BIOINFORMATICS

Vol. 00 no. 00 2005
Pages 1–7

A max-margin model for predicting residue–base contacts in protein–RNA interactions

Kengo Sato [1,*], Shunya Kashiwagi [1] and Yasubumi Sakakibara [1]

[1]Department of Biosciences and Informatics, Keio University, 3–14–1 Hiyoshi, Kohoku-ku,
Yokohama 223–8522, Japan

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation: Protein–RNA interactions (PRIs) are essential for many biological processes, so understanding aspects of the sequence and structure in PRIs is important for understanding those processes. Due to the expensive and time-consuming processes required for experimental determination of complex protein–RNA structures, various computational methods have been developed to predict PRIs. However, most of these methods focus on predicting only RNA-binding regions in proteins or only protein-binding motifs in RNA. Methods for predicting entire residue–base contacts in PRIs have not yet achieved sufficient accuracy. Furthermore, some of these methods require 3D structures or homologous sequences, which are not available for all protein and RNA sequences.

Results: We propose a prediction method for residue–base contacts between proteins and RNAs using only sequence information and structural information predicted from only sequences. The method can be applied to any protein–RNA pair, even when rich information such as 3D structure is not available. Residue–base contact prediction is formalized as an integer programming problem. We predict a residue–base contact map that maximizes a scoring function based on sequence-based features such as k-mer of sequences and predicted secondary structure. The scoring function is trained by a max-margin framework from known PRIs with 3D structures. To verify our method, we conducted several computational experiments. The results suggest that our method, which is based on only sequence information, is comparable with RNA-binding residue prediction methods based on known binding data.

Availability: The source code of our algorithm is available at `https://github.com/satoken/practip`.

Contact: satoken@bio.keio.ac.jp

1 INTRODUCTION

Recent studies have been unraveling the mechanisms of biological processes involving functional non-coding RNAs, most of which play essential roles in interacting with RNA-binding proteins (RBPs), such as splicing, transport, localization and translation. These interactions involve sequence- and structure-specific recognition between proteins and RNAs. Therefore, understanding aspects of the sequence and structure in protein–RNA interactions (PRIs) is important for understanding biological processes. To that end, several works have focused on the analysis and discussion of PRIs (Kondo and Westhof, 2011; Iwakiri *et al.*, 2012, 2013).

Compared with deciphering genomic sequences by using high-throughput sequencing technology, experimental determination of protein–RNA joint structures is more expensive and time consuming. Therefore, rapid computational prediction of PRIs from only sequence information is desirable. Existing methods for computational prediction of PRIs can be roughly classified into four groups. The first group predicts whether a given protein–RNA pair interacts or not (Pancaldi and Bahler, 2011; Muppirala *et al.*, 2011; Bellucci *et al.*, 2011; Wang *et al.*, 2013). A prediction algorithm for this approach can be simply designed from interacting protein–RNA pairs alone, so 3D structures and residue–base contacts are not necessary for use in model training. However, this approach cannot predict binding sites of proteins and RNAs that should be biologically and structurally essential for PRIs. The second group aims at predicting RNA-binding residues from protein information. DR_bind1 (Chen *et al.*, 2014), KYG (Kim *et al.*, 2006), and OPRA (Perez-Cano and Fernandez-Recio, 2010) are structure-based methods that use 3D structures from PDB to extract descriptors for prediction. BindN+ (Wang *et al.*, 2010) and Pprint (Kumar *et al.*, 2008) are sequence-based methods that employ evolutionary information instead of 3D structures. This approach ignores the binding partners of target proteins although some of RNA-binding domains in RBPs recognize sequence- and structure-specific motifs in RNA sequences. The third group computes RNA structural motifs recognized by RNA-binding domains in certain proteins and contains MEMERIS (Hiller *et al.*, 2006), RNAcontext (Kazan *et al.*, 2010), CapR (Fukunaga *et al.*, 2014), and GraphProt (Maticzka *et al.*, 2014). This approach focuses on a certain RBP, and extracts RNA motifs as consensus sequences and/or secondary structures of the RBP-binding RNAs. The final group predicts intermolecular joint structures between proteins and RNAs such as residue–base contacts. To the best of our knowledge, (Hayashida *et al.*, 2013) is the only method of this type. However, it is unfortunately not sufficiently accurate.

We propose a prediction method for residue–base contacts between proteins and RNAs with using only sequence information and structural information predicted from only sequences. Our method can be applied to any protein–RNA pair, even when rich information such as 3D structure is unavailable. Residue–base contact prediction is formalized as an integer programming (IP)

[*]to whom correspondence should be addressed

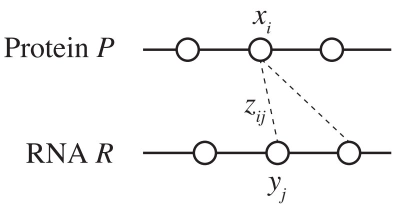

Fig. 1. An illustration of binary-valued variables used in the IP formulation.

problem. We predict a residue–base contact map that maximizes a scoring function based on sequence features such as k-mer of sequences and predicted secondary structure. The scoring function is trained by a max-margin framework from known PRIs with 3D structures. To verify our method, we performed several computational experiments. The results suggest that our method based on only sequence information is comparable with RNA-binding residue prediction methods based on known binding data.

