A Neural Network Method for Identification of RNA-Interacting Residues in Protein

Euna Jeong **I-Fang Chung** **Satoru Miyano**
eajeong@ims.u-tokyo.ac.jp cif@ims.u-tokyo.ac.jp miyano@ims.u-tokyo.ac.jp
Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Abstract

Identification of the most putative RNA-interacting residues in protein is an important and challenging problem in a field of molecular recognition. Structural analysis of protein-RNA complexes reveals a strong correlation between interaction residues and their structure. Building on this viewpoint, we have developed a neural network predictor to correctly identify residues involved in protein-RNA interactions from protein sequence and its structural information. The system has been exhaustedly cross-validated with various strategies differing in input encoding, amount of input information, and network architectures. In addition, we have evaluated performance among functional subsets of complexes. Finally, to reflect the properties of protein-RNA complexes in our dataset, two kinds of post-processing method are adopted. The experimental result shows that our system yields not-trivial performance although the residues in interaction sites are too scarce.

Keywords: protein-RNA interaction; neural networks; interaction sites prediction

1 Introduction

In recognition of RNA's functional importance in living molecules and the close association with protein in its activities, structural studies of protein-RNA complexes have been substantially increased [1, 4, 5, 10, 11, 19]. These studies are mostly focused on discovering the specific mechanisms of protein-RNA interactions by analyzing intra- and inter-molecular interactions in diverse aspects. Many studies indicate that there is a strong correlation between interaction residues and their compositions. There are two canonical contacts at the secondary structure level: 1) binding between α-helix or loop and the groove of RNA pockets, and 2) binding between β-sheet surface and unpaired RNA bases [5].

The interaction patterns discovered from the analysis give us useful information to understand how protein interacts with RNA with specificity. Identifying interaction patterns is often based on statistical methods in terms of propensity value, which represents the frequency of the occurrences of amino acids in the interaction sites [10, 11]. The interaction propensity is defined as the proportion of a given amino acid in interaction sites divided by the proportion of the amino acid in the dataset. Because the propensity does not consider the influence of adjacent residues and structural features, it is suitable for estimation of the tendency that a given amino acid occurs more frequently or less in the protein-RNA interaction sites than other amino acids, not for prediction of interacting residues given a protein involved in interactions

Based on this viewpoint, we have developed a system to correctly identify residues involved in protein-RNA interactions from protein sequence and secondary structure information. Very few studies have been addressed so far to the important problem of predicting RNA-interacting sites in the protein as an critical goal in the field of molecular recognition. One of the main reasons for this is that there were a small number of protein-RNA interactions available for training a predictor. As

the number of protein-RNA complex structures is gradually increased in recent years, we made an attempt to identify RNA-interacting residues in protein. There are several types of interactions to give an effect to macromolecular structure and interactions. In this study, we consider three kinds of interaction including hydrogen bonds, van der Waals interactions, and electrostatic interactions between protein and RNA. To discriminate RNA-interacting sites in protein is of direct relevance to not only understanding a wide range of biological processes, but also the design of RNA drugs for binding or blocking unwanted protein function [6, 8, 18].

Artificial neural networks are selected as tools for our prediction system since it can rapidly classify differences and patterns in dispersed data like protein-RNA interactions. Given the structure of a protein that is supposed to interact with an unknown RNA, our problem is to predict the residues of the protein that will interact with the RNA. The interaction sites between protein and RNA are defined based on the distance between atoms of them. Our predictor uses secondary structure information about each residue and information about amino acid types. We had trained and tested our system with a multiple cross-validation on a data base of the most representative protein-RNA complexes. As an extension of our previous study [9], in this paper, we experimented our system with various strategies differing in the method of input encoding, the amount of input information, and the number of network parameters. We have divided the complexes into several subsets based on protein's function and also evaluated performance among functional subsets of complexes. There are two characteristics caused by the dataset. Because our dataset is fairly unbalanced, the neural network system is apt to learn more representative pattern with high probability of occurrence, rather than less one. In addition, the analysis of protein-RNA complexes reveals that most interacting residues tend to have neighbor interacting residues. From these observations, we add two kinds of post-processing method to obtain more improvement of performance. The network approach has been adopted as tools in predicting protein secondary structures [13, 14, 16] and protein-protein interaction [7, 12, 20].

