Discriminative Random Field Approach to Prediction of Protein Residue Contacts

Mayumi Kamada*, Morihiro Hayashida*, Jiangning Song†‡ and Tatsuya Akutsu*
*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
Email: {kamada, morihiro, 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 of proteins is important to reveal networks and functions of molecules. Many investigations have been conducted to analyze interactions and contacts between residues. It is supported that residues at interacting sites have co-evolved with those at the corresponding residues in the partner protein to keep the interactions between the proteins. Therefore, mutual information (MI) between residues calculated from multiple sequence alignments of homologous proteins is considered to be useful for identifying contact residues in interacting proteins. In our previous work, we proposed a prediction method for protein-protein interactions using mutual information and conditional random fields (CRFs), and confirmed its usefulness. The discriminative random field (DRF) is a special type of CRFs, and can recognize some specific characteristic regions in an image. Since the matrix consisted of mutual information between residues in two interacting proteins can be regarded as an image, we propose a prediction method for protein residue contacts using DRF models with mutual information. To validate our method, we perform computational experiments for several interactions between Pfam domains. The results suggest that the proposed DRF-based method with MI is useful for predicting protein residue contacts compared with that using the corresponding Markov random field (MRF) model.**

I. Introduction

Analyses of molecular recognition and specific interactions of proteins are important for understanding construction and evolution of molecular networks and cellular systems. Several investigations of amino acid residues of proteins have been conducted to reveal interactions and contacts between residues [1]–[4]. In evolutionary processes of organisms, it can be considered that protein residues at important sites for interactions have been simultaneously mutated to keep their interactions. Otherwise, such mutated proteins might lose the interactions, and the individual would receive selection pressure. In fact, it was confirmed from comparison of putatively orthologous proteins between *S. cerevisiae* and *C. elegans* that interacting proteins evolve at similar evolutionary rates [5]. It means that interacting residues have been mutated at the same time. Therefore, mutual information between 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.

Several methods for predicting interactions of protein residues have been developed based on the idea of coevolution of interacting residues. Little and Chen proposed a normalized mutual information, called ZRes, to remove the biases associated with mutual information, and analyzed the PDZ domain and chorismate synthase family [6]. Weigt et al. proposed Direct Information (DI) that is an improvement of mutual information, and estimated direct residue contacts between sensor kinase and response regulator proteins from the DI calculated by using message passing [4]. Burger and van Nimwegen developed a prediction method based on a Bayesian network method by constructing a dependence tree where a node corresponds to a position of protein sequence alignments [2]. However, predicting protein residue contacts is one of challenging tasks, and it is unknown whether or not the above methods can be applied to various protein pairs. Cheng and Baldi proposed the prediction method using support vector machines for finding contact residues inside of a protein for fold recognition and 3-dimensional structure prediction [7].

In the field of image analysis, Markov random fields (MRFs) have been well studied, for instance, for texture segmentation, a deformable contour model, called EigenSnake, and matching to multiple overlapping objects [8]–[10]. Also in the field of bioinformatics, MRFs have been used for protein function prediction from protein-protein interaction networks [11], [12]. In our previous work, we modeled protein-protein interactions based on domain-domain interactions using conditional random fields (CRFs), and developed prediction methods, which outperformed existing methods based on probabilistic models with domains [13]. Kumar and Hebert proposed discriminative random fields (DRFs) to model spatial interactions in images based on CRFs [14]. They argued that DRFs have several advantages compared to conventional MRFs. For instance, DRFs allow to relax the assumption of conditional independence of observed data, and have higher discriminative ability than that of MRFs. The matrix that consists of all mutual information between two positions in multiple sequence alignments can be considered as an image. Therefore, in this paper, we make use of mutual information, and propose a DRF-based method for predicting residue-residue interactions. Furthermore, we perform computational experiments, and the results suggest that the DRF-based method is useful compared with that using the corresponding MRF model.

2011 IEEE International Conference on Systems Biology (ISB)
978-1-4577-1666-9/11/$26.00 ©2011 IEEE

Zhuhai, China, September 2–4, 2011

II. METHOD

In this section, we propose a discriminative random field (DRF)-based method for predicting contact residues. The input data are two amino acid sequences. Then, homologous sequences are collected for each sequence, mutual information between two residues is calculated, and the probability that two residues interact with each other is calculated according to our proposed DRF model. For training parameters of the DRF model, several pairs of protein sequences and the interacting residues are given.

A. Mutual Information

In our proposed method, mutual information for the distribution of amino acids at two positions of protein sequence alignments is one of important inputs. In this section, we briefly review mutual information for such distributions.