2 METHODS

We present a novel algorithm for predicting PRIs using integer programming. Our algorithm consists of the following two parts: predicts a residue–base contact map given a protein and an RNA by solving an integer programming problem; and learns a scoring function from a given training dataset by a max-margin framework.

2.1 Preliminaries

Let Σ_p be the set of 20 amino acid residues and let Σ_p^* denote the set of all finite amino acid sequences consisting of residues in Σ_p. Similarly, let Σ_r be the set of four ribonucleotide bases (A, C, G, and U) and let Σ_r^* denote the set of all finite RNA sequences consisting of bases in Σ_r. Given a protein $P = \{p_1, \dots, p_{|P|}\} \in \Sigma_p^*$ consisting of $|P|$ residues and an RNA $R = \{r_1, \dots, r_{|R|}\} \in \Sigma_r^*$ consisting of $|R|$ bases, let $\mathcal{CM}(P, R)$ be a space of all possible residue–base contact maps between P and R. An element $z \in \mathcal{CM}(P, R)$ is represented as an $|P| \times |R|$ binary-valued matrix, where $z_{ij} = 1$ indicates that the residue p_i interacts with the base r_j (Fig. 1). We define the problem of predicting PRI as follows: given a protein P and an RNA R , predict a residue–base contact map $z \in \mathcal{CM}(P, R)$.

2.2 Scoring model

A scoring model f is a function that assigns real-valued scores to protein–RNA pairs (P, R) and residue–base contact maps $z \in \mathcal{CM}(P, R)$. Our aim is to find a residue–base contact map $z \in \mathcal{CM}(P, R)$ that maximizes the scoring function $f(P, R, z)$ for a given protein–RNA pair (P, R). The scoring function $f(P, R, z)$ is computed on the basis of various local features of $P, R,$ and z. These features correspond to residue features, base features, and residue–base contact features that describe local contexts around residue–base contacts.

Residue features describe the binding preference in the amino acid sequences by local contexts around residue–base contacts. For this purpose, we employ the k-mer of the amino acids centered on the interacting ith residue. For each k-mer of the amino acids, $p_{kmer} \in \Sigma_p^k$, we define a binary-valued local feature of the ith residue as

$$\phi_{p_{kmer}}(P, z, i) = I(kmer(P, i) = p_{kmer})I(x_i = 1),$$

where $I(condition)$ is an indicator function that takes a value of 1 or 0 depending on whether the $condition$ is true or false, $kmer(P, i)$ is the k-mer of the substring of P centered on the ith residue p_i, that is, $kmer(P, i) = p_{i-(k-1)/2} \cdots p_i \cdots p_{i+(k-1)/2}$, and x_i is a binary-valued variable such that $x_i = 1$ if and only if the residue p_i is a binding site (Fig. 1), that is, $\sum_{j=1}^{|R|} z_{ij} \geq 1$. We use $k = 3$ and 5 for the k-mer features.

Table 1. Groups of amino acids (Murphy *et al.*, 2000)

	#	groups
Σ_{g10}	10	LVIM, C, A, G, ST, P, FYW, EDNQ, KR, H
Σ_{g4}	4	LVIMC, AGSTP, FYW, EDNQKRH

Table 2. A summary of residue features

Type	Context len.	# of features
Residues	3	20^3
	5	20^5
Simplified alphabets (10 groups)	5	10^5
	7	10^7
Simplified alphabets (4 groups)	5	4^5
	7	4^7
Secondary structures	3	3^3
	5	3^5

To reduce the sparsity of amino acid contexts, we consider the k-mers of simplified alphabets of amino acids proposed in (Murphy *et al.*, 2000), which calculated groups of simplified alphabets based on the BLOSUM50 matrix (Henikoff and Henikoff, 1992). Note that Murphy *et al.* (2000) have shown that the simplified alphabets are correlated with physiochemical properties such as hydrophobic, hydrophilic and polar that may be important for PRIs. We employ the simplified alphabets of 10 groups, Σ_{g10}, and those of 4 groups, Σ_{g4} (Table 1). For each string $sa_{kmer} \in \Sigma_{g10}^k$ (or Σ_{g4}^k), we define a binary-valued local feature of the ith residue as

$$\phi_{sa_{kmer}}(P, z, i) = I(kmer(P_{sa}, i) = sa_{kmer})I(x_i = 1),$$

where P_{sa} is the string of simplified alphabets Σ_{g10} (or Σ_{g4}) converted from P according to Table 1. We use $k = 5$ and 7 for the k-mers of simplified alphabets.

To consider structural preference of RNA-binding residues, we employ secondary structures predicted by PSIPRED (Jones, 1999). We predict one structural element (α helix, β sheet, or coil) for each residue. For each string sp_{kmer} of structural elements of length k, we define a binary-valued local feature of the ith residue as

$$\phi_{sp_{kmer}}(P, z, i) = I(kmer(P_{sp}, i) = sp_{kmer})I(x_i = 1),$$

where P_{sp} is the string of structural elements predicted from P. We use structural contexts with lengths $k = 3$ and 5.

