The Fourth International Conference on Computational Systems Biology (ISB2010)
Suzhou, China, September 9–11, 2010

Conditional Random Field Approach to Prediction of Protein-Protein Interactions Using Mutual Information Between Domains

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

[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

Abstract Analysis of functions and interactions of proteins and domains is important for understanding cellular systems and biological networks. Many methods for predicting protein-protein interactions have been developed. It is known that mutual information between residues at interacting sites can be higher than that at non-interacting sites. It is based on the thought that amino acid residues at interacting sites have coevolved with those at the corresponding residues in the partner proteins. Several studies have shown that such mutual information is useful for identifying contact residues in interacting proteins. Therefore, we focus on the mutual information, and propose a novel method using conditional random fields combined with mutual information between residues. In the method, protein-protein interactions are modeled using domain-domain interactions. We perform computational experiments, and calculate AUC (Area Under the Curve) score. The results suggest that our proposed model with mutual information is useful.

Keywords Conditional Random Field; Mutual Information; Protein Domain; Protein-Protein Interaction

1 Introduction

Understanding of protein functions and protein-protein interactions is one of important topics in fields of molecular biology and bioinformatics. Recently, many researchers have focused on coevolution of amino acid residues of proteins to investigate interactions and contacts between residues [18, 2, 9, 17]. If residues at important sites for interactions between proteins are substituted in one protein, corresponding residues in interacting partner proteins are expected to be also substituted by selection pressure. Otherwise, such mutated proteins may lose the interactions. Fraser et al. confirmed that interacting proteins evolve at similar evolutionary rates by comparing putatively orthologous protein

*Email: morihiro@kuicr.kyoto-u.ac.jp

sequences between *S. cerevisiae* and *C. elegans* [8]. It means that substitutions for contact residues occur in both interacting proteins as long as the proteins keep interacting with each other. Therefore, mutual information (MI) between residues is useful for predicting protein-protein interactions for proteins of unknown function. MI is calculated from multiple sequence alignments for homologous protein sequences. Weigt et al. identified direct residue contacts between sensor kinase and response regulator proteins by message passing, which is an improvement of MI [17]. Burger and van Nimwegen used a dependence tree where a node corresponds to a position of amino acid sequences, and predicted interactions using a Bayesian network method [2].

On the other hand, Markov random field and conditional random field models have been well studied in fields of natural language processing [14, 15]. Also in bioinformatics, protein function prediction methods from protein-protein interaction network and other biological networks were developed using Markov random fields [6, 4]. On the other hand, several prediction methods have been developed based on domain-domain interactions. Deng et al. proposed a domain-based probabilistic model of protein-protein interactions, and developed EM (Expectation Maximization) method [5]. Based on this probabilistic model, LP (Linear Programming)-based methods were developed [10], and Chen et al. improved the accuracy of interaction strength prediction [3]. In this paper, we propose a prediction method based on domain-domain interactions and the mutual information using conditional random fields.

2 Mutual Information Between Domains

In order to investigate the relationship between two positions of proteins, MI for distributions of amino acids at the positions is used. Such distributions can be obtained from multiple alignments of protein sequences and domain sequences. In this section, we briefly review MI for distributions of amino acids, and explain MI between domains.

We assume that multiple sequence alignments for domains D_m and D_n are obtained, respectively (see Figure 1). Let $\mathscr{A}$ be a set of amino acids, $f_i(A)$ be the appearance frequency of amino acid A at position i in domains D_m and D_n, and $f_{ij}(A,B)$ be the joint appearance frequency of a pair of amino acid A at position i in D_m and B at position j in D_n, where each frequency is divided by the number of sequence pairs of multiple alignments, M such that $\sum_{A\in\mathscr{A}} f_i(A) = \sum_{A,B\in\mathscr{A}} f_{ij}(A,B) = 1$. However, we cannot see which sequence in the multiple alignment of domain D_m corresponds to a specified sequence in that of D_n. Therefore, we assume that sequences contained in the same organism can be paired. In the example of Figure 1, the second sequence of D_m is paired with the first one of D_n, the third one of D_m is paired with the second one of D_n, and so on. The first sequence of D_m is not counted into the appearance frequencies because it is not paired with any sequence of D_n although it may be paired with sequences of other domains than D_n.

