Predicting Protein-RNA Residue-base Contacts Using Two-dimensional Conditional Random Field

Morihiro Hayashida*, Mayumi Kamada*, Jiangning Song†‡ and Tatsuya Akutsu*
*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
Email: {morihiro, kamada, takutsu}@kuicr.kyoto-u.ac.jp
†Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia
Email: Jiangning.Song@monash.edu
‡Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China

***Abstract*—Understanding of interactions between proteins and RNAs is essential to reveal networks and functions of molecules in cellular systems. Many studies have been done for analyzing and investigating interactions between protein residues and RNA bases. For interactions between protein residues, it is supported that residues at interacting sites have co-evolved with the corresponding residues in the partner protein to keep the interactions between the proteins. In our previous work, on the basis of this idea, we calculated mutual information (MI) between residues from multiple sequence alignments of homologous proteins for identifying interacting pairs of residues in interacting proteins, and combined it with the discriminative random field (DRF), which is useful to extract some characteristic regions from an image in the field of image processing, and is a special type of conditional random fields (CRFs). In a similar way, in this paper, we make use of mutual information for predicting interactions between protein residues and RNA bases. Furthermore, we introduce labels of amino acids and bases as features of a simple two-dimensional CRF instead of DRF. To evaluate our method, we perform computational experiments for several interactions between Pfam domains and Rfam entries. The results suggest that the CRF model with MI and labels is more useful than the CRF model with only MI.**

I. Introduction

Analyzing molecular recognition and specific interactions between proteins and RNAs is important for understanding construction and evolution of molecular networks and cellular systems. Protein-RNA interactions are involved with regulatory mechanisms such as RNA splicing, translation, and post-transcriptional control. Several studies have investigated tertiary structures of some complexes of proteins with specific RNAs for analyzing how proteins selectively interact with specific sites on nucleic acids [1], [2]. The U1A protein, which is a part of ribosomes, recognizes the same RNA subsequence consisting of seven bases, AUUGCAC, either in the context of a hairpin loop or internal loop [3]. Most protein-RNA complexes are formed by some degree of mutual accommodation between the protein binding surfaces and RNA. A loop of the L11 RNA binding domain is absolutely unstructured without the partner RNA, but becomes ordered on binding [4]. In protein-RNA, protein-single(double)-stranded DNA complexes, van der Waals contacts are more commonly used than hydrogen bond contacts. In protein-RNA interactions, proteins prefer to contact the purine guanine and the pyrimidine uracil using van der Waals contacts and hydrogen bonds, and prefer for the residues arginine, tyrosine and phenylalanine presented in the RNA binding site [2].

In our previous work, we proposed a method for predicting residue-residue contacts between proteins [5]. Also for interactions between amino acid residues, several investigations have been done to reveal detailed interactions between residues [6]–[9]. It can be considered that protein residues at important sites for interactions have been simultaneously mutated to keep their interactions through evolutionary processes. Otherwise, such mutated proteins might lose the interactions, and the individual would disappear by the selection pressure. Thus, interacting residues have been mutated at the same time. Mutual information (MI) between protein residues is useful for predicting interacting residues, which is a quantity representing dependent relationship between two residues, and is calculated from the distribution of amino acids in multiple sequence alignments for homologous proteins. For interactions between protein amino acid residues and RNA bases as well as for those between residues, it can be considered that interacting residues and bases have a tendency to be mutated at the same time. Therefore, we make use of mutual information for predicting residue-base contacts.

Several methods for predicting RNA binding sites in protein sequences have been developed. Kim et al. performed computational analyses of tertiary structures of protein-RNA complexes, and introduced the residue doublet interface propensity, which is a measure of residue pairing preferences in the RNA interface of a protein [10]. Kumar et al. proposed a prediction method using support vector machine (SVM) and evolutionary information, position-specific scoring matrix (PSSM) profiles of protein sequences generated by PSI-BLAST [11]. Liu et al. proposed a new interaction propensity that represents a binding specificity of a residue to the interacting RNA nucleotide by taking its two-side neighborhood in a residue triplet into account, combined with other sequence and structure-based features, and used the random forest technique for the prediction [12].