Table 2 shows a summary of the residue features. The collection of occurrences of the residue features are calculated as

$$\mathbf{\Phi}_p(P, z) = \sum_{i=1}^{|P|} \boldsymbol{\phi}_p(P, z, i), \tag{1}$$

where $\boldsymbol{\phi}_p(P, z, i)$ is a vector whose elements are the residue features of the ith residue mentioned above.

Base features describe the binding preference in the ribonucleotide sequences by local contexts around residue–base contacts. In addition to the residue features, we employ the k-mer contexts of the ribonucleotides centered on the interacting jth base. For each k-mer of the ribonucleotides $r_{kmer} \in \Sigma_r^k$, we define a binary-valued local feature of the jth base as

$$\phi_{r_{kmer}}(R, z, j) = I(kmer(R, j) = r_{kmer})I(y_j = 1),$$

where y_j is a binary-valued variable such that $y_j = 1$ if and only if the residue r_j is a binding site (Fig. 1), that is, $\sum_{i=1}^{|P|} z_{ij} \geq 1$. We use $k = 3$ and 5 for the k-mer features.

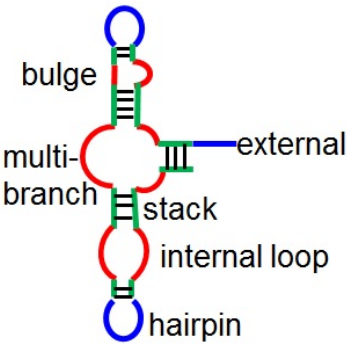

Fig. 2. Structural elements in RNA secondary structures.

Table 3. A summary of base features

Type	Context len.	# of features
Bases	3	4^3
	5	4^5
Secondary structures	3	6^3
	5	6^5

To consider structural preference of binding sites, we employ secondary structures predicted by CentroidFold (Hamada *et al.*, 2009). We assign a structural element (one of external loop, hairpin loop, internal loop, bulge, multibranch loop, or stack, as shown in Fig. 2) for each base. Note that to encode secondary structures as a sequence, this encoding of structural profiles loses part of structural information, e.g. base-pairing partners for stacking bases. However, it is still efficient for describing structural information (Hiller *et al.*, 2006; Kazan *et al.*, 2010; Fukunaga *et al.*, 2014). For each k-length string sr_{kmer} of structural elements, we define a binary-valued local feature of the jth base as

$$\phi_{sr_{kmer}}(R, z, j) = I(kmer(R_{sr}, j) = sr_{kmer})I(y_j = 1),$$

where R_{sr} is the string of structural elements predicted from R. We use structural contexts with lengths $k = 3$ and 5.

Table 3 shows a summary of the base features. The collection of occurrences of the base features are calculated as

$$\boldsymbol{\Phi}_r(R, z) = \sum_{j=1}^{|R|} \boldsymbol{\phi}_r(R, z, j), \tag{2}$$

where $\boldsymbol{\phi}_r(R, z, j)$ is a vector whose elements are the base features of the jth base mentioned above.

Residue–base contact features describe the binding affinity between the local contexts of amino acids and ribonucleotides. For this purpose, we employ combinations of the residue features and the base features mentioned above. For example, for each pair of k-mer of amino acids p_{kmer} and ribonucleotides r_{kmer}, we define a binary-valued local feature of the ith residue and the jth base:

$$\begin{aligned}&\phi_{p_{kmer}, r_{kmer}}(P, R, z, i, j) = \\ &\qquad I(kmer(P, i) = p_{kmer})I(kmer(R, j) = r_{kmer})I(z_{ij} = 1).\end{aligned}$$

Table 4 shows a summary of the residue–base contact features. The collection of occurrences of the residue–base contact features are calculated as

$$\boldsymbol{\Phi}_c(P, R, z) = \sum_{i=1}^{|P|}\sum_{j=1}^{|R|} \boldsymbol{\phi}_c(P, R, z, i, j), \tag{3}$$

where $\boldsymbol{\phi}_c(P, R, z, i, j)$ is a vector whose elements are the residue–base contact features of the ith residue and the jth base mentioned above.