Multiple alignments often include some gaps. Weigt et al. counted the frequencies of gaps as well as amino acids [17]. Therefore, we also consider gaps to be a kind of amino acids, that is, the number of distinct amino acids is $|\mathscr{A}| = 21$. Then, mutual information for position i in D_m and j in D_n is defined as the Kullback-Leibler divergence between the multiplication of appearance frequencies, $f_i(A)f_j(B)$, and the joint appearance frequen-

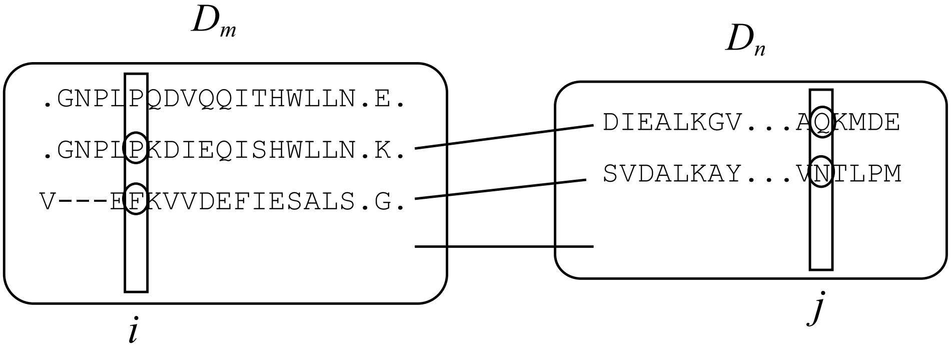

Figure 1: Illustration on the calculation of mutual information from multiple alignments of domains. Domains D_m and D_n have multiple alignments of sequences from several organisms, respectively. Mutual information is calculated for each pair of positions i and j.

cies, $f_{ij}(A,B)$, as follows.

$$MI_{ij} = \sum_{A,B\in\mathscr{A}} f_{ij}(A,B)\log\frac{f_{ij}(A,B)}{f_i(A)f_j(B)}. \tag{1}$$

If frequency distributions of amino acids at positions i and j are independent from each other, $f_{ij}(A,B) \approx f_i(A)f_j(B)$, and MI_{ij} approaches to zero. This means that the two positions are not related with each other in the evolutionary process. If domains D_m and D_n interact at the positions, it is considered that MI_{ij} becomes high because the positions have coevolved through the evolutionary process in order to keep the interaction. It should be noted that two positions i and j do not always directly interact even if MI_{ij} is high [17]. However, such proteins having high values of MI have a possibility to directly interact with each other at other positions in the proteins.

However, we need to reduce MI_{ij} because it can be unnecessarily high depending on distributions of $f_i(A)$ and $f_j(B)$. For that purpose, we make use of $MI_{ij}^{(random)}$, which is the mutual information MI_{ij} from the joint frequency, $f_{ij}(A,B)$, obtained by shuffling at random the combinations of sequences in multiple alignments. In this paper, we repeat the procedure 100 times, and take the average.

For practical uses of MI, $f_i(A)$, $f_j(B)$ and $f_{ij}(A,B)$ should be positive values. Otherwise, we cannot calculate MI_{ij} by using computers. Therefore, we use the following pseudocount as in [17],

$$f_i^{(pseudo)}(A) = \frac{\eta + f_i(A)M}{|\mathscr{A}|\eta + M} \tag{2}$$

$$f_{ij}^{(pseudo)}(A,B) = \frac{\eta/|\mathscr{A}| + f_{ij}(A,B)M}{|\mathscr{A}|\eta + M}, \tag{3}$$

where η is a constant value, in this paper we use $\eta = 1$. It should be noted that the sum

