Hayashida et al. BMC Systems Biology 2013, 7(Suppl 2):S15
http://www.biomedcentral.com/1752-0509/7/S2/S15

BMC Systems Biology

RESEARCH Open Access

Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso

Morihiro Hayashida[1*], Mayumi Kamada[1], Jiangning Song[2,3], Tatsuya Akutsu[1]

From The 6th International Conference on Computational Systems Biology (ISB2012)
Xi'an, China. 18-20 August 2012

Abstract

Background: To uncover molecular functions and networks in biological cellular systems, it is important to dissect interactions between proteins and RNAs. Many studies have been performed to investigate and analyze interactions between protein amino acid residues and RNA bases. In terms of interactions between residues in proteins, it is generally accepted that an amino acid residue at interacting sites has coevolved together with the partner residue in order to keep the interaction between residues in proteins. Based on this hypothesis, in our previous study to identify residue-residue contact pairs in interacting proteins, we made calculations of mutual information (*M I*) between amino acid residues from some multiple sequence alignment of homologous proteins, and combined it with a discriminative random field (DRF) approach, which is a special type of conditional random fields (CRFs) and has been proved useful for the purpose of extracting distinguishing areas from a photograph in the image processing field. Recently, the evolutionary correlation of interactions between residues and DNA bases has also been found in certain transcription factors and the DNA-binding sites.

Results: In this paper, we employ more generic two-dimensional CRFs than such DRFs to predict interactions between protein amino acid residues and RNA bases. In addition, we introduce labels representing kinds of amino acids and bases as local features of a CRF. Furthermore, we examine the utility of L_1-norm regularization (lasso) for the CRF. For evaluation of our method, we use residue-base interactions between several Pfam domains and Rfam entries, conduct cross-validation, and calculate the average AUC (Area under ROC Curve) score. The results suggest that our CRF-based method using mutual information and labels with the lasso is useful for further improving the performance, especially provided that the features of CRF are successfully reduced by the lasso approach.

Conclusions: We propose simple and generic two-dimensional CRF models using labels and mutual information with the lasso. Use of the CRF-based method in combination with the lasso is particularly useful for predicting the residue-base contacts in protein-RNA interactions.

Introduction

It is essential to understand the organization and evolution of cellular systems and molecular networks through the analysis of interactions and molecular recognition. Protein-RNA interactions are related with regulatory mechanisms including RNA splicing, post-transcriptional control, protein translation, and so on. Many researchers have focused on tertiary structures of complexes consisting of specific proteins and RNAs, and have analyzed how proteins selectively make physical contacts with specific sites on nucleic acids [1,2]. Some degree of mutual accommodation between the protein binding surfaces and RNA causes the formulation of most protein-RNA complexes. Markus et al. reported that a loop of the L11 RNA binding domain becomes ordered on binding although the loop is absolutely unstructured without the

* Correspondence: morihiro@kuicr.kyoto-u.ac.jp
[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
Full list of author information is available at the end of the article

partner RNA [3]. Scherly et al. reported that the same RNA subsequence containing seven bases, AUUGCAC, is recognized by the U1A protein, a part of ribosomes, under the context of an internal loop or hairpin loop [4]. Jones et al. reported that van der Waals contacts are more widely used rather than hydrogen bond contacts in protein-single(double)-stranded DNA and protein-RNA complexes. They pointed out that proteins are likely to use van der Waals contacts and hydrogen bonds in interactions to the pyrimidine uracil and the purine guanine, and prefer phenylalanine, arginine, tyrosine residues in the RNA binding site [2]. Thus, in this paper, we focus on prediction of such residue-base contacts in interacting protein-RNA pairs.

In our previous study, we proposed a prediction method for protein residue-residue contacts [5]. In order to uncover details of interactions between protein amino acid residues, several investigations have been done [6-9]. It is generally accepted that interacting residues in a protein have a pressure to be simultaneously mutated with each other through evolutionary processes to keep their interactions. Under the selection pressure, otherwise, mutations at such interacting sites might lead to loss of the interactions and disappearance of individual. Thus, interacting residues are required to be mutated in a coordinated manner in order to maintain their interactions. Since mutual information (MI) is defined as a quantity representing dependent relationship between two random variables, MI between positions in a protein, which is obtained from the distribution of amino acids in multiple sequence alignments for its homologous proteins, is useful for predicting interacting residues.

For interactions between protein amino acid residues and DNA bases, Yang et al. showed that the evolutions of the transcription factors and the DNA binding sites of the basic helix-loop-helix family, homeo family, high-mobility group family, and transient receptor potential channels family are significantly correlated across eukaryotes [10]. Accordingly, a mutual information-based method was developed for identifying coevolved protein residues and DNA bases. From analogy to interactions between residues, and between residues and DNA bases, it can be concluded that interacting residues and RNA bases tend to be simultaneously mutated. We therefore utilize MI for prediction of residue-base contacts between proteins and RNAs.

Some researchers have developed methods to predict RNA-binding regions in protein sequences. Kumar et al. proposed utilization of evolutionary information and position-specific scoring matrix (PSSM) profiles that PSI-BLAST generates, and predicted using support vector machine (SVM) approach [11]. Furthermore, they developed different hybrid approaches, and improved the prediction accuracy [12]. Kim et al. introduced some propensity in the RNA interface of a protein to measure residue pairing preferences by computationally analyzing tertiary structures of protein-RNA complexes [13]. Muppirala et al. developed a prediction method from only sequence information for interactions between RNAs and proteins, called RPISeq [14]. Liu et al. proposed a novel interaction propensity representing a binding selectivity of a residue to the interacting RNA nucleotide by considering its two-side neighborhood in a residue triplet with combination of other sequence, features based on structures, and the random forest technique [15]. These methods, however, do not predict contacts between specific bases and residues in RNAs and proteins, and only detect RNA-binding regions in proteins.

Markov random fields (MRFs) have been widely used in fields of pattern recognition, image processing, and so on. For modeling of spatial interactions in images, Kumar and Hebert proposed the discriminative random field (DRF) that is defined as a special type of conditional random fields (CRFs), and applied their method to detection of regions of non-natural, artificial buildings from photographs [16]. They maintained that their DRFs have some advantages in comparison with general MRFs. For instance, DRFs are able to discriminate in higher accuracies than MRFs, and can be constructed without the assumption of conditional independence for observed data. It should be noted that such DRFs might not represent actual structures. MRFs and CRFs have been also used in the field of computational biology. Deng et al. proposed an MRF-based method to predict protein functions from protein-protein interaction networks [17,18]. Hayashida et al. proposed a CRF-based method to predict protein-protein interactions using protein domain information [19]. Kamada et al. proposed a DRF approach to predict protein residue contacts [5]. On the other hand, the DRF proposed by Kumar and Hebert [16] is strongly associated with images, and the interaction potential works to smooth borders of regions. Thus, DRFs may not be directly applicable to prediction of protein residue contacts. Hence, instead of DRFs, we propose simple and generic two-dimensional CRF models that accept more interaction structures. In our previous study, we provided ordinary mutual information between two positions obtained from multiple alignments as an input to CRFs [20]. Dunn et al. proposed an improvement of MI, called MI_p, and claimed that it dramatically improved residue contact prediction [21]. We therefore examine MI_p as well as MI. In addition, we introduce labels representing kinds of amino acids and bases as local features of our CRF models. However, inclusion of more parameters in CRF models may cause overfitting. Hence, we examine L_1-norm regularization, or the least absolute shrinkage and selection operator (lasso) [22] for the purpose of avoidance of overfitting. We perform computational experiments, and the

results suggest that the CRF-based method using mutual information and labels with the lasso is useful.

