Quantitative Biology 2018, 6(2): 155–162
https://doi.org/10.1007/s40484-018-0136-7

RESEARCH ARTICLE

Improving conditional random field model for prediction of protein-RNA residue-base contacts

Morihiro Hayashida[1,*], Noriyuki Okada[1], Mayumi Kamada[2] and Hitoshi Koyano[3]

[1] Department of Electrical Engineering and Computer Science, National Institute of Technology, Matsue College, Shimane 690-8518, Japan
[2] Graduate School of Medicine, Kyoto University, Kyoto 606-8507, Japan
[3] Riken Quantitative Biology Center, Hyogo 650-0047, Japan
* Correspondence: morihiro@matsue-ct.jp

Received September 1, 2017; Revised December 4, 2017; Accepted December 13, 2017

***Background*: For understanding biological cellular systems, it is important to analyze interactions between protein residues and RNA bases. A method based on conditional random fields (CRFs) was developed for predicting contacts between residues and bases, which receives multiple sequence alignments for given protein and RNA sequences, respectively, and learns the model with many parameters involved in relationships between neighboring residue-base pairs by maximizing the pseudo likelihood function.**
***Methods*: In this paper, we proposed a novel CRF-based model with more complicated dependency relationships between random variables than the previous model, but which takes less parameters for the sake of avoidance of overfitting to training data.**
***Results*: We performed cross-validation experiments for evaluating the proposed model, and took the average of AUC (area under receiver operating characteristic curve) scores. The result suggests that the proposed CRF-based model without using L_1-norm regularization (lasso) outperforms the existing model with and without the lasso under several input observations to CRFs.**
***Conclusions*: We proposed a novel stochastic model for predicting protein-RNA residue-base contacts, and improved the prediction accuracy in terms of the AUC score. It implies that more dependency relationships in a CRF could be controlled by less parameters.**

Keywords: protein-RNA interaction; residue-base contact; conditional random field

Author summary: A life system is realized by many interactions between various biomolecules including proteins and RNAs. The proposed stochastic model can be a clue to reveal the mechanism of interactions between protein and RNA molecules, which is based on a conditional random field (CRF) with dependency relationships between neighboring residue-base pairs on the sequences. In this study, we made the CRF-based model more realistic by introducing several dependencies and reducing parameters to be trained, as shown in higher prediction accuracies for several actual protein-RNA complexes of which tertiary structures were experimentally determined.

INTRODUCTION

It is important to uncover biological cellular systems from a molecular point of view. Interactions between proteins and RNAs play essential roles in the regulation of gene expression, the stabilization of protein complexes, maturation of mRNA to the trafficking [1]. Therefore, some disruption to RNA-binding proteins can lead various diseases. In many interactions between proteins and RNAs, its protein and RNA recognize specific sites of each other. It was reported that DNA-protein interactions are different from RNA-protein interactions, and RNA bases make more direct contacts with proteins than do DNA bases [2]. As RNA-binding regions of proteins, the K-homology (KH) domains [3], double-stranded RNA-binding domains (dsRBD) [4,5], DEAD-box domains [6], Pumilio repeat domain [7], zinc fingers [8] and so on, are known. In contrast, binding regions of RNAs have not been well investigated. Gupta and Gribskov reported that different bases are preferred in base-specific and base-nonspecific interactions, and RNA structures in protein-binding regions can be sufficiently distinguished from non-binding regions [9].

Several computational methods for detecting RNA-binding sites and protein-RNA interactions have been developed. Peled *et al*. proposed a *de-novo* function prediction approach based on identifying biophysical features [10]. In their method, random forest (RF) [11] was employed because it yielded better results than neural networks and support vector machines (SVMs). Kumar *et al*. made use of evolutionary information and position-specific scoring matrix (PSSM) profiles, and employed support vector machine (SVM) [12,13]. Perez-Cano and Fernandez-Recio developed an ad hoc algorithm using protein-RNA interface propensities calculated from non-redundant X-ray structures of protein-RNA complexes [14]. Liu *et al*. combined a new interaction propensity with features based on sequences and structures, and achieved an accuracy of 84.5% [15].