In the fields of image processing and pattern recognition, Markov random fields (MRFs) have been well studied. Kumar and Hebert proposed discriminative random fields (DRFs) to model spatial interactions in images based on conditional random fields (CRFs) [13]. They claimed that DRFs have

2012 IEEE 6th International Conference on Systems Biology (ISB)
978-1-4673-4398-5/12/$31.00 ©2012 IEEE

Xi'an, China, August 18–20, 2012

several advantages compared to conventional MRFs. For example, DRFs allow to relax the assumption of conditional independence of observed data, and have higher discriminative ability than that of MRFs. Also in the field of bioinformatics, MRFs and CRFs have been used for protein function prediction from protein-protein interaction networks [14], [15], for protein-protein interaction prediction based on protein domain information [16], and for protein residue contacts prediction [5]. However, DRFs have strong associations with images, and thus DRFs may not necessarily be appropriate for predicting residue contacts. Therefore, we propose simple two-dimensional CRF models instead of DRFs. As in the previous work, we give the matrix that consists of all mutual information between two positions in multiple sequence alignments as an input of CRFs. Furthermore, we introduce labels of amino acids and bases as features to our CRF model. We perform computational experiments, and the results suggest that the CRF model with MI and labels is more useful than the CRF model with only MI.

II. Method

In this section, we propose a prediction method based on simple conditional random fields (CRFs) for residue-base contacts between protein-RNA pairs. A protein with an amino acid sequence and an RNA with a base sequence are given as input data. Then, homologous sequences for each sequence are collected, mutual information between two positions of the amino acid and base sequences is calculated, and the probability that a residue interacts with another base is estimated using our proposed CRF models. For training parameters of the CRF model, several pairs of protein and RNA sequences and the interacting pairs of residues and bases are given.

A. Mutual Information

In our proposed method, mutual information for the distribution of amino acids and bases at two positions of protein and RNA sequence alignments is one of important inputs as in our previous work. In this section, we briefly review mutual information for such distributions used in this paper.

Fig. 1 shows an illustration on calculation of mutual information between two positions in two multiple sequence alignments. Suppose that protein amino acid sequence A, RNA base sequence B and the information of residue-base contacts in a protein-RNA complex are obtained. Then, several homologous sequences for sequences A and B are collected, respectively, and a multiple alignments for each set of sequences is calculated in some appropriate way. After that, gaps inserted to sequences A and B by the calculation of the alignment are removed because only residues contained in sequence A and bases in B are the target of our contact prediction. Thus, the length of each multiple alignment becomes the length of the target sequence. Fig. 1 shows such multiple alignments, where the sequences at the first lines denote sequence A and B, respectively. Let $\mathcal{A}$ be the set of 20 distinct amino acids and 1 character that represents a gap, and $\mathcal{B}$ be the set of 4 distinct bases and 1 gap character. Let $p_i(a), p_j(b), p_{ij}(a,b)$ be the observed frequency of amino acid $a \in \mathcal{A}$ at position i, that of base $b \in \mathcal{B}$ at position j, and that of amino acid $a \in \mathcal{A}$ and base $b \in \mathcal{B}$ at positions i and j, respectively, where the frequency is divided by the total number, that is, the number of sequences in an alignment. It should be noted that amino acid a and base b are regarded to simultaneously appear in this paper if both a sequence including a and one including b belong to the same species. Therefore, each sequence in a multiple alignment is needed to be assigned to a sequence in another alignment (see Fig. 1). Then, mutual information m_{ij} between two positions i in protein sequence A and j in RNA sequence B is calculated as follows.

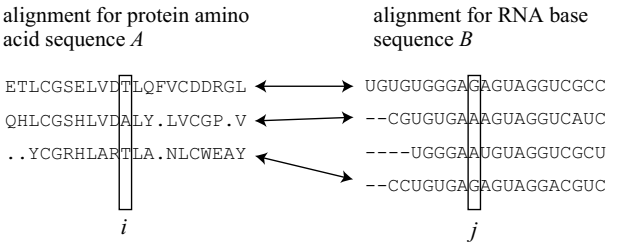

Fig. 1. Illustration on calculation of mutual information between position i in a multiple sequence alignment for protein amino acid sequence A and j in an alignment for RNA base sequence B, where sequences belonging to the same species are connected by arrows. Sequences A and B are shown at the first line of multiple sequence alignments, respectively, and gaps inserted by alignment algorithms are removed.

$$m_{ij} = \sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p_{ij}(a,b)\log\frac{p_{ij}(a,b)}{p_i(a)p_j(b)} \quad (1)$$

$$= H_i + H'_j - H_{ij}, \quad (2)$$

where H_i and H'_j denote the marginal entropies at positions i and j, respectively, that is, $H_i = -\sum_{a\in\mathcal{A}} p_i(a)\log p_i(a)$, $H'_j = -\sum_{b\in\mathcal{B}} p_j(b)\log p_j(b)$, and H_{ij} denotes the joint entropy $H_{ij} = -\sum_{a\in\mathcal{A}}\sum_{b\in\mathcal{B}} p_{ij}(a,b)\log p_{ij}(a,b)$.

B. Two-dimensional Conditional Random Field Models for Residue-base Contacts

In this section, we propose simple two-dimensional CRF models for residue-base contacts.

