Vol.2014-MPS-98 No.26
Vol.2014-BIO-38 No.26
2014/6/27

Prediction of Heterotrimeric Protein Complexes by Two-phase Learning Using Neighboring Kernels

Peiying Ruan[1,a)] Morihiro Hayashida[1,b)] Osamu Maruyama[2,c)] Tatsuya Akutsu[1,d)]

Abstract: In biological systems, protein complexes are one of important molecules to perform as transcription factors and enzymes. Protein complexes with size more than three have been focused by most prediction methods. It, however, is known that protein complexes with smaller sizes occupy a large part of whole complexes for several species. In our previous work, we developed a method with several feature space mappings and the domain composition kernel for prediction of heterodimeric protein complexes, which outperforms existing methods.
In this technical report, we propose methods for prediction of heterotrimeric protein complexes by extending the previous prediction method on the basis of some ability that heterotrimeric protein complexes are not likely to share the same protein with each other. We make use of the discriminant function in support vector machines (SVMs), and design novel feature space mappings for the second phase. As the second classifier, we examine SVMs and relevance vector machines (RVMs).
We perform ten-fold cross-validation computational experiments. The results suggest that our proposed two-phase methods and SVM with the extended features and the domain composition kernel outperform the existing method NWE in terms of F-measure, which was reported to outperform other existing methods for prediction of heterotrimeric protein complexes.

1. Introduction

In biological systems, protein complexes are one of important molecules to perform as transcription factors and enzymes. Identification of functional protein complexes is essential for understanding molecular systems in living cells. Several proteins form a complex and work as a transcription factor, whereas there exist another type of proteins that work as enzymes. Hence, to identify proteins that constitute such transcription factors is useful for uncovering gene regulatory networks. Also for metabolic pathways, it is important to find enzymes that consist of several proteins.

Many computational methods have been developed for predicting protein complexes from protein-protein interaction networks [1], [2]. Enright *et al.* developed the Markov cluster (MCL) algorithm [3], which repeatedly executes two operators called expansion and inflation to a matrix whose element represents the transition probability from a protein to another. The expansion operation takes the power of the matrix, and the inflation operation takes the Hadamard power of the matrix. MCL is fast and efficient because of these operations. Macropol *et al.* developed the repeated random walks (RRW) method [4], which iteratively expands a cluster depending on the probabilities in steady states of random walks with restarts. Maruyama and Chihara improved the RRW method by weighting the restart probabilities and proposed the node-weighted expansion (NWE) method [5]. Bader and Hogue developed the molecular complex detection (MCODE) method [6], which uses a modified clustering coefficient defined by edge density in a subset of the original and adjacent vertices to find densely connected regions. King *et al.* developed the restricted neighborhood search clustering (RNSC) method [7], which selects clusters generated by a cost function according to the cluster size, density and functional homogeneity. Altaf-Ul-Amin *et al.* developed DPClus [8], which tries to find densely connected regions. Chua *et al.* developed the protein complex prediction (PCP) method [9], which finds maximal cliques using the functional similarity weight based on indirect interactions. Liu *et al.* developed the clustering based on maximal cliques (CMC) method [10], which generates all maximal cliques from the protein-protein interaction networks, and assembles highly overlapped clusters based on their interconnectivity. Wu *et al.* developed the core-attachment based (COACH) method [11]. Most methods basically focus on finding densely connected subgraph in protein-protein interaction networks. Hence, it is considered to be difficult that they detect small protein complexes because, for instance, the edge density of two interacting proteins is always 1.0 even if the proteins do not form a complex.

Protein complexes with small sizes, however, occupy a large part of whole known protein complexes. CYC2008 is a comprehensive catalogue of 408 manually curated yeast protein complexes [12]. In the catalogue, 172 complexes (42%) are heterodimeric, and 87 complexes (21%) are heterotrimeric as re-

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611–0011, Japan
[2] Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819–0395, Japan
[a)] ruan@kuicr.kyoto-u.ac.jp
[b)] morihiro@kuicr.kyoto-u.ac.jp
[c)] om@imi.kyushu-u.ac.jp
[d)] takutsu@kuicr.kyoto-u.ac.jp

ported also in [13]. In our previous study, hence, we developed a method using our proposed kernel for predicting heterodimeric protein complexes that consist of two distinct proteins [14], which outperforms an existing method using the naive Bayes classifier [15].

