Inferring Strengths of Protein-Protein Interactions Using Support Vector Regression

Yusuke Sakuma[1] Mayumi Kamada[1] Morihiro Hayashida[1,a)] Tatsuya Akutsu[1,b)]

Abstract: Due to the importance of protein-protein interactions (PPIs) in living organisms, many efforts have been made to investigate and predict PPIs. Analysis of strengths of PPIs is important as well as PPIs because such strengths are involved in functionality of proteins. In this technical report, we propose several feature space mappings from protein pairs, which make use of protein domain information, and perform five-fold cross-validation for data obtained from biological experiments. The result of average root mean square error (RMSE) using support vector regression (SVR) with our proposed feature was better than that by the best existing method, APM proposed by Chen et al.

1. Introduction

Protein-protein interactions (PPIs) play various important roles in cellular systems. Many investigations and analyses have been done for PPIs, and many prediction methods have been developed. As well as studies of PPIs, analyses of *strengths* of PPIs are important because such strengths are involved in functionality of proteins. In terms of transcription factor complexes, if a member protein has a weak binding affinity, target genes may not be transcribed depending on intracellular circumstance. For example, it is known that multi-subunit complex NuA3 in *Saccharomyces Cerevisiae* consists of five proteins, Sas3, Nto1, Yng1, Eaf6, and Taf30, acetylates lysine 14 of histone H3, and activates gene transcription. However, Yng1 and Nto1 are often found in the complex, and interactions with other member proteins are difficult to be observed by biological experiments. Hence, Byrum et al. proposed a biological methodology for identifying transient and unstable protein interactions recently [1].

Although many biological experiments have been conducted for protein-protein interactions [2], [3], strengths of PPIs have not been always provided. Ito et al. conducted large-scale yeast two-hybrid experiments for whole yeast proteins. In their experiments, yeast two-hybrid experiments were conducted for each protein pair multiple times, and the number of experiments that interactions were observed, or the number of interaction sequence tags (ISTs), was counted. Consequently, they decided that protein pairs having three or more ISTs should interact, and reported interacting protein pairs.

The ratio of the number of ISTs to the total number of experiments for a protein pair can be regarded as the interaction strength between their proteins. On the basis of this consideration, several prediction methods for strengths of PPIs have been developed. LPNM [4] is a linear programming-based method, and ASNM [4] is a modified method from the association method [5] for predicting PPIs. Chen et al. proposed association probabilistic method (APM) [6], which is the best existing method for predicting strengths of PPIs as far as we know. These methods make use of protein domain information. Domains are known as structural and functional units in proteins, and are stored in several databases such as Pfam [7] and InterPro [8]. The same domain can be identified in several different proteins. In these prediction methods, interaction strengths between domains are estimated from known interaction strengths between proteins, and interaction strengths for target protein pairs are predicted from estimated strengths of domain-domain interactions.

In this technical report, we also make use of domain information, and propose several feature space mappings from protein pairs. We use support vector regression (SVR), perform five-fold cross-validation for data from biological experiments by Ito et al. [3] and WI-PHI dataset [9], and take the average root mean square error (RMSE). The average RMSE by our proposed method was smaller than that by the best existing method, APM [6].

2. Method

In this section, we briefly review a probabilistic model and related methods, the association method [5], ASNM (association method for numerical interaction data) [4], APM (association probabilistic method) [6], and propose several feature space mappings using domain information.