Conditional random fields (CRFs) were proposed by Lafferty et al. [17]. Let $G(V,E)$ be a graph with a set of vertices V and a set of edges E, where each vertex is related with a random variable x_v, and y_v is observed from the corresponding vertex $v \in V$. Then, $(\boldsymbol{x},\boldsymbol{y})$ is a conditional random field if the random variables x_v follow the Markov property under the conditions y_v according to the graph G, that is, $P(x_v|\boldsymbol{x}_{\{v'\in V|v'\neq v\}},\boldsymbol{y}) = P(x_v|\boldsymbol{x}_{\mathcal{N}_v},\boldsymbol{y})$, where $\mathcal{N}_v$ denotes the set of vertices adjacent to the vertex v in the graph G. CRFs require $P(\boldsymbol{x}|\boldsymbol{y}) > 0$ for all $\boldsymbol{x}$, and are represented as

$$P(x_v|\boldsymbol{x}_{\mathcal{N}_v},\boldsymbol{y}) = \frac{1}{Z_v}\exp\left\{-U_v(\boldsymbol{x},\boldsymbol{y})\right\}, \quad (3)$$

where $U_v(\boldsymbol{x},\boldsymbol{y})$ is a potential function concerning the vertex v, and Z_v is the normalization constant defined by $\sum_{x_v}\exp\{-U_v(\boldsymbol{x},\boldsymbol{y})\}$.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

In our previous work, we used the discriminative random field (DRF) proposed by Kumar and Hebert [13], which is a special type of CRFs, and the potential function is defined as follows.

$$U_v(\boldsymbol{x}, \boldsymbol{y}) = A(x_v, \boldsymbol{y}) + \beta \sum_{v' \in \mathcal{N}_v} I(x_v, x_{v'}, \boldsymbol{y}), \quad (4)$$

where $A(x_v, \boldsymbol{y})$ and $I(x_v, x_{v'}, \boldsymbol{y})$ are called the association and interaction potentials, respectively, $x_v \in \{1, -1\}$, and β is a constant. The potential functions are defined as $A(x_v, \boldsymbol{y}) = -\log\left(\sigma\left(x_v \boldsymbol{w}_f^T \boldsymbol{f}_v(\boldsymbol{y})\right)\right)$, and $I(x_v, x_{v'}, \boldsymbol{y}) = Kx_v x_{v'} + (1-K)\left(2\sigma\left(x_v x_{v'} \boldsymbol{w}_g^T \boldsymbol{g}_{vv'}(\boldsymbol{y})\right) - 1\right)$, respectively, where $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ are parameter vectors, $\boldsymbol{f}_v$ and $\boldsymbol{g}_{vv'}$ are vector-valued functions that map observations $\boldsymbol{y}$ to feature vectors, $\sigma(x) = \frac{1}{1+e^{-x}}$, K $(0 \leq K \leq 1)$ is a constant, and $\boldsymbol{w}^T$ denotes the transpose of $\boldsymbol{w}$. In the field of image processing, the DRF is useful for extracting specific characteristic regions from images. The association potential $A(x_v, \boldsymbol{y})$ can be considered as a gain obtained only from the vertex v and the observations $\boldsymbol{y}$. The interaction potential $I(x_v, x_{v'}, \boldsymbol{y})$ can be considered as a gain obtained from the relationship between vertices v and v', and plays a role of smoothing the truth assignment for random variables $\boldsymbol{x}$ because neighboring pixels in images tend to have similar values to each other. However, the smoothing property is not considered to be desirable for predicting residue-residue and residue-base contacts. Therefore, we propose the following potential function for random variables $r_{ij} \in \{0, 1\}$ that represent whether or not the residue and base at positions i and j interact with each other, where $r_{ij} = 1$ means there exists some contact between i and j, otherwise $r_{ij} = 0$.

$$U_{ij}(\boldsymbol{r}, \boldsymbol{y}) = \boldsymbol{w}_f^T \boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{y}) + \boldsymbol{w}_g^T \sum_{(k,l) \in \mathcal{N}_{ij}} \boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}), \quad (5)$$

where terms in the right-hand side are corresponding to the association and interaction potentials in the DRF, respectively. It should be noted that the set of parameters θ in our CRF model consists of $\boldsymbol{w}_f$, and $\boldsymbol{w}_g$.