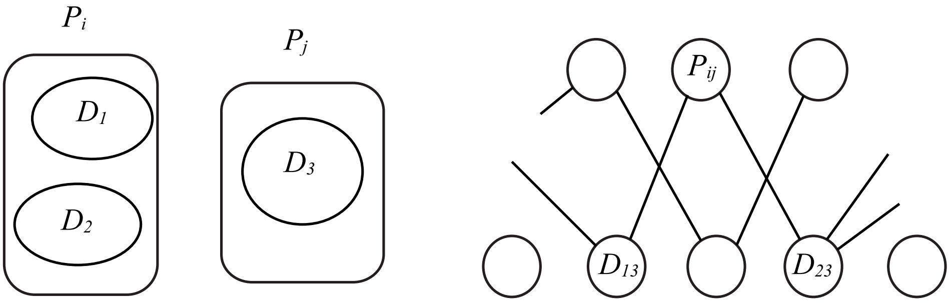

Figure 2: Markov random field model for protein-protein interactions. Left: Example of proteins P_i and P_j. P_i consists of domains D_1 and D_2, and P_j consists of domain D_3, respectively. Right: Factor graph $G(U,V,E)$. A vertex corresponds to $P_{ij} \in U$ or $D_{mn} \in V$, and there exists an edge between P_{ij} and D_{mn} if and only if $D_{mn} \in P_{ij}$.

over all amino acids $\mathscr{A}$, $\sum_{A \in \mathscr{A}} f_i^{(pseudo)}(A) = 1$ and $\sum_{A,B \in \mathscr{A}} f_{ij}^{(pseudo)}(A,B) = 1$ because $\sum_{A \in \mathscr{A}} f_i(A) = \sum_{A,B \in \mathscr{A}} f_{ij}(A,B) = 1$.

In order to investigate interactions between proteins, we need MI between domains included in the proteins. Thus, we define MI between domains D_m and D_n, m_{mn}, to be the maximum of MI over all positions i and j as follows.

$$m_{mn} = \max_{i,j}(MI_{ij} - \left\langle MI_{ij}^{(random)} \right\rangle), \tag{4}$$

where $\langle v \rangle$ means the average of v, i and j are positions of D_m and D_n, respectively.

3 Conditional Random Field Model for PPI

In this section, we propose a probabilistic model for protein-protein and domain-domain interactions using conditional random fields [14, 15] because it can be considered that two domains D_m and D_n do not always interact even if m_{mn} is large. For example, Weight et al. improved MI and proposed direct information (DI) because residues do not always contact with each other even if the MI is large [17].

Most proteins contain domains as is well known. If two proteins do not interact with each other, any two domains contained in the proteins must not interact with each other. In the left example of Figure 2, protein P_i consists of domains D_1 and D_2, and protein P_j consists of domain D_3, respectively. If P_i and P_j do not interact, any pair of (D_1, D_3) and (D_2, D_3) does not interact. Deng et al. proposed a probabilistic model for a pair of proteins as follows [5] by assuming that proteins P_i and P_j interact if and only if at least a pair of domains included in the proteins interacts, and events that domains interact are independent from each other:

$$Pr(P_{ij} = 1) = 1 - \prod_{D_{mn} \in P_{ij}} (1 - Pr(D_{mn} = 1)), \tag{5}$$

where $P_{ij} = 1$ means that proteins P_i and P_j interact, $D_{mn} = 1$ means that domains D_m and D_n interact, and $D_{mn} \in P_{ij}$ means that domain D_m is included in protein P_i and D_n is included in P_j, the product in the right hand side is calculated for all domain pair (D_m, D_n) included in the protein pair (P_i, P_j).

By transforming Eq. 5, we have

$$
\begin{aligned}
1 - Pr(P_{ij} = 1) &= \prod_{D_{mn} \in P_{ij}} (1 - Pr(D_{mn} = 1)) &(6)\\
&= \exp\left(\sum_{D_{mn} \in P_{ij}} \lambda^{(mn)} \right), &(7)
\end{aligned}
$$

where $\lambda^{(mn)} = \log(1 - Pr(D_{mn} = 1))$.

From this equation, we can consider the following Markov random field model for protein pair (P_i, P_j) (see Figure 2).

$$
Pr(P_{ij} = p_{ij}, \boldsymbol{d}) = \frac{1}{Z_{ij}} \exp\left(\sum_{D_{mn} \in P_{ij}} \sum_{t,u \in \{0,1\}} \lambda_{t,u}^{(ij,mn)} f_{t,u}^{(ij,mn)}(p_{ij}, d_{mn}) \right), \quad (8)
$$

where $p_{ij} \in \{0,1\}$, $\boldsymbol{d}$ means a set of events on domain-domain interactions, $D_{mn} = d_{mn}$ ($d_{mn} \in \{0,1\}$), $f_{t,u}^{(ij,mn)}(p_{ij}, d_{mn})$ denotes a local feature, $\lambda_{t,u}^{(ij,mn)}$ is the corresponding weight parameter, and Z_{ij} denotes the normalization constant. For instance, Eq. 8 for $p_{ij} = 0$ is equivalent to Eq. 7 in the case that $f_{t,u}^{(ij,mn)}(p_{ij}, d_{mn}) = 1$ if $t = p_{ij}$ and $d_{mn} = 0$, otherwise 0.

