Artificial Neural Network Approach to Prediction of Protein-RNA Residue-base Contacts

Morihiro Hayashida¹, Jose Nacher² and Hitoshi Koyano³

¹Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, Matsue, Shimane, Japan
²Department of Information Science, Faculty of Science, Toho University, Funabashi, Chiba, Japan
³School of Life Science and Technology, Tokyo Institute of Technology, Meguro-ku, Tokyo, Japan

Keywords: Fully Connected Neural Network, Protein-RNA Interaction, Residue-base Contact.

Abstract: Protein-RNA complexes play essential roles in a cell, and are involved in the post-transcriptional regulation of gene expression. Therefore, it is important to analyze and elucidate structures of protein-RNA complexes and also contacts between residues and bases in their interactions. A method based on conditional random fields (CRFs) was developed for predicting residue-base contacts using evolutionary relationships between individual positions of a residue and a base. Further, the probabilistic model was modified to improve the prediction accuracy. Recently, many researchers focus on deep neural networks due to its classification performance. In this paper, we develop a neural network with five layers for predicting residue-base contacts. From computational experiments, in terms of the area under the receiver operating characteristic curve (AUC), the predictive performance of our proposed method was comparable or better than those of the CRF-based methods.

1 INTRODUCTION

Interactions between proteins and RNAs are involved in the post-transcriptional regulation of gene expression including alternative splicing, polyadenylation, localization and translation (Glisovic et al., 2008). For instance, a ribosome is formed with multiple proteins and RNAs, and synthesizes proteins from messenger RNAs. Disruption of protein-RNA interactions may cause various diseases including cancers. Therefore, it is needed to obtain precise knowledge of contact positions between proteins and RNAs to understand their molecular function.

It is known that there are several RNA-binding domains such as RNA recognition motif (RRM), heterogeneous nuclear ribonucleoproteins (hnRNPs) (Beyer et al., 1977), the K-homology domains (Siomi et al., 1993), double-stranded RNA-binding domains (dsRBD) (Feng et al., 1992), TIA-1 (Kedersha et al., 1999) and zinc fingers (Hall, 2005). The sequence and structural properties of RNA-protein interaction sites in 211 RNA-protein chain pairs were investigated, and it was reported that 78% of hydrogen bonds involve amino acid side chains, and the remaining involve the protein backbone (Gupta and Gribskov, 2011). Several computational methods have been developed for predicting RNA-binding amino acid residues in proteins. Sun et al. proposed RNAProSite, which utilizes the random forest (Breiman, 2001) with electrostatic feature, triplet interface propensity, position-specific scoring matrices (PSSM) profile, geometrical characteristic and physicochemical property (Sun et al., 2016). Sharan et al. developed an integrated pipeline, called APRICOT, that identifies functional motifs in protein sequences using PSSMs, hidden Markov models of RNA-binding domains, and sequence-based features (Sharan et al., 2017). It was reported that APRICOT achieved sensitivities higher than or as good as RNAProSite and other high performing predictors. Tang et al. proposed PredRBR that utilizes gradient tree boosting and many kinds of sequence and structural site features (Tang et al., 2017).

On the other hand, concerning binding sites of RNAs, 24.6% of hydrogen bonds involve nucleobase-specific interactions, and the remaining involve the RNA backbone (Gupta and Gribskov, 2011). Alipanahi et al. proposed DeepBind that uses many RNA sequences with binding scores determined from protein binding microarray experiments, and extracts motif features by a deep learning technique, which outperformed 26 other methods such as FeatureRE-
DUCE (Weirauch et al., 2013) and BEEML-PBM (Zhao et al., 2011) when the correlation between the predicted and actual probe intensities was evaluated (Alipanahi et al., 2015). Zeng et al. systematically examined various models of convolutional neural networks to further improve DeepBind, and observed that deploying more convolutional kernels was always beneficial, but the local-pooling and additional convolutional layers were useful only in the motif occupancy task when higher-order features existed, and they did not achieve significant improvement (Zeng et al., 2016). These methods were developed for identifying RNA and DNA motifs bounded by transcription factors. However, both positions in binding RNA and protein were not identified. Methods for predicting residue-base contacts in protein-RNA interactions were proposed using conditional random fields (CRFs), which used evolutionary measurements obtained from multiple sequence alignments (Hayashida et al., 2013; Hayashida et al., 2018). For analyzing mechanisms of protein-RNA interactions in detail, we also deal with the problem of predicting residue-base contacts.