2.1 Probabilistic Model of Protein-Protein Interactions Based on Domain-Domain Interactions

Many strength prediction methods are based on the probabilistic model of protein-protein interactions proposed by Deng et al. [10]. This model utilizes domain-domain interactions, and assumes that two proteins interact with each other if and only if at least one pair of domains contained in the respective proteins

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
[a)] morihiro@kuicr.kyoto-u.ac.jp
[b)] takutsu@kuicr.kyoto-u.ac.jp

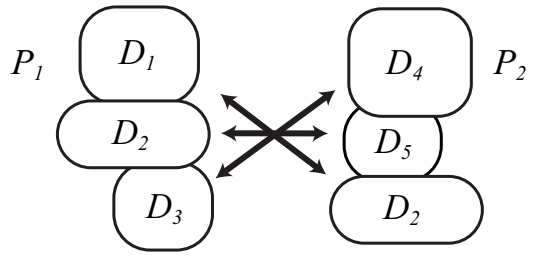

Fig. 1 Illustration of protein-protein interaction model based on domain-domain interactions

interacts. Fig. 1 illustrates this interaction model between two proteins P_1 and P_2, which consist of domains D_1, D_2, D_3, and domains D_2, D_4, D_5, respectively. As in this case, two proteins can contain the same domain. According to this model, if P_1 and P_2 interact, at least one pair among (D_1, D_2), (D_1, D_4), (D_1, D_5), (D_2, D_2), (D_2, D_4), (D_2, D_5), (D_3, D_2), (D_3, D_4), and (D_3, D_5) interacts. Conversely, if a pair, for instance (D_3, D_4), interacts, P_1 and P_2 interact.

From the assumption of this model, we can derive the following simple probability that two proteins P_i and P_j interact with each other.

$$Pr(P_{ij} = 1) = 1 - \prod_{D_m \in P_i, D_n \in P_j} (1 - Pr(D_{mn} = 1)), \quad (1)$$

where $P_{ij} = 1$ indicates the event that proteins P_i and P_j interact (otherwise $P_{ij} = 0$), $D_{mn} = 1$ indicates the event that domains D_m and D_n interact (otherwise $D_{mn} = 0$), P_i and P_j also represent the sets of domains contained in P_i and P_j, respectively. Deng et al. applied the EM (expectation maximization) algorithm to the problem of maximizing log-likelihood functions, estimated probabilities that two domains interact, $Pr(D_{mn} = 1)$, and proposed a method for predicting PPIs using the estimated probabilities of domain-domain interactions [10]. Actually, they calculated $Pr(P_{ij} = 1)$ using Eq. (1), and determined whether or not P_i and P_j interact by introducing a threshold θ, that is, P_i and P_j interact if $Pr(P_{ij} = 1) \geq \theta$, otherwise the proteins do not interact. Since interacting sites may not be always included in some known domain region, it can cause the decrease of prediction accuracy in this framework.

2.2 Association Method

Let $\mathcal{P}$ be a set of protein pairs that have been observed to interact or not to interact. The association method [5] gives the following simple score for two domains D_m and D_n using proteins that include the domains.

$$ASSOC(D_m, D_n) = \frac{|\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j, P_{ij} = 1\}|}{|\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j\}|}, \quad (2)$$

where $|S|$ indicates the number of elements contained in the set S. This score represents the ratio of the number of interacting protein pairs including D_m and D_n to the total number of protein pairs including D_m and D_n. Hence, it can be considered as the probability that D_m and D_n interact.

2.3 Association Method for Numerical Interaction Data (ASNM)

The association method for numerical interaction data (ASNM) [4] is a modified method for predicting strengths of PPIs from the original association method [5]. This method takes strengths of PPIs as input data. Let ρ_{ij} represent the interaction strength between P_i and P_j, and we suppose that ρ_{ij} is defined for all $(P_i, P_j) \in \mathcal{P}$. Then, the ASNM score for domains D_m and D_n is defined as the average strength over protein pairs including D_m and D_n by

$$ASNM(D_m, D_n) = \frac{\sum_{\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j\}} \rho_{ij}}{|\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j\}|}. \quad (3)$$

If ρ_{ij} always takes only 0 or 1, $ASNM(D_m, D_n)$ becomes $ASSOC(D_m, D_n)$.