In order to determine a CRF model, we must design vector-valued functions $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ijkl}$ and the set $\mathcal{N}_{ij}$ of vertices adjacent with the vertex (i, j) corresponding to positions i and j. In this paper, we use $\mathcal{N}_{ij} = \{(i-1, j), (i, j-1), (i, j+1), (i+1, j)\}$ as adjacent vertices to (i, j) (see Fig. 2). Furthermore, we use mutual information m_{ij} between positions i and j as observations $\boldsymbol{y}$. Then, we define vector-valued functions $\boldsymbol{f}_{ij}^{(1)}$ and $\boldsymbol{g}_{ijkl}^{(1)}$ that give local features as follows.

$$\boldsymbol{f}_{ij}^{(1)}(\boldsymbol{r}, \boldsymbol{m}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{ij} \end{pmatrix}, \quad (6)$$

$$\boldsymbol{g}_{ijkl}^{(1)}(\boldsymbol{r}, \boldsymbol{m}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{kl} \end{pmatrix}, \quad (7)$$

where $\bar{r}$ represents the negation of r, and $\otimes$ denotes the Kronecker product, that is, $A \otimes B = \begin{pmatrix} a_1 B \\ a_2 B \end{pmatrix}$ for matrix $A = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix}$, for example, $\boldsymbol{f}_{ij}^{(1)}(\boldsymbol{r}, \boldsymbol{m}) = (r_{ij}, r_{ij} m_{ij}, \bar{r}_{ij}, \bar{r}_{ij} m_{ij})^T$.

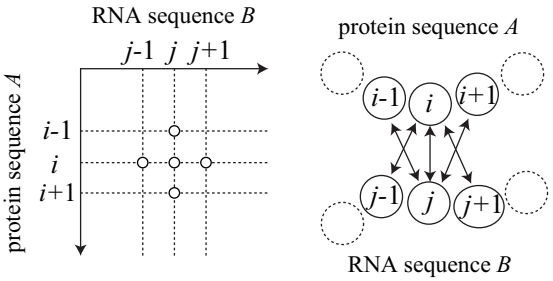

Fig. 2. Adjacent residue-base pairs for (i, j) in two-dimensional random fields.

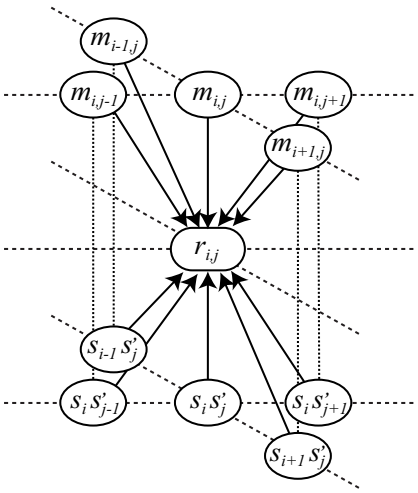

Fig. 3. Relationship between random variable r_{ij} and observations, mutual information m_{ij}, and the pair (s_i, s'_j) of the i-th amino acid in protein sequence A and the j-th base in RNA sequence B, in our CRF model.

In addition to mutual information, we use the protein and RNA sequences as observations. Let s_i and s'_j be the i-th amino acid in protein sequence A and the j-th base in RNA sequence B, respectively. Then, we define other functions $\boldsymbol{f}_{ij}^{(2)}$ and $\boldsymbol{g}_{ijkl}^{(2)}$ as follows.

$$\boldsymbol{f}_{ij}^{(2)}(\boldsymbol{r}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \delta_{(s_i, s'_j)} \otimes \begin{pmatrix} 1 \\ m_{ij} \end{pmatrix}, \quad (8)$$

$$\boldsymbol{g}_{ijkl}^{(2)}(\boldsymbol{r}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \delta_{(s_k, s'_l)} \otimes \begin{pmatrix} 1 \\ m_{kl} \end{pmatrix}, \quad (9)$$

where $\delta_{(a,b)}$ $(a \in \mathcal{A}, b \in \mathcal{B})$ denotes a 0-1 vector with size $20 \times 4 = 80$, the element of which corresponds to (a, b) is 1 and the remaining is 0. The relationship between random variable r_{ij} and observations, mutual information m_{ij}, amino acids s_i, and bases s'_j, is represented in our CRF model as Fig. 3, that is, r_{ij} is related with multiple observations m_{ij} and (s_i, s'_j).

C. Parameter Estimation of Two-dimensional CRFs

We estimate parameters $\theta = \{\boldsymbol{w}_f, \boldsymbol{w}_g\}$ by maximizing pseudo-likelihood function as in [5], [13]. Suppose that N

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

pairs of multiple alignments for protein and RNA sequences and residue-base contacts $\boldsymbol{r}^{(n)}(n = 1, \ldots, N)$ for each pair of proteins and RNAs are given. We calculate mutual information $\boldsymbol{m}^{(n)}$ for each pair. Then, the logarithm of pseudo-likelihood function is given as

$$L(\theta) = \log \prod_{n=1}^{N} \prod_{i} \prod_{j} P(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)}, \boldsymbol{m}^{(n)}, \theta) \tag{10}$$

$$= \sum_{n=1}^{N} \sum_{i} \sum_{j} \Big\{ -U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}) - \log \sum_{r_{ij}^{(n)} \in \{0,1\}} \exp\Big\{ -U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}) \Big\} \Big\}. \tag{11}$$