Method

We propose a prediction method based on simple and generic conditional random fields (CRFs) with L_1-norm regularization (lasso) for amino acid residue-base contacts between RNAs and proteins. It takes the amino acid sequence of a protein and the base sequence of an RNA as input data. Then, a sufficient number of homologous sequences for each sequence is gathered in some adequate manner, and mutual information between a position of the protein and one of the RNA is computed. Our method estimates the probability that the residue at a position and the base at another position interact with each other according to our probability formulation of CRFs. To determine parameters of the CRF model for training data, the method takes several protein-RNA pairs with their sequences, and known pairs of positions that a residue and a base interact.

Mutual information

In this section, we briefly review mutual information for distributions of amino acids and bases, and one of its improvements, MI_p, proposed by Dunn et al. [21]. Let A and B be a protein amino acid sequence and an RNA base sequence, respectively. The calculation of mutual information between two positions in two multiple sequence alignments is illustrated as in Figure 1. A sufficient number of homologous sequences for each of sequences A and B is gathered, and multiple sequence alignments are constructed in some appropriate manner. Then, gaps inserted to sequences A and B in the construction of alignments are deleted with the columns because the target of our contact prediction is not such gaps, but amino acid residues in protein A and bases in RNA B. After the deletion, the length of each multiple alignment becomes the same as that of the original sequence. The example in Figure 1 shows such multiple alignments, in which the first sequence in each alignment indicates sequence A or B. Let Σ_a and Σ_b be the set of twenty distinct amino acids and one character representing a gap, and the set of four distinct bases and one gap character, respectively. Let P_i (a) and P_j (b) be the observed frequencies of amino acid a $(\in \Sigma_a)$ at position i, that of base b $(\in \Sigma_b)$ at position j, respectively. Let P_{ij} (a, b) be the joint frequency of amino acid a $(\in \Sigma_a)$ and base b $(\in \Sigma_b)$ at positions i and j. These frequencies are divided by the total number of sequences in a multiple alignment. We assume that the sequence containing amino acid a and the sequence containing base b belong to the same organism for each pair (a, b). Hence, each sequence in a multiple alignment must have a corresponding sequence in another alignment (see Figure 1). Then, mutual information m_{ij} between two positions i in protein A and j in RNA B is defined by

$$m_{ij} = \sum_{a \in \sum_a} \sum_{b \in \sum_b} P_{ij}(a,b) \log \frac{P_{ij}(a,b)}{P_i(a) P_j(b)}. \quad (1)$$

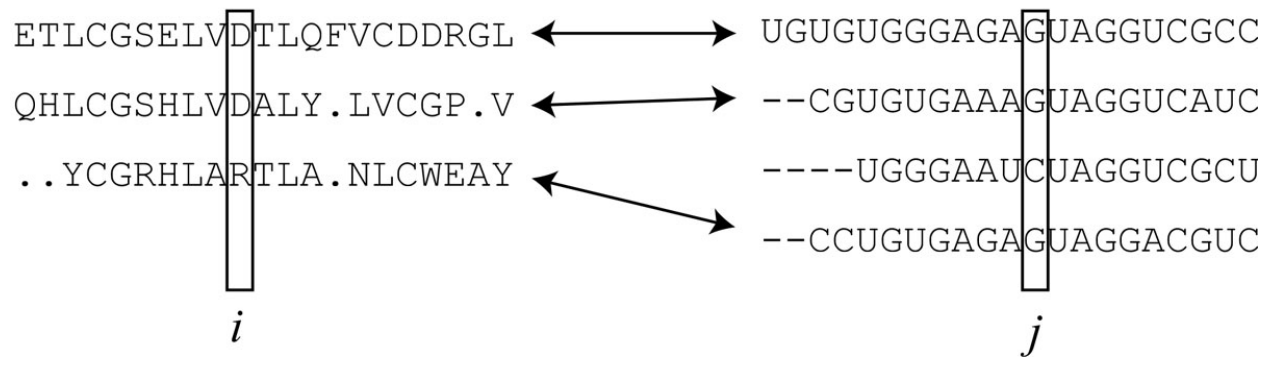

Figure 1 Illustration on calculation of mutual information. Illustration on calculation of mutual information between positions i and j in multiple sequence alignments for protein amino acid sequence A and RNA base sequence B. In this figure, an arrow indicates that sequences connected with each other by the arrow belong to the same organism, and the third sequence in the alignment for RNA B is ignored in calculation of mutual information because it does not have a partner protein sequence of the same organism. Sequences A and B are shown at the first line of multiple sequence alignments, respectively, and gaps inserted by alignment algorithms are deleted with the columns.

However, it has been reported that in some cases it is difficult to identify residue-residue contacts in a protein by MI and thus the usefulness is limited [21]. Dunn et al. proposed a metric, MI_p, by removing background noise of MI. MI_p for residues at positions i and j in a protein is defined by

$$m_{ij} - \frac{\left(\frac{1}{N_p-1}\sum_{k\neq i} m_{ik}\right)\left(\frac{1}{N_p-1}\sum_{k\neq j} m_{jk}\right)}{\frac{2}{N_p(N_p-1)}\sum_{i<j} m_{ij}}, \quad (2)$$

where N_p indicates the number of amino acid residues in the protein. For our purpose of the prediction of residue-base contacts, MI_p is modified to $m_{ij}^{(p)}$ for a pair of a residue at position i and a base at position j as follows:

$$m_{ij}^{(p)} = m_{ij} - \frac{\left(\frac{1}{N_r}\sum_{k=1}^{N_r} m_{ik}\right)\left(\frac{1}{N_p}\sum_{k=1}^{N_p} m_{kj}\right)}{\frac{1}{N_pN_r}\sum_{i=1}^{N_p}\sum_{j=1}^{N_r} m_{ij}}, \quad (3)$$

$$= m_{ij} - \frac{\sum_{k=1}^{N_r} m_{ik}\sum_{k=1}^{N_p} m_{kj}}{\sum_{i=1}^{N_p}\sum_{j=1}^{N_r} m_{ij}}, \quad (4)$$

where N_p and N_r are the number of residues in protein A and that of bases in RNA B, respectively.

Two-dimensional conditional random field (CRF) for residue-base contact prediction

In this section, we show our simple and generic two-dimensional CRFs for prediction of residue-base contacts.