2.4 Association Probabilistic Method (APM)

Although ASNM is a simple average of strengths of PPIs, Chen et al. proposed the association probabilistic method (APM) by replacing the strength with an improved strength [6]. It is based on the idea that the contribution of one domain pair to the strength of a PPI should vary depending on the number of domain pairs included in a protein pair. They assumed that the interaction probability of each domain pair is equivalent in a protein pair, and transformed Eq. (1) as follows:

$$Pr(D_{mn} = 1) = 1 - (1 - Pr(P_{ij} = 1))^{\frac{1}{|P_i||P_j|}}. \quad (4)$$

Thus, by substituting the numerator of ASNM, APM is defined by

$$APM(D_m, D_n) = \frac{\sum_{\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j\}} (1 - (1 - \rho_{ij})^{\frac{1}{|P_i||P_j|}})}{|\{(P_i, P_j) \in \mathcal{P} | D_m \in P_i, D_n \in P_j\}|}. \quad (5)$$

They conducted some computational experiments, and reported that APM outperforms existing prediction methods such as ASNM and LPNM.

2.5 Feature Based on Number of Domains (DN)

We propose a feature space mapping based on the number of domains (DN) from two proteins. It can be considered that the probability that two proteins interact increases with a larger number of domains included in the proteins. Thus, the feature vector of DN for two proteins P_i and P_j is defined by

$$f_{ij}^{(m)} = M(D_m, P_i) \ (\text{for } D_m \in P_i), \quad (6)$$

$$f_{ij}^{(T+n)} = M(D_n, P_j) \ (\text{for } D_n \in P_j), \quad (7)$$

$$f_{ij}^{(l)} = 0 \ (\text{for } D_l \notin P_i \cup P_j), \quad (8)$$

where T indicates the total number of domains over all proteins, and $M(D_m, P_i)$ indicates the number of domains identified as D_m in protein P_i.

2.6 Feature by Restriction of Spectrum Kernel to Domain Region (SPD)

Furthermore, we propose a feature space mapping by restricting the application of the spectrum kernel [11] to domain regions

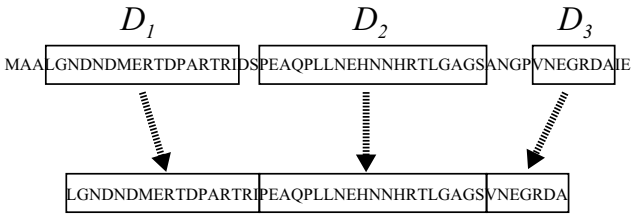

Fig. 2 Illustration of restricting an amino acid sequence to which the spectrum kernel is applied to the domain regions

(SPD). Let $\mathcal{A}$ be the set of alphabets representing twenty types of amino acids. Then, $\mathcal{A}^k$ ($k \geq 1$) means the set of all strings with length k generated from $\mathcal{A}$. The k-spectrum kernel for sequences x and y is defined by

$$K_k(x, y) = \langle \Phi_k(x), \Phi_k(y) \rangle, \tag{9}$$

where $\Phi_k(x) = (\phi_s(x))_{s \in \mathcal{A}^k}$ and $\phi_s(x)$ indicates the number of times that s occurs in x.

To make use of domain information, we restrict an amino acid sequence to which the k-spectrum kernel is applied to the domain regions. Fig. 2 illustrates the restriction. In this example, the protein consists of domains D_1, D_2, D_3, and each domain region is surrounded by a square. Then, the subsequence in each domain is extracted, and all the subsequences in the protein are concatenated in the same order as domains. We apply the k-spectrum kernel to the concatenated sequence. Let $\phi_s^{(r)}(x)$ be the number of times that string s occurs in the sequence restricted to the domain regions in protein x in the above manner. The feature vector of SPD for proteins P_i and P_j is defined by

$$f_{ij}^{(l)} = \phi_{s_l}^{(r)}(P_i) \ \text{(for } s_l \in \mathcal{A}^k\text{)}, \tag{10}$$

$$f_{ij}^{(20^k+l)} = \phi_{s_l}^{(r)}(P_j) \ \text{(for } s_l \in \mathcal{A}^k\text{)}. \tag{11}$$

It should be noted that $\phi_s^{(r)}$ for proteins having the same composition of domains can vary depending on the amino acid sequences of their proteins. That is, even if P_i and P_j have the same compositions as P_k and P_l, respectively, and the feature vector of DN for P_i and P_j is the same as that for P_k and P_l, then the feature vector of SPD for P_i and P_j can be different from that for P_k and P_l.