There are several types of residue-residue interactions, interactions between proteins having the same amino acid sequence, homodimers, and interactions between different proteins, heterodimers. Fig. 1 shows an illustration on calculation of mutual information between two positions in multiple sequence alignments. Suppose that protein sequence A and the information of interactions of residues in the homodimer are obtained. Then, several homologous sequences for sequence A are collected, and a multiple alignment is calculated in some adequate way. After that, gaps added to sequence A by the calculation of the alignment are deleted because only residues in sequence A are the target of our prediction of interactions. The length of the multiple alignment becomes the length of sequence A. The left figure in Fig. 1 shows such a multiple alignment, where the sequence at the first line denotes sequence A. Let $\mathcal{A}$ be the set of 20 amino acids and 1 character that represents undetermined amino acids. Let $p_i(a), p_{ij}(a,b)$ be the observed frequency of amino acid $a \in \mathcal{A}$ at position i and that of amino acids $a, b \in \mathcal{A}$ at positions i and j, respectively, where the frequency is divided by the total number. Then, mutual information m_{ij} between two positions i and j is calculated as follows.

$$m_{ij} = H_i + H_j - H_{ij}, \quad (1)$$

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

In a similar way, mutual information m_{ij} for a heterodimer is calculated as well as Eq. (1), where the joint frequency $p_{ij}(a,b)$ between positions i and j for amino acids a and b is calculated after each sequence in a multiple alignment is assigned to a sequence in another alignment according to organisms that the two sequences belong to (see Fig. 1).

Fig. 2 shows an example of the matrix of mutual information between two sequences, where in fact, both sequences A and B are Pfam domains of PF05269 with length of 91, which is contained in regulatory protein CII, and we used the homodimer in computational experiments. There are 194 pairs of interacting residues among total 4,095 residue pairs between the domains, and 25 homologous protein sequences in Pfam database were used for the calculation of mutual information between residues of PF05269. The matrix can be considered as an image. Therefore, we make use of an image processing technique, discriminative random fields, for prediction of interacting residues.

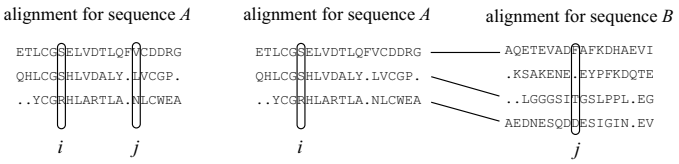

Fig. 1. Illustration on calculation of mutual information between two positions in multiple alignments for sequences A and B. Left) mutual information between positions i and j in a multiple alignment for sequence A. Right) mutual information between positions i in a multiple alignment for sequences A and j in a multiple alignment for sequence B, where sequences belonging to the same organism are connected. Sequences A and B are shown at the first line of alignments, and respectively.

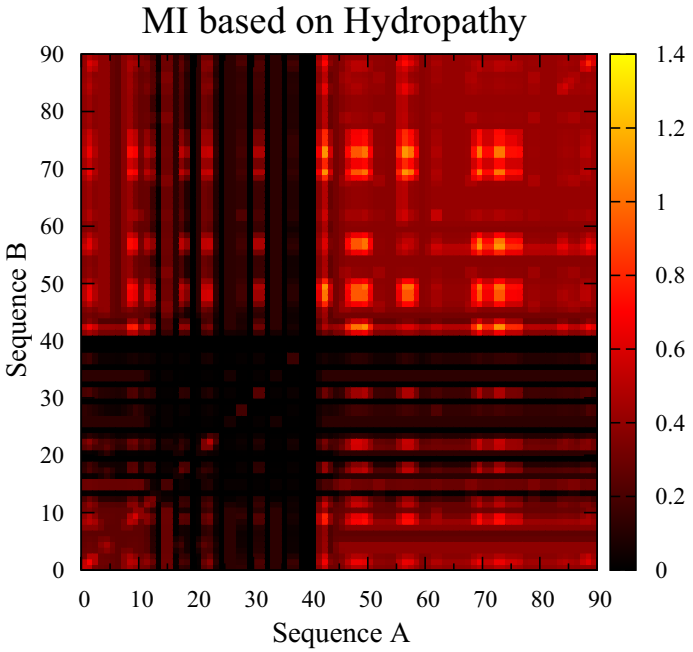

Fig. 2. Example of matrix of mutual information between two residues, where both sequences A and B are Pfam domains of PF05269 with length of 91, and hydropathy classification of amino acids was used for the calculation of MI. The brighter the color of (i,j) is, the higher their mutual information is.

B. Discriminative Random Field Models for Residue-Residue Interactions

In this section, we describe the discriminative random field (DRF) proposed by Kumar and Hebert [14], and propose DRF models for residue-residue interactions.