Lafferty et al. proposed conditional random fields (CRFs) by extending Markov random fields (MRFs) [23]. Let $G(V, E)$ be a graph that consists of a set of vertices V and a set of edges E. In these random fields, each vertex v ($\in V$) is related with a random variable x_v. Then, $(\boldsymbol{x}, \boldsymbol{y})$ is a conditional random field if random variables x_v ($\in \boldsymbol{x}$) follow the Markov property under observations $\boldsymbol{y}$ according to the graph G. It means that $Pr(x_v|\boldsymbol{x}_{\{v'\in V|v'\neq v\}}, \boldsymbol{y}) = Pr(x_v|\boldsymbol{x}_{\mathcal{N}_v}, \boldsymbol{y})$, where $\mathcal{N}_v$ indicates the set of vertices neighboring with the vertex v in G. This property requires $Pr(\boldsymbol{x}'|\boldsymbol{y}) > 0$ for all subsets $\boldsymbol{x}'$ of random variables $\boldsymbol{x}$. Thus, CRFs can be represented as

$$Pr(x_v|\boldsymbol{x}_{n_v}, \boldsymbol{y}) = \frac{1}{Z_v}\exp\left\{-U_v(\boldsymbol{x}, \boldsymbol{y})\right\}, \quad (5)$$

where $U_v(\boldsymbol{x}, \boldsymbol{y})$ indicates a potential function with respect to the vertex v, and Z_v indicates the normalization constant defined as $\sum_{x_v}\exp\left\{-U_v(\boldsymbol{x}, \boldsymbol{y})\right\}$.

The discriminative random field (DRF) proposed by Kumar and Hebert [16] is a special type of CRFs. In our previous study [5], we applied the DRF to prediction of residue-residue contacts. The potential function $U_v(\boldsymbol{x}, \boldsymbol{y})$ is defined by

$$U_v(\boldsymbol{x}, \boldsymbol{y}) = A(x_v, \boldsymbol{y}) + \beta\sum_{v'\in\mathcal{N}_v} I(x_v, x_{v'}, \boldsymbol{y}), \quad (6)$$

where β is a constant, and random variable x_v takes 1 or -1. The association potential $A(x_v, \boldsymbol{y})$ and interaction potential $I(x_v, x_{v'}, \boldsymbol{y})$ are defined by

$$A(x_v, \boldsymbol{y}) = -\log\left(\sigma\left(x_v\boldsymbol{w}_f^T\boldsymbol{f}_v(\boldsymbol{y})\right)\right), \quad (7)$$

$$I(x_v, x_{v'}, \boldsymbol{y}) = \alpha x_v x_{v'} + (1-\alpha)\left(2\sigma\left(x_v x_{v'}\boldsymbol{w}_g^T\boldsymbol{g}_{vv'}(\boldsymbol{y})\right) - 1\right) \quad (8)$$

respectively, where $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ indicate vectors of parameters, $\boldsymbol{f}_v$ and $\boldsymbol{g}_{vv'}$ indicate vector-valued functions of mapping $\boldsymbol{y}$ to feature vectors, α ($0 \le \alpha \le 1$) is a constant, $\sigma(x) = \frac{1}{1+e^{-x}}$, and $\boldsymbol{w}^T$ indicates the transpose of $\boldsymbol{w}$. It has been shown that the DRF is effective to extraction of distinguishing areas from photo images. The association potential $A(x_v, \boldsymbol{y})$ represents a gain obtained only from v and $\boldsymbol{y}$, and the interaction potential $I(x_v, x_v, \boldsymbol{y})$ represents a gain obtained from some relationship of v with v', and works to smooth the truth assignment for random variables $\boldsymbol{x}$ because adjacent pixels in photographs are likely to have similar colors to each other. The smoothing property, however, is not desired for predicting contacts between protein residues and RNA bases. Hence, we use the following potential for random variables $r_{ij} \in \{0, 1\}$ representing whether or not the residue and the base at positions i and j interact with each other, that is to say, $r_{ij} = 1$ if they interact, otherwise $r_{ij} = 0$.

$$U_{ij}(\boldsymbol{r}, \boldsymbol{y}) = \boldsymbol{w}_f^T\boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{y}) + \boldsymbol{w}_g^T\sum_{(k,l)\in\mathcal{N}_{ij}}\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}) \quad (9)$$

Here, it should be noted that the first and second terms in the right-hand side are corresponding to the association and interaction potentials in DRF, respectively. In our CRF model, each vertex in graph G is associated with a position pair (i, j), and the parameter set θ consists of $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$.

To decide a CRF model, vector-valued functions $\boldsymbol{f}_{ij}$, $\boldsymbol{g}_{ijkl}$ that give local features, and a set $\mathcal{N}_{ij}$ of vertices neighboring with vertex (i, j) must be designed. In this paper, we define neighboring vertices with (i, j) as $\mathcal{N}_{ij} = \{(i \pm 1, j), (i, j \pm 1)\}$ (see Figure 2). In addition, we consider MI (m_{ij}) and MI_p ($m_{ij}^{(p)}$) between positions i and j as observations $\boldsymbol{y}$.

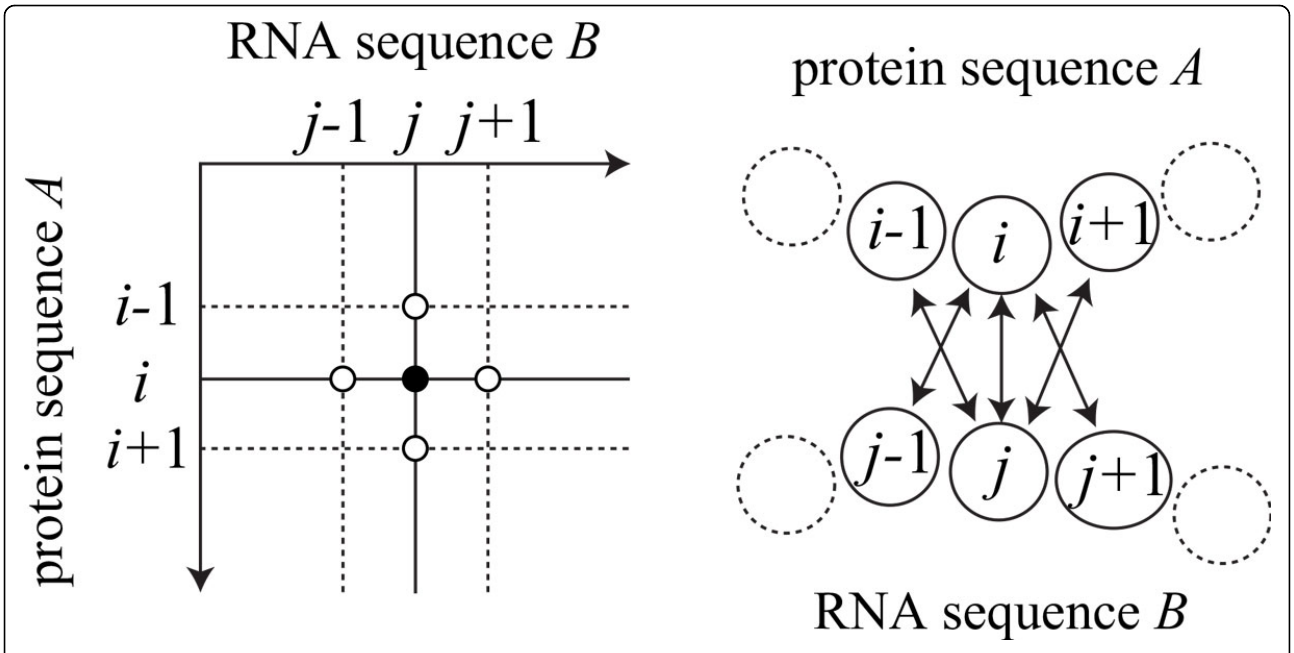

Figure 2 Neighboring residue-base pairs with (*i, j*) in our two-dimensional conditional random fields. Neighboring pairs with (*i, j*) are defined as ($i \pm 1, j$), and ($i, j \pm 1$).