It is no simple matter to predict RNA-interacting sites given a protein chain without prior knowledge about the corresponding RNA, even though the protein chain is known to interact with RNA. In this paper, we present the preliminary results about the performance of a predictor for interacting residues between protein and RNA. The experimental result shows that our system yields not-trivial performance though the residues in interaction sites are too scarce.

2 Materials and Methods

2.1 The Dataset Generation

The 130 protein-RNA complexes in the Protein Data Bank (PDB; May 2003 release) [3] were examined, which were solved by X-ray crystallography with better than 3.0 Å resolution. The redundant protein structures in these 130 protein-RNA complexes are eliminated by running PSI-BLAST program [2] with default parameters. Two protein sequences are defined as homologous if 90% of sequences is matched and the sequence identity over the matched region is greater than 70%.

Consequently we yielded 96 representative, non-homologous protein chains paired with RNA chains from 58 protein-RNA complexes. Each protein chain forms at least one interaction with one RNA chain which contains equal to and more 4 nucleotides in length. A residue is considered to be in interaction sites if the closest distance between atoms of the protein and the partner RNA is less than 6.0Å. As a matter of course, one protein chain may interact with several different RNA chains. In this case, the protein chain from a pair which has the largest number of interaction residues is selected. It is done with the intention to have more interaction patterns in our dataset. The redundancy of RNA chains is not taken into account because our prediction method considers only protein information in this study. As a result, the total number of residues in our dataset is 21990 and 21.7% (4782) of residues is found in interaction sites.

The corresponding protein secondary structure of given sequence is obtained from PHD pro-

Table 1: Statistical analysis of protein-RNA complexes classified into five subsets. Notation: c, a, and i for the number of chains, amino acids, and interaction residues, respectively. Each number is followed by the ratio of it in percentage.

Type	Chains		Animo acids		Interactions	
	c	$c/\Sigma c$	a	$a/\Sigma a$	i	$i/\Sigma i$
A	21	21.9	4579	17.5	610	12.8
B	47	49.0	6599	30.0	2705	56.6
C	15	15.6	8607	33.3	1117	23.4
D	9	9.4	1441	6.6	141	2.9
E	4	4.1	764	12.6	209	4.4
Total	96		21990		4782	

gram [16, 15]. The three kinds of secondary structure are considered for this study, such as helix, strand, and coil. From the PHDsec output [21], 'H' and 'E' corresponds to helix and strand, respectively and all residues forming neither 'H' nor 'E' are considered as coil throughout this paper.

In order to view functional characteristics of protein-RNA complexes, 96 complexes are classified into 5 subsets dependent on their function. Although protein by itself would not be expected differences in some kinds of subsets, this classification is intended to further analysis of the partner RNA, as well as protein in future. They can be summarized as follows: (A) Proteins involved in RNA modification and binding messenger RNA, (B) those binding ribosomal RNA, (C) those involved in protein synthesis and binding transfer RNA, (D) those binding viral RNA, and (E) others, i.e., those belonging to neither of other subsets. The statistical analysis of each subset is shown in Table 1 in the order of the number and its percentage of chains, amino acids, and interacting residues. The ribosomal binding proteins, type B in the table, occupy about 49% of chains and 56% of interacting residues in our dataset.

2.2 Neural Network Modeling

Feed-forward neural networks have been implemented to predict protein-RNA interaction sites with a back-propagation algorithm and momentum term as learning procedure. We consider information of a segment of consecutive protein residues as input and associated protein-RNA interaction information of the central residue of the segment as output for the neural network predictor.

Information of each residue includes its amino acid type and corresponding secondary structure information. To prevent an artificial ordering among 20 amino acids, each amino acid type is encoded as a vector of 20 units in unary format [13], which all elements set to 0 but one sets to 1. The position of value one in the vector identifies a particular amino acid type. In fact, totally 21 binary bits are used to represent an amino-acid type, because an additional input is required per residue for windows overlapped either end of the chain. The secondary structure information is examined by combining each type, i.e., helix, strand and coil, and corresponding *reliability index* provided by the PHDsec program. It is encoded similar way with that of each residue, but in this case three units are used for each of three types. The reliability index is treated as an indicator of probability strengths for each of secondary structures represented by an integer in the range from 0 to 9. Each index is encoded as a real number between 0 to 1 dividing by factor 10. For example, if a helix residue has index value 3, then the three units are '0.3 0 0'. In the case of strand and coil, '0 0.3 0' and '0 0 0.3' are be used, respectively. Under the consideration of all the factors described above, for the input of neural networks, each residue is encoded with 24 units, i.e., 20 for each amino acid type, 1 for overlapped terminal, and 3 for secondary structure. Here we adopt the concept of 'window size' w to consider the

influence of a segment of consecutively neighbored residues. Hence totally $w \times 24$ units are needed for one input pattern, where w is tuned to find best performance.