The discriminative random field was developed based on the conditional random field (CRF) proposed by Lafferty et al. [15]. 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_s, and y_s is observed from the corresponding

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

vertex. Then, $(\boldsymbol{x}, \boldsymbol{y})$ is a conditional random field if the random variables x_s follow the Markov property under the conditions y_s according to the graph G, that is, $P(x_s|\boldsymbol{x}_{\{t\in V|t\neq s\}}, \boldsymbol{y}) = P(x_s|\boldsymbol{x}_{\mathcal{N}_s}, \boldsymbol{y})$, where $\mathcal{N}_s$ denotes the set of vertices adjacent to the vertex s in the graph G. As well as CRFs, DRFs require $P(\boldsymbol{x}|\boldsymbol{y}) > 0$ for all $\boldsymbol{x}$, and are represented by the following formula

$$P(x_s|\boldsymbol{x}_{\mathcal{N}_s}, \boldsymbol{y}) = \frac{1}{Z_s} \exp\left\{-U_s(\boldsymbol{x}, \boldsymbol{y})\right\}, \tag{2}$$

where $U_s(\boldsymbol{x}, \boldsymbol{y})$ is a potential function concerning the vertex s, and Z_s is the normalization constant defined by $\sum_{x_s} \exp\left\{-U_s(\boldsymbol{x}, \boldsymbol{y})\right\}$. In the framework of DRFs, it is assumed that only up to pairwise clique potentials are nonzero, and the potential function is defined as follows.

$$U_s(\boldsymbol{x}, \boldsymbol{y}) \;=\; \alpha A(x_s, \boldsymbol{y}) + \beta \sum_{t\in\mathcal{N}_s} I(x_s, x_t, \boldsymbol{y}), \tag{3}$$

where $A(x_s, \boldsymbol{y})$ and $I(x_s, x_t, \boldsymbol{y})$ are the unary and binary potential functions, and called the association potential and the interaction potential, respectively, each random variable x_s takes 1 or -1, $\alpha \in \{0, 1\}$, and β is a variable. Let $\boldsymbol{w}$ and $\boldsymbol{v}$ be parameter vectors, and $\boldsymbol{f}_s$ and $\boldsymbol{g}_{st}$ be vector-valued functions that map observations $\boldsymbol{y}$ to feature vectors with the same size as parameter vectors. Then, the association potential $A(x_s, \boldsymbol{y})$ can be considered as a gain obtained only from the vertex s and the observations $\boldsymbol{y}$, and is defined as

$$A(x_s, \boldsymbol{y}) \;=\; -\log\left(\sigma\left(x_s \boldsymbol{w}^T \boldsymbol{f}_s(\boldsymbol{y})\right)\right), \tag{4}$$

where $\sigma(x)$ is the logistic function defined by $\frac{1}{1+e^{-x}}$, and $\boldsymbol{w}^T$ denotes the transpose of $\boldsymbol{w}$. It means that the DRF model includes generalized linear models (GLM), where other functions such as the probit function can be used as the link function of the DRF. On the other hand, the interaction potential $I(x_s, x_t, \boldsymbol{y})$ can be considered as a gain obtained from the relationship between vertices s and t, and is defined as

$$\begin{aligned} I_1(x_s, x_t, \boldsymbol{y}) \;&=\; K x_s x_t \\ &\quad + (1-K)\left(2\sigma\left(x_s x_t \boldsymbol{v}^T \boldsymbol{g}_{st}(\boldsymbol{y})\right) - 1\right), \end{aligned} \tag{5}$$

where $0 \leq K \leq 1$, or simply defined as

$$I_2(x_s, x_t, \boldsymbol{y}) \;=\; x_s x_t \boldsymbol{v}^T \boldsymbol{g}_{st}(\boldsymbol{y}). \tag{6}$$

It should be noted that the set of parameters θ in DRF models consists of $\boldsymbol{w}, \boldsymbol{v}, \beta$, and K.

In order to determine a DRF model, we must design vector-valued functions $\boldsymbol{f}_s$ and $\boldsymbol{g}_{st}$. Kumar and Hebert used histograms of luminance values ($\boldsymbol{y}$) in neighbor pixels at some scales for recognition of man-made structures in an image [14]. For our purpose, we use random variables $r_{ij} (\in \{1, -1\})$ that represent residue contacts instead of x_s, where $r_{ij} = 1$ means residues between position i and j interact with each other, otherwise $r_{ij} = -1$. Here, the set of vertices V consists of pairs of positions (i, j), and 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. 3). Furthermore, we use mutual information m_{ij} between

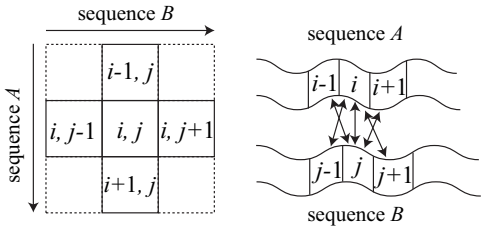

Fig. 3. Adjacent residue pairs for (i, j).