Then, as a formulation of $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ijkl}$, $\boldsymbol{f}_{ij}^{(1)}$ and $\boldsymbol{g}_{ijkl}^{(1)}$ are defined by

$$f_{ij}^{(1)}(\boldsymbol{r}, \boldsymbol{m}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{ij} \end{pmatrix}, \quad (10)$$

$$g_{ijkl}^{(1)}(\boldsymbol{r}, \boldsymbol{m}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \begin{pmatrix} 1 \\ m_{kl} \end{pmatrix}, \quad (11)$$

where $\bar{r}$ indicates the negation of r, that is, $\bar{1} = 0$, $\bar{0} = 1$, and $\otimes$ indicates the Kronecker product, for instance, $X \otimes Y = \begin{pmatrix} x_1 Y \\ x_2 Y \end{pmatrix}$ for matrices $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ and Y, and then $f_{ij}^{(1)}(\boldsymbol{r}, \boldsymbol{m})$ can be also written as $\left(r_{ij}, r_{ij}m_{ij}, \bar{r}_{ij}, \bar{r}_{ij}m_{ij}\right)^T$.

In addition to mutual information, we introduce labels representing kinds of amino acids and bases in the target protein and RNA sequences as observations. Suppose protein sequence A and RNA sequence B are represented by $a_1a_2 \ldots a_{Np}$ and $b_1b_2 \ldots b_{Nr}$, respectively. Then, As another formulation, $\boldsymbol{f}_{ij}^{(2)}$ and $\boldsymbol{g}_{ijkl}^{(2)}$ are defined by

$$f_{ij}^{(2)}(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \delta_{(a_i, b_j)} \otimes \begin{pmatrix} 1 \\ m_{ij} \end{pmatrix}, \quad (12)$$

$$g_{ijkl}^{(2)}(\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \delta_{(a_k, b_l)} \otimes \begin{pmatrix} 1 \\ m_{kl} \end{pmatrix}, \quad (13)$$

respectively, where $\delta_{(a,\ b)}$ ($a \in \Sigma_A$, $b \in \Sigma_B$) without grouping amino acids indicates a 0-1 constant vector having size $20 \times 4 = 80$ that the element corresponding to (a, b) is 1 and the others are 0. Figure 3 shows the relationship of the random variable r_{ij} at sequence positions (i, j) with observations including mutual information m_{ij}, amino acids a_i, and bases b_j, in our CRF model. It means that r_{ij} is related with observations m_{ij} and (a_i, b_j) at multiple neighboring positions, which is an important property of CRFs different from MRFs. Besides, we consider another model witout mutual information for the purpose of model comparison as follows:

$$f_{ij}^{(3)}(\boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \delta_{(a_i, b_j)}, \quad (14)$$

$$g_{ijkl}^{(3)}(\boldsymbol{r}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \bar{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \bar{r}_{kl} \end{pmatrix} \otimes \delta_{(a_k, b_l)}. \quad (15)$$

Estimation of parameters in two-dimensional CRFs

We can estimate parameters $\theta = \{\boldsymbol{w}_f, \boldsymbol{w}_g\}$ from training data by maximizing a pseudo-likelihood function as described in [5,16]. Let N be the number of pairs of given protein and RNA sequences. Let $\boldsymbol{a}^{(n)}$ and $\boldsymbol{b}^{(n)}$(n = 1, ..., N) be the n-th protein and RNA sequences, respectively. Let $\boldsymbol{r}^{(n)}$ be the residue-base contacts for the n-th protein-RNA pair. Then, MI (and also MI_p) $\boldsymbol{m}^{(n)}$ is calculated for the n-th pair. The logarithm pseudo-likelihood function $L(\theta)$ is defined by

$$L(\theta) = \log \prod_{n=1}^{N} \prod_{i} \prod_{j} Pr(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}ij}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta) \quad (16)$$

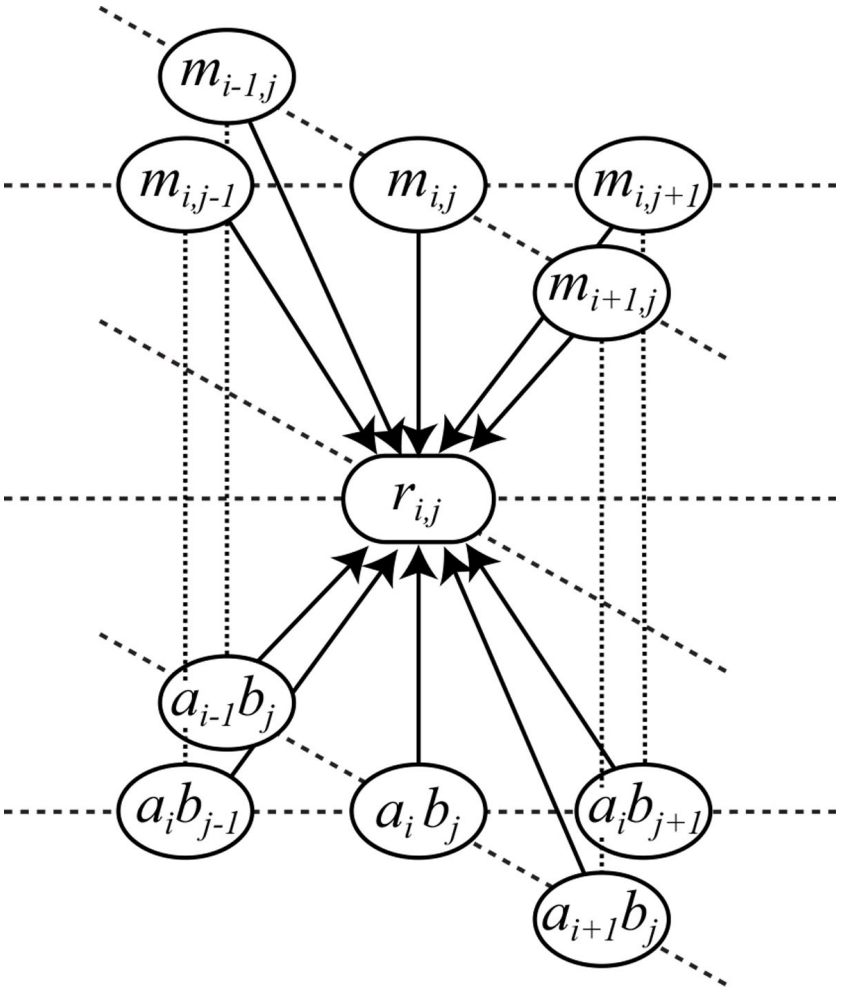

Figure 3 Relationship between the random variable r_{ij} and observations in our CRF model. Relationship between the random variable r_{ij} and the observations of mutual information m_{ij}, and the pair (a_i, b_j) of the i-th amino acid in protein sequence A and the j-th base in RNA sequence B, in our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

We employ the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [24] to find parameters θ maximizing $L(\theta)$, which is a quasi-Newton method that approximates the Hessian matrix by some efficient method using partial differentials. For our problem, the following formulae of $L(\theta)$ partially differentiated by each parameter vector $\boldsymbol{w}$ ($\in \{\boldsymbol{w}_f, \boldsymbol{w}_g\}$) are required.