Zhang *et al*. proposed a hidden Markov model (HMM)-based algorithm to predict clustered functional RNA-binding sites of proteins by integrating the number and spacing of individual motif sites, the accessibility in RNA secondary structures, and cross-species conservation [16]. Zhao *et al*. developed a method based on structural alignment to known protein-RNA complex structures [17]. Ren and Shen proposed new structural features based on accumulated distances from template patches extracted from RNA-binding interfaces [18]. Wang *et al*. proposed an extended naive Bayes classifier for *de novo* prediction of protein-RNA interactions [19]. Sun *et al*. proposed structural features of residue electrostatic surface potential and triplet interface propensity according to the statistical and structural analysis of protein-RNA complexes [20]. These methods predict RNA-binding sites of proteins and interactions of proteins and RNAs. In this paper, we focus on interactions between both sites of amino acid residues and bases in protein-RNA interactions.

Lafferty *et al*. developed conditional random fields (CRFs) to segment and label sequence data [21]. CRFs have been applied to many problems in the fields of image recognition, natural language processing, and bioinformatics [22–24]. Statistical models based on CRFs have been developed for predicting protein-protein interactions [25], protein residue-residue contacts [26], and protein-RNA residue-base contacts [27,28]. CRFs require evidences that another event has occurred, and mutual information (MI) between residues and bases was introduced, which is calculated from multiple sequence alignments. In general, it is considered that an amino acid residue at an interacting site has coevolved together with its partner RNA base to keep the interaction. MIp was developed to improve residue-residue contact prediction, and is calculated by subtracting a bias value from MI [29]. A prediction method for residue-base contacts in protein-RNA complexes was developed using a CRF-based model [27]. In the model, relationships between neighboring residue-base pairs were considered. Since the model has many parameters, L_1-norm regularization (lasso) [30] was applied to improve the prediction accuracy [28]. In this study, we proposed a novel CRF-based model with more complicated dependency relationships and less parameters than the existing one. As well as MIp, we examined the pseudolikelihood maximization direct-coupling analysis (plmDCA) [31], which was developed to infer a protein tertiary structure from its protein sequence, and try to separate direct interactions from indirect ones between residues. For evaluating the proposed CRF-based model, we performed cross-validation computational experiments, and showed that the proposed model without using the lasso regularization outperforms the existing model with and without the lasso under both input observations of MIp and plmDCA to CRFs.

RESULTS

To evaluate the proposed CRF-based model, we used the same dataset as that in the previous paper, which was extracted from tertiary structures of protein-RNA complexes in PDB [32], and consists of the residue-base pairs included in thirteen protein-RNA pairs as shown in Table 1.

Here, the sequences stored in PDB for these proteins and RNAs were the same as those included in multiple sequence alignments of the corresponding Pfam [33] and Rfam [34] entries, respectively, and the sequence in a PDB entry was the same as that in UniProt [35]. Table 1 shows the followings: the identifier of UniProt of a protein sequence, its length, the identifier of GenBank [36] of an RNA sequence, its length, the identifiers of Pfam and Rfam of alignments, the identifier of PDB, and the number of contacts. It was assumed that a residue and a base interact with each other if the Euclidean distance between an atom of the residue and one of the base is less than or equal to 3 Å 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.

To calculate MIp and plmDCA, we used the file

Table 1 Dataset of the residue-base pairs of protein-RNA pairs [28]

Protein sequence		RNA sequence		Alignment		PDB	#contacts
UniProt	length	GenBank	length	Pfam	Rfam		
RL18_THETH	110	X01554	1543	PF00861	RF00001	2 hgu	28
RL27_THET8	81	X12612	1356	PF01016	RF01118	2 hgu	20
RL27_ECOLI	77	J01695	1356	PF01016	RF01118	3 kcr	18
RL33_THET8	48	X12612	1445	PF00471	RF01118	2 hgu	18
RL35_ECOLI	61	J01695	1337	PF01632	RF01118	3 kcr	12
RS5_ECOLI	67	J01695	1701	PF00333	RF00177	3 kc4	13
RS7_ECOLI	147	J01695	1941	PF00177	RF00177	3 kc4	25
RS8_THET8	135	M26923	1889	PF00410	RF00177	1 yl4	29
RS10_THET8	97	M26923	1711	PF00338	RF00177	1 yl4	20
RS12_THET8	122	M26923	1972	PF00164	RF00177	1 yl4	45
RS15_ECO57	83	J01695	1821	PF00312	RF00177	3 kc4	21
RS17_ECOLI	69	J01695	1690	PF00366	RF00177	3 kc4	18
RS17_THET8	69	M26923	1690	PF00366	RF00177	1 yl4	29

"Pfam-A.full" of Pfam database (release 26.0) and "Rfam.full" of Rfam database (release 10.1) for getting multiple sequence alignment data of proteins and RNAs, respectively. We used an implementation of plmDCA available from https://github.com/pagnani/PlmDCA. In counting the frequencies of amino acids and bases, we also examined several classifications of amino acids with 8, 10, and 15 groups proposed by Murphy *et al.* [37] as shown in Table 2.