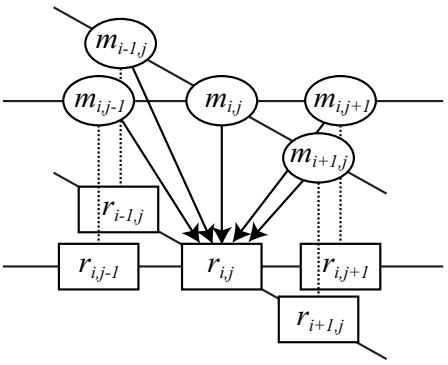

Fig. 4. Relationship between mutual information m_{ij} and random variable r_{ij} in the DRF framework.

positions i and j as observations $\boldsymbol{y}$. Then, we define vector-valued functions $\boldsymbol{f}_{ij}$ and $\boldsymbol{g}_{ij,kl}$ that map $\boldsymbol{m}$ to feature vectors as follows.

$$\begin{aligned} \boldsymbol{f}_{ij}(\boldsymbol{m}) \;&=\; \Big(1, m_{i,j}, \frac{1}{2}(m_{i,j-1} + m_{i,j+1}), \\ &\qquad \frac{1}{2}(m_{i-1,j} + m_{i+1,j})\Big)^T, \end{aligned} \tag{7}$$

$$\boldsymbol{g}^{(h)}_{ij,kl}(\boldsymbol{m}) \;=\; \begin{cases} 1 & (h = 1) \\ |\boldsymbol{f}^{(h)}_{ij} - \boldsymbol{f}^{(h)}_{kl}| & (h = 2, 3, 4) \end{cases}, \tag{8}$$

where $\boldsymbol{g}^{(h)}$ denotes the h-th element of vector $\boldsymbol{g}$, and $|x|$ denotes the absolute value of x. The relationship between mutual information m_{ij} and random variable r_{ij} is represented in the DRF framework as Fig. 4, that is, r_{ij} is related with multiple observations m_{ij}. It should be noted that adjacent vertices used in feature vectors are allowed to be different from $\mathcal{N}_{ij}$. Therefore, we also consider the following feature vector,

$$\begin{aligned} \boldsymbol{f}'_{ij}(\boldsymbol{m}) \;&=\; \Big(1, m_{i,j}, \frac{1}{2}(m_{i,j-1} + m_{i,j+1}), \\ &\quad \frac{1}{2}(m_{i-1,j} + m_{i+1,j}), \frac{1}{2}(m_{i-1,j-1} + m_{i+1,j+1}), \\ &\quad \frac{1}{2}(m_{i-1,j+1} + m_{i+1,j-1})\Big)^T. \end{aligned} \tag{9}$$

On the other hand, in the MRF framework, r_{ij} is related with only an observation m_{ij}. We define the following feature vector for comparison of random fields.

$$\boldsymbol{f}^0_{ij}(\boldsymbol{m}) \;=\; \Big(1, m_{i,j}\Big)^T \tag{10}$$

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

C. Parameter Estimation

We estimate parameters $\theta = \{\boldsymbol{w}, \boldsymbol{v}, \beta, K\}$ by maximizing pseudo-likelihood function as in [14]. Suppose that N pairs of multiple alignments for protein sequences and interacting residues $\boldsymbol{r}^{(n)}(n = 1, \ldots, N)$ for each pair of proteins are given. We calculate mutual information $\boldsymbol{m}^{(n)}$ for each pair. Then, the logarithm of pseudo-likelihood function is given as

$$L(\theta) = \log \prod_{n=1}^{N} \prod_{i} \prod_{j} P(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)}, \boldsymbol{m}^{(n)}, \theta) \quad (11)$$

$$= \sum_{n=1}^{N} \sum_{i} \sum_{j} \Big\{ -U_{ij}(\boldsymbol{r}^{(n)}) - \log \sum_{r_{ij}^{(n)} \in \{1,-1\}} \exp\Big\{ -U_{ij}(\boldsymbol{r}^{(n)}) \Big\} \Big\} \quad (12)$$