The target output is designed to describe whether a residue in the center of a window interacts with RNA or not. This information is encoded as two output units, 'y_1 y_2', with value '1 0' for interacting residues and '0 1' for not-interacting ones. In the testing process, the output is interpreted qualitatively in a sense that bigger one exclusively scores while smaller one has no score at all. For example, if two output units are '0.4 0.3', the residue is predicted to be in interaction sites because y_1 is larger than y_2.

The architecture of networks we used basically consists of a multi-layer perceptrons with a single hidden layer. Various feed-forward network architectures were explored to find a network with optimal performance which effectively generalizes from the training data to the test data. Input patterns are presented to the networks in randomized order during training process. In order to make sure that every residue is at least chosen once as a center of an input pattern, the procedure of selecting patterns is repeated about five times as many as the number of total residues. It lets the network errors to be converged in a stable state. Evaluating a prediction method is done by a 10-fold cross-validation where the dataset is divided into 10 subsets. Among 96 protein chains, 86 chains are taken for training, 10 for testing (for four sets splitted into 85 and 9). This is repeated ten times with different sets until all proteins have been used for testing exactly once. For all simulations, the learning rate was set to 0.1 for back-propagation and 0.001 for momentum term.

2.3 Measures of Performance

The performance of a particular prediction is assessed by total accuracy, accuracy, coverage, and correlation coefficient, as follows:

- Total accuracy as the percentage of all correctly predicted interaction and not-interaction sites, given by
$$\frac{tp + tn}{tp + tn + fp + fn} \times 100,$$
- Accuracy as the ratio of the number of residues correctly predicted to be in interaction sites and the number of residues predicted to be in interaction sites, given by
$$\frac{tp}{tp + fp} \times 100,$$
- Coverage as the ratio of the number of residues correctly predicted to be in interaction sites to the number of actual residues to be in interaction sites, defined by
$$\frac{tp}{tp + fn} \times 100, \text{ and}$$
- Matthews correlation coefficient (MCC) as a more complicated measure indicating the magnitude of correlation between actual and predicted values, using
$$\frac{(tp \times tn - fp \times fn)}{\sqrt{(tp + fp) \times (tp + fn) \times (tn + fp) \times (tn + fn)}},$$

where tp is the number of correctly predicted interacting residues (true positive), tn is that of correctly predicted not-interacting residues (true negative), fp is that of actually not-interacting residues predicted as interacting one (false positive), and fn is that of actually interacting residues predicted as not-interacting one (false negative). Accuracy is the probability that how many of the predicted interaction sites are correct, whereas the coverage is the probability that how many of the correct interaction sites are predicted. The correlation coefficient value is ranged between 1 and -1, which corresponds to a perfect and a completely wrong prediction, respectively.

2.4 State-Shifting and Filtering

As the training set is quite unbalanced, i.e., only around one residue out of five is in interaction sites, the neural network is too often disposed to learn not-interaction patterns, rather than interaction patterns. In our method, since the output value is credited by a *winner-takes-all* way, a single residue that has a little score has no credit at all. To alleviate this problem, we add an additional measure *state-shifting*. If an actual interacting residue r_i is predicted to be in not-interaction sites, the system searches whether there are any interacting residues within up to two residues (r_{i-2} to r_{i+2}) in all direction. The prediction is considered to correctly predict if there is at least one interacting residue within the range [17].

On the other hand, our analysis of RNA-interacting proteins presented that interacting residues have a tendency to consecutively occur. The proportion which two to nine consecutive residues are all in interaction sites accounts for above 56%, while isolated interacting residues have only 12% of total interacting residues. Isolated interacting residues are those which have no interacting residues next to them, To reflect this property of interacting residues, we *filtered* isolated interacting residues from the prediction. Precisely, incorrectly predicted interacting residues which have no interacting residues within up to two residues are considered to correctly predict as not-interacting residues.