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}} = \sum_n \sum_i \sum_j \left\{ -\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta)}{\partial \boldsymbol{w}} + \sum_{r_{ij}^{(n)}} Pr(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta) \frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta)}{\partial \boldsymbol{w}} \right\}, \quad (17)$$

where

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta)}{\partial \boldsymbol{w}_f} = \boldsymbol{f}_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}), \quad (18)$$

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}, \theta)}{\partial \boldsymbol{w}_g} = \sum_{(k,l) \in \mathcal{N}_{ij}} \boldsymbol{g}_{ijkl}(\boldsymbol{r}^{(n)}, \boldsymbol{m}^{(n)}, \boldsymbol{a}^{(n)}, \boldsymbol{b}^{(n)}). \quad (19)$$

It should be noted that parameters θ to be estimated are not included in $\frac{\partial U_{ij}}{\partial w}$.

In addition, we propose to use L_1-norm regularization, or the least absolute shrinkage and selection operator (lasso) [22]. That is, we maximize the following function.

$$L(\theta) - C(||\boldsymbol{w}_f||_1 + ||\boldsymbol{w}_g||_1), \quad (20)$$

where C is a positive constant, and $||\boldsymbol{w}||_1$ indicates L_1 norm of $\boldsymbol{w}$, $\sum_{i=1}^{n} |w_i|$ for $\boldsymbol{w} = (w_1, ..., w_n)^T$.

Contact inference

We determine whether or not a new residue-base pair forms a contact depending on the CRF with the parameters

estimated by the method described in the previous section. Although we used the iterated conditional modes (ICM) [25] in our previous study, it has been recognized that ICM often converges to local solutions in image processing benchmark problems [26]. In this paper, therefore, we apply an improved algorithm of the tree-reweighted message passing (TRW) algorithm [27], the sequential tree-reweighted message passing (TRW-S) algorithm [28]. These method iteratively update messages $M_{v v';x}$ from a vertex v to another v' with state x, and iteratively replace edge weights $\boldsymbol{w}$ for all trees decomposed from the original graph, to minimize the upper bound of the objective function for a maximization problem. In our two-dimensional CRF model, the vertex v and the state x mean a position pair (i, j) and a random variable r_{ij}, respectively, and then $v' \in \mathcal{N}_{ij}$.

Computational experiments

Data and implementation

For the evaluation of our method, we used tertiary structures of protein-RNA complexes in the PDB databank [29], and prepared thirteen protein-RNA pairs, (RL18_THETH, X01554), (RL27_ECOLI, J01695), (RL27_THET8, X12612), (RL33_THET8, X12612), (RL35_ECOLI, J01695), (RS5_ECOLI, J01695), (RS7_ECOLI, J01695), (RS8_HET8, M26923), (RS10_THET8, M26923), (RS12_THET8, M26923), (RS15_ECO57, J01695), (RS17_ECOLI, J01695), and (RS17_THET8, M26923), which are contained in ribosomes, '1yl4', '2hgu', '3kc4' and '3kcr' in PDB code. It should be noted that to get contacts between residues and bases, the sequences stored in PDB for these proteins and RNAs must be the same as those included in multiple sequence alignments of the corresponding Pfam [30] and Rfam [31] entries, respectively, and the sequence in a PDB entry is not always the same as that in UniProt [32] entry referred from the PDB entry. We used only the PDB entries in which the sequence is the same as that in UniProt. For each protein-RNA pair of the dataset, Table 1 shows the followings: the identifiers of UniProt, Pfam, and the chain in PDB, the length of protein sequence A, the identifiers of GenBank [33], Rfam, and the chain, the length of RNA sequence B, the PDB code, the number of sequences in the multiple alignment combined on the basis of the organisms, and the number of contacts within 3 Å and that within 5Å. We supposed that a residue and a base form a contact if the Euclidean distance between an atom of the residue and one of the base is less than or equal to some threshold. In this paper, we examined 3 Å and 5 Å as the threshold of contacts because the distances of hydrogen bonds between oxygen and nitrogen atoms, OH-O, OH-N, NH-O, and NH-N, are about 2.7 to 2.9 Å. For instance, protein RS12_THET8 (chain 'O' of '1yl4') and the atoms of RNA_M26923 (chain 'A') within 3 Å of the protein are shown in Figure 4A, and on the other hand, the protein and the atoms of the RNA within 5 Å of the protein is shown in Figure 4B.

In order to calculate $M\ I$ and $M\ I_p$, we used the file 'Pfam-A.full' of Pfam database (release 26.0) [30] and 'Rfam.full' of Rfam database (release 10.1) [31] for getting multiple sequence alignment data of proteins and RNAs, respectively. In counting the frequencies of amino acids and bases, we also examined several classifications of amino acids with 2, 4, 8, 10, and 15 groups proposed by Murphy et al. [34] as shown in Table 2.

For the parameter estimation of our CRF, as an implementation of BFGS methods, libLBFGS (version 1.10),

Table 1 Dataset of thirteen interacting protein-RNA pairs

protein sequence A				RNA sequence B				PDB code	# sequences in MSA	# contacts	
UniProt	Pfam	chain	length	GenBank	Rfam	chain	length			≤ 3 Å	≤ 5 Å
RL18_THETH	PF00861	R	110	X01554	RF00001	B	117	2hgu	1543	28	85
RL27_THET8	PF01016	Z	81	X12612	RF01118	A	108	2hgu	1356	20	67
RL27_ECOLI	PF01016	W	77	J01695	RF01118	8	108	3kcr	1356	18	69
RL33_THET8	PF00471	5	48	X12612	RF01118	A	108	2hgu	1445	18	40
RL35_ECOLI	PF01632	3	61	J01695	RF01118	8	108	3kcr	1337	12	38
RS5_ECOLI	PF00333	E	67	J01695	RF00177	A	1530	3kc4	1701	13	57
RS7_ECOLI	PF00177	G	147	J01695	RF00177	A	1530	3kc4	1941	25	127
RS8_THET8	PF00410	K	135	M26923	RF00177	A	1515	1yl4	1889	29	93
RS10_THET8	PF00338	M	97	M26923	RF00177	A	1515	1yl4	1711	20	84
RS12_THET8	PF00164	O	122	M26923	RF00177	A	1515	1yl4	1972	45	161
RS15_ECO57	PF00312	O	83	J01695	RF00177	A	1530	3kc4	1821	21	89
RS17_ECOLI	PF00366	Q	69	J01695	RF00177	A	1530	3kc4	1690	18	85
RS17_THET8	PF00366	T	69	M26923	RF00177	A	1515	1yl4	1690	29	93

For each protein-RNA pair, the identifiers of UniProt, Pfam, and the chain in PDB, the length of protein sequence A, the identifiers of GenBank, Rfam, and the chain, the length of RNA sequence B, the PDB code, the number of sequences in the multiple sequence alignment (MSA) combined on the basis of the organisms, and the number of contacts within 3 Å and that within 5 Å are shown.