Recently, artificial neural network techniques (LeCun et al., 2015; Glorot et al., 2011) have been applied in many research fields due to its high classification performance. We explore neural network architectures to overcome the prediction accuracy by the CRF-based method. For evaluating the proposed neural network, we perform computational experiments as done in the previous study. The results show that in terms of the area under the receiver operating characteristic curve (AUC), the predictive performance of our proposed method was comparable or better than those of the CRF-based methods.

\[ \text{Pr}(x(n)_{i j} | m(n), p(n), r(n), \theta) \]
given \( x(n), m(n) \) for each sequence pair \( p(n), r(n) \).

### 2.2 Conditional Random Field (CRF)-based Method

Conditional random fields (CRFs) were proposed by extending Markov random fields (Lafferty et al., 2001). The CRF with a strictly positive density in the previous study was defined by

\[ \text{Pr}(x_{ij}) = \sum_{z_{ij}} \exp \left( \sum_{l} w_l x_{ij} \right) \]

where \( x_{ij} = 1 \) means that \( i \)-th residue and \( j \)-th base interact with each other, \( x_{ij} = -1 \) otherwise.

\[ f_{ij}(x, m, p, r) = x_{ij} \left( \max_{k \in \mathbb{N}_j} m_{ik} \right) \]

\[ g_{kl}(x, m, p, r) = x_{ik} x_{kl} \left( \frac{|m_{ij} - m_{kl}|}{m_{ij} m_{kl}} \right) \delta(p, r) \]

\( \delta(p, r) \) denotes a \(|A| \times |B|\) dimensional vector which only the element corresponding to the amino acid \( p \) and base \( r \) is one, \( \mathbb{N}_j \) for \( i = 1, \ldots, |A| \), and \( j = 1, \ldots, |B| \) is a set of adjacent pairs of \((i, j)\), and was defined by \((i \pm 1, j), (i, j \pm 1)\) (see Figure 1). In this method, \( m_{ij} \) at the gray square in the figure was given as one of input data for a residue-base pair \((i, j)\), that is, \((i', j') \in \{(i, j)\} \cup \mathbb{N}_j \cup \bigcup_{k \in \mathbb{N}_j} \mathbb{N}_k\). The probability model contains dependency relationships between positions \((i, j)\) and \((i', j') \in \mathbb{N}_j\).

Parameters \( w_l \) and \( w_k \) in the CRF model are determined by maximizing the pseudo-likelihood function,

\[ L(\theta) = \prod_{(i, j) \in \mathbb{N}} \prod_{p, r} \text{Pr}(x_{ij}(n) | m(n), p(n), r(n), \theta) \]

given \( x(n), m(n) \) for each sequence pair \( p(n), r(n) \).
2.3 Neural Network Architecture

We propose a neural network with five layers as shown in Figure 2. The input layer has \((2k_m + 1)^2 + (|A| + |B|)(2k_s + 1)\) neurons having evolutionary measurement values of pairs \((i', j')\) for all \(i - k_m \leq i' \leq i + k_m, j - k_m \leq j' \leq j + k_m\), and one-hot vectors corresponding to subsequences from \(r_{i-k_s}^{i+k_s}\) to \(r_{j-k_s}^{j+k_s}\) of protein \(p\) and from \(r_{j-k_s}^{j+k_s}\) to \(r_{i-k_s}^{i+k_s}\) of RNA \(r\), where \(k_m\) and \(k_s\) are constant integers, and a one-hot vector for a constant number \(c\) is defined as a vector that only \(c\)-th element is one and the others are zero. For example, when \(k_m = 2, m_i, j\) at the gray and white square in Figure 1 is given with respect to \((i, j)\) to the input layer. In addition, amino acids can be classified, and \(|A|\) can be equal or less than 20. The neural network has three hidden layers with \(n_1, n_2,\) and \(n_3\) neurons, respectively. The output layer has two neurons corresponding to the two classification results. All successive layers are fully connected and the rectified linear unit (ReLU) (Glorot et al., 2011) defined by \(f(x) = \max\{0, x\}\) is applied to the output of each neuron except the final layer. The softmax function is applied in the final layer. In addition, bias variables are added to each neuron except the input layer. Then, the total number of parameters is \((2k_m + 1)^2 + (|A| + |B|)(2k_s + 1)n_1 + n_1 + \sum_{i=1}^{3} \{n_i(n_{i+1} + n_{i+1}) + 2n_3 + 2\}.