In order to find parameters maximizing $L(\theta)$, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [18], which is one of quasi-Newton methods, uses partial differentials, and approximates the Hessian matrix by some efficient method. To apply the optimization method, by partially differentiating $L(\theta)$ with respect to each parameter vector $\boldsymbol{w}$, we have

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}} = \sum_{n} \sum_{i} \sum_{j} \Big\{ -\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)})}{\partial \boldsymbol{w}} + \sum_{r_{ij}^{(n)}} P(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)}, \boldsymbol{m}^{(n)}, \theta) \frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)})}{\partial \boldsymbol{w}} \Big\}, \tag{12}$$

where

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)})}{\partial \boldsymbol{w}_f} = \boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{m}), \tag{13}$$

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)})}{\partial \boldsymbol{w}_g} = \sum_{(k,l) \in \mathcal{N}_{ij}} \boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{m}). \tag{14}$$

D. Contact Inference

After estimating parameters, for new pairs of residues and bases, we decide whether or not each pair interacts with each other. In our previous work, we used Iterated Conditional Modes (ICM) [19], which repeatedly updates random variables by maximizing conditional probabilities until each variable is not changed. However, the ICM method often converges to local solutions, for example, also for image processing benchmark problems drawn from energy functions used for stereo, image stitching, and denoising [20]. Therefore, in this paper, we use the sequential tree-reweighted message passing (TRW-S) algorithm [21], which is an improved algorithm of the tree-reweighted message passing (TRW) algorithm [22]. The TRW algorithms try to minimize the upper bound of energy functions for maximization problems by iteratively updating messages $M_{vv';x}$, that vertex v sends to its neighbor v' with state x, and weights $\boldsymbol{w}$ for all decomposed trees.

III. Computational Experiments

A. Data and Implementation

We used seven protein-RNA pairs of chains included in ribosomes, PDB code '1yl4', '2hgu', and '3kcr' from the PDB databank [23], (RS12_THET8, M26923), (RS17_THET8, M26923), (RS8_THET8, M26923), (RL33_THET8, X12612), (RL18_THETH, X01554), (RL27_ECOLI, J01695), and (RL35_ECOLI, J01695), to get residue-base contact data. In addition to the dataset, we used four pairs of chains included in PDB code '3kc4', (RS5_ECOLI, J01695), (RS7_ECOLI, J01695), (RS15_ECO57, J01695), and (RS17_ECOLI, J01695). Tables I shows details of the datasets, for each protein-RNA pair, the PDB code, the identifiers of the chain, UniProt [24], and Pfam [25], and the length of protein sequence A, the identifiers of the chain, GenBank [26], and Rfam [27], and the length of RNA sequence B, and the number of contacts. We supposed that there exists a contact between a residue and a base if the Euclidean distance between an atom of the residue and one of the base is less than or equal to 3 Å. Figure 4 shows protein RS12_THET8 (chain 'O' of '1yl4') and the atoms of RNA M26923 (chain 'A' of '1yl4') within 3 Å of the protein.

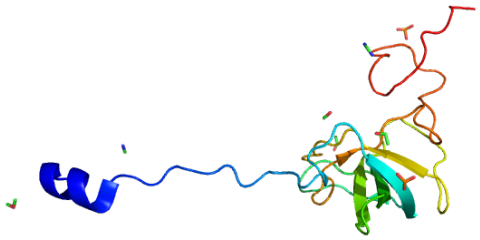

Fig. 4. Protein RS12_THET8, chain 'O' of PDB code '1yl4', and the atoms of RNA M26923, chain 'A' of '1yl4', within 3 Å of the protein.

For the calculation of mutual information between two positions of a residue and a base, we used multiple sequence alignment data provided in the file 'Pfam-A.full' from Pfam database (release 26.0) [25] for protein sequences, and in the file 'Rfam.full' from Rfam database (release 10.1) [27] for RNA sequences. For the calculation of marginal entropies and joint entropies, we used amino acids and bases without classification, and supposed $0 \log 0 = 0$ for $p_i(a) = 0$, $p_j(b) = 0$, or $p_{ij}(a, b) = 0$ because $p \log p \to 0$ for $p \to 0$.

We used libLBFGS (version 1.10) with default parameters to estimate the parameters θ, which is a C implementation of the limited memory BFGS method [28], and is available on the web page, http://www.chokkan.org/software/liblbfgs/. For inferring contacts, we used MRF energy minimization software (version 2.1), which provides a C++ implementation of the TRW-S method [21], available on http://vision.middlebury.edu/MRF/code/, and modified it depending on our energy function formulation.