The notation $\boldsymbol{\Phi}(P, R, z)$ denotes the feature representation of protein–RNA pair (P, R) and its residue–base contact map $z \in \mathcal{CM}(P, R)$, that is, the collection of occurrences of local features in P, R, and z defined as follows:

$$\boldsymbol{\Phi}(P, R, z) = \begin{pmatrix} \boldsymbol{\Phi}_p(P, z) \\ \boldsymbol{\Phi}_r(R, z) \\ \boldsymbol{\Phi}_c(P, R, z) \end{pmatrix}. \tag{4}$$

Each feature in $\boldsymbol{\Phi}$ is associated with a corresponding parameter, and the score for the feature is defined as the value of the occurrence multiplied by the corresponding parameter. We define the scoring model $f(P, R, z)$ as a linear function

$$\begin{aligned} f_{\boldsymbol{\lambda}}(P, R, z) &= \langle \boldsymbol{\lambda}, \boldsymbol{\Phi}(P, R, z)\rangle \\ &= \langle \boldsymbol{\lambda}_p, \boldsymbol{\Phi}_p(P, z)\rangle + \langle \boldsymbol{\lambda}_r, \boldsymbol{\Phi}_r(R, z)\rangle + \langle \boldsymbol{\lambda}_c, \boldsymbol{\Phi}_c(P, R, z)\rangle, \end{aligned} \tag{5}$$

where $\langle\cdot,\cdot\rangle$ is the inner product, and $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_p^T, \boldsymbol{\lambda}_r^T, \boldsymbol{\lambda}_c^T)^T$ is the corresponding parameter vector trained from training data as described in Sec. 2.4.

2.3 IP formulation

To formulate the problem as an integer programming (IP) problem, we rewrite the scoring function (5) as

$$f_{\boldsymbol{\lambda}}(P, R, z) = \sum_{i=1}^{|P|} u_i x_i + \sum_{j=1}^{|R|} v_j y_j + \sum_{i=1}^{|P|}\sum_{j=1}^{|R|} w_{ij} z_{ij}, \tag{6}$$

where u_i, v_i, and w_{ij} mean binding preferences for x_i, y_j, and z_{ij}, calculated as

$$\begin{aligned} u_i &= \langle \boldsymbol{\lambda}_p, \boldsymbol{\Phi}_p(P, z, i)\rangle \\ v_j &= \langle \boldsymbol{\lambda}_r, \boldsymbol{\Phi}_r(R, z, j)\rangle \\ w_{ij} &= \langle \boldsymbol{\lambda}_c, \boldsymbol{\Phi}_c(P, R, z, i, j)\rangle. \end{aligned}$$

We find a $z \in \mathcal{CM}(P, R)$ that maximizes the objective function (6) under the following constraints to satisfy the consistency in all the variables x_i, y_j, and z_{ij}:

$$x_i + y_j \geq 2z_{ij} \qquad (1 \leq \forall i \leq |P|, 1 \leq \forall j \leq |R|) \tag{7}$$

$$x_i \leq \sum_{j=1}^{|R|} z_{ij} \qquad (1 \leq \forall i \leq |P|) \tag{8}$$

$$y_j \leq \sum_{i=1}^{|P|} z_{ij} \qquad (1 \leq \forall j \leq |R|) \tag{9}$$

$$y_{j-1} + (1 - y_j) + y_{j+1} \geq 1 \qquad (1 \leq \forall j \leq |R|) \tag{10}$$

$$\sum_{j=1}^{|R|} z_{ij} \leq X_i \qquad (1 \leq \forall i \leq |P|) \tag{11}$$

$$\sum_{i=1}^{|P|} z_{ij} \leq Y_j \qquad (1 \leq \forall j \leq |R|) \tag{12}$$

The constraints (7)–(9) describe the relation between contacts z_{ij} and binding sites x_i, y_j. The constraint (10) disallows any isolated interacting bases, which are rare in PRIs. The constraints (11) and (12) define the upper bound on the number of contacts X_i and Y_j for each residue and base, respectively. As shown in Table 5, X_i and Y_j for each residue and base depend on its structural element, which were determined from the dataset described in Sec. 3.2 (see Supplementary Material for details).

2.4 Learning algorithm

To optimize the feature parameter $\boldsymbol{\lambda}$, we employ a max-margin framework called structured support vector machines (Tsochantaridis *et al.*, 2005).

Table 4. A summary of residue–base contact features

Type			
Residue	Base	Context len.	# of features
Residues	Bases	3	$20^3 \times 4^3$
		5	$20^5 \times 4^5$
Secondary structures	Secondary structures	3	$3^3 \times 6^3$
		5	$3^5 \times 6^5$
Simplified alphabets (10 groups)	Bases	3	$10^3 \times 4^3$
		5	$10^5 \times 4^5$
Simplified alphabets (10 groups)	Secondary structures	3	$10^3 \times 6^3$
		5	$10^5 \times 6^5$
Simplified alphabets (4 groups)	Bases	3	$4^3 \times 4^3$
		5	$4^5 \times 4^5$
Simplified alphabets (4 groups)	Secondary structures	3	$4^3 \times 6^3$
		5	$4^5 \times 6^5$

Table 5. The maximum number of contacts for each residue and base.