To estimate the parameters of the CRF-based models, we employed the limited memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) method [38,39] implemented by libLBFGS (version 1.10), available from http://www.chokkan.org/software/liblbfgs/, with default options, which is a quasi-Newton method approximating the Hessian matrix to maximize the likelihood function. For the contact inference, an implementation of the sequential tree-reweighted message passing (TRW-S) algorithm [40], MRF energy minimization software (version 2.1), available from http://vision.middlebury.edu/MRF/code/, was modified for use, which iteratively update messages from a node to another in the graph, and replace edge weights to minimize the upper bound of the objective function for a maximization problem.

We performed cross-validation procedures, and took the average of AUC (area under ROC curve) scores as in the previous work, where 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. For the previous model, the lasso with coefficient $C = 2$ of the regularization term was applied to the parameter estimation of θ also in this study because it output the best result among $C = 0,1,2$ in the previous study.

Table 3 shows the result on the average AUC scores for test data by the proposed and previous CRF-based models using MIp and plmDCA as input observations in 8, 10, 15, and 20 groups of amino acids. In both input

Table 2 Classification of amino acids proposed by Murphy *et al.* [37]

#groups	Classification of amino acids
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}

Table 3 Result on average AUC scores by proposed and previous CRF-based models using MIp and plmDCA as input observations in 8, 10, 15, and 20 groups of amino acids

#groups	MIp		plmDCA	
	previous [28]	proposed	previous [28]	proposed
8	0.618	0.651	0.609	0.692
10	0.633	0.663	0.623	0.699
15	0.645	0.660	0.647	0.699
20	0.642	0.661	0.632	0.693

observations of MIp and plmDCA, the average AUC score by the proposed model was larger than that by the previous model. The average AUC score by the proposed model with plmDCA in 10 and 15 groups of amino acids was larger than those by the others.

Figure 1 shows the result on the average ROC (receiver operating characteristic) curves by the proposed and previous CRF-based models in 15 groups of amino acids. The curve of the proposed model with plmDCA was above the other curves, and the prediction accuracy, which is the ratio of the number of truely predicted residue-base pairs to the total number of residue-base pairs, was 0.997. These results suggest that the proposed CRF-based model outperforms the existing model even if the lasso regularization is not applied to the proposed model.

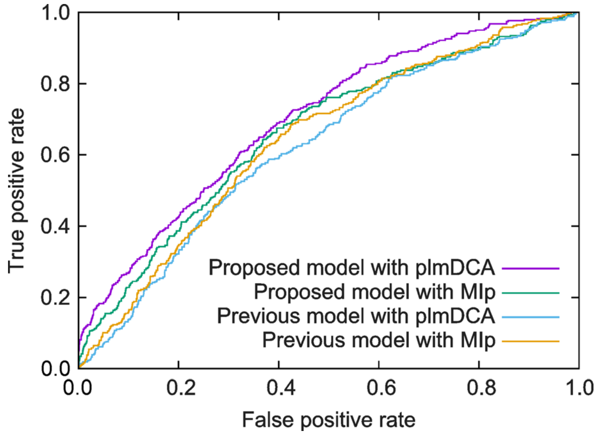

Figure 1. Result on average ROC curves by proposed and previous CRF-based models in 15 groups of amino acids.

CONCLUSION

We improved the existing model for predicting residue-base contacts between proteins and RNAs, and developed a novel model with more complicated dependency relationships and less parameters based on conditional random fields. For evaluation of our proposed model, we performed cross-validation computational experiments, and took the average of AUC scores. The results suggest that the proposed CRF-based model without using L_1-norm regularization (lasso) outperforms the existing model with and without the lasso under both input observations of MIp and plmDCA to CRFs. The number of parameters of the proposed model is 86 without using any classification of amino acids, whereas that of the existing model is 960. It can be considered that the lasso regularization increased the average AUC score for the existing model by automatically selecting effective parameters. In contrast, the proposed model did not need the lasso and obtained the better result because it has a sufficiently small number of parameters and rich dependency relationships between a target residue-base pair and its neighboring pairs. As future work, we would like to further improve the prediction accuracy for understanding detailed mechanisms of protein-RNA interactions. For instance, we can take other features in our model than evolutionary relationships calculated from multiple sequence alignments such as structural and biophysical features.