3 Experimental Results

Several tests have been implemented in an attempt to improve the performance of neural networks. The strategies we considered are the method of input encoding, the dimension of each input pattern, and the number of the network parameters. Firstly, two kinds of input encoding are explored: (i) no 2D encoding – protein sequence information only, and (ii) 2D real encoding – protein sequence information plus secondary structure information. Secondly, to evaluate the effect of the amount of input information, the neural network with the window length ranging from 7 to 57 has been investigated for each of two input encodings. Lastly, from the viewpoint of network architectures, the number of hidden units has been tuned to 5, 10, and 30. The combination of three strategies is exhaustedly performed to find optimal networks. A neural network with the optimal performance is selected by comparing MCCs averaged over the 10 different training sets. In addition to the above three strategies, state-shifting and filtering, which are designed to reflect characteristics of protein-RNA complexes in our dataset are tested to raise further the performance of the predictor.

3.1 Optimal Network Structure

We first decide an optimal network architecture by adjusting the number of hidden units. The criteria to select the optimal one is to measure the average of MCCs and its variance. Table 2 shows the MCC values of two encoding methods with different number of hidden units h=5, 10, and 30. In the case of 2D real encoding, the MCC values are 0.261, 0.264, and 0.274 in the third line, while those of no 2D encoding are around 0.241, 0.239, and 0.257 in the second line, with h=5, 10, and 30, respectively. As a whole, 2D real encoding obtains a little bit better performance than no 2D encoding. When compared between different numbers of hidden units in 2D real encoding, the MCC value is improved a little according to the increase of the hidden units. We can view that the network structure with h=5 is an optimal one with respect to the computational complexity, though the result of h=30 yields more about 0.01 than others in 2D real encoding.

3.2 Alternative Input Encoding and Input Pattern

In order to investigate the effect of the dimension of input pattern to the network performance, we have tried to adjust the input window length w ranging from 7 to 57 with h=5. To assess the performance of the neural networks, we take the result of no 2D encoding with w=7 as a baseline, which obtains

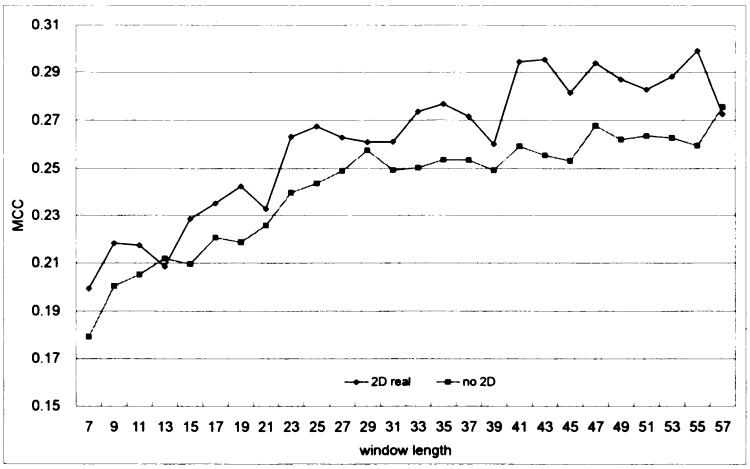

Figure 1: Distribution of MCC values with the increase of window length.

Table 2: Comparison of MCCs and their variances of neural networks along with different number of hidden units.

Encoding	h=5	h=10	h=30
no 2D	0.241±0.020	0.239±0.024	0.257±0.034
2D real	0.261±0.029	0.264±0.031	0.274±0.037

Table 3: Distribution of the performance with 2D real encoding.

Window size	Total accuracy	Accuracy	Coverage	MCC
11	73.7	38.3	38.5	0.217
21	75.1	40.6	37.2	0.233
31	76.4	43.6	38.2	0.261
41	77.5	46.7	40.3	0.294
51	77.2	45.7	39.3	0.283

Table 4: Distribution of the performance with no 2D encoding.

Window size	Total accuracy	Accuracy	Coverage	MCC
7	72.9	35.8	34.3	0.179
17	74.4	39.3	37.0	0.221
27	75.8	42.3	37.8	0.249
47	76.8	44.7	37.8	0.268
57	76.7	44.6	39.6	0.275

Table 5: Performance of the predictor with state-shifting and filtering.