3 RESULTS

For evaluation of the proposed method, we used the same dataset as that in the previous study, which consists of residue-base pairs included in thirteen protein-RNA pairs in four complexes identified by ‘1yl4’, ‘2hgu’, ‘3kc4’, and ‘3kcr’ in PDB (Rose et al., 2017) as shown in Table 1.

Each line in the table shows a protein-RNA pair, the PDB identifier, the chain identifier, the protein and RNA sequences \(p, r\) in UniProt and GenBank databases (The UniProt Consortium, 2017; Benson et al., 2011), the length, and the number of contacts between residues and bases in the protein and the RNA. It was assumed that \(i\)-th residue and \(j\)-th base interact with each other if the Euclidean distance between atoms of the residue and base is less than or equal to 3 Å because the distances of hydrogen bonds are about 2.7 to 2.9 Å.

We used multiple sequence alignments of Pfam and Rfam databases (Finn et al., 2016; Kalvari et al., 2018) for the protein \(p\) and RNA \(r\) to calculate \(M_{lp}\). We examined classifications of amino acids with 8, 10, and 15 groups according to the study (Murphy et al., 2000), that is, \(|A|\) took 8, 10, 15, and 20, and \(M_{lp}\) was calculated based on each case of \(A\) (see Table 2).

We performed cross-validation procedures, and took the average of AUC scores as well as the previous study, where among thirteen protein-RNA pairs in Table 1, all residue-base pairs included in one protein-RNA pair were used for test, and the others were used for training. We examined the window size by varying the values of \(k_m\) and \(k_s\), and set \(n_1 = 750, n_2 = 680,\) and \(n_3 = 250\). We used the tensorflow-gpu library (version 1.4.1) to minimize the cross entropy of the output and to calculate the AUC score (Abadi et al., 2015).

Table 3 shows the results on average AUC scores by the CRF-based method and the proposed fully-connected neural network method in the cases \((k_m, k_s) = (1, 2), (2, 1), (2, 2)\). In the three cases, the AUC scores in the case of \((2, 1)\) were better than those in other cases in the same classification of amino acids. It implies that amino acids and bases at positions \((i \pm 2, j \pm 2)\) far from the position of interest \((i, j)\) were not effective to enhance the prediction accuracy. On the other hand, the evolutionary measurement \(M_{lp}\) between positions \((i \pm 2, j')\) and \((i', j \pm 2)\) for \(i - 2 \leq i' \leq i + 2, j - 2 \leq j' \leq j + 2\) were useful. It is expected that the neural network with \((k_m, k_s) = (2, 2)\) would be equivalent to the neural network with \((2, 1)\) if appropriate parameters are estimated as zero, and the AUC score with \((2, 2)\) would.
Table 1: Dataset of the residue-base pairs in protein-RNA complexes.