B. Results

In order to evaluate the proposed CRF-based method, we performed computational experiments using two types of feature vectors $\{\boldsymbol{f}_{ij}^{(1)}, \boldsymbol{g}_{ijkl}^{(1)}\}$, and $\{\boldsymbol{f}_{ij}^{(2)}, \boldsymbol{g}_{ijkl}^{(2)}\}$, and five types of classification of amino acids of 2, 4, 8, 10, and 15 groups proposed by Murphy et al. [29] (see Table II). We performed cross-validation procedures, where a procedure

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE I
DATASET OF SEVEN INTERACTING PROTEIN-RNA PAIRS.

		protein sequence A				RNA sequence B			
PDB	chain	UniProt	Pfam	length	chain	GenBank	Rfam	length	# contacts ($\leq$ 3 Å)
1yl4	O	RS12_THET8	PF00164	122	A	M26923	RF00177	1515	45
1yl4	T	RS17_THET8	PF00366	69	A	M26923	RF00177	1515	43
1yl4	K	RS8_THET8	PF00410	135	A	M26923	RF00177	1515	34
2hgu	5	RL33_THET8	PF00471	48	A	X12612	RF01118	108	21
2hgu	R	RL18_THETH	PF00861	110	B	X01554	RF00001	117	28
3kcr	W	RL27_ECOLI	PF01016	77	8	J01695	RF01118	108	50
3kcr	3	RL35_ECOLI	PF01632	61	8	J01695	RF01118	108	39
3kc4	E	RS5_ECOLI	PF00333	67	A	J01695	RF00177	1530	18
3kc4	G	RS7_ECOLI	PF00177	147	A	J01695	RF00177	1530	25
3kc4	O	RS15_ECO57	PF00312	83	A	J01695	RF00177	1530	25
3kc4	Q	RS17_ECOLI	PF00366	69	A	J01695	RF00177	1530	20

TABLE II
CLASSIFICATION OF AMINO ACIDS BY MURPHY ET AL. [29]

#	amino acid
2	(LVIMCAGSTPFYW),(EDNQKRH)
4	(LVIMC),(AGSTP),(FYW),(EDNQKRH)
8	(LVIMC),(AG),(ST),(P),(FYW),(EDNQ),(KR),(H)
10	(LVIM),(C),(A),(G),(ST),(P),(FYW),(EDNQ),(KR),(H)
15	(LVIM),(C),(A),(G),(S),(T),(P),(FY),(W),(E),(D),(N),(Q),(KR),(H)

'#' denotes the number of groups in each classification.

TABLE III
RESULTS ON AUC SCORES FOR TEST PAIRS OF THE FIRST DATASET USING MUTUAL INFORMATION, LABELS OF AMINO ACIDS AND BASES, AND THE CLASSIFICATION OF AMINO ACIDS.

test pair	MI	MI+2	MI+4	MI+8	MI+10	MI+15	MI+20
(RS12_THET8, M26923)	0.584414	0.434911	0.479028	0.511932	0.535143	0.555171	0.471759
(RS17_THET8, M26923)	0.520389	0.422199	0.465945	0.541198	0.564818	0.604454	0.577107
(RS8_THET8, M26923)	0.448519	0.637362	0.639753	0.618562	0.655294	0.673612	0.611092
(RL33_THET8, X12612)	0.458122	0.634749	0.598589	0.648907	0.678755	0.573224	0.750755
(RL18_THETH, X01554)	0.497109	0.372135	0.484518	0.520543	0.492414	0.445005	0.565814
(RL27_ECOLI, J01695)	0.554078	0.414698	0.51501	0.610803	0.717221	0.614431	0.57308
(RL35_ECOLI, J01695)	0.56244	0.683728	0.559783	0.672034	0.670191	0.685277	0.745234
average	0.517867	0.514254	0.534660	0.589139	0.616262	0.593024	0.613548

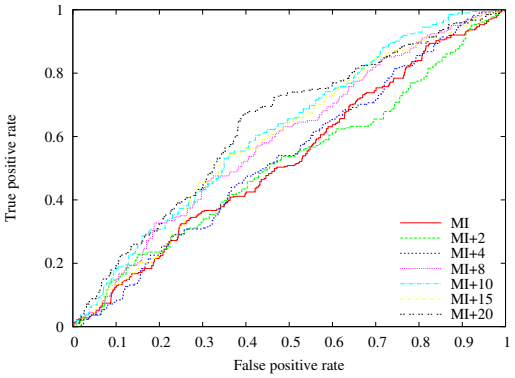

Fig. 5. Average ROC curves for test pairs of the first dataset using mutual information, labels of amino acids and bases, and the classification of amino acids. 'MI' denotes the CRF model with only features of mutual information, $\boldsymbol{f}_{ij}^{(1)}, \boldsymbol{g}_{ijkl}^{(1)}$, and 'MI+$d$' denotes the CRF model with mutual information and the amino acid classification by d groups, $\boldsymbol{f}_{ij}^{(2)}, \boldsymbol{g}_{ijkl}^{(2)}$.