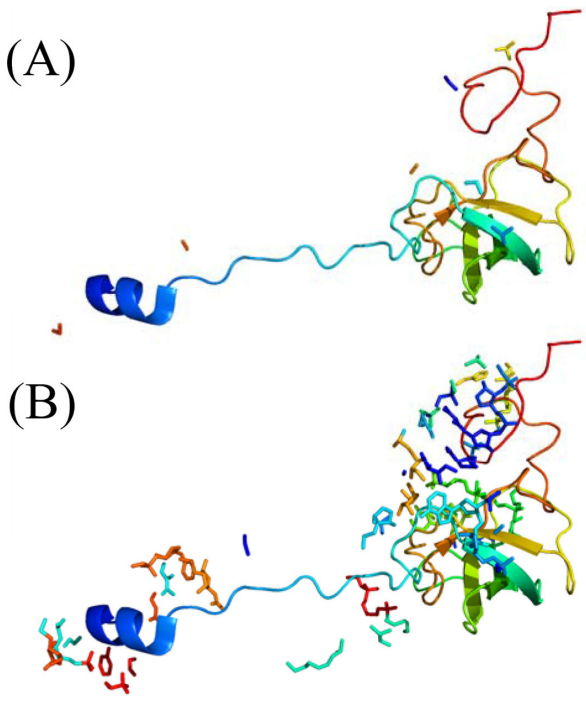

Figure 4 Example of residue-base contacts. (A) Protein RS12_THET8, chain 'O' of PDB code '1yl4', and the atoms of RNA M26923, chain 'A' within 3 Å of the protein. (B) Protein RS12_THET8 and the atoms of RNA M26923 within 5 Å of the protein. It should be noted that for the RNA molecule, only atoms within 3 Å/5Å of the protein are shown.

available from http://www.chokkan.org/software/liblbfgs/, was used with default options, which carries out the limited memory BFGS method [35]. For the contact inference, as an implementation of the TRW-S method [28], MRF energy minimization software (version 2.1), available from http://vision.middlebury.edu/MRF/code/, was modified for use depending on our pseudo-likelihood function formulation.

Results

For the evaluation of our proposed CRF, computational experiments were performed in both contact definitions of 3 Å and 5 Å. Three types of local features $\left\{f_{ij}^{(1)}, g_{ijkl}^{(1)}\right\}$, $\left\{f_{ij}^{(3)}, g_{ijkl}^{(3)}\right\}$, and $\left\{f_{ij}^{(3)}, g_{ijkl}^{(3)}\right\}$, five types of grouping amino acids as 2, 4, 8, 10, and 15 groups [34] as shown in Table 2, and lasso parameter C = 0, 1, and 2 were examined. We performed cross-validation procedures, in which each procedure used all residue-base pairs contained in one protein-RNA pair of the dataset for test, and those in the other protein-RNA pairs for training. The conditional probability $Pr(r_{ij} = 1|\mathbf{r}_{\mathcal{N}_{ij}}, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{b}, \theta)$ and the average AUC (Area Under ROC Curve) score were calculated.

Tables 3 and 4 show results on the average AUC scores for test protein-RNA pairs using the contact definitions of 3 Å and 5 Å, respectively, under several conditions. '$M\ I$' ('$M\ I_p$') indicates the CRF model having only features of $M\ I$ ($M\ I_p$), that is, the feature vectors are $\left\{f_{ij}^{(1)}, g_{ijkl}^{(1)}\right\}$, '$M\ I$ + label' ('$M\ I_p$ + label') indicates the model having $M\ I$ ($M\ I_p$) and labels representing kinds of bases and classified amino acids, $\left\{f_{ij}^{(2)}, g_{ijkl}^{(2)}\right\}$, and 'label' indicates the model having only labels, $\left\{f_{ij}^{(3)}, g_{ijkl}^{(3)}\right\}$. It should be noted that the same grouping of amino acids was used in the calculation of $M\ I$ and $M\ I_p$ and in the labels of features for each case of our experiments. The average AUC score using both of the improved mutual information and labels '$M\ I_p$+label' with the grouping of 15 groups with lasso parameter C = 2

Table 2 Classification of amino acids

# groups	classification of amino acids
2	MLVICGATSPFYW/DENQRKH
4	MLVIC/GATSP/FYW/DENQRKH
8	MLVIC/GA/TS/P/FYW/DENQ/RK/H
10	MLVI/C/G/A/TS/P/FYW/DENQ/RK/H
15	MLVI/C/G/A/T/S/P/FY/W/D/E/N/Q/RK/H

Classification of amino acids by Murphy et al. [34]. The two groups are classified by the hydrophobic and hydrophilic properties of amino acid side-chains. The group of (FYW) is aromatic hydrophobic, (TS), (DENQ), and (RK) are polar.

Table 3 Results on average AUC scores for test pairs using the contact definition of 3 Å

# groups	$M\ I$	$M\ I_p$	label	$M\ I$+label	$M\ I_p$+label
without lasso (C = 0)					
2	0.550	0.557	0.503	0.511	0.502
4	0.534	0.517	0.547	0.505	0.502
8	0.541	0.555	0.535	0.512	0.521
10	0.528	0.557	0.519	0.529	0.536
15	0.538	**0.579**	0.533	0.498	0.523
20	0.539	0.574	0.546	0.561	0.557
lasso (C = 1)					
2	0.556	0.570	0.505	0.520	0.492
4	0.525	0.542	0.611	0.615	0.596
8	0.509	0.562	0.610	0.603	0.600
10	0.525	0.553	0.634	0.633	0.629
15	0.510	0.569	**0.635**	0.634	0.621
20	0.510	0.579	0.625	0.631	0.622
lasso (C = 2)					
2	0.533	0.521	0.510	0.504	0.508
4	0.533	0.543	0.620	0.623	0.620
8	0.550	0.529	0.632	0.624	0.618
10	0.525	0.527	0.625	0.628	0.633
15	0.516	0.524	0.640	0.640	**0.645**
20	0.514	0.546	0.626	0.641	0.642

Results on average AUC scores for test pairs using the contact definition of 3 Å, $M\ I$, $M\ I_p$, labels representing kinds of amino acids and bases, and the grouping of amino acids with lasso parameter C = 0, 1, and 2.