Method		Total	Accuracy	Coverage	MCC
shifting	filtering	accuracy			
0	n/a	76.9	46.5	40.6	0.291
±1	n/a	81.0	55.9	59.1	0.452
±2	n/a	82.8	59.1	67.5	0.521
±1	±1	84.5	65.9	59.1	0.527
±1	±2	83.0	61.4	59.1	0.495
±2	±1	86.3	68.8	67.5	0.594
±2	±2	84.9	64.5	67.5	0.563

total accuracy 72.9%, accuracy 35.8%, coverage 34.3%, and MCC 0.179. The distribution of MCCs along with the increase of the window length is shown in Figure 1. The detail of the performance with 2D real and no 2D encodings is represented in Table 3 and 4, respectively. Though the highest MCC value is obtained at w=55 as 0.299 in 2D real encoding, the ability of generalization of patterns is not too much increased and even unstable. Therefore it is viewed that the best performance is obtained at w=41 and increased up to total accuracy 77.5%, accuracy 46.7%, coverage 40.3%, and MCC 0.294 with 2D real encoding. No 2D encoding achieves 76.7%, 44.6%, 39.6%, and 0.275 for each of measures with w=57. Compared with those of the baseline, we are aware that the overall performance is improved very well, i.e., increased by total accuracy 4.6% and MCC 0.11 in 2D real encoding, and 3.8% and 0.09 in no 2D encoding, respectively. If the results of two encodings are examined with the respect to the same window size (for example w=41) in Figure 1, it is obvious that secondary structure information with the reliability index plays an important role to improve the predictor's capability.

Our experiments to some extent show a tendency that the larger the window size is, the better the performance is, especially in no 2D encoding. In addition, we also observe that the differences between two encodings may be seemingly trivial if the size of the input information and hidden units is extremely increased. For example, in Figure 1, the MCC values of both two encodings at w=57 look like being similar. This phenomenon is also shown at h=30 in Table 2, i.e., the difference between MCC values of two encodings became smaller than h=5 and 10.

3.3 Performance Effect by State-Shifting and Filtering

Table 5 shows the performance achieved by *state-shifting* and *filtering* with 2D real encoding. The performance measures indicate that the introduction of state-shifting remarkably improved the overall prediction score, since state-shifting reduces only the number of false-negatives and leaves true-positive and false-positive untouched. MCC values are increased from 0.29 to 0.45 and to 0.52 when a displacement in prediction is allowed by up to one and two residues, in the third and the fourth line, respectively. The total accuracy is up to above 80% with both displacements.

We had experimented with 4 different combinations of state-shifting and filtering by changing the distance from one to two residues. State-shifting followed by filtering reveals more improvement, because filtering reduces overpredictions by eliminating incorrectly predicted isolated interacting residues. It is clarified from the fact that after filtering accuracy is improved by 5.4 to 10%, while coverage is not changed and the same as that of only state-shifting. The result shows that the process of filtering ±1 after shifting ±2 obtains the best performance as MCC value 0.59 among of 4 combinations as shown in the Table 5. Since two processes state-shifting and filtering are exclusive, we can view that adding filtering ±1 obtains more 0.07 in MCC when compared with only state-shifting ±2 (MCC 0.52).