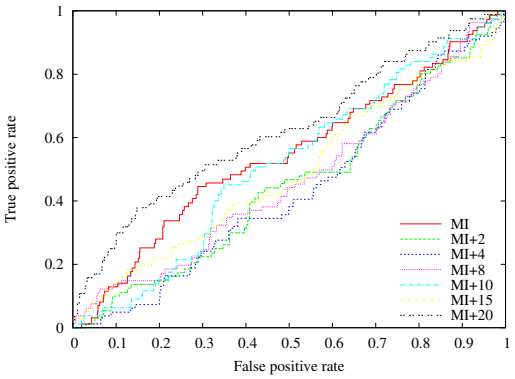

Fig. 6. Average ROC curves for test pairs of the second dataset using mutual information, labels of amino acids and bases, and the classification of amino acids.

used one protein-RNA pair as test data and the remaining pairs as training data. We calculated the conditional probabilities $P(r_{ij} = 1|\boldsymbol{r}_{\mathcal{N}_{ij}}, \boldsymbol{m}, \theta)$ and AUC (Area Under ROC Curve) scores, and took the average.

Table III shows the results on the AUC scores for test pairs of the first dataset using mutual information m_{ij}, labels of amino acids and bases (s_i, s'_j), and the classification of amino acids. 'MI' denotes the CRF model with only features of mutual information, that is, $\{\boldsymbol{f}_{ij}^{(1)}, \boldsymbol{g}_{ijkl}^{(1)}\}$, and 'MI+$d$' denotes the CRF model with mutual information and labels of bases and amino acids classified in d groups, that is, $\{\boldsymbol{f}_{ij}^{(2)}, \boldsymbol{g}_{ijkl}^{(2)}\}$. We can see from the table that the average AUC score using both of mutual information and labels without classification denoted by 'MI+20' was better than that using only mutual information denoted by 'MI'. Furthermore, the average AUC scores of the classifications of 8, 10, and 15 groups were better than those of 2 and 4 groups. It might be considered that a classification of amino acids of a few groups is not able to discriminate whether or not a residue and a base

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

interact with each other. The average ROC (Receiver Operating Characteristic) curves for test pairs of datasets using MI and labels of bases and amino acids classified in d groups are shown in Fig. 5 and 6. These results suggest that the CRF model with MI and labels of amino acids and bases is more useful than the CRF model with only MI.

IV. Conclusion

We proposed a simple two-dimensional conditional random field (CRF)-based method for predicting protein-RNA residue-base contacts, and introduced labels of amino acids and bases as features of the CRF in addition to mutual information. We performed computational experiments for eleven protein-RNA pairs from PDB to evaluate our models, and calculated the average AUC scores for test datasets. The results suggest that the CRF model with MI and labels of amino acids and bases is more useful than the CRF model with only MI. In our previous work, the BFGS method for parameter estimation of the discriminative random field (DRF) did not converge if the potential function includes interaction potentials, which represent relationships between neighbor vertices. Our simple CRF in this paper improved it, and we were able to deal with interaction potentials for predicting residue-base contacts. However, the problem of predicting residue-base contacts is difficult, and the prediction accuracy by our method was still not good. Although we supposed that a residue and a base interact if the distance is at most 3 Å, we may need to decide the contact condition according to more biological meanings. Furthermore, it is necessary to compare our method with other existing methods. However, there is room to improve our method. We can use other correlation values between residues and bases than mutual information, and modify the feature vectors and potential functions of the CRF.

Acknowledgment

This work was partially supported by Grants-in-Aid #22240009 and #24500361 from MEXT, Japan. JS would like to thank the National Health and Medical Research Council of Australia (NHMRC) and the Chinese Academy of Sciences (CAS) for financially supporting this research via the NHMRC Peter Doherty Fellowship and the Hundred Talents Program of CAS.

References

[1] D. Draper, "Themes in RNA-protein recognition," *Journal of Molecular Biology*, vol. 293, pp. 255–270, 1999.

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

[3] D. Scherly, W. Boelens, W. Venrooij, N. Dathan, J. Hamm, and I. Mattaj, "Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA," *EMBO J.*, vol. 8, pp. 4163–4170, 1989.