In this technical report, we introduce prediction methods for heterotrimeric protein complexes by extending techniques in our previous method on the basis of some ability that heterotrimeric protein complexes are not likely to share the same protein with other heterotrimeric protein complexes [16]. For that purpose, we apply supervised learning methods twice such as support vector machine (SVM) [17] and relevance vector machine (RVM) [18].

Tatsuke and Maruyama developed the proteins' partition sampler (PPSampler) method based on the Metropolis-Hastings algorithm, which generates clusters whose sizes follow a power-law distribution, and outperforms other existing methods in F-measure for whole protein complexes [13]. For prediction of heterotrimeric protein complexes, they reported that the F-measure of NWE was better than those of the existing methods, MCL, MCODE, DPClus, CMC, COACH, RRW, and PPSampler.

We perform ten-fold cross-validation, and calculate the average F-measure. The results suggest that our proposed methods outperform the existing method NWE.

2. Methods

In this section, we introduce prediction methods for heterotrimeric protein complexes. More accurately, we consider the following problem: Given a network of protein-protein interactions weighted by some reliability, determine whether or not three distinct proteins that are connected in the protein-protein interaction network form a protein complex.

Let $G(V, E)$ be an undirected graph with a set V of vertices and a set E of edges, representing the protein-protein interaction network. Here, a vertex represents a protein, an edge (i, j) represents an interaction between proteins P_i and P_j, and the weight w_{ij} represents reliability and strength of the interaction between P_i and P_j. In this technical report, we use the WI-PHI database [1] as edge weights, which has been calculated from heterogeneous biological experimental data. We say that P_i is a *neighboring* protein to P_j if $(i, j) \in E$. Then, our proposed methods use the support vector machine (SVM), its discriminant function, and the relevance vector machine (RVM).

2.1 Support and relevance vector machine

We briefly review the support and relevance vector machines [17], [18]. Suppose that N training data $\{\boldsymbol{x}_i, t_i\}$ with target $t_i \in \{-1, 1\}$ are given. For our purpose, $\boldsymbol{x}_i$ corresponds to a set of three distinct proteins, $t_i = 1$ corresponds to the case that the set forms a heterotrimeric protein complex. Then, we consider linear models represented by the form

$$y(\boldsymbol{x}) = \sum_{i=1}^{M} a_i \phi_i(\boldsymbol{x}) + b, \tag{1}$$

where ϕ_i denotes a basis function, M denotes the number of basis functions, a_i denotes the coefficient, and b denotes the bias parameter. In the SVM, $\phi_i(\boldsymbol{x})$ is implicitly defined as $K(\boldsymbol{x}_i, \boldsymbol{x})$ with a positive semidefinite kernel function K, M is equal to N, and a_i and b are determined by maximizing the margin. New sets $\boldsymbol{x}$ of proteins are classified according to the sign of $y(\boldsymbol{x})$. We make use of this discriminant function $y(\boldsymbol{x})$ in our proposed methods.

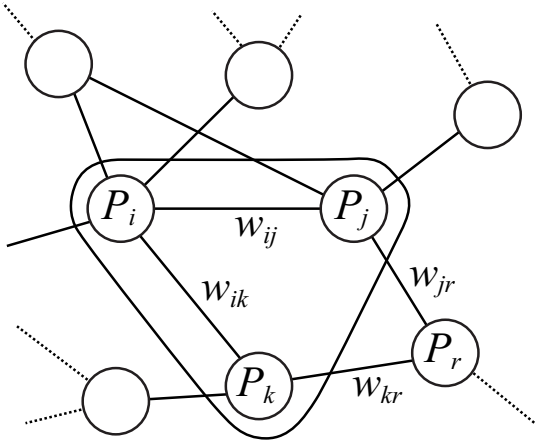

Fig. 1 Example of a subgraph including three focused proteins P_i, P_j, P_k and their neighboring proteins. In this example, protein P_r is neighboring to both of P_j and P_k.

The RVM is a Bayesian sparse kernel technique for classification and regression, and shares some characteristics of the SVM. As well as the SVM, the basis functions of the RVM are given by kernels, which are not required to be positive semidefinite. It, however, is known that training time of the RVM is in general longer than that of the SVM. In the RVM, a hyperparameter γ_i for each parameter a_i and a prior distribution over parameters a_i are introduced to obtain a sparse model.