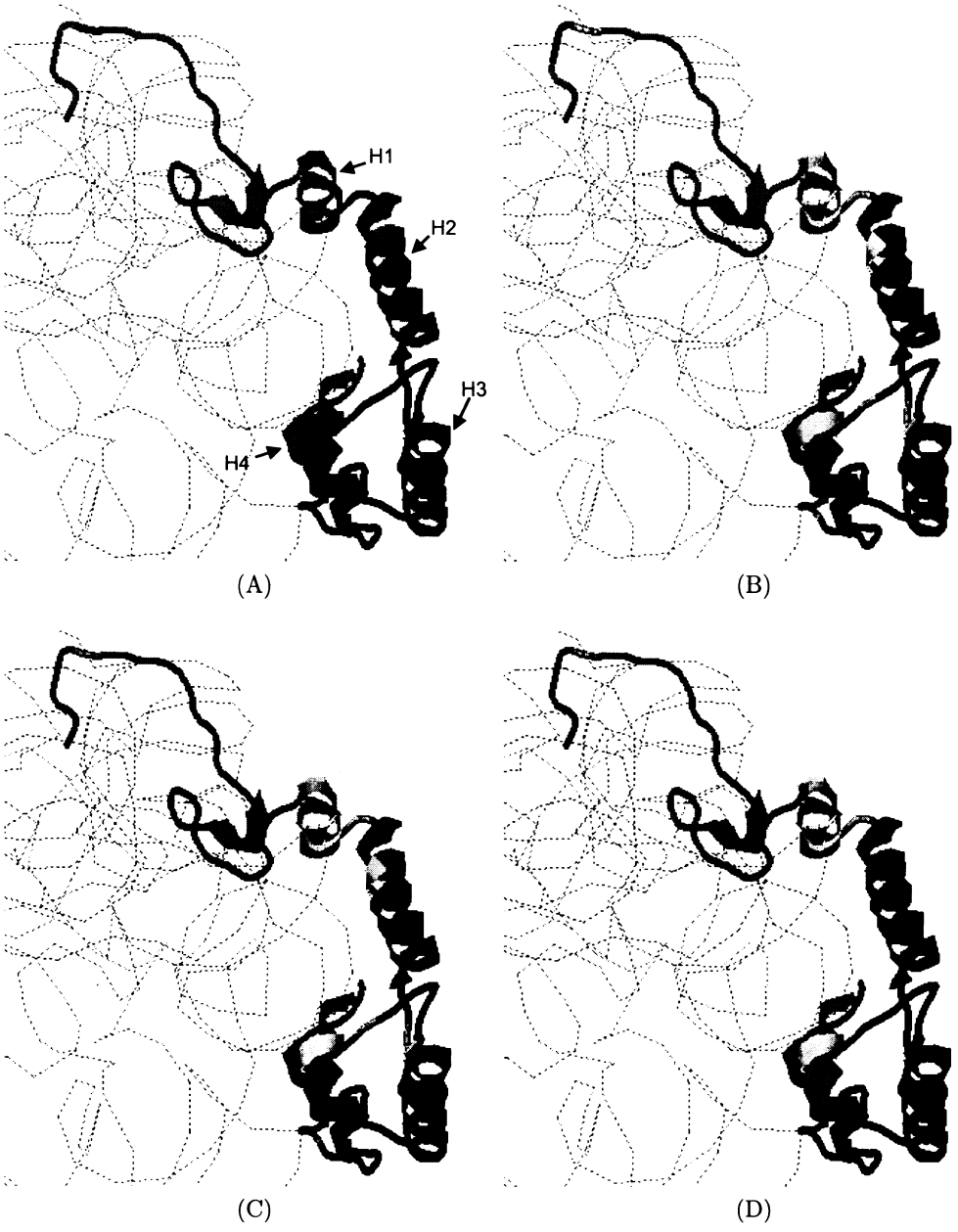

Figure 2: Example for prediction for 30S ribosomal protein (protein chain id: M in PDB code:1FJG). (A) for the actual interactions, (B) for the predicted interactions, (C) for the result of state-shifting up to two residues, and (D) for the result of state-shifting up to two followed by filtering up to one. Format: ribbon for protein molecules and backbone-dash for the RNA molecules. Color for proteins: green for interaction sites, red for not-interaction sites, yellow for overpredictions, and blue for underpredictions.

Figure 2 depicts example of predictions in pictorial representations. To focus on the result of our prediction system, the protein molecule is centered and described in ribbon format, while the RNA molecule is drawn in backbone-dash format. Different colors are used to indicate the result of predictions such as green for interaction sites (true-positive), red for not-interaction sites (true-negative), yellow for overpredictions (false-positive), and blue for underpredictions (false-negative). The figures show protein chain M which interacts with RNA chain A in 1FJG (PDB code). Each of four figures describes the actual interactions in (A), the predicted result in (B), the result of state-shifting ± 2 in (C), and lastly that of combining state-shifting ± 2 and filtering ± 1 in (D), respectively. It is known that interactions between the helix residues and the base-paired nucleotides are one of the canonical hydrogen bonds in protein-RNA complexes [11]. The groove of RNA is formed by pairing between bases and then backbones are relatively free to contact with protein. In protein, the helix residues towards RNA are easily inserted into the groove of RNA, but the opposite residues in helix rarely interact with RNA, which is caused by the shape of helix. It is well depicted in Figure 2.A shown the actual interactions of the complex. There are four helices in the protein chain and all helices are involved in interacting with RNA except H3. In the three helices H1, H2, and H4, it is clearly shown that around half of helix residues near RNA interacts with RNA, while the other half is not. Compared with Figure 2.A and 2.B, in our prediction, most underpredictions (blue colored) occurred at residues in helix involved in RNA interactions. On the other side, about fifty percent of overpredictions (yellow colored) took place in coil residues and the rest equally occurred in helix and sheet residues. In Figure 2.B, it is interesting that the neural network prefers to predict not-interacting residues as interacting ones in H1, while interacting residues are mostly predicted as not-interacting ones in H2. The reason is seemingly that H1 is more close to interaction sites in the upper part of it, while H2 has adjacent not-interaction sites in the lower part of it. After state-shifting, underpredictions appeared in helix residues are extremely decreased as shown in Figure 2.C. Figure 2.D shows the effect of adding filtering that overpredictions are reduced evenly by a small quantity from each of three secondary structures.