In order to maximize $L(\theta)$, we use the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [16] method, which is one of quasi-Newton methods that uses partial differentials and approximates the Hessian matrix by some efficient method. For that purpose, by partially differentiating $L(\theta)$ with respect to each parameter, we have

$$\frac{\partial L(\theta)}{\partial \boldsymbol{w}} = \sum_{n} \sum_{i} \sum_{j} \Big\{ -\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \boldsymbol{w}} + \sum_{r_{ij}^{(n)}} P(r_{ij}^{(n)} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(n)}, \boldsymbol{m}^{(n)}, \theta) \frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \boldsymbol{w}} \Big\}, \quad (13)$$

where

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \boldsymbol{w}} = -\alpha \sigma\Big(-r_{ij}^{(n)} \boldsymbol{w}^T \boldsymbol{f}_{ij} \Big) r_{ij}^{(n)} \boldsymbol{f}_{ij}. \quad (14)$$

In a similar way, for β, we have

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \beta} = \sum_{(k,l) \in \mathcal{N}_{ij}} I(r_{ij}^{(n)}, r_{kl}^{(n)}, \boldsymbol{m}). \quad (15)$$

If I_1 is used as the interaction potential, we have

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \boldsymbol{v}} = 2\beta(1-K) \sum_{(k,l) \in \mathcal{N}_{ij}} \sigma\Big(-r_{ij}^{(n)} r_{kl}^{(n)} \boldsymbol{v}^T \boldsymbol{g}_{ij,kl} \Big) r_{ij}^{(n)} r_{kl}^{(n)} \boldsymbol{g}_{ij,kl}, \quad (16)$$

$$\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial K} = \beta \sum_{(k,l) \in \mathcal{N}_{ij}} \Big(r_{ij}^{(n)} r_{kl}^{(n)} - 2\sigma\Big(r_{ij}^{(n)} r_{kl}^{(n)} \boldsymbol{v}^T \boldsymbol{g}_{ij,kl} \Big) + 1 \Big). \quad (17)$$

If I_2 is used, we have $\frac{\partial U_{ij}(\boldsymbol{r}^{(n)})}{\partial \boldsymbol{v}} = \beta \sum_{(k,l) \in \mathcal{N}_{ij}} r_{ij}^{(n)} r_{kl}^{(n)} \boldsymbol{g}_{ij,kl}$.

D. Contact Decision

After estimating parameters, for new pairs of residues, we decide whether or not each pair interacts with each other. For that purpose, we use Iterated Conditional Modes (ICM) [17], which iteratively updates random variables $r_{ij} \in \{1, -1\}$ until each variable cannot be changed using the following.

$$r_{ij}^{(t+1)} = \text{argmax}_{r_{ij} \in \{1,-1\}} P(r_{ij} | \boldsymbol{r}_{\mathcal{N}_{ij}}^{(t)}, \boldsymbol{m}, \theta), \quad (18)$$

where $r_{ij}^{(t)}$ denotes the value of random variable r_{ij} at step t.

TABLE I
CLASSIFICATION OF AMINO ACIDS BASED ON HYDROPATHY AND CHEMICAL STRUCTURE.

Hydropathy	Amino acid
hydrophobic	G,A,P,V,L,I,M,W,E
hydrophilic	R,N,D,E,Q,H,K,S,T,C,Y
Chemical structure	**Amino acid**
only hydrogen atom	G
hydroxyl group	S,T
sulfur atom	C,M
aliphatic hydrocarbon	A,V,L,I,P
carboxylic structure	D,E
amidated carboxyl group	N,Q
nitrogen atom	K,R,H
aromatic ring	F,Y,W

TABLE II
DETAILS OF 12 INTERACTING DOMAIN PAIRS FOR EVALUATION.

	sequence A			sequence B		
PDB	acc	Pfam	#	acc	Pfam	#
1ylf	Q81EX1	PF02082	125	Q81EX1	PF02082	125
1mkm	Q9WXS0	PF09339	88	Q9WXS0	PF09339	88
1zpq	P03042	PF05269	91	P03042	PF05269	91
1r1v	O85142	PF01022	47	O85142	PF01022	47
1l3l	P33905	PF00196	58	P33905	PF00196	58
1rio	P03034	PF01381	56	Q9EZJ8	PF04545	54
1zzb	Q56185	PF01381	56	Q56185	PF01381	56
1z7u	Q838C3	PF01638	91	Q838C3	PF01638	91
1s7o	P67253	PF04297	101	P67253	PF04297	101
1hw2	P0A8V6	PF00392	64	P0A8V6	PF00392	64
1xcb	Q9X2V5	PF06971	50	Q9X2V5	PF06971	50
1b4a	O31408	PF01316	70	O31408	PF01316	70

'acc' denotes the accession number of the protein. '#' denotes the number of residues.