[4] M. Markus, A. Hinck, S. Huang, D. Draper, and D. Torchia, "High resolution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding RNA," *Nature Struct. Biol.*, vol. 4, pp. 70–77, 1997.

[5] M. Kamada, M. Hayashida, J. Song, and T. Akutsu, "Discriminative random field approach to prediction of protein residue contacts," in *Proc. 2011 IEEE International Conference on Systems Biology*, 2011, pp. 285–291.

[6] R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Features of protein-protein interactions in two-component signaling deduced from genomic libraries," *Methods Enzymol*, vol. 422, pp. 75–101, 2007.

[7] L. Burger and E. van Nimwegen, "Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method," *Molecular Systems Biology*, vol. 4, p. 165, 2008.

[8] N. Halabi, O. Rivoire, S. Leibler, and R. Ranganathan, "Protein sectors: Evolutionary units of three-dimensional structure," *Cell*, vol. 138, pp. 774–786, 2009.

[9] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa, "Identification of direct residue contacts in protein-protein interaction by message passing," *Proc. Natl. Acad. Sci. USA*, vol. 106, pp. 67–72, 2009.

[10] O. Kim, K. Yura, and N. Go, "Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction," *Nucleic Acids Research*, vol. 34, no. 22, pp. 6450–6460, 2006.

[11] M. Kumar, M. Gromiha, and G. Raghava, "Prediction of RNA binding sites in a protein using SVM and PSSM profile," *Proteins: Structure, Function, and Bioinformatics*, vol. 71, pp. 189–194, 2008.

[12] Z.-P. Liu, L.-Y. Wu, Y. Wang, X.-S. Zhang, and L. Chen, "Prediction of protein-RNA binding sites by a random forest method with combined features," *Bioinformatics*, vol. 26, pp. 1616–1622, 2010.

[13] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–201, 2006.

[14] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun, "Prediction of protein function using protein-protein interaction data," *Journal of Computational Biology*, vol. 10, no. 6, pp. 947–960, 2003.

[15] M. Deng, T. Chen, and F. Sun, "An integrated probabilistic model for functional prediction of proteins," *Journal of Computational Biology*, vol. 11, pp. 463–475, 2004.

[16] M. Hayashida, M. Kamada, J. Song, and T. Akutsu, "Conditional random field approach to prediction of protein-protein interactions using domain information," *BMC Systems Biology*, vol. 5, no. Suppl 1, p. S8, 2011.

[17] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. Int. Conf. on Machine Learning*, 2001.

[18] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.

[19] J. Besag, "On the statistical analysis of dirty pictures," *Journal of Royal Statistical Soc.*, vol. B-48, pp. 259–302, 1986.

[20] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother, "A comparative study of energy minimization methods for Markov random fields with smoothness-based priors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 1068–1080, 2008.

[21] V. Kolmogorov, "Convergent tree-reweighted message passing for energy minimization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1568–1583, 2006.

[22] M. Wainwright, T. Jaakkola, and A. Willsky, "MAP estimation via agreement on trees: message-passing and linear programming," *IEEE Transactions on Information Theory*, vol. 51, pp. 3697–3717, 2005.

[23] P. W. Rose, B. Beran, C. Bi, W. F. Bluhm, D. Dimitropoulos, D. S. Goodsell, A. Prlic, M. Quesada, G. B. Quinn, J. D. Westbrook, J. Young, B. Yukich, C. Zardecki, H. M. Berman, and P. E. Bourne, "The RCSB Protein Data Bank: redesigned web site and web services," *Nucleic Acids Research*, vol. 39, pp. D392–D401, 2011.

[24] The UniProt Consortium, "The Universal Protein Resource (UniProt) in 2010," *Nucleic Acids Research*, vol. 38, pp. D142–D148, 2010.

[25] M. Punta, P. Coggill, R. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. Sonnhammer, S. Eddy, A. Bateman, and R. Finn, "The Pfam protein families database," *Nucleic Acids Research*, vol. 40, pp. D290–D301, 2012.

[26] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and E. Sayers, "Genbank," *Nucleic Acids Research*, vol. 39, pp. D32–D37, 2011.

[27] P. Gardner, J. Daub, J. Tate, B. Moore, I. Osuch, S. Griffiths-Jones, R. Finn, E. Nawrocki, D. Kolbe, S. Eddy, and A. Bateman, "Rfam: Wikipedia, clans and the "decimal" release," *Nucleic Acids Research*, 2011.

[28] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

[29] L. Murphy, A. Wallqvist, and R. Levy, "Simplified amino acid alphabets for protein fold recognition and implications for folding," *Protein Engineering*, vol. 13, pp. 149–152, 2000.

978-1-4673-4398-5/12/$31.00 ©2012 IEEE