2.7 Support Vector Regression (SVR)

We employ support vector regression (SVR) [12] with our proposed features to predict strengths of PPIs. In the case of linear functions, SVR finds parameters w and b for $f(x) = \langle w, x \rangle + b$ by solving the following optimization problem.

$$\begin{aligned} \text{minimize} \quad & \tfrac{1}{2}\|w\|^2 + C \textstyle\sum_i (\xi_i + \xi_i'), \\ \text{subject to} \quad & y_i - \langle w, x_i \rangle - b \leq \epsilon + \xi_i, \\ & y_i - \langle w, x_i \rangle - b \geq -\epsilon - \xi_i', \\ & \xi_i \geq 0, \ \ \xi_i' \geq 0, \end{aligned}$$

where C and ϵ are positive constants, and (x_i, y_i) is a training data. Here, the penalty is added only if the difference between $f(x_i)$ and y_i is larger than ϵ. In our problem, x_i means a protein pair, and y_i means the corresponding interaction strength.

3. Computational Experiments

To evaluate our proposed features, DN and SPD, we conducted computational experiments, and compared them with the existing method, APM.

Table 1 Results of the average RMSE by SVR with our proposed features, DN and SPD ($k = 1, 2$), and by the existing method, APM, for training and test data

method	RMSE for training	RMSE for test
SVR with DN	0.0927	0.0831
SVR with SPD (k=1)	0.0289	0.0516
SVR with SPD (k=2)	**0.0242**	**0.0282**
APM	0.0265	0.0331

3.1 Data and Implementation

It is difficult to directly measure actual strengths of PPIs for many protein pairs by biological and physical experiments. Hence, we used Ito's yeast two-hybrid data with 1586 interacting protein pairs [3] and WI-PHI dataset with 50000 protein pairs [9]. For each protein-protein interaction, WI-PHI contains a weight that is considered to represent some reliability of the PPI, and is calculated from several different kinds of PPI datasets in some statistical manner. As strengths of PPIs, we used the value dividing the number of ISTs by the total number of yeast two-hybrid experiments for Ito's data, and used the value dividing the weight of PPI by the maximum weight for WI-PHI. Since these datasets do not include protein pairs with interaction strength 0, we randomly selected 100 protein pairs that were not included in the datasets, and added them as protein pairs with strength 0. We used UniProt database [13] to get amino acid sequences and information of domain compositions and domain regions in proteins. We used SVM light [14] for executing support vector regression, and used the polynomial kernel $K(x, y) = (s\langle x, y \rangle + c)^d$.

3.2 Root Mean Square Error (RMSE)

The root mean square error (RMSE) is a measure of differences between predicted values $\hat{y}_i$ and actually observed values y_i, and is defined by

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}, \tag{12}$$

where N is the number of test data.

3.3 Result

We conducted five-fold cross-validation, and calculated the average RMSE. We examined various values of parameters of the polynomial kernel in the range of $1 \leq s, c, d \leq 50$. Table 1 shows the results of the average RMSE by SVR with our proposed features, DN and SPD of $k = 1, 2$, and by APM [6], for training and test data, where parameters (s, c, d) for the polynomial kernel were $(1, 1, 3)$ in DN, $(28, 7, 17)$ in SPD of $k = 1$, and $(19, 4, 23)$ in SPD of $k = 2$. Although the average RMSEs by SVR with DN and by SVR with SPD of $k = 1$ were larger than those by APM for both training and test data, those by SVR with SPD of $k = 2$ were smaller than those by APM.