For the classification, the model in Eq. (1) is transformed as $\sigma(y(\boldsymbol{x}))$, where $\sigma(y)$ denotes the logistic sigmoid function $1/(1 + e^{-y})$, and a_i and b are determined by maximizing the marginal log-likelihood with respect to $\boldsymbol{\gamma}$.

2.2 Extension of feature space mapping

In our previous study, we proposed seven feature space mappings for prediction of heterodimeric protein complexes [14]. These are based on the idea that the reliability of the interaction in a heterodimer should be high and conversely the reliability of the interaction between a protein in a heterodimer and a protein not in the heterodimer should be low. We extend the feature space mappings for two interacting proteins to mappings for three distinct proteins P_i, P_j, and P_k that are connected in the protein-protein interaction network as follows:

(F1) $\max_{\{(p,q)\in E | p,q\in\{i,j,k\}\}} w_{pq}$

(F2) $\min_{\{(p,q)\in E | p,q\in\{i,j,k\}\}} w_{pq}$

(F3) $\max_{\{(p,r)\in E | p\in\{i,j,k\}, r\notin\{i,j,k\}\}} w_{pr}$

(F4) $\min_{\{(p,r)\in E | p\in\{i,j,k\}, r\notin\{i,j,k\}\}} w_{pr}$

(F5) $\max_{\{(p,r),(q,r)\in E | p,q\in\{i,j,k\}, p\neq q, r\notin\{i,j,k\}\}} \min\{w_{pr}, w_{qr}\}$

(F6) $\max\{\#\text{ domains of } P_i, \#\text{ domains of } P_j, \#\text{ domains of } P_k\}$

(F7) $\min\{\#\text{ domains of } P_i, \#\text{ domains of } P_j, \#\text{ domains of } P_k\}$

Here, the fifth mapping in the previous study is eliminated be-

cause more neighboring proteins increase the maximum of differences close to the maximum of neighboring weights denoted by (F3).

(F1) and (F2) denote the maximum and minimum of the weights of interactions between P_i, P_j, and P_k, respectively. The first feature in the previous study is the weight of the interaction between two proteins. Since there are at least two interactions for three focused proteins and we cannot use all the weights as elements of our feature vector without changes, we take the maximum and minimum of the weights (see Fig. 1). In addition, the proteins in a heterotrimer should interact with each other, and (F2), which is the minimum of the weights, is expected to be high.

(F3) and (F4) denote the maximum and minimum of the weights of interactions between either of P_i, P_j, P_k and a neighboring protein P_r, respectively, where $r \neq i, j, k$ and $(i, r) \in E$, $(j, r) \in E$, or $(k, r) \in E$. It is considered that (F3), which is the maximum of the neighboring weights of a heterotrimer, should be lower than the weights of interactions in the heterotrimer. Consider the case that a protein P_r interacts with two of proteins P_i, P_j, and P_k, where P_r is not any of P_i, P_j, and P_k (see Fig. 1). If the weights of both interactions are large, these proteins including P_r may form a complex. We introduce the maximum of smaller weights of interactions with neighboring proteins P_r denoted by (F5).

(F6) and (F7) denote the maximum and the minimum of the numbers of domains contained in P_i, P_j, and P_k, respectively. The number of domains in a protein complex is expected to be large because domains are considered as mediators of protein-protein interactions.

In addition to the extended features, we examine the domain composition kernel developed in our previous study [14]. We defined equivalence $=_d$ between two proteins P_i and P_j as the condition that P_i consists of the same domains of P_j, and defined equivalence $=_c$ between two sets $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ that consist of $\{P_{i_1}, \cdots, P_{i_n}\}$ and $\{P_{j_1}, \cdots, P_{j_n}\}$, respectively, as

$$(\exists \sigma \in \mathfrak{S}_n) \forall k (P_{i_k} =_d P_{j_{\sigma(k)}}), \tag{2}$$

where $\mathfrak{S}_n$ denotes the symmetric group of degree n on the set $\{1, \cdots, n\}$. Then, the domain composition kernel K_c was defined by

$$K_c(\boldsymbol{x}_i, \boldsymbol{x}_j) = \delta(\boldsymbol{x}_i =_c \boldsymbol{x}_j), \tag{3}$$

where $\delta(A) = 1$ if condition A holds, otherwise 0.

2.3 Two-phase learning approach

Our proposed methods take two-phase learning approach. The basic idea for designing our methods is based on some ability that heterotrimeric protein complexes share the same protein with other heterotrimeric protein complexes.