<table>
<thead>
<tr>
<th>PDB</th>
<th>Protein sequence</th>
<th>RNA sequence</th>
<th># contacts</th>
</tr>
</thead>
<tbody>
<tr>
<td>1yl4</td>
<td>K RS8.THET8</td>
<td>A M26923</td>
<td>135 29</td>
</tr>
<tr>
<td>1yl4</td>
<td>M RS10.THET8</td>
<td>A M26923</td>
<td>97 20</td>
</tr>
<tr>
<td>1yl4</td>
<td>O RS12.THET8</td>
<td>A M26923</td>
<td>122 45</td>
</tr>
<tr>
<td>1yl4</td>
<td>T RS17.THET8</td>
<td>A M26923</td>
<td>69 29</td>
</tr>
<tr>
<td>2hgu</td>
<td>R RL18.THETH</td>
<td>B X01554</td>
<td>110 28</td>
</tr>
<tr>
<td>2hgu</td>
<td>Z RL27.THET8</td>
<td>A X12612</td>
<td>81 20</td>
</tr>
<tr>
<td>2hgu</td>
<td>5 RL33.THET8</td>
<td>A X12612</td>
<td>48 18</td>
</tr>
<tr>
<td>3kc4</td>
<td>E RS5_ECOLI</td>
<td>A J01695</td>
<td>67 13</td>
</tr>
<tr>
<td>3kc4</td>
<td>G RS7_ECOLI</td>
<td>A J01695</td>
<td>147 25</td>
</tr>
<tr>
<td>3kc4</td>
<td>O RS15_ECOLI</td>
<td>A J01695</td>
<td>83 21</td>
</tr>
<tr>
<td>3kc4</td>
<td>Q RS17_ECOLI</td>
<td>A J01695</td>
<td>69 18</td>
</tr>
<tr>
<td>3kc4</td>
<td>W RL27_ECOLI</td>
<td>8 J01695</td>
<td>77 18</td>
</tr>
<tr>
<td>3kc4</td>
<td>3 RL35_ECOLI</td>
<td>8 J01695</td>
<td>61 12</td>
</tr>
</tbody>
</table>

Table 2: Grouping of amino acids (Murphy et al., 2000).

<table>
<thead>
<tr>
<th>#groups</th>
<th>groups of amino acids</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>(MLVIC) (GA) (TS) (P) (FYW) (DENQ) (RK) (H)</td>
</tr>
<tr>
<td>10</td>
<td>(MLVI) (C) (G) (A) (TS) (P) (FYW) (DENQ) (RK) (H)</td>
</tr>
<tr>
<td>15</td>
<td>(MLVI) (C) (G) (A) (T) (S) (P) (FY) (W) (D) (E) (N) (Q) (RK) (H)</td>
</tr>
</tbody>
</table>

Table 3: Results on average AUC scores by the CRF-based method and the proposed neural network method with \((k_m, k_s) = (1, 2), (2, 1), (2, 2)\).

<table>
<thead>
<tr>
<th>#groups</th>
<th>CRF</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1,2)</td>
<td>(2,1)</td>
</tr>
<tr>
<td>8</td>
<td>0.692</td>
<td>0.671</td>
</tr>
<tr>
<td>10</td>
<td>0.699</td>
<td>0.673</td>
</tr>
<tr>
<td>15</td>
<td>0.699</td>
<td>0.670</td>
</tr>
<tr>
<td>20</td>
<td>0.693</td>
<td>0.665</td>
</tr>
</tbody>
</table>

For classification of amino acids, as reported in the previous study, the AUC score with 15 groups was better than others in almost all cases. In addition, the AUC score by our method with \((k_m, k_s) = (2, 1)\) and 15 groups was better than that by the CRF-based method.

4 CONCLUSIONS

We proposed a neural network approach to prediction of protein-RNA residue-base contacts. In the neural network, neurons between successive layers were fully connected, and the ReLU activation function was applied. From the cross-validation computational experiments to evaluate the proposed method, the results show that in terms of the area under the receiver operating characteristic curve (AUC), the predictive performance of our proposed method was comparable or better than those of the CRF-based method.

For future work, other types of advanced neural networks should be examined for further improvement of prediction accuracy, and for understanding interactions between residues and bases in detail. In our method, subsequences as long as motifs cannot be dealt. Therefore, we would like to improve our method to deal with longer subsequences.

ACKNOWLEDGEMENTS

This work was partially supported by Grants-in-Aid #16K00392 and #16KT0020 from JSPS, Japan.

REFERENCES


Alipanahi, B., Delong, A., Weirauch, M., and Frey, B. (2015). Predicting the sequence specificities of DNA-