METHODS

In this section, we briefly reviewed the existing CRF-based model, coevolution measures, MIp and plmDCA, which are input observations to CRFs, calculated from multiple sequence alignments of given protein amino acid and RNA base sequences. In addition, we described the proposed CRF-based model with more complicated dependency relationships and less parameters.

Conditional random field (CRF)-based models

Conditional random fields were developed by extending Markov random fields (MRFs) [21]. Suppose that $G(V, E)$ is a graph with a set V of nodes and a set E of edges, and for a subgraph $G_x(V_x, E_x)$ of G, x_v and $y_{v'}$ are random variables corresponded to nodes $v(\in V_x)$ and $v'(\in V - V_x)$, respectively. Let $\mathcal{N}_v$ be a set of neighboring nodes to v, that is, $\mathcal{N}_v = \{v' | (v, v') \in E_x\}$. Then, $(\boldsymbol{x} = \{x_v\}, \boldsymbol{y} = \{y_{v'}\})$ is a conditional random field if all x_v s follow the Markov property under observations $\boldsymbol{y}$ according to the graph G_x. It means that the probability of x_v given x_v for all $v' \in V_x - \{v\}$ and $\boldsymbol{y}$ is equal to the probability of x_v given $x_{v'}$ for only neighboring nodes $v' \in \mathcal{N}_v$ and $\boldsymbol{y}$, that is, $Pr(x_v | x_{v'}(v' \in V_x - \{v\}), \boldsymbol{y}) = Pr(x_v | x_{v'}(v' \in \mathcal{N}_v), \boldsymbol{y})$. A conditional random field with a strictly positive density can be written by

$$Pr(x_v | x_{v'}(v' \in \mathcal{N}_v), \boldsymbol{y}) = \frac{1}{Z_v} \exp\{-U_v(\boldsymbol{x}, \boldsymbol{y})\}, \quad (1)$$

where Z_v denotes the normalization constant as $\Sigma_{x_v} \exp\{-U_v(\boldsymbol{x}, \boldsymbol{y})\}$, and $U_v(\boldsymbol{x}, \boldsymbol{y})$ denotes a potential function concerning the node v.

For our purpose, given a protein sequence $\boldsymbol{a} = a_1, \ldots, a_{n_p}$ and an RNA sequence $\boldsymbol{b} = b_1, \ldots, b_{n_r}$, a node v in G_x is corresponding to a residue-base position pair (i, j) $(i = 1, \ldots, n_p, j = 1, \ldots, n_r)$. Figure 2 illustrates residue-base pairs around (i, j). A set of neighboring nodes of (i, j) is defined as $\mathcal{N}_{ij} = \{(i \pm 1, j), (i, j \pm 1)\}$. r_{ij} is a random variable, and $r_{ij} = 1$ if residue a_i and base b_j at positions i and j interact with each other, $r_{ij} = 0$ otherwise. Suppose that $\boldsymbol{r} = \{r_{ij}\}$, $\boldsymbol{r}_{\mathcal{N}_{ij}} = \{r_{kl} | (k, l) \in \mathcal{N}_{ij}\}$, and

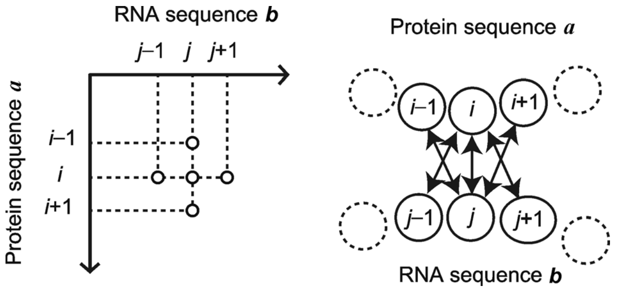

Figure 2. Illustration of residue-base pairs around (i,j).

$\delta_{(a_i,\, b_j)}$ is a $0-1$ constant vector with size $20 \times 4 = 80$ that the element of the amino acid-base pair corresponding to (a_i, b_j) is 1 and the others are 0.