We estimate model parameters of SVM using training data in the first phase, and predict whether or not the training data and the neighboring sets sharing at least one protein with the training data are heterotrimeric protein complexes, respectively. Then, the second phase predictor makes use of the discriminant values obtained by the first phase predictor. It is expected that the discriminant values for a target set of proteins and its neighboring set do not become large together if heterotrimeric protein complexes do not share the same protein.

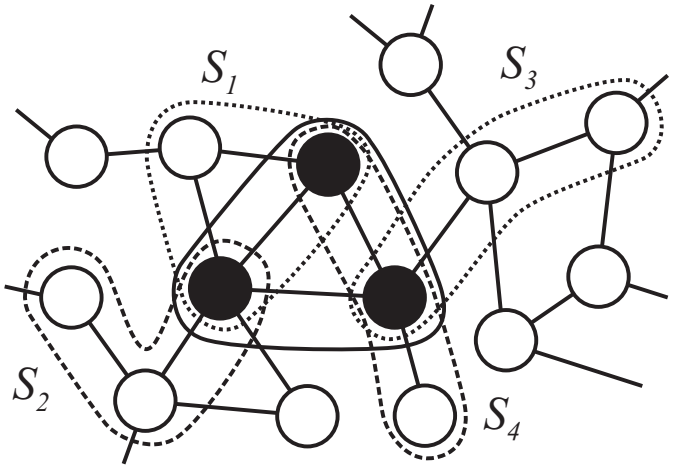

Fig. 2 Example of a subgraph including a focused set of proteins and neighboring sets of proteins. Each neighboring set of three proteins shares at least one protein with the focused set (black circle). In this example, sets S_1 and S_4 of three proteins share two proteins with the focused set, and S_2, S_3 share one protein, respectively.

Suppose that the training data set comprises N sets $\boldsymbol{x}_i$ of three distinct proteins with the corresponding label $t_i \in \{-1, 1\}$. For each $\boldsymbol{x}_i$, we calculate seven-dimensional feature vector $\boldsymbol{f}^{(1)}(\boldsymbol{x}_i)$ using (F1),...,(F7), and the combination kernel matrix whose (i, j)-th element is

$$\langle \boldsymbol{f}^{(1)}(\boldsymbol{x}_i), \boldsymbol{f}^{(1)}(\boldsymbol{x}_j) \rangle + \alpha K_c(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{4}$$

where α is a constant and $\langle \cdot, \cdot \rangle$ denotes the inner product. Then, we obtain the model parameters in Eq. (1) by applying the SVM to the training data set.

Let $\mathcal{N}(\boldsymbol{x})$ be all sets of three distinct proteins that are neighboring to $\boldsymbol{x}$ and connected in the protein-protein interaction network, where we say that $\boldsymbol{x}_i$ is a *neighboring* set to $\boldsymbol{x}_j$ if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ share the same protein and $\boldsymbol{x}_i$ is not $\boldsymbol{x}_j$ (see Fig. 2). For each $\boldsymbol{x}_i$, we calculate the discriminant values $y(\boldsymbol{x}_i)$ and $y(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i)$. Since the discriminant values may include outliers, by taking the averages of positive and negative discriminant values separately, we define four feature space mappings for $\boldsymbol{x}_i$,

$$f^{(2s)}(\boldsymbol{x}_i) = y(\boldsymbol{x}_i), \tag{5}$$

$$f^{(2p)}(\boldsymbol{x}_i) = \frac{1}{|\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) > 0\}|} \sum_{\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) > 0\}} y(\boldsymbol{x}), \tag{6}$$

$$f^{(2n)}(\boldsymbol{x}_i) = \frac{1}{|\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) < 0\}|} \sum_{\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) < 0\}} y(\boldsymbol{x}), \tag{7}$$

$$f^{(2a)}(\boldsymbol{x}_i) = \frac{1}{|\mathcal{N}(\boldsymbol{x}_i)|} \sum_{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i)} y(\boldsymbol{x}), \tag{8}$$

where $|S|$ denotes the number of elements in the set S. Here, we define $f^{(2p)}(\boldsymbol{x}_i) = 0$, $f^{(2n)}(\boldsymbol{x}_i) = 0$, and $f^{(2a)}(\boldsymbol{x}_i) = 0$ if $|\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) > 0\}| = 0$, $|\{\boldsymbol{x} \in \mathcal{N}(\boldsymbol{x}_i) | y(\boldsymbol{x}) < 0\}| = 0$, and $|\mathcal{N}(\boldsymbol{x}_i)| = 0$, respectively.