Residue	α helix	β sheet	Coil			
X_i	3	3	3			
Base	External	Hairpin	Internal	Bulge	Multibranch	Stack
Y_j	7	5	4	4	4	4

Given a training dataset $\mathcal{D} = \{(P^{(k)}, R^{(k)}, z^{(k)})\}_{k=1}^{K}$, where $P^{(k)}$ and $R^{(k)}$ are respectively the protein and RNA sequences and $z^{(k)} \in \mathcal{CM}(P^{(k)}, R^{(k)})$ is their corresponding contact map for the kth data, we aim to find $\boldsymbol{\lambda}$ that minimizes the objective function

$$\mathcal{L}(\boldsymbol{\lambda}) = \sum_{(P,R,z)\in\mathcal{D}} \Big(\max_{\hat{z}\in\mathcal{CM}(P,R)} [f_{\boldsymbol{\lambda}}(P,R,\hat{z}) + \Delta(z,\hat{z})] - f_{\boldsymbol{\lambda}}(P,R,z) + C\|\boldsymbol{\lambda}\|_1 \Big), \quad (13)$$

where $||.||_1$ is the ℓ_1 norm and C is a weight for the ℓ_1 regularization term to avoid overfitting to training data. Here, $\Delta(z, \hat{z})$ is a loss function of $\hat{z}$ for z defined as

$$\begin{aligned} \Delta(z,\hat{z}) = & \delta^{\text{FN residue}}(\text{\# of false negative residues}) \quad (14) \\ & + \delta^{\text{FP residue}}(\text{\# of false positive residues}) \\ & + \delta^{\text{FN base}}(\text{\# of false negative bases}) \\ & + \delta^{\text{FP base}}(\text{\# of false positive bases}) \\ & + \delta^{\text{FN contact}}(\text{\# of false negative contacts}) \\ & + \delta^{\text{FP contact}}(\text{\# of false positive contacts}), \end{aligned}$$

where $\delta^{\text{FN residue}}$, $\delta^{\text{FP residue}}$, $\delta^{\text{FN base}}$, $\delta^{\text{FP base}}$, $\delta^{\text{FN contact}}$, and $\delta^{\text{FP contact}}$ are hyperparameters to control the trade-off between sensitivity and specificity for learning the parameters. In this case, we can calculate the first term of Eq. (13) by replacing scores u_i, v_j, and w_{ij} in Eq. (6) as

$$\bar{u}_i = \begin{cases} u_i - \delta^{\text{FN residue}} & (\text{if } x_i\text{=1}) \\ u_i + \delta^{\text{FP residue}} & (\text{if } x_i\text{=0}) \end{cases}$$

$$\bar{v}_i = \begin{cases} v_i - \delta^{\text{FN base}} & (\text{if } y_j\text{=1}) \\ v_i + \delta^{\text{FP base}} & (\text{if } y_j\text{=0}) \end{cases}$$

$$\bar{w}_{ij} = \begin{cases} w_{ij} - \delta^{\text{FN contact}} & (\text{if } w_{ij}\text{=1}) \\ w_{ij} + \delta^{\text{FP contact}} & (\text{if } w_{ij}\text{=0}) \end{cases}$$

See Sec. S1 in Supplementary Material for the derivation.

```
$\lambda_k \leftarrow 0$ for $\forall\lambda_k \in \boldsymbol{\lambda}$
repeat
    for all $(P, R, z) \in \mathcal{D}$ do
        $\hat{z} \leftarrow \arg\max_{\hat{z}} [f_{\boldsymbol{\lambda}}(P, R, \hat{z}) + \Delta(z, \hat{z})]$
        for all $\lambda_k \in \boldsymbol{\lambda}$ do
            $\lambda_k \leftarrow \lambda_k - \eta(\phi_k(P, R, \hat{z}) - \phi_k(P, R, z) + C\mathrm{sgn}\lambda_k)$
        end for
    end for
until all the parameters converge
```

Fig. 3. The stochastic subgradient descent algorithm for structured SVMs. sgn is the sign function. $\eta > 0$ is the predefined learning rate.

To minimize the objective function (13), we can apply stochastic subgradient descent (Fig. 3) or forward-backward splitting (Duchi and Singer, 2009).

3 RESULTS

3.1 Implementation

Our method was implemented using the IBM CPLEX optimizer[1] for solving integer programming problems (6)–(10). To extract the structural feature elements described in Sec. 2.2, we employed PSIPRED (Jones, 1999) and CENTROIDFOLD (Hamada *et al.*, 2009) to predict secondary structures of protein and RNA sequences, respectively. We empirically chose the hyperparameters: the penalty for positives $\delta^{\text{FN}\,*} = 4.0$, the penalty for negatives $\delta^{\text{FP}\,*} = 1.0$, and the weight for ℓ_1 regularization term $C = 0.125$ (see Supplementary Material for details).We implemented AdaGrad (Duchi *et al.*, 2011) to control the learning rate η in Fig. 3. The source code of our algorithm is available at `https://github.com/satoken/practip`.

3.2 Dataset

We prepared our dataset in accordance with (Chen *et al.*, 2014) and extracted RNA-bound proteins with X-ray resolution of $\leq$ 3.0Å from the Protein Data Bank (PDB) (Rose *et al.*, 2011). To reduce

[1] `http://www.ibm.com/software/integration/optimization/cplex-optimizer/`

dataset redundancy, we discarded some extracted data such that the dataset contains no protein pairs whose sequence identity is $> 30\%$. As a result, we collected 101 protein–RNA interacting pairs from 81 protein–RNA complexes from the PDB. We considered a residue to bind RNA if at least 1 non-hydrogen atom is contained within the van der Waals contact (4.0Å) or hydrogen-bonding distance (3.5Å) to the non-hydrogen atom of its binding partner. We employed HBPLUS (McDonald and Thornton, 1994) to detect the hydrogen bonds and van der Waals contacts. Among the 101 protein–RNA pairs in our dataset, we found 5,794 residue–base contacts from 3,055 residues and 2,207 bases. See Sec. S4 in Supplementary Material for the list of PDB structures we used.