In Markov random fields, random variables have Markov properties represented as an undirected graph [11]. The factor graph for our model is represented to be a bipartite graph $G(U, V, E)$ with a set of vertices U corresponding to protein-protein interactions P_{ij}, a set of vertices V corresponding to domain-domain interactions D_{mn}, and a set of edges E between a vertex in U and one in V as the right figure of Figure 2. There exists an edge between $P_{ij} \in U$ and $D_{mn} \in V$ if and only if $D_{mn} \in P_{ij}$. For the left example of Figure 2, protein pair (P_i, P_j) includes domain pairs (D_1, D_3) and (D_2, D_3). Then, in the factor graph, the vertex of P_{ij} is connected with vertices of D_{13} and D_{23}, respectively. Although the vertex of P_{ij} does not have other adjacent vertices than the vertices of D_{13} and D_{23}, those of D_{13} and D_{23} can be connected with other vertices than that of P_{ij}.

In order to simplify the model, we substitute $\lambda_{t,u}^{(ij,mn)} = \lambda_{t,u}^{(mn)}$ and $f_{t,u}^{(ij,mn)} = f_{t,u}^{(mn)}$ for all protein pairs P_{ij}. Then, we have the following joint probability,

$$
Pr(\boldsymbol{p}, \boldsymbol{d}) = \frac{1}{Z} \exp\left(\sum_{P_{ij}} \sum_{D_{mn} \in P_{ij}} \sum_{t,u \in \{0,1\}} \lambda_{t,u}^{(mn)} f_{t,u}^{(mn)}(p_{ij}, d_{mn}) \right), \quad (9)
$$

where $\boldsymbol{p}$ means a set of events on protein-protein interactions, $P_{ij} = p_{ij}$.

We here introduce mutual information between domains as given conditional data in order to combine it with the probabilistic model. Then, Eq. 8 can be written as

$$
Pr(p_{ij} | \boldsymbol{m}) = \frac{1}{Z_{ij}(\boldsymbol{m})} \exp\left(\sum_{D_{mn} \in P_{ij}} \sum_{t \in \{0,1\}} \lambda_t^{(mn)} f_t^{(mn)}(p_{ij}, m_{mn}) \right), \quad (10)
$$

where

$$Z_{ij}(\boldsymbol{m}) \quad = \quad \sum_{p_{ij}\in\{0,1\}} \exp\left(\sum_{D_{mn}\in P_{ij}} \sum_{t\in\{0,1\}} \lambda_t^{(mn)} f_t^{(mn)}(p_{ij}, m_{mn})\right), \tag{11}$$

$$f_t^{(mn)}(p_{ij}, m_{mn}) = \begin{cases} m_{mn} & (\text{if } p_{ij} = 1 \text{ and } t = 1) \\ 1/m_{mn} & (\text{if } p_{ij} = 0 \text{ and } t = 0) \\ 0 & (\text{if } p_{ij} = 1 \text{ and } t = 0) \\ -1 & (\text{if } p_{ij} = 0 \text{ and } t = 1) \end{cases}. \tag{12}$$

It should be noted that a negative value, -1, is given to $f_t^{(mn)}$ because it is undesired that a pair of domains interact although proteins having the pair do not interact. In this way, the local feature $f_t^{(mn)}$ correlates protein-protein interactions P_{ij} with domain-domain interactions D_{mn} (see Figure 2).