We compose eleven-dimensional feature vector $\boldsymbol{f}^{(2)}(\boldsymbol{x}_i)$ using $\boldsymbol{f}^{(1)}$, $f^{(2s)}$, $f^{(2p)}$, $f^{(2n)}$ and $f^{(2a)}$, calculate the combination kernel matrix with the (i, j)-th element

$$\langle \boldsymbol{f}^{(2)}(\boldsymbol{x}_i), \boldsymbol{f}^{(2)}(\boldsymbol{x}_j) \rangle + \alpha K_c(\boldsymbol{x}_i, \boldsymbol{x}_j), \tag{9}$$

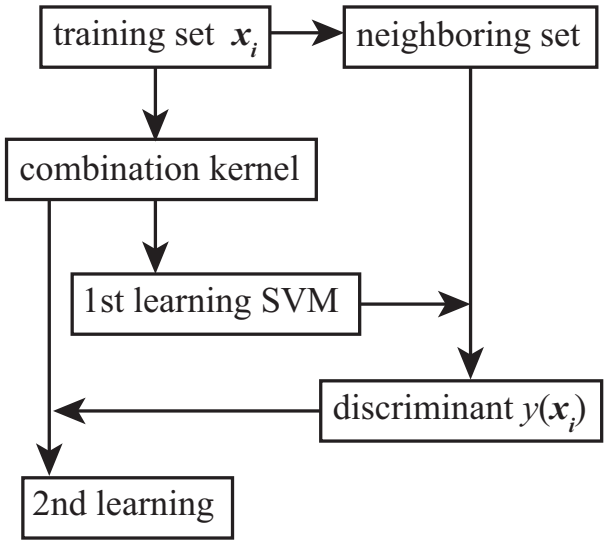

Fig. 3 Illustration of two-phase learning for training set $\boldsymbol{x}_i$.

and we apply some supervised learning method. It should be noted that our methods use only training data to estimate model parameters. Fig. 3 illustrates the two-phase learning for training set $\boldsymbol{x}_i$.

For test data $\boldsymbol{x}$, we calculate $\langle \boldsymbol{f}^{(2)}(\boldsymbol{x}_i), \boldsymbol{f}^{(2)}(\boldsymbol{x}) \rangle + \alpha K_c(\boldsymbol{x}_i, \boldsymbol{x})$ for training data $\boldsymbol{x}_i$, and determine whether or not $\boldsymbol{x}$ is a heterotrimeric protein complex according to the second classifier.

3. Computational Experiments

3.1 Data

To evaluate our proposed methods, we performed computational experiments and compared them with the existing method NWE [5]. We used the WI-PHI database [1] containing 49607 interacting protein pairs except self interactions as input weights of interactions, which is available at the supporting information web page of the paper. The weights were obtained from high-throughput yeast two-hybrid data [19], [20] and several biological databases such as BioGRID [2] and BIND [21] by using a log-likelihood score (LLS) to each dataset and the socioaffinity (SA) index [22] that measures the log-odds score of the number of times that two proteins are observed to interact to the expectation value from the dataset.

We prepared datasets using heterotrimeric protein complexes in CYC2008 protein complex catalogue [12], which contains 87 heterotrimeric protein complexes, and is available at http://wodaklab.org/cyc2008/. We restricted positive and negative examples to sets of three distinct proteins that form a single connected component in the input protein-protein interaction network. Thus, 7 heterotrimers were eliminated, and we used 80 heterotrimers as positive examples. For negative examples, we extracted 32647 sets of three proteins included in protein complexes with size more than three of CYC2008, and we selected uniquely at random 100 examples from the sets because our methods require many neighboring sets of three proteins for an example in the second phase. It is considered that negative examples selected from such sets are more difficult to be classified than those selected from all sets of three proteins except heterotrimers.

For NWE, we set some options related with the size of complexes so that NWE output protein complexes with size two or more from the WI-PHI protein-protein interaction network in the same way as [13], and extracted only protein complexes with size three from the result.