III. COMPUTATIONAL EXPERIMENTS

A. Data and Implementation

To get protein residue interaction data, we used the files, 'int_pfamA.txt' and 'interaction.txt', from Pfam database (version 21.0) [18]. The former includes 6,079 interacting domain pairs, and the latter includes information of interacting residue pairs between domains. There were 26 interacting domain pairs that both domains belong to CL0123 group, which is called helix-turn-helix clan and contains a diverse range of mostly DNA-binding domains including a helix-turn-helix motif, and we selected 12 interacting domain pairs at random from the pairs, (PF02082, PF02082), (PF09339, PF09339), (PF05269, PF05269), (PF01022, PF01022), (PF00196, PF00196), (PF01381, PF04545), (PF01381, PF01381), (PF01638, PF01638), (PF04297, PF04297), (PF00392, PF00392), (PF06971, PF06971), and (PF01316, PF01316), where we excluded pairs that contain less than 2 interacting residues and contain less than 5 sequences for multiple alignments. Table II shows the details of the datasets. For each pair of domains, PDB ID [19] is shown,

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

TABLE III
RESULTS ON AVERAGE AUC SCORES FOR TRAINING AND TEST DATASETS USING MUTUAL INFORMATION FOR MRF MODEL WITH FEATURE VECTOR $\boldsymbol{f}^0_{ij}$, DRF MODEL WITH $\boldsymbol{f}_{ij}$, AND DRF MODEL WITH $\boldsymbol{f}'_{ij}$.

Alphabet	MRF ($\boldsymbol{f}^0_{ij}$)	DRF ($\boldsymbol{f}_{ij}$)	DRF ($\boldsymbol{f}'_{ij}$)
for training dataset			
20 amino acids	0.671846	0.690184	0.704212
hydropathy	0.684646	0.72279	0.725438
chemical structure	0.668725	0.6983	0.720599
for test dataset			
20 amino acids	0.629492	0.643685	0.621458
hydropathy	0.630355	0.637544	0.630765
chemical structure	0.62826	0.642325	0.624627

TABLE IV
RESULTS ON AVERAGE AUC SCORES FOR TRAINING AND TEST DATASETS USING ZRES FOR MRF MODEL WITH FEATURE VECTOR $\boldsymbol{f}^0_{ij}$, DRF MODEL WITH $\boldsymbol{f}_{ij}$, AND DRF MODEL WITH $\boldsymbol{f}'_{ij}$.

Alphabet	MRF ($\boldsymbol{f}^0_{ij}$)	DRF ($\boldsymbol{f}_{ij}$)	DRF ($\boldsymbol{f}'_{ij}$)
for training dataset			
20 amino acids	0.570207	0.604435	0.624169
hydropathy	0.480385	0.546194	0.543357
chemical structure	0.550335	0.620711	0.637602
for test dataset			
20 amino acids	0.546399	0.572829	0.594989
hydropathy	0.503718	0.486512	0.501379
chemical structure	0.501243	0.569000	0.569988

and for each domain included in the PDB ID, the accession number of the protein containing the domain, Pfam ID, and the number of residues are shown. Since each sequence included from 47 to 125 residues and the number of residue pairs was more than 47×47=2,209, it is considered to be enough for estimating parameters. However, the number of interacting residues (positive examples) is too few in a pair of domains compared with that of non-interacting residues (negative examples). Therefore, we selected uniformly at random the same number of negative examples as that of positive examples.

For the calculation of mutual information between residues, we used multiple alignment data provided in the file 'Pfam-A.full' in Pfam database. For the calculation of marginal entropies and joint entropies, we used three types of classification of amino acids. One is not classified, that is, each group has a distinct amino acid, and the number of groups is 20. Another is hydropathy-based classification. It classifies 20 amino acids into 2 groups, hydrophobic (G, A, P, V, L, I, M, W and E) and hydrophilic amino acids (R, N, D, E, Q, H, K, S, T, C and Y). The other is classification by chemical structures of amino acids, which has eight groups. Table I shows the details for hydropathy-based and chemical structure-based classification.

Furthermore, we calculated ZRes [6] from mutual information, and we also used $ZRes_{i,j} = Z_i(j)Z_j(i)$ instead of $m_{i,j}$ for the feature vectors, where $Z_i(j)$ denotes the z-score for $Res_{i,j}$ to $Res_{i,*}$, and $Res_{i,j}$ is obtained by taking the residual of $m_{i,j}$ from $\overline{m}_{i,*}\overline{m}_{*,j}$ after the linear least squares regression.