3.3 Prediction of residue–base contacts

To verify our method, we conducted computational experiments on our dataset, comparing the accuracy under several conditions related to the maximum number of contacts for each residue and base, which restrict at most 1, 2, or 3 contacts, namely, $X_i = Y_i = 1, 2,$ or 3 in Eqs. (11) and (12), or depend on structural profiles (SP) on each residue and base as described in Sec. 2.3.

We evaluated the accuracy of predicting residue–base contacts between proteins and RNAs through three measures: predicted residue–base contacts, binding residues in proteins, and binding bases in RNA sequences. The accuracy of residue–base contacts is assessed by the positive predictive value (PPV) and the sensitivity (SEN), defined as

$$PPV = \frac{TP}{TP + FP}, \quad SEN = \frac{TP}{TP + FN},$$

where TP is the number of correctly predicted contacts (true positives), FP is the number of incorrectly predicted contacts (false positives), and FN is the number of contacts in the true contact map that were not predicted (false negatives). We also used the F-value as the balanced measure between PPV and SEN, which is defined as their harmonic mean:

$$F = \frac{2 \times PPV \times SEN}{PPV + SEN}.$$

The accuracy of binding residues and binding bases is defined in the same way.

We performed 10-fold cross validation. We first divided the dataset into ten subsets, then evaluated the accuracy for each subset following parameter tuning using the other nine subsets. We averaged the accuracy over ten subsets.

Table 6 shows the accuracy of predicting residue–base contacts in PRIs, binding residues in proteins, and binding bases in RNA sequences. As can be seen, more accurate predictions were achieved with larger upper bounds on the number of contacts for each residue and base. Furthermore, when we adapted the upper bound on the number of contacts for each residue and base depending on its structural profile, more accurate predictions were achieved than in the case of a constant upper bound.

It should be noted that in this experiment we could not compare our method with (Hayashida *et al.*, 2013), which is the only method for predicting reside–base contacts in PRIs. This is because we could not conduct an experiment for the Hayashida's method on the same dataset since the software is not available yet, and it requires homologous sequences with accurate alignments for calculating evolutionary information. In addition, Hayashida *et al.* (2013) have reported that it is unfortunately not sufficiently accurate.

3.4 Prediction of binding residues compared with existing methods

We compared our method with existing methods for predicting RNA-binding residues in proteins. DR_bind1 (Chen *et al.*, 2014), KYG (Kim *et al.*, 2006), and OPRA (Perez-Cano and Fernandez-Recio, 2010) are structure-based methods that use 3D structures from the PDB to extract descriptors for prediction. BindN+ (Wang *et al.*, 2010) and Pprint (Kumar *et al.*, 2008) are sequence-based methods that employ evolutionary information instead of 3D structures. Table 7 indicates that our method is comparable with equal or slightly less accuracy than the other methods. Recall that our method employs only sequence information and structural information predicted from only sequences as well as the partner RNAs bound to RNA-binding proteins, instead of 3D structures and evolutionary information.

4 DISCUSSION

We employ ℓ_1 regularization for the weight of features. It is known that ℓ_1 regularization not only avoids overfitting to training data, but also leads to a compact model, that is, fewer features have non-zero weights. Thus, after training the model only 10,594 features have non-zero weights ($>$ 0: 2,870 and $<$ 0: 7,724), as shown in Tables S3–S5 in the Supplementary Material, while the number of potential features is more than 4 billion. This serves as the feature selection, which chooses the features that contribute to the scoring function. As described in Sec. 2.2, our scoring model is a linear combination of feature weights appearing in a given protein–RNA pair. Therefore, we suggest that the larger the weight of a feature after training the model, the more preferable it is for residue–base contacts. This analysis indicated that the weight for long continuous coil regions in protein sequences have large positive values (Table S6 in the Supplementary Material). In other words, such regions preferably interact with RNAs, supporting the result in (Zhang *et al.*, 2010).

Several existing methods for predicting PRIs utilized evolutionary information from homologous sequences, (Wang *et al.*, 2010; Kumar *et al.*, 2008) for protein sequences and (Hayashida *et al.*, 2013) for both protein and RNA sequences. To obtain homologous sequences of target sequences, homologous sequences are typically searched for in large databases using a highly sensitive homology search engine such as PSI-BLAST (Altschul *et al.*, 1997). Furthermore, to extract evolutionary information, homologous sequences must be aligned before predicting PRIs. Homology searches are employed in a wide range of analyses, such as functional analysis of proteins, because if homologous proteins can be found in curated databases we can easily infer the function of the target protein. However, as described above and in (Zhang *et al.*, 2010), the secondary structures of proteins play an essential role in residue–base contacts. Similarly, structural elements of RNA secondary structures also work as key descriptors for residue–base contact prediction (Hiller *et al.*, 2006; Kazan *et al.*, 2010; Fukunaga *et al.*, 2014; Maticzka *et al.*, 2014). This means that structure-based homology searches are needed for PRI prediction based on evolutionary information. Although efficient structural alignment

Table 6. Accuracy under varying conditions on the maximum number of contacts for each residue and base.