Then, the conditional probability of r_{ij} given $\boldsymbol{r}_{\mathcal{N}_{ij}}$ and $\boldsymbol{y}$, and sequences in the previous work [28] was defined using parameter vectors $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ by

$$Pr(r_{ij}|\boldsymbol{r}_{\mathcal{N}_{ij}}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = \frac{1}{Z_{ij}} \exp\left\{ \boldsymbol{w}_f^T \boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) + \boldsymbol{w}_g^T \sum_{(k,l)\in\mathcal{N}_{ij}} \boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) \right\}, \qquad (2)$$

$$\boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \overline{r}_{ij} \end{pmatrix} \otimes \delta_{(a_i,\, b_j)} \otimes \begin{pmatrix} 1 \\ y_{ij} \end{pmatrix}, \qquad (3)$$

$$\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = \begin{pmatrix} r_{ij} \\ \overline{r}_{ij} \end{pmatrix} \otimes \begin{pmatrix} r_{kl} \\ \overline{r}_{kl} \end{pmatrix} \otimes \delta_{(a_k,\, b_l)} \otimes \begin{pmatrix} 1 \\ y_{kl} \end{pmatrix}, \qquad (4)$$

where Z_{ij} denotes the normalization constant, $\boldsymbol{w}^T$ denotes the transpose of $\boldsymbol{w}$, $\overline{0} = 1$, $\overline{1} = 0$, and $\otimes$ denotes the Kronecker product, for example, $\begin{pmatrix} a \\ b \end{pmatrix} \otimes \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} ac \\ ad \\ bc \\ bd \end{pmatrix}$. The number of parameters is equal to the sum of dimensions of $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$, that is, $2 \times 80 \times 2 + 2 \times 2 \times 80 \times 2 = 960$. Mutual information (MI) and the improved MI calculated from multiple sequence alignments were used as input observations y.

Figure 3 illustrates the dependency relationship between random variables by an element of $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$. In this model, the number of parameters $\boldsymbol{w}_f$, $\boldsymbol{w}_g$ to be estimated is large, and the L_1-norm regularization (lasso) was utilized by improving the prediction accuracy. In addition, r_{ij} depends on only y_{ij}, a_i, and b_j in $\boldsymbol{f}_{ij}$. Hence, we propose the potential function with more complicated dependency relationships and less parameters having the following local features $\boldsymbol{f}_{ij}$, $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$ by adding other association with r_{ij}.

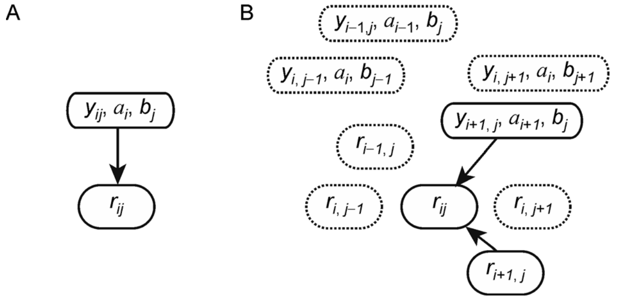
A
B
y_{ij}, a_i, b_j
r_{ij}
$y_{i-1,j}, a_{i-1}, b_j$
$y_{i,j-1}, a_i, b_{j-1}$
$y_{i,j+1}, a_i, b_{j+1}$
$y_{i+1,j}, a_{i+1}, b_j$
$r_{i-1,j}$
$r_{i,j-1}$
r_{ij}
$r_{i,j+1}$
$r_{i+1,j}$

Figure 3. Dependency relationship between random variables in the previous CRF-based model. (A) By an element of $\boldsymbol{f}_{ij}$. (B) By an element of $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$, where the case of $k = i + 1$ and $i = j$ is shown.

$$\boldsymbol{f}_{ij}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = r_{ij} \left(\begin{pmatrix} 1 \\ y_{ij} \\ \max\limits_{(k,\,l)\in\mathcal{N}_{ij}} y_{kl} \\ \min\limits_{(k,\,l)\in\mathcal{N}_{ij}} y_{kl} \end{pmatrix} \otimes \delta_{(a_i,\, b_j)} \right), \qquad (5)$$

$$\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b}) = r_{ij} r_{kl} \begin{pmatrix} |y_{ij} - y_{kl}| \\ y_{ij} y_{kl} \end{pmatrix}, \qquad (6)$$

where $r_{ij} = 1$ if residue a_i and base b_j at positions i and j interact with each other, $r_{ij} = -1$ otherwise, the conditional probability is written by Equation (2), and $\oplus$ denotes the direct sum, for example, $\begin{pmatrix} a \\ b \end{pmatrix} \oplus \begin{pmatrix} c \\ d \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix}$. The number of parameters $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ to be estimated in the training phase is $4 + 80 + 2 = 86$.

