ and specificity = $TN/(TN + FP)$ are computed. Furthermore, the average of sensitivity and specificity may provide a fair measure of prediction strength [4], [5].

The Receiver Operating Characteristic (ROC) curve is probably the most robust approach for classifier evaluation [13]. The ROC curve is drawn by plotting the true positive rate (*i.e.*, sensitivity) against the false positive rate, which equals to (1 – specificity). The different points on the ROC curve represent the tradeoffs between sensitivity and specificity. When a classifier's sensitivity increases, its specificity often drops. In this work, the ROC curve has been generated by using different threshold values for the output of a classifier and plotting the true positive rate against false positive rate for each threshold value. The area under the ROC curve (AUC) can be used as a reliable measure of classifier performance [14]. Since the ROC plot is a unit square, the maximum value of AUC is 1, which is achieved by a perfect classifier. Weak classifiers and random guessing have AUC values close to 0.5.

III. RESULTS

As described in Methods, five different sequence features have been used to encode an amino acid residue in this study. Each data instance consists of eleven residues including the target residue in the middle and its five neighboring residues on each side. The prediction is made for the target residue, and the neighboring residues provide context information for the target residue.

It is important to note that each feature captures certain information about a specific aspect of RNA-binding. For example, the *B* feature is an index to the conservation of a position in homologous sequences. It is likely that RNA-binding sites as well as other functional positions tend to be conserved among homologous proteins. Although the *B* feature does not appear to be sufficient to define a RNA-binding residue, it captures some relevant information that is not present in the other features. Thus, combination of the different features may enhance the prediction accuracy.

Table I shows the performance of the SVM classifiers in five-fold cross validations. The results have been obtained using the training parameters, C = 0.5 and γ = 0.1, which gave slightly better performance than other values. The SVM classifier constructed using all the five features achieves 74.25% overall accuracy with 65.78% sensitivity and 75.70% specificity. The prediction strength (average of sensitivity and specificity) reaches 70.74%. The ROC curve of the classifier is shown in Fig. 1 (the curve indicated as 'SVM'), and the area under the ROC curve (AUC) is 0.7538.

To determine whether all the five sequence features are needed for accurate prediction of RNA-binding residues, different feature subsets have been used to train SVMs. It is found that removing any one or more of the five features reduces the prediction strength and ROC AUC (Table I and

TABLE I

PERFORMANCE OF THE SUPPORT VECTOR MACHINE CLASSIFIERS

Features	Accuracy (%)	Sensitivity (%)	Specificity (%)	Strength (%)	ROC AUC
A,B,H,K,M	74.25	65.78	75.70	70.74	0.7538
B, H, K, M	73.70	66.24	74.89	70.57	0.7467
H, K, M	69.32	66.28	69.84	68.06	0.7308
H, K	67.64	66.01	67.92	66.96	0.7259
H	59.05	69.67	57.24	63.46	0.6894

TABLE II

PERFORMANCE OF THE NEURAL NETWORK CLASSIFIERS

# Hidden Units	Accuracy (%)	Sensitivity (%)	Specificity (%)	Strength (%)	ROC AUC
4	60.76	77.77	57.36	67.56	0.7418
8	63.03	74.92	60.66	67.79	0.7444
12	62.43	75.05	59.90	67.48	0.7353
16	64.45	68.74	63.59	66.17	0.7206

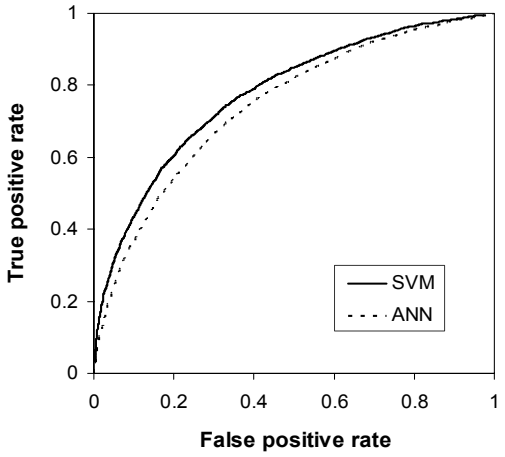

Fig. 1. ROC curves of the best support vector machine (SVM) and neural network (ANN) classifiers.

data not shown). The best classifier using four features (*B*, *H*, *K*, *M*) achieves the prediction strength at 70.57% and AUC = 0.7467. Removal of the *A* feature results in only a slight decrease of the prediction strength, suggesting that the *A* feature captures information mostly overlapping with the other sequence features. This may be explained by the fact that the feature *A* values have been predicted from sequence data (see Methods). The best single feature for prediction of RNA-binding residues appears to be *H* (hydrophobicity index), which gives rise to relatively high sensitivity (69.67%) but low specificity (57.24%). The result is consistent with the observation that hydrophobic amino acids are rarely found at protein-RNA interfaces [2], [3].

To compare the performance of SVMs versus artificial neural networks (ANNs) for prediction of RNA-binding residues, the same dataset that has been used for constructing the SVM models is also used to train and evaluate the ANN classifiers. Various numbers of hidden units have been tested to optimize the neural network settings. As shown in Table II, the ANN with eight hidden units achieves the highest prediction strength (67.79%) and AUC value (0.7444). In Fig. 1, the ROC curves of the best SVM and ANN classifiers are compared. Clearly, the SVM is the better classifier for RNA-binding residues.

IV. Conclusion

We have described a new method for prediction of RNA-binding residues in protein sequences. Five relevant features have been selected for input encoding, and the most accurate classifier has been obtained by training a support vector machine with all the five features. Our method appears to be better than the previous neural network-based approaches developed for prediction of DNA-binding residues [4], [5]. The results from this work have been used to develop the BindN web server (http://bioinformatics.ksu.edu/bindn/) for online prediction of nucleic acid-binding residues.

References

[1] K. L. McKnight and B. A. Heinz, "RNA as a target for developing antivirals," *Antivir. Chem. Chemother.*, vol. 14, pp. 61–73, 2003.

[2] D.E. Draper, "Themes in RNA-protein recognition," *J. Mol. Biol.*, vol. 293, pp. 255–270, 1999.

[3] S. Jones, D. T. A. Daley, N. M. Luscombe, H. M. Berman and J. M. Thornton, "Protein-RNA interactions: a structural analysis," *Nucleic Acids Res.*, vol. 29, pp. 943–954, 2001.

[4] S. Ahmad, M. M. Gromiha and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, pp. 477–486, 2004.

[5] S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC Bioinformatics*, vol. 6, paper 33, 2005.

[6] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. M. Chung and Y. Z. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *RNA*, vol. 10, pp. 355–368, 2004.

[7] V. N. Vapnik, *Statistical Learning Theory*. New York: John Wiley and Sons, 1998.

[8] W. S. Noble, "Support vector machine applications in computational biology," in *Kernel Methods in Computational Biology*, B. Scholkopf, K. Tsuda and J. P. Vert, Ed. Cambridge: MIT Press, 2004.

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

[10] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, pp. 105–132, 1982.

[11] D. L. Nelson and M. M. Cox, *Lehninger Principles of Biochemistry*, 3rd ed. New York: Worth Publishers, 2000.

[12] T. Joachims, "Making large scale SVM learning practical," in *Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf, C. Burges and A. Sola, Ed. Cambridge: MIT Press, 1999.

[13] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, pp. 1285–1293, 1988.

[14] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms," *Pattern Recognition*, vol. 30, pp. 1145–1159, 1997.