3.1 Parameter Estimation

In this section, we discuss how to estimate the parameters $\boldsymbol{\lambda} = \{\lambda_t^{(mn)}\}$. We assume that protein-protein interaction data $\boldsymbol{p} = \{p_{ij}\}$ are given. Then, the likelihood function is represented by

$$P(\boldsymbol{p}|\boldsymbol{m}) \;=\; \prod_{P_{ij}} P(p_{ij}|\boldsymbol{m}) \;=\; \frac{1}{Z(\boldsymbol{m})} \exp\left(\sum_{P_{ij}} \sum_{D_{mn}\in P_{ij}} \sum_{t\in\{0,1\}} \lambda_t^{(mn)} f_t^{(mn)}(p_{ij}, m_{mn})\right), \tag{13}$$

where $Z(\boldsymbol{m}) = \prod_{P_{ij}} Z_{ij}(\boldsymbol{m})$. By taking the logarithm, we have

$$l(\boldsymbol{\lambda}) = \log P(\boldsymbol{p}|\boldsymbol{m}) = \sum_{P_{ij}} \left(\sum_{D_{mn}\in P_{ij}} \sum_{t\in\{0,1\}} \lambda_t^{(mn)} f_t^{(mn)}(p_{ij}, m_{mn}) - \log Z_{ij}(\boldsymbol{m})\right). \tag{14}$$

We estimate the parameters by maximizing the log-likelihood function, $l(\boldsymbol{\lambda})$. Since $\log(e^x + e^y)$ is a convex function for variables x and y, that is, $l(\boldsymbol{\lambda})$ is a concave function, we are able to obtain a global maximum. For maximizing such functions, various methods such as the steepest descent method, Newton's method, and the Broyden-Fletcher-Goldfarb-Shanno (BFGS) [1] method have been developed. Newton's method calculates the inverse of the Hessian matrix for the objective function. However, the computational cost is high. Therefore, the quasi-Newton method approximates the matrix by some efficient method using the first derivatives, the gradient. In this paper, we use the BFGS method, which is one of the quasi-Newton methods.

By differentiating Eq. 14 partially with respect to each parameter $\lambda_t^{(mn)}$, we have

$$\frac{\partial l(\boldsymbol{\lambda})}{\partial \lambda_t^{(mn)}} = \sum_{P_{ij}:D_{mn}\in P_{ij}} \left(f_t^{(mn)}(p_{ij}, m_{mn}) - \sum_{p_{ij}\in\{0,1\}} P(p_{ij}|\boldsymbol{m}) f_t^{(mn)}(p_{ij}, m_{mn})\right). \tag{15}$$

Table 1: Result on AUC for training and test datasets using CRF model with MI and that without MI.

iteration	with MI		without MI	
	training	test	training	test
1st	0.999384	0.731399	0.999439	0.742315
2nd	0.999584	0.727377	0.999511	0.731686
3rd	0.998895	0.706693	0.998823	0.703821
4th	0.999511	0.672795	0.999638	0.665613
5th	0.999497	0.763377	0.999533	0.736623
average	0.999374	0.720328	0.999389	0.716012

4 Computational Experiments

4.1 Data and Implementation

We used human protein-protein interaction data from the DIP database [13], the file name is 'dip20091230.txt'. We used the UniProt Knowledgebase database (version 15.4) [16] as protein domain inclusion data. We deleted proteins that do not have any domain, and obtained 294 interacting protein pairs as positive data that include 300 distinct proteins and 320 domains. We used the Pfam database (version 24.0) [7] to obtain multiple sequence alignments for domains, and calculated MI, m_{mn}, for each pair of domains. We selected 294 non-interacting protein pairs uniformly at random as negative data such that negative data do not overlap the positive data.

We used libLBFGS (version 1.9) with default parameters to estimate the parameters $\lambda_t^{(mn)}$, which is a C implementation of the limited memory BFGS method [12], and is available on the web page, http://www.chokkan.org/software/liblbfgs/.

4.2 Result

In order to evaluate our method, we compared two models, the proposed CRF model with MI, and that without MI (i.e., $m_{mn} = 1$ is given for each D_{mn}). We performed five-fold cross-validation, that is, split the data into 5 datasets (4 for training and 1 for test), estimated $\lambda_t^{(mn)}$ from the training datasets, and calculated $Pr(P_{ij} = 1|\boldsymbol{m})$ of Eq. 10 for each protein pair in the test dataset and AUC (Area Under the Curve) score. We repeated 5 times, and took the average. Table 1 shows the results on AUC for training and test datasets using the CRF model with MI and that without MI. We can see from this table that the results using the model with MI are better than those without MI. Figure 3 shows the average ROC (Receiver Operating Characteristic) curves for training and test datasets using CRF model with MI and that without MI. For training datasets, the results using both CRF models were almost perfect. For test datasets, CRF model with MI outperformed that without MI in low false positive rates. Since getting higher true positive rates in low false positive rate regions is practically important, CRF model with MI is considered to be better than that without MI.