Figure 4 illustrates the dependency relationship between random variables by an element of $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$. r_{ij} depends on input observations of all the neighboring nodes according to the maximum and

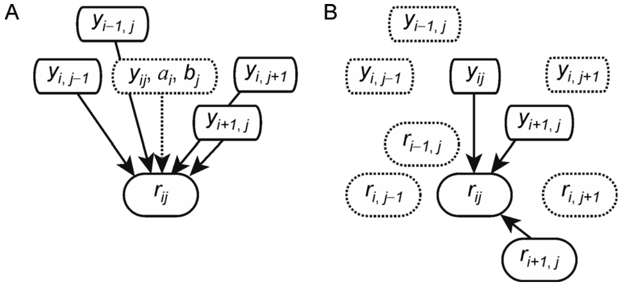

Figure 4. Dependency relationship between random variables in the proposed CRF-based model. (A) By an element of $\boldsymbol{f}_{ij}$. (B) By an element of $\boldsymbol{g}_{ijkl}(\boldsymbol{r}, \boldsymbol{y}, \boldsymbol{a}, \boldsymbol{b})$, where the case of $k = i + 1$ and $l = j$ is shown.

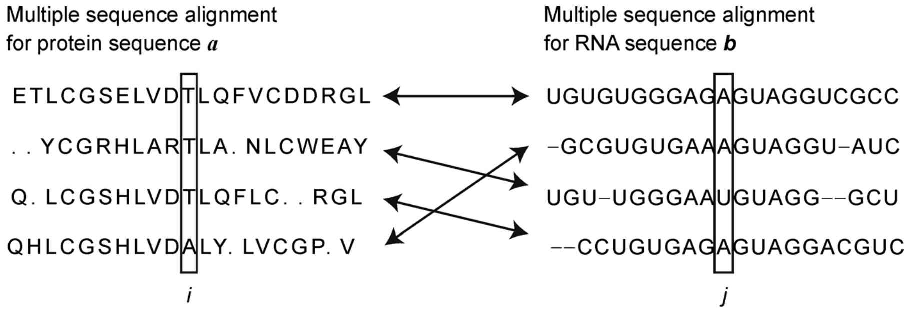

Figure 5. Illustration of multiple sequence alignments for protein sequence *a* and RNA sequence *b*. Arrows denote that two sequences belong to the same species.

minimum of y_{kl} for all $(k,l)\in\mathcal{N}_{ij}$ in $\boldsymbol{f}_{ij}$.

For both CRF-based models, parameters $\theta=\{\boldsymbol{w}_f,\boldsymbol{w}_g\}$ can be estimated from training data of N protein-RNA sequence pairs $\boldsymbol{a}^{(n)}$, $\boldsymbol{b}^{(n)}$, and contacts $\boldsymbol{r}^{(n)}\,(n=1,\ldots,\mathrm{N})$ by maximizing the following pseudo-likelihood function.

$$L(\theta)=\prod_{n=1}^{N}\prod_{i=1}^{n_p^{(n)}}\prod_{j=1}^{n_r^{(n)}} Pr(r_{ij}^{(n)}|\boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)},\boldsymbol{y}^{(n)},\boldsymbol{a}^{(n)},\boldsymbol{b}^{(n)},\theta) \qquad (7)$$

For the sake of reducing redundant parameters of $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$, in the previous model, we used the lasso, and maximized $L(\theta)-C(||\boldsymbol{w}_f||_1+||\boldsymbol{w}_g||_1)$, where C is a positive constant, and $||\boldsymbol{w}||_1$ denotes the L_1 norm of $\boldsymbol{w}$.

In the prediction phase, r_{ij} is determined for test data using the estimated parameters θ and input observations. Then, the problem of finding $r_{ij}\in\{1,-1\}$ maximizing $L(\theta)$ for all i,j under trained parameters $\boldsymbol{w}_f$ and $\boldsymbol{w}_g$ is NP-hard as generally discussed in [40].

Coevolution measure

We examine the improved mutual information MIp [29] and the pseudolikelihood maximization direct-coupling analysis (plmDCA) [31] as input observations $\boldsymbol{y}$ to CRFs. We have two multiple sequence alignments for protein and RNA sequences $\boldsymbol{a}$, $\boldsymbol{b}$ (see Figure 5).