	Contacts			Binding residues			Binding bases		
# of contacts	PPV	SEN	F	PPV	SEN	F	PPV	SEN	F
At most 1 contact ($X_i = Y_j = 1$)	0.2281	0.2917	0.2210	0.6241	0.4104	0.4355	0.8140	0.3898	0.4854
At most 2 contacts ($X_i = Y_j = 2$)	0.3633	0.4082	0.3617	0.5199	0.5045	0.4808	0.6864	0.4700	0.5322
At most 3 contacts ($X_i = Y_j = 3$)	0.4379	0.4590	0.4323	0.5941	0.5463	0.5405	0.6251	0.5327	0.5498
Depending on each SP (Table 5)	0.4797	0.5759	0.5051	0.5766	0.6555	0.5861	0.5653	0.6192	0.5585

Table 7. Comparison with other existing methods on our dataset.

	Our method	DR_bind1	KYG	OPRA	BindN+	Pprint
SEN	0.66	0.05	0.60	0.33	0.73	0.82
PPV	0.58	0.69	0.38	0.50	0.54	0.42
F	0.59	0.09	0.47	0.40	0.62	0.56

algorithms for proteins (e.g., Deng and Cheng (2011)) and RNAs (e.g., Sato *et al.* (2012)) have recently been developed, they have not yet been successfully applied to large-scale homology searches.

To the best of our knowledge, (Hayashida *et al.*, 2013) is the only existing method that predicts intermolecular joint structures between proteins and RNAs such as residue–base contacts. However, it is unfortunately not sufficiently accurate. The Hayashida's method is similar to our method in the approach that is based on the machine learning technique with the ℓ_1 regularization. The main difference between our method and the Hayashida's method is that our method employs the large number of features including the structural information of proteins and RNAs, which has been shown to work as key descriptors in PRIs as mentioned above.

We calculated RNA structural profiles from RNA secondary structures predicted by CentroidFold (Hamada *et al.*, 2009), which is one of the most accurate tools for RNA secondary structure prediction. However, it is suspicious that use of RNA secondary structure prediction tools for single RNA molecules is applicable for this purpose, because they do not consider conformational changes induced by interacting with proteins, which may frequently occur in environments *in vivo*. To tackle this problem, we plan to develop an algorithm for simultaneously predicting residue–base contact maps and secondary structures of proteins and RNAs, which may employ a similar approach to RactIP (Kato *et al.*, 2010) for RNA–RNA interaction prediction.

We utilized the structural profiles of predicted RNA secondary structures, which lose important part of structural information, such as base-pairing partners for stacking bases. Most of the existing RBP-binding RNA motif finding methods (Hiller *et al.*, 2006; Kazan *et al.*, 2010; Fukunaga *et al.*, 2014) have also utilized similar encoding, which may not be suitable for dealing with the recognition sites of double-stranded RNA-binding proteins. GraphProt (Maticzka *et al.*, 2014) is an exceptional algorithm that utilized graph-based encoding of RNA secondary structures. Our method should be extended by utilizing another structural profile with no loss of base pairing information like the graph-based encoding of GraphProt.

As shown in Sec. 2.3, we formulated the residue–base contact prediction as an IP problem, which enables us to build a flexible model, such as the constraints on the upper bound on the number of contacts for each residue and base. In contrast to the RNA–RNA interaction model in which each base interacts with at most one base by hydrogen bonds such as Watson–Crick and wobble base-pairs, PRIs contain diverse patterns of residue–base contacts. For example, Kondo and Westhof (2011) have classified the residue–base contacts with respect to three interaction edges on nucleotides (Watson–Crick, Hoogsteen and Sugar) with side-chains and backbones of their partner residues, and have analyzed their propensity. Thus, there is room for further improvement on our model, which can be extended using other constraints for each contact between a residue and a base to consider such observations.

RNA-related high-throughput sequencing technologies have been actively developed, such as Structure-seq (Ding *et al.*, 2014) and hiCLIP (Sugimoto *et al.*, 2015). The large-scale sequencing data produced by these techniques will help us improve our algorithm, especially for training the model. Here we employed complete joint 3D structures of proteins and RNAs as the training dataset, which is not sufficiently large. We cannot build from large-scale sequencing data a complete dataset with residue–base contact maps, but can partially calculate structural profiles and binding bases from *in vivo* chemical probing such as Structure-seq. This information will significantly help us improve our model.