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 [20], and is available on the web page, http://www.chokkan.org/software/liblbfgs/.

B. Results

In order to evaluate the proposed DRF-based method, we performed computational experiments using three types of vector-valued functions $\boldsymbol{f}^0_{ij}$, $\boldsymbol{f}_{ij}$, and $\boldsymbol{f}'_{ij}$, and three types of classification of amino acids, 20 amino acids, hydropathy-based, and chemical structure-based classification. We performed leave-one-out cross validation, where one dataset was used for test and the remaining datasets were for training, this process was repeated, and the number of repeated times was the number of datasets, that is, 12. 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.

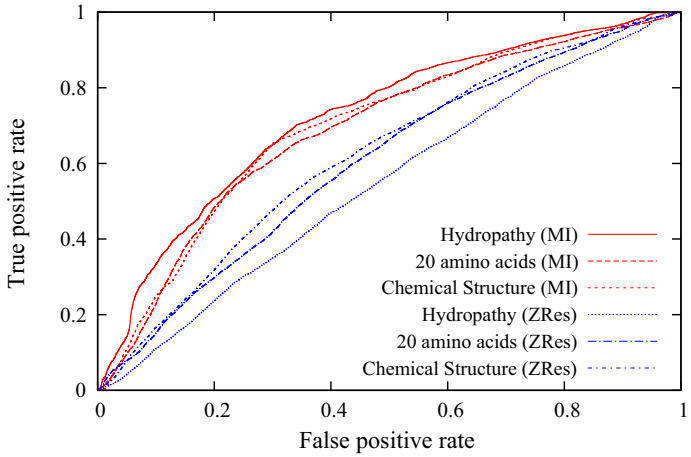

Fig. 5. Average ROC curves for training datasets using mutual information and ZRes for DRF model with feature vector $\boldsymbol{f}_{ij}$.

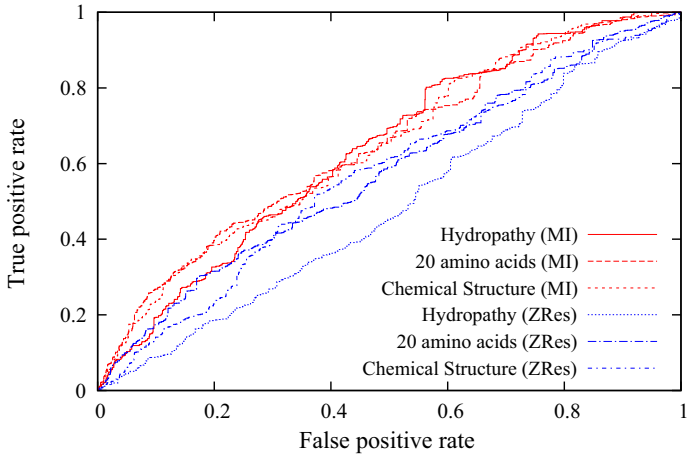

Fig. 6. Average ROC curves for test datasets using mutual information and ZRes for DRF model with feature vector $\boldsymbol{f}_{ij}$.

First, we set $\alpha = 1$ and $\beta = 0$. It means that DRF models contained only the association potential $A(r_{ij}, \boldsymbol{m})$. Tables III and IV show the results on the average AUC scores for training and test datasets using mutual information and ZRes for the

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

MRF model with feature vector $\boldsymbol{f}_{ij}^0$, the DRF model with $\boldsymbol{f}_{ij}$, and the DRF model with $\boldsymbol{f}'_{ij}$, respectively. It should be noted that only a small fraction of training datasets was used for the parameter estimation of the models. We can see from these tables that the average AUC scores using mutual information for the feature vectors were better than those using ZRes. Furthermore, the average AUC scores of the MRF model were smaller than almost all those of the DRF models for training and test dataset. It is considered because the MRF model can use only an observation m_{ij} although DRFs can use multiple observations. The average AUC scores of the DRF model with $\boldsymbol{f}'_{ij}$ were better than those of the DRF model with $\boldsymbol{f}_{ij}$ for training set, while for test set those of the DRF model with $\boldsymbol{f}_{ij}$ were better than those of the DRF model with $\boldsymbol{f}'_{ij}$. For training datasets, the average AUC scores using mutual information for the DRF models by the hydropathy-based classification was better than those by others. For test datasets, the average AUC score using mutual information for the DRF model with $\boldsymbol{f}_{ij}$ by 20 amino acids was better than those by others. The average ROC (Receiver Operating Characteristic) curves for training and test datasets using mutual information and ZRes for the DRF model with $\boldsymbol{f}_{ij}$ are shown in Figures 5 and 6. The average computation times of parameter estimation using mutual information and ZRes were about 0.49 and 0.47 seconds, respectively. Although we added a heterodimer to the datasets that consists of 11 homodimers and a heterodimer, and performed similar experiments, the average AUC score became smaller. It may suggest that the parameters of our DRF models should be estimated for homodimers and heterodimers independently.