Let $P_i(a)$ and $P_j(b)$ be the observed frequencies of amino acid a at position i, and that of base b at position j, respectively. Let $P_{ij}(a,b)$ be the joint frequency of amino acid a and base b at positions i and j, where the sequence that a_i appears must belong to the same species as the sequence that b_j appears. These frequencies are divided by the total number of sequences in a multiple alignment. Then, mutual information m_{ij} between positions i and j is defined by $\Sigma_a\Sigma_b P_{ij}(a,b)\log\dfrac{P_{ij}(a,b)}{P_i(a)P_j(b)}$. For removing background noise of MI, MIp was proposed to be $m_{ij}-\dfrac{\left(\frac{1}{n_p-1}\Sigma_{k\neq i}m_{ik}\right)\left(\frac{1}{n_p-1}\Sigma_{k\neq j}m_{jk}\right)}{\frac{2}{n_p(n_p-1)}\Sigma_{i<j}m_{ij}}$ for protein residue-residue contacts. For our purpose of predicting residue-base contacts, MIp is modified to

$$m_{ij}-\frac{\Sigma_{i=1}^{n_p}m_{ij}\Sigma_{j=1}^{n_r}m_{ij}}{\Sigma_{i=1}^{n_p}\Sigma_{j=1}^{n_r}m_{ij}}. \qquad (8)$$

Ekeberg *et al.* developed plmDCA for predicting the tertiary structure of a protein by solving the inverse Potts problem. A generalized Potts model can reproduce the empirically observed amino acid frequencies $P_i(a)$ and $P_{ij}(a,b)$, and is defined as

$$Pr(\boldsymbol{a})=\frac{1}{z}\exp(\sum_{i=1}^{n_p}h_i(a_i)+\sum_{1\leqslant i<j\leqslant n_p}J_{ij}(a_i,a_j)), \qquad (9)$$

where $h_i(a_i)$ and $J_{ij}(a_i,a_j)$ are parameters to be determined by the constraints, $Pr(a_i=a)=P_i(a)$ and $Pr(a_i=a,a_j=b)=P_{ij}(a,b)$. From a multiple sequence alignment of a given protein sequence, $\boldsymbol{J}_{ij}$ is determined. Then, the score of plmDCA between amino acid residues is defined by

$$S_{ij}-\frac{\Sigma_{i=1}^{n_p}S_{ij}\Sigma_{j=1}^{n_p}S_{ij}}{\Sigma_{i=1}^{n_p}\Sigma_{j=1}^{n_p}S_{ij}}, \qquad (10)$$

where S_{ij} denotes the Frobenius norm of $\boldsymbol{J}_{ij}'$, which is the zero-sum gauge of $\boldsymbol{J}_{ij}$. For our purpose, we concatenate two multiple sequence alignments of protein and RNA sequences into one alignment such that the species of a protein sequence is the same as that of an RNA sequence.

ACKNOWLEDGEMENTS

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

COMPLIANCE WITH ETHICS GUIDELINES

The authors Morihiro Hayashida, Noriyuki Okada, Mayumi Kamada and Hitoshi Koyano declare that they have no conflict of interests.