Table 4 Results on average AUC scores for test pairs using the contact definition of 5 Å

# groups	*M I*	*M* I_p	label	*M I*+label	*M* I_p+label
without lasso (C = 0)					
2	0.550	0.520	0.568	0.547	0.565
4	0.543	0.506	0.584	0.563	0.581
8	0.541	0.576	0.584	0.578	0.570
10	0.527	**0.588**	0.545	0.528	0.560
15	0.527	0.587	0.539	0.526	0.518
20	0.530	0.570	0.539	0.506	0.508
lasso (C = 1)					
2	0.527	0.570	0.564	0.575	0.562
4	0.552	0.555	0.582	0.571	0.575
8	0.510	0.559	0.581	0.584	0.590
10	0.511	0.567	0.587	0.579	0.590
15	0.523	0.571	0.571	0.578	0.574
20	0.514	0.572	0.581	0.587	**0.592**
lasso (C = 2)					
2	0.543	0.585	0.581	0.567	0.566
4	0.513	0.557	0.582	0.584	0.580
8	0.509	0.568	0.576	0.574	0.579
10	0.500	0.563	0.594	0.588	0.590
15	0.505	0.591	0.583	0.576	0.582
20	0.502	0.566	0.594	0.598	**0.602**

Results on average AUC scores for test pairs using the contact definition of 5 Å, *M I*, *M* I_p, labels representing kinds of amino acids and bases, and the grouping of amino acids with lasso parameter C = 0, 1, and 2.

using the contact definition of 3 Å was best for the tested residue-base pairs. Figure 5 shows the average ROC (Receiver Operating Characteristic) curves for training and test pairs in that case, where the average AUC score for training pairs was 0.673. In many cases, the average AUC scores of '*M* I_p' were better than those of '*M I*'. It suggests that *M* I_p is useful also for prediction of residue-base contacts. However, the AUC scores of '*M* I_p+ label' were comparable with those of '*M I*+label'. It is considered because in $f_{ij}^{(2)}$ and $g_{ijkl}^{(2)}$ features of labels largely affected the

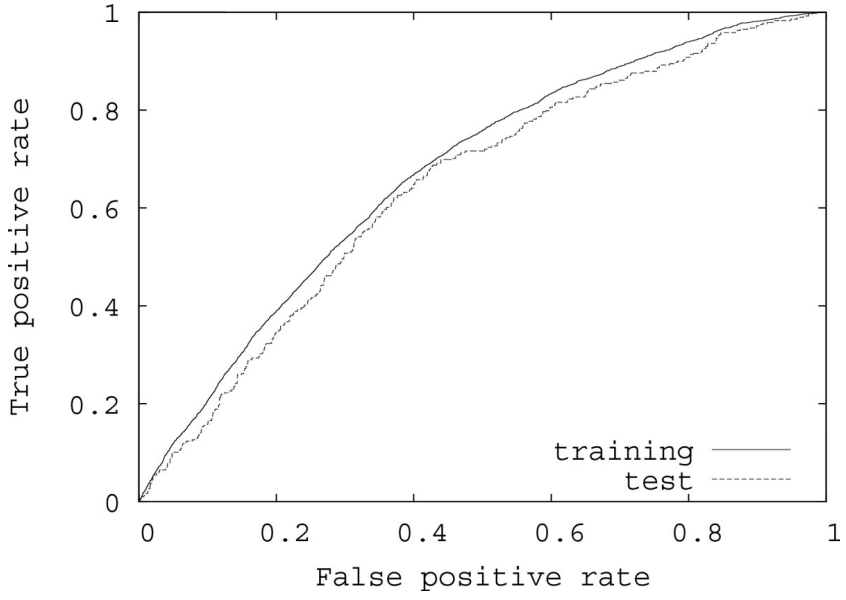

Figure 5 Average ROC curves of the best case in our experiments for training and test pairs. Average ROC curves for training and test pairs using both of *M* I_p and labels with the classification of 15 groups with lasso parameter C = 2 using the contact definition of 3 Å.

results. On the other hand, for the CRF models having features of labels, the AUC scores with the lasso were better than those without the lasso in most cases. It means that the lasso was able to reduce the dimension of parameters concerning labels well. However, the reduction using the contact definition of 5 Å was smaller than that using the contact definition of 3 Å. This might be that false positives increase with the relaxation of contact definitions, which restricted the reduction by the lasso. In such a case, it may be necessary to prepare interacting residue-base pairs manually.

Table 5 shows results on average elapsed time (sec) for an iteration of the cross validation using the contact definition of 3 Å, $M\ I_p$, labels representing kinds of amino acids and bases, and the grouping of amino acids with lasso parameter C = 0, 1, and 2. It should be noted that in an iteration, about 1140000 residue-base pairs on average were used as training data for parameter estimation and about 95000 residue-base pairs were used as test data. Each computational experiment was conducted using a Xeon CPU 3.47GHz. The average elapsed times by '$M\ I_p$ +label' were longer than those by '$M\ I_p$' and 'label' because '$M\ I_p$+label' uses more parameters. For the methods using labels, the average elapsed times with the lasso were shorter than those without the lasso in most cases. It means that parameter reduction by the lasso contributed to the decrease of execution time. All together, these results suggest that the CRF-based method using mutual information and labels representing kinds of amino acids and bases with the lasso is very useful for further improving the prediction performance.

Conclusion

We addressed residue-base contacts between proteins and RNAs, and developed the conditional random field (CRF)-based prediction method, which used labels representing kinds of classified amino acids and bases as local features of the CRF combined with mutual information. In addition, we applied L_1-norm regularization (lasso) to our CRF-based method for avoiding overfitting. For the evaluation of our proposed method, thirteen protein-RNA pairs included in PDB were used in computational experiments, and the average AUC score for test datasets was calculated. From the results, it is seen that the CRF-based method using mutual information and labels representing kinds of amino acids and bases with the lasso is very useful. Furthermore, our proposed CRFs have another advantage. In the previous study [5], the optimization method to the discriminative random field (DRF) with interaction potentials representing relationships between neighboring vertices did not converge. On the other hand, in this paper, our generic two-dimensional CRFs improved this aspect, and was able to deal with interaction potentials for prediction of residue-base contacts. The problem of predicting residue-base contacts, however, is still difficult, and the prediction accuracy was not satisfying. Hence, high-quality datasets of residue-base contacts may need to be prepared with the assistance of biological experts although in this paper contact data were generated depending on only distances between atoms included in a residue and a base. Besides, we can consider use of other measures representing the correlation of a residue with a base instead of mutual information to further improve our prediction method. Modifying local features and potentials in the CRF is also another future work.

Table 5 Results on average elapsed time

# groups	$M\ I_p$			label			$M\ I_p$+label		
	C = 0	1	2	C = 0	1	2	C = 0	1	2
2	55.5	43.5	41.5	46.2	46.9	46.2	80.8	64.6	62.6
4	57.9	51.9	42.8	50.6	47.7	48.0	127.7	63.8	62.5
8	56.9	55.8	55.6	54.3	50.5	50.8	194.9	68.2	67.1
10	54.2	57.4	52.5	57.1	52.2	51.8	235.1	73.0	72.8
15	55.6	57.2	55.2	65.2	55.5	55.1	342.5	79.8	79.2
20	57.8	60.4	55.2	68.1	58.2	58.3	320.8	84.6	82.9

Results on average elapsed time (sec) for an iteration of the cross validation using the contact definition of 3 Å, $M\ I_p$, labels representing kinds of amino acids and bases, and the grouping of amino acids with lasso parameter C = 0, 1, and 2.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
MH developed and implemented the methods. MH drafted the manuscript. MK, JS and TA participated in the discussions during the development of the methods and helped draft the manuscript. All authors read and approved the final manuscript.

Acknowledgements
A preliminary version of this paper was published in the proceedings of IEEE ISB2012.