3.4 Comparison of Performance between Functional Subsets of Complexes

We evaluate the performance of each subset of protein-RNA complexes which is classified dependent on their function. For details of classification of complexes, see Section 2.1. Table 6 and 7 show the performance of 4 subsets at before state-shifting and after state-shifting, respectively. The analysis of complexes with no main functional feature (type E) is omitted. The ribosomal proteins (type B) show the best performance among 4 subsets. It may be caused from the fact that the ribosomal proteins form around 50 percent of complexes in our dataset as pointed out in Table 1. Our prediction system represents remarkably high accuracy (>77%), coverage (>74%) and MCC (>0.52) for some ribosomal proteins such as 1FJG and 1JJ2 (data not shown). Those two complexes also shows great improvement in all measures after state-shifting ± 2, , which are increased up to over 86%, 74%, and 0.71, for accuracy, coverage, and MCC, respectively. From the observation, a prediction system trained by the dataset with similar molecular features can achieve significant improvement in performance. It also demonstrates that neural networks can show their strengths with an abundant data.

4 Conclusion

We presented neural network methods for discovering putative RNA-interacting sites given a protein chain. Our system was experimented with various strategies which are different in the method of input encoding, the amount of input information, and the number of network parameters. Analysis of performance in each of functional subsets is also accomplished. In addition, to improve the performance of the predictor, two post-processing methods state-shifting and filtering are tested.

From the preliminary results, it has been examined that a prediction system obtains improvement

Table 6: Performance of protein-RNA complexes classified dependent on their function. Each type is abbreviated to A for proteins binding messenger RNA, B for those binding ribosomal RNA, C for those binding transfer RNA, D for those binding viral RNA. The analysis of complexes with no functional feature (type E) is omitted. For details see Section 2.1.

Type	Total accuracy	Accuracy	Coverage	MCC
A	80.4	31.7	30.9	0.20
B	69.9	67.6	51.1	0.36
C	79.4	20.7	20.7	0.09
D	83.8	26.2	36.2	0.22
shifting 0	76.9	46.5	40.6	0.29

Table 7: Performance of protein-RNA complexes classified dependent on their function after state-shifting up to two residues. The notation for each type is the same as Table 6.

Type	Total accuracy	Accuracy	Coverage	MCC
A	84.8	47.9	61.2	0.45
B	80.8	76.0	77.6	0.60
C	82.7	36.7	45.8	0.31
D	86.7	39.5	66.7	0.44
shifting ± 2	82.8	59.1	67.5	0.52

by adding secondary structure information to protein sequence, rather than only sequence. Our predictor shows a little weakness, for example, in the case of residues in helix involved in RNA interactions. Because of helix shape in interaction sites, it is difficult for the predictor to distinguish interacting residues and not-interacting ones existing by turns in helix (Figure 2). By execution of state-shifting and filtering, the number of underpredictions and overpredictions caused by the structural characteristics of protein-RNA complexes is decreased very much to some degree. The experimental result shows that our system yields not-trivial performance although the residues in interaction sites are too scarce.

We had pointed that the propensity value based on statistical methods does not give enough information to predict residues in interaction sites in Section 1, because it just reflects the tendency of interactions of a given amino acid. From the calculation of MCC values for 20 amino acids (data not shown), we found that there is no strong correlation between the MCC values and the propensity values. Those two values are not proportional to each other and in some case an amino acid with a low propensity value achieves a high MCC value. In another experiments, a simple prediction method which predicts all residues to have high propensity values predicted to be interaction sites did not outperform our predictor as MCC value 0.11. It confirms that our predictor can more correctly predict interacting residues by considering the effects of consecutively neighbored residues and their structural features.