Next, we set $\alpha = 0$ and $\beta = 1$. It means that DRF models contained only the interaction potential $I(r_{ij}, r_{kl}, \boldsymbol{m})$. However, the BFGS method for parameter estimation did not converge for the potentials I_1 and I_2. It can be considered because colors of neighbor pixels are often similar to each other, and the interaction potentials in DRFs were originally developed for smoothing images. However, pairs of neighbor residues are not always similar, that is, even if residues at positions (i, j) interact, it might be difficult to determine whether or not residues at $(k, l) \in \mathcal{N}_{ij}$ interact. On the other hand, it is considered from the results that the association potential in DRFs is useful for predicting interacting residues, and mutual information between neighbor residues is useful.

IV. Conclusion

We proposed a method for predicting protein residue contacts using the discriminative random field proposed by Kumar and Hebert, which is a special type of conditional random fields and is able to recognize characteristic sub-images from an image. In order to make use of DRFs, mutual information between residues was given as observations in the potential of DRFs, where mutual information was calculated from multiple sequence alignments of homologous proteins. To validate the proposed DRF-based method with mutual information, we performed computational experiments using leave-one-out cross validation and calculated the average AUC scores. The results suggest that our proposed DRF-based method with mutual information is useful for prediction of protein residue contacts compared with that based on the corresponding Markov random field model. It means that mutual information between neighbor residues is useful for the contact prediction. On the other hand, interaction potentials were not useful because DRFs have been originally developed for image analyses. The problem of predicting residue contacts is one of difficult problems, and it cannot be said that the prediction accuracy by our method was good. However, there are some possibilities to improve our method, for instance, the modification of the observation and the potential function. We can use other observations than mutual information and ZRes [6] from distributions of amino acids, for instance, Direct Information (DI) [4], that are correlation values calculated in different ways. In addition, we can introduce some parameters with respect to each amino acid in the potential function that represent properties for each amino acid because the results imply that the number of parameters was not sufficient for explaining protein residue contacts.

Acknowledgment

This work was partially supported by Grants-in-Aid #22240009 and #21700323 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] 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.

[2] 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.

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

[4] 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.

[5] H. B. Fraser, A. E. Hirsh, L. M. Steinmetz, C. Scharfe, and M. W. Feldman, "Evolutionary rate in the protein interaction network," *Science*, vol. 296, pp. 750–752, 2002.

[6] D. Y. Little and L. Chen, "Identification of coevolving residues and co-evolution potentials emphasizing structure, bond formation and catalytic coordination in protein evolution," *PLoS One*, vol. 4, p. e4762, 2009.

[7] J. Cheng and P. Baldi, "Improved residue contact prediction using support vector machines and a large feature set," *BMC Bioinformatics*, vol. 8, p. 113, 2007.

[8] S. Z. Li, *Markov random field modeling in image analysis*, 3rd ed., S. Singh, Ed. Springer-Verlag London, 2009.

[9] H. Derin and H. Elliott, "Modeling and segmentation of noisy and textured images using gibbs random fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 9, no. 1, pp. 39–55, 1987.

[10] S. Z. Li and J. Lu, "Modeling Bayesian estimation for deformable contours," in *Proc. Seventh IEEE International Conference on Computer Vision*, 1999, pp. 991–996.

[11] 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.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE

[12] 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.
[13] 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.
[14] S. Kumar and M. Hebert, "Discriminative random fields," *International Journal of Computer Vision*, vol. 68, no. 2, pp. 179–201, 2006.
[15] 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.
[16] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.
[17] J. Besag, "On the statistical analysis of dirty pictures," *Journal of Royal Statistical Soc.*, vol. B-48, pp. 259–302, 1986.
[18] R. D. Finn, J. Mistry, J. Tate, P. Coggill, A. Heger, J. E. Pollington, O. L. Gavin, P. Gunasekaran, G. Ceric, K. Forslund, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, and A. Bateman, "The Pfam protein families database," *Nucleic Acids Research*, vol. 38, pp. D211–D222, 2010.
[19] 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.
[20] J. Nocedal, "Updating quasi-Newton matrices with limited storage," *Mathematics of Computation*, vol. 35, no. 151, pp. 773–782, 1980.

978-1-4577-1666-9/11/$26.00 ©2011 IEEE