Declarations
The publication of this article has been funded by Grants-in-Aid #22240009 and #24500361 from MEXT, Japan. This work was also supported by grants from the Hundred Talents Program of the Chinese Academy of Sciences (CAS), the National Natural Science Foundation of China (61202167), the Knowledge Innovation Program of CAS (KSCX2-EW-G-8), Tianjin Municipal Science & Technology Commission (10ZCKFSY05600) and the National Health and Medical Research Council of Australia (NHMRC) (490989). JS is an NHMRC Peter Doherty Fellow, and a Recipient of the Hundred Talents Program of CAS and the Japan Society for the Promotion of Science (JSPS) Short-term Invitation Fellowship to the Bioinformatics Center, Kyoto University, Japan.
This article has been published as part of *BMC Systems Biology* Volume 7 Supplement 2, 2013: Selected articles from The 6th International Conference of Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/7/S2.

Authors' details
[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan. [2]Department of Biochemistry and Molecular Biology, Monash University, Clayton, VIC 3800, Australia. [3]Tianjin Institute of Industrial Biotechnology, Chinese Academy of Sciences, Tianjin 300308, China.

Published: 17 December 2013

References

1. Draper D: **Themes in RNA-protein recognition.** *Journal of Molecular Biology* 1999, **293**:255-270.
2. Jones S, Daley D, Luscombe N, Berman H, Thornton J: **Protein-RNA interactions: a structural analysis.** *Nucleic Acids Research* 2001, **29**:943-954.
3. Markus M, Hinck A, Huang S, Draper D, Torchia D: **High resolution structure of ribosomal protein L11-C76, a helical protein with a flexible loop that becomes structured upon binding RNA.** *Nature Struct. Biol* 1997, **4**:70-77.
4. Scherly D, Boelens W, Venrooij W, Dathan N, Hamm J, Mattaj I: **Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA.** *EMBO J* 1989, **8**:4163-4170.
5. Kamada M, Hayashida M, Song J, Akutsu T: **Discriminative random field approach to prediction of protein residue contacts.** *Proc. 2011 IEEE International Conference on Systems Biology* 2011, 285-291.
6. White RA, Szurmant H, Hoch JA, Hwa T: **Features of protein-protein interactions in two-component signaling deduced from genomic libraries.** *Methods Enzymol* 2007, **422**:75-101.
7. Burger L, van Nimwegen E: **Accurate prediction of protein-protein interactions from sequence alignments using a Bayesian method.** *Molecular Systems Biology* 2008, **4**:165.
8. Halabi N, Rivoire O, Leibler S, Ranganathan R: **Protein sectors: Evolutionary units of three-dimensional structure.** *Cell* 2009, **138**:774-786.
9. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T: **Identification of direct residue contacts in protein-protein interaction by message passing.** *Proc Natl Acad Sci USA* 2009, **106**:67-72.
10. Yang S, Yalamanchili H, Li X, Yao K, Sham P, Zhang M, Wang J: **Correlated evolution of transcription factors and their binding sites.** *Bioinformatics* 2011, **27**:2972-2978.
11. Kumar M, Gromiha M, Raghava G: **Prediction of RNA binding sites in a protein using SVM and PSSM profile.** *Proteins: Structure, Function, and Bioinformatics* 2008, **71**:189-194.
12. Kumar M, Gromiha M, Raghava G: **SVM based prediction of RNA-binding proteins using binding residues and evolutionary information.** *Journal of Molecular Recognition* 2010, **24**:303-313.
13. Kim O, Yura K, Go N: **Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction.** *Nucleic Acids Research* 2006, **34(22)**:6450-6460.
14. Muppirala U, Honavar V, Dobbs D: **Predicting RNA-protein interactions using only sequence information.** *BMC Bioinformatics* 2011, **12**:489.
15. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen L: **Prediction of protein-RNA binding sites by a random forest method with combined features.** *Bioinformatics* 2010, **26**:1616-1622.
16. Kumar S, Hebert M: **Discriminative random fields.** *International Journal of Computer Vision* 2006, **68(2)**:179-201.
17. Deng M, Zhang K, Mehta S, Chen T, Sun F: **Prediction of protein function using protein-protein interaction data.** *Journal of Computational Biology* 2003, **10(6)**:947-960.
18. Deng M, Chen T, Sun F: **An integrated probabilistic model for functional prediction of proteins.** *Journal of Computational Biology* 2004, **11**:463-475.
19. Hayashida M, Kamada M, Song J, Akutsu T: **Conditional random field approach to prediction of protein-protein interactions using domain information.** *BMC Systems Biology* 2011, **5(Suppl 1)**:S8.
20. Hayashida M, Kamada M, Song J, Akutsu T: **Predicting protein-RNA residue-base contacts using two-dimensional conditional random field.** *Proc. 2012 IEEE International Conference on Systems Biology* 2012.
21. Dunn S, Wahl L, Gloor G: **Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction.** *Bioinformatics* 2008, **24**:333-340.
22. Tibshirani R: **Regression shrinkage and selection via the lasso.** *Journal of the Royal Statistical Society. Series B (Methodological)* 1996, **58**:267-288.
23. Lafferty J, McCallum A, Pereira F: **Conditional random fields: Probabilistic models for segmenting and labeling sequence data.** *Proc. Int. Conf. on Machine Learning* 2001.
24. Bertsekas DP: *Nonlinear Programming* Athena Scientific; 1999.
25. Besag J: **On the statistical analysis of dirty pictures.** *Journal of Royal Statistical Soc* 1986, **B-48**:259-302.
26. Szeliski R, Zabih R, Scharstein D, Veksler O, Kolmogorov V, Agarwala A, Tappen M, Rother C: **A comparative study of energy minimization methods for Markov random fields with smoothness-based priors.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2008, **30**:1068-1080.
27. Wainwright M, Jaakkola T, Willsky A: **MAP estimation via agreement on trees: message-passing and linear programming.** *IEEE Transactions on Information Theory* 2005, **51**:3697-3717.
28. Kolmogorov V: **Convergent tree-reweighted message passing for energy minimization.** *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2006, **28**:1568-1583.
29. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE: **The RCSB Protein Data Bank: redesigned web site and web services.** *Nucleic Acids Research* 2011, **39**:D392-D401.
30. Punta M, Coggill P, Eberhardt R, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer E, Eddy S, Bateman A, Finn R: **The Pfam protein families database.** *Nucleic Acids Research* 2012, **40**:D290-D301.
31. Gardner P, Daub J, Tate J, Moore B, Osuch I, Griffiths-Jones S, Finn R, Nawrocki E, Kolbe D, Eddy S, Bateman A: **Rfam: Wikipedia, clans and the "decimal" release.** *Nucleic Acids Research* 2011.
32. The UniProt Consortium: **The Universal Protein Resource (UniProt) in 2010.** *Nucleic Acids Research* 2010, **38**:D142-D148.
33. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Sayers E: **GenBank.** *Nucleic Acids Research* 2011, **39**:D32-D37.
34. Murphy L, Wallqvist A, Levy R: **Simplified amino acid alphabets for protein fold recognition and implications for folding.** *Protein Engineering* 2000, **13**:149-152.
35. Nocedal J: **Updating quasi-Newton matrices with limited storage.** *Mathematics of Computation* 1980, **35(151)**:773-782.

doi:10.1186/1752-0509-7-S2-S15
Cite this article as: Hayashida *et al.*: Prediction of protein-RNA residue-base contacts using two-dimensional conditional random field with the lasso. *BMC Systems Biology* 2013 **7**(Suppl 2):S15.