In the analysis of functional subsets of complexes, we found that ribosomal proteins which occupy the majority of the dataset achieve a considerable performance. It suggests that a neural network trained by a plentiful data with similar features can obtain better performance. Some ribosomal proteins have very huge RNA chains which interact with many protein chains. In 1FJG, a RNA chain interacts with six different protein chains. Around five complexes out of those 6 show good performance in MCC value over 0.32. It seems that one large RNA molecule interacts with different proteins according to some generic rules. In conclusion, a combination of protein information adopted

here and the partner RNA information not only may provide insight into specificity of protein-RNA interactions, but also can obtain progress of the prediction ability of neural network systems.

References

[1] Allers, J. and Shamoo, Y., Structure-based analysis of protein-RNA interactions using the program ENTANGLE, *Journal of Molecular Biology*, 311:75–86, 2001.

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

[3] Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., The protein data bank, *Nucleic Acids Research*, 28:235–242, 2000.

[4] Cheng, A.C., Chen, W.W., Fuhrmann, C.N., and Frankel, A.D., Recognition of nucleic acid bases and base pairs by hydrogen bonding to amino acid side chains, *Journal of Molecular Biology*, 327:781–796, 2003.

[5] Draper, D.E., Themes in RNA-protein recognition, *Journal of Molecular Biology*, 293:255–270, 1999.

[6] Ecker, D.J. and Griffey, R.H., RNA as a small-molecule drug target: Doubling the value of genomics, *Drug Discovery Today*, 4:420–429, 1999.

[7] Fariselli, P., Pazos, F., Valencia, A., and Casadio, R., Prediction of protein-protein interaction sites in heterocomplexes with neural networks, *European Journal of Biochemistry*, 269:1356–1361, 2002.

[8] Hermann, T. and Westhof, E., RNA as a drug target: Chemical, modelling, and evolutionary tools, *Current Opinion in Biotechnology*, 9:66–73, 1998.

[9] Jeong, E., Chung, I., and Miyano, S., Prediction of residues in protein-RNA interaction sites by neural networks, *Genome Informatics*, 14:506–507, 2003.

[10] Jones, S., Daley, D.T.A., Luscombe, N.M., Berman, H.M., Thornton, J.M., Protein-RNA interaction: A structural analysis, *Nucleic Acids Research*, 29(4):943–954, 2001.

[11] Kim, H., Jeong, E., Lee, S.W., and Han, K., Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns, *FEBS Letters*, 552:231–239, 2003.

[12] Ofran, Y. and Rost, B., Predicted protein-protein interaction sites from local sequence information, *FEBS Letters*, 544:236–239, 2003.

[13] Qian, N. and Sejnowski, T.J., Predicting the secondary structure of globular proteins using neural network models, *Journal of Molecular Biology*, 202:865–884, 1988.

[14] Riis, S.K. and Krogh, A., Improving predicting the secondary structure of globular proteins using structured neural networks and multiple sequence alignments, *Journal of Computational Biology*, 3:163–183, 1996.

[15] Rost, B. and Sander, C., Combining evolutionary information and neural networks to predict protein secondary structure, *Proteins: Structure, Function, and Genetics*, 19:55–72, 1994.

[16] Rost, B. and Sander, C., Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology*, 232:584–599, 1993.

[17] Shepherd, A.J., Gorse, D., and Thornton, J.A., Prediction of the location and type of β-turns in proteins using neural networks, *Protein Science*, 8:1045–1055, 1999.

[18] Sucheck, S.J. and Wong, C.-H., RNA as a target for small molecules, *Current Opinion in Chemistry Biology*, 4:678–686, 2000.

[19] Treger, M. and Westhof, E., Statistical analysis of atomic contacts at RNA-protein interface, *Journal of Molecular Recognition*, 14:199–214, 2001.

[20] Zhou, H.X. and Shan, Y., Prediction of protein interaction sites from sequence profile and residue neighbor list, *Proteins: Structure, Function, and Genetics*, 44:336–343, 2001.

[21] `http://cubic.bioc.columbia.edu/predictprotein/`