4. Conclusion

We proposed feature space mappings, DN and SPD, for predicting strengths of protein-protein interactions. DN is based on

the number of domains in a protein. SPD is based on the spectrum kernel, and is defined using the amino acid subsequences in domain regions. We employed support vector regression (SVR) with polynomial kernel, and conducted five-fold cross-validation using Ito's yeast two-hybrid data and WI-PHI dataset. For both training and test data, the average RMSEs by SVR with SPD of $k = 2$ were smaller than those by APM, which is the best existing method. It implies that the use of amino acid sequences in domain regions enhanced the prediction accuracy comparing with only information of domain compositions.

It is desired that additional datasets of accurate interaction strengths for many proteins are provided. However, to further enhance the prediction accuracy, we can improve kernel functions combining physical characteristics of domains and amino acids.

Acknowledgments This work was partially supported by Grants-in-Aid #22240009 and #24500361 from MEXT, Japan.

References

[1] Byrum, S., Smart, S., Larson, S. and Tackett, A.: Analysis of stable and transient protein-protein interactions, *Methods in Molecular Biology*, Vol. 833, pp. 143–152 (2012).

[2] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. and Rothberg, J.: A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Nature*, Vol. 403, pp. 623–627 (2000).

[3] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proceedings of the National Academy of Sciences of USA*, Vol. 98, pp. 4569–4574 (2001).

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

[5] Sprinzak, E. and Margalit, H.: Correlated sequence-signatures as markets of protein-protein interaction, *Journal of Molecular Biology*, Vol. 311, pp. 681–692 (2001).

[6] Chen, L., Wu, L.-Y., Wang, Y. and Zhang, X.-S.: Inferring protein interactions from experimental data by association probabilistic method, *Proteins: Structure, Function, and Bioinformatics*, Vol. 62, pp. 833–837 (2006).

[7] Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J., Gavin, O., Gunesekaran, P., Ceric, G., Forslund, K., Holm, L., Sonnhammer, E., Eddy, S. and Bateman, A.: The Pfam protein families database, *Nucleic Acids Research*, Vol. 38, pp. D211–D222 (2010).

[8] Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., Bernard, T., Binns, D., Bork, P., Burge, S., de Castro, E., Coggill, P., Corbett, M., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Fraser, M., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., McMenamin, C., Mi, H., Mutowo-Muellenet, P., Mulder, N., Natale, D., Orengo, C., Pesseat, S., Punta, M., Quinn, A. F., Rivoire, C., Sangrador-Vegas, A., Selengut, J. D., Sigrist, C. J. A., Scheremetjew, M., Tate, J., Thimmajanarthanan, M., Thomas, P. D., Wu, C. H., Yeats, C. and Yong, S.-Y.: InterPro in 2011: new developments in the family and domain prediction database, *Nucleic Acids Research*, Vol. 40, pp. D306–D312 (2012).

[9] Kiemer, L., Costa, S., Ueffing, M. and Cesareni, G.: WI-PHI: A weighted yeast interactome enriched for direct physical interactions, *Proteomics*, Vol. 7, pp. 932–943 (2007).

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

[11] Leslie, C., Eskin, E. and Noble, W.: The spectrum kernel: a string kernel for SVM protein classification, *Proceedings of Pacific Symposium on Biocomputing 2002*, pp. 564–575 (2002).

[12] Vapnik, V.: *The Nature of Statistical Learning Theory*, Springer, New York (1995).

[13] The UniProt Consortium: Reorganizing the protein space at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, Vol. 40, pp. D71–D75 (2012).

[14] Joachims, T.: *Advances in Kernel Methods – Support Vector Learning*, chapter Making large-scale SVM learning practical, MIT-Press (1999).