5 CONCLUSION

We developed a max-margin framework for predicting residue–base contacts between proteins and RNAs based on integer programming. To verify our method, we performed several computational experiments. The results suggest that our method based on only sequence information and structural information predicted from only sequences is comparable with RNA-binding residue prediction methods based on known binding data. Further improvements are needed, such as adding informative features, developing a joint prediction model that simultaneously predicts RNA secondary structures and protein contact maps, and using high-throughput sequencing data that can deal with PRI with no residue–base contact information as training data.

ACKNOWLEDGEMENT

The authors thank Professor Kiyoshi Asai and Dr. Junichi Iwakiri at the University of Tokyo, Professor Michiaki Hamada at Waseda University, and participants in the RNA Informatics Workshop 2014 in Sapporo for fruitful discussions. The supercomputer system was provided by the National Institute of Genetics (NIG), Research Organization of Information and Systems (ROIS).

Funding: This work was supported in part by a Grant-in-Aid for Scientific Research (A) (KAKENHI) (No. 23241066) from the Japan Society for the Promotion of Science (JSPS) to K.S. and Y.S., a Grant-in-Aid for Scientific Research (C) (KAKENHI) (No. 25330348) from JSPS to K.S., and a Grant-in-Aid for Scientific Research on Innovative Areas (KAKENHI) (No. 221S0002) from the Ministry of Education, Culture, Sports, Science and Technology (MEXT) to Y.S.

REFERENCES

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**(17), 3389–3402.

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding RNAs. *Nat. Methods*, **8**(6), 444–445.

Chen, Y. C., Sargsyan, K., Wright, J. D., Huang, Y. S., and Lim, C. (2014). Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. *Nucleic Acids Res.*, **42**(3), e15.

Deng, X. and Cheng, J. (2011). MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. *BMC Bioinformatics*, **12**, 472.

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., and Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, **505**(7485), 696–700.

Duchi, J. and Singer, Y. (2009). Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, **10**, 2899–2934.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *The Journal of Machine Learning Research*, **12**, 2121–2159.

Fukunaga, T., Ozaki, H., Terai, G., Asai, K., Iwasaki, W., and Kiryu, H. (2014). CapR: revealing structural specificities of RNA-binding protein target recognition using CLIP-seq data. *Genome Biol.*, **15**(1), R16.

Hamada, M., Kiryu, H., Sato, K., Mituyama, T., and Asai, K. (2009). Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, **25**(4), 465–473.

Hayashida, M., Kamada, M., Song, J., and Akutsu, T. (2013). Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso. *BMC Syst Biol*, **7** **Suppl 2**, S15.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**(22), 10915–10919.

Hiller, M., Pudimat, R., Busch, A., and Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Res.*, **34**(17), e117.

Iwakiri, J., Tateishi, H., Chakraborty, A., Patil, P., and Kenmochi, N. (2012). Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Res.*, **40**(8), 3299–3306.

Iwakiri, J., Kameda, T., Asai, K., and Hamada, M. (2013). Analysis of base-pairing probabilities of RNA molecules involved in protein-RNA interactions. *Bioinformatics*, **29**(20), 2524–2528.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**(2), 195–202.

Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K., and Akutsu, T. (2010). RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, **26**(18), i460–466.

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput. Biol.*, **6**, e1000832.

Kim, O. T., Yura, K., and Go, N. (2006). Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **34**(22), 6450–6460.

Kondo, J. and Westhof, E. (2011). Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide-protein complexes. *Nucleic Acids Res.*, **39**(19), 8628–8637.

Kumar, M., Gromiha, M. M., and Raghava, G. P. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**(1), 189–194.

Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**(1), R17.

McDonald, I. K. and Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**(5), 777–793.

Muppirala, U. K., Honavar, V. G., and Dobbs, D. (2011). Predicting RNA-protein interactions using only sequence information. *BMC Bioinformatics*, **12**, 489.

Murphy, L. R., Wallqvist, A., and Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Eng.*, **13**(3), 149–152.

Pancaldi, V. and Bahler, J. (2011). In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Res.*, **39**(14), 5826–5836.

Perez-Cano, L. and Fernandez-Recio, J. (2010). Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*, **78**(1), 25–35.

Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlic, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B., Zardecki, C., Berman, H. M., and Bourne, P. E. (2011). The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**(Database issue), 392–401.

Sato, K., Kato, Y., Akutsu, T., Asai, K., and Sakakibara, Y. (2012). DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. *Bioinformatics*, **28**(24), 3218–3224.

Sugimoto, Y., Vigilante, A., Darbo, E., Zirra, A., Militti, C., D'Ambrogio, A., Luscombe, N. M., and Ule, J. (2015). hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. *Nature*, **519**(7544), 491–494.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. (2005). Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, **6**, 1453–1484.

Wang, L., Huang, C., Yang, M. Q., and Yang, J. Y. (2010). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol*, **4** **Suppl 1**, S3.

Wang, Y., Chen, X., Liu, Z. P., Huang, Q., Wang, Y., Xu, D., Zhang, X. S., Chen, R., and Chen, L. (2013). De novo prediction of RNA-protein interactions from sequence information. *Mol Biosyst*, **9**(1), 133–142.

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., and Kurgan, L. (2010). Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci.*, **11**(7), 609–628.