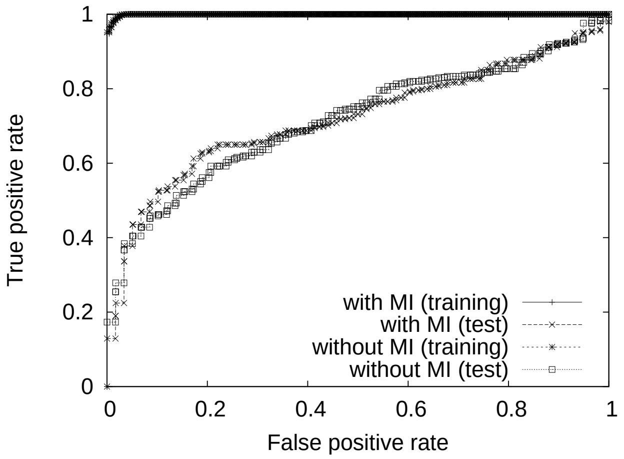

Figure 3: Average ROC curves for training and test datasets using CRF model with MI and that without MI.

5 Conclusion

We proposed a novel method which combines conditional random fields with the domain based model of protein-protein interactions. Furthermore, we improved the method by introducing mutual information. In the improved CRF model, mutual information between domains is given as conditions, where MI between domains is defined as the maximum of MIs between residues in the domains. It is considered that amino acid residues at important sites for interactions have coevolved with each other, and MI has been used for identifying contact residues in interactions.

We performed five-fold cross-validation experiments, and calculated AUC for probabilities that two proteins interact. The results suggested that our proposed model with mutual information is useful. However, the results of AUC for training datasets implied that estimated parameters were overfitting to training datasets. For avoiding that problem, we can improve the model, for instance, by adding regularization terms, l_1-norm of parameters to the log-likelihood function. Since CRF has an advantage to be able to incorporate large number of features, it remains as a future work to modify the model itself to obtain better accuracy for instance, modifying the local feature and adding new features. Comparison of our method with other existing methods and evaluation of our method by using more datasets are left as other future work.

Acknowledgements

This work was partially supported by Grants-in-Aid #22240009 and #21700323 from MEXT, Japan.

References

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

[2] L. Burger and E. van Nimwegen. Accurate prediction of protein-protein interactions from sequence alignments using a bayesian method. *Molecular Systems Biology*, 4:165, 2008.

[3] L. Chen, L. Y. Wu, Y. Wang, and X. S. Zhang. Inferring protein interactions from experimental data by association probabilistic method. *Proteins*, 62(4):833–837, 2006.

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

[5] M. Deng, S. Mehta, F. Sun, and T. Chen. Inferring domain-domain interactions from protein-protein interactions. *Genome Research*, 12:1540–1548, 2002.

[6] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein interaction data. *Journal of Computational Biology*, 10(6):947–960, 2003.

[7] 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*, 38:D211–D222, 2010.

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

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

[10] M. Hayashida, N. Ueda, and T. Akutsu. Inferring strengths of protein-protein interactions from experimental data using linear programming. *Bioinformatics*, 19(suppl 2):ii58–ii65, 2003.

[11] J. Moussouri. Gibbs and markov random systems with constraints. *Journal of Statistical Physics*, 10(1):11–33, 1974.

[12] J. Nocedal. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.

[13] L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, 32:D449–D451, 2004.

[14] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proc. HLT-NAACL 2003*, pages 134–141, 2003.

[15] C. Sutton and A. McCallum. *Introduction to statistical relational learning*, chapter An introduction to conditional random fields for relational learning, pages 93–128. MIT Press, 2006.

[16] The UniProt Consortium. The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Research*, 38:D142–D148, 2010.

[17] 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*, 106:67–72, 2009.

[18] 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*, 422:75–101, 2007.