This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Re, A., Joshi, T., Kulberkyte, E., Morris, Q. and Workman, C. T. (2014) RNA-protein interactions: an overview. Methods Mol. Biol., 1097, 491–521
2. Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A. and Brasseur, R. (2005) Protein-nucleic acid recognition: statistical analysis of atomic interactions and influence of DNA structure. Proteins, 61, 258–271
3. Siomi, H., Matunis, M. J., Michael, W. M. and Dreyfuss, G. (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res., 21, 1193–1198
4. Feng, G. S., Chong, K., Kumar, A. and Williams, B. R. (1992) Identification of double-stranded RNA-binding domains in the interferon-induced double-stranded RNA-activated p68 kinase. Proc. Natl. Acad. Sci. USA, 89, 5447–5451
5. St Johnston, D., Brown, N. H., Gall, J. G. and Jantsch, M. (1992) A conserved double-stranded RNA-binding domain. Proc. Natl. Acad. Sci. USA, 89, 10979–10983
6. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. and Blinov, V. M. (1989) Two related superfamilies of putative helicases involved in replication, recombination, repair and expression of DNA and RNA genomes. Nucleic Acids Res., 17, 4713–4730
7. Parisi, M. and Lin, H. (2000) Translational repression: a duet of Nanos and Pumilio. Curr. Biol., 10, R81–R83
8. Hall, T. M. (2005) Multiple modes of RNA recognition by zinc finger proteins. Curr. Opin. Struct. Biol., 15, 367–373
9. Gupta, A. and Gribskov, M. (2011) The role of RNA sequence and structure in RNA–protein interactions. J. Mol. Biol., 409, 574–587
10. Peled, S., Leiderman, O., Charar, R., Efroni, G., Shav-Tal, Y. and Ofran, Y. (2016) *De-novo* protein function prediction using DNA binding and RNA binding proteins as a test case. Nat Commun, 7, 13424
11. Ho, T. (1995) Random decision forests. Proc. Third Int. Con. on Document Analysis and Recognition, 1, 278–282
12. 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, 189–194
13. Kumar, M., Gromiha, M. M. and Raghava, G. P. (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J. Mol. Recognit., 24, 303–313
14. Pérez-Cano, L. and Fernández-Recio, J. (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins, 78, 25–35
15. Liu, Z. P., Wu, L. Y., Wang, Y., Zhang, X. S. and Chen, L. (2010) Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics, 26, 1616–1622
16. Zhang, C., Lee, K. Y., Swanson, M. S. and Darnell, R. B. (2013) Prediction of clustered RNA-binding protein motif sites in the mammalian genome. Nucleic Acids Res., 41, 6793–6807
17. Zhao, H., Yang, Y. and Zhou, Y. (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Res., 39, 3017–3025
18. Ren, H. and Shen, Y. (2015) RNA-binding residues prediction using structural features. BMC Bioinformatics, 16, 249
19. 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, 133–142
20. Sun, M., Wang, X., Zou, C., He, Z., Liu, W. and Li, H. (2016) Accurate prediction of RNA-binding protein residues with two discriminative structural descriptors. BMC Bioinformatics, 17, 231
21. Lafferty, J., McCallum, A. and Pereira, F. (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In Proc. Int. Conf. on Machine Learning 2001, pp. 282–289
22. Sha, F. and Pereira, F. (2003) Shallow parsing with conditional random fields. In: Proc. HLT-NAACL 2003, pp. 134–141
23. Yao, K., Peng, B., Zweig, G., Yu, D., Li, X. and Gao, F. (2014) Recurrent conditional random field for language understanding. In 2014 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP), pp. 4077–4081
24. Vemulapalli, R., Tuzel, O., Liu, M. Y. and Chella, R. (2016) Gaussian conditional random field network for semantic segmentation. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3224–3233
25. Hayashida, M., Kamada, M., Song, J. and Akutsu, T. (2011) Conditional random field approach to prediction of protein-protein interactions using domain information. BMC Syst. Biol., 5, S8
26. Kamada, M., Hayashida, M., Song, J. and Akutsu, T. (2011) Discriminative random field approach to prediction of protein residue contacts. In IEEE International Conference on Systems Biology, pp. 285–291
27. Hayashida, M., Kamada, M., Song, J. and Akutsu, T. (2012) Predicting protein-RNA residue-base contacts using two-dimensional conditional random field. In 2012 IEEE International Conference on Systems Biology
28. 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, S15
29. Dunn, S. D., Wahl, L. M. and Gloor, G. B. (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics, 24, 333–340
30. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. J. R. Stat. Soc. B, 58, 267–288
31. Ekeberg, M., Lövkvist, C., Lan, Y., Weigt, M. and Aurell, E. (2013) Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 87, 012707
32. 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., *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res., 39, D392–D401

33. Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., *et al.* (2012) The Pfam protein families database. Nucleic Acids Res., 40, D290–D301
34. Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R., *et al.* (2011) Rfam: Wikipedia, clans and the “decimal” release. Nucleic Acids Res., 39, D141–D145
35. The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res., 38, D142–D148
36. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. and Sayers, E. W. (2011) GenBank. Nucleic Acids Res., 39, D32–D37
37. 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, 149–152
38. Bertsekas, D. P. (1999) Nonlinear Programming. Nashua: Athena Scientific
39. Nocedal, J. (1980) Updating quasi-Newton matrices with limited storage. Math. Comput., 35, 773–782
40. Kolmogorov, V. (2006) Convergent tree-reweighted message passing for energy minimization. IEEE Trans. Pattern Anal. Mach. Intell., 28, 1568–1583