For measuring the performance, we used precision, recall, and F-measure defined by

$$precision = \frac{TP}{TP + FP}, \tag{10}$$

$$recall = \frac{TP}{TP + FN}, \tag{11}$$

$$F\text{-}measure = \frac{2 \cdot precision \cdot recall}{precision + recall}, \tag{12}$$

where TP, FP, and FN mean the number of true positive, false positive, false negative examples, respectively.

We used 'libsvm' (version 3.11) [23] and 'SparseBayes' package (version 2.0) [24] as implementations of SVM and RVM, respectively.

3.2 Results

We performed ten-fold cross-validation, and took the average of precision, recall, and F-measure. Furthermore, we repeated this procedure 10 times for other datasets with randomly selected negative examples, and took the average.

Table 1 shows the results on the average of precision, recall, and F-measure by our proposed methods and NWE. 'SVM+SVM' and 'SVM+RVM' denote two-phase methods using SVM and RVM as the second classifier, respectively. 'SVM' denotes usual SVM using only features $\boldsymbol{f}^{(1)}$. α denotes the coefficient of the domain composition kernel K_c. We examined $\alpha = 0.5$ because the case was best for prediction of heterodimeric protein complexes in our previous study [14]. NWE predicted 54 protein complexes with size three from the WI-PHI protein-protein interaction network, and 19 of them were actual heterotrimeric protein complexes in the CYC2008 protein complex catalogue.

We can see from the table that the F-measures by SVM+SVM, SVM+RVM, SVM for both $\alpha = 0$, and 0.5 were higher than those by NWE, respectively. Furthermore, the F-measure by the two-phase method SVM+SVM was higher than those by usual SVM with $\boldsymbol{f}^{(1)}$. The F-measure by SVM+RVM, however, was lower than those by SVM. It implies that RVMs may be less useful than SVMs for these problems that SVMs can be applied. Thus, the results suggest that our proposed methods SVM+SVM, SVM+RVM, and SVM outperform the existing method NWE. The results also suggest the usefulness of the second phase.

4. Conclusions

We proposed prediction methods by two-phase learning for heterotrimeric protein complexes. In the methods, we extended the feature space mappings in our previous study for prediction of heterodimeric protein complexes, and made use of the discriminant function for neighboring sets of three proteins.

To validate our proposed methods, we performed ten-fold cross-validation computational experiments. The results suggest that our two-phase prediction methods and SVM with the extended features outperform the existing method NWE, which

Table 1 Results on the average of precision, recall, and F-measure by our proposed methods and NWE. 'SVM+SVM' and 'SVM+RVM' denote two-phase methods using SVM and RVM as the second classifier, respectively. 'SVM' denotes usual SVM using only features $\boldsymbol{f}^{(1)}$. α denotes the coefficient of the domain composition kernel K_c. Note that NWE is unsupervised, and predicts protein complexes of various sizes. The precision and recall for NWE were calculated as TP divided by the numbers of predicted and known heterotrimers, respectively.

	SVM+SVM		SVM+RVM		SVM		NWE
α	0	0.5	0	0.5	0	0.5	
precision	**0.936**	0.869	0.847	0.899	0.909	0.873	0.352
recall	0.840	**0.926**	0.770	0.766	0.819	0.862	0.218
F-measure	0.880	**0.891**	0.767	0.810	0.854	0.862	0.270

was reported to outperform many other existing methods such as MCL, MCODE, DPClus, CMC, COACH, RRW, and PPSampler, although our methods are limited to prediction of heterotrimeric protein complexes. For further evaluation, we would like to perform computational experiments for other datasets if such data become available.

We have some possibility to further improve the prediction accuracy. For instance, we can use sequence information for designing feature space mappings as well as domains contained in proteins. In addition, we can introduce some probabilistic model such as conditional random fields to neighboring sets of three proteins although in this technical report we considered kernels between neighboring sets.

Acknowledgements

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

References

[1] 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).

[2] Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M.: BioGRID: a general repository for interaction datasets, *Nucleic Acids Research*, Vol. 34, pp. D535–D539 (2006).

[3] Enright, A., Dongen, S. V. and Ouzounis, C.: An efficient algorithm for large-scale detection of protein families, *Nucleic Acids Research*, Vol. 30, pp. 1575–1584 (2002).

[4] Macropol, K., Can, T. and Singh, A.: Repeated random walks on genome-scale protein networks for local cluster discovery, *BMC Bioinformatics*, Vol. 10, p. 283 (2009).

[5] Maruyama, O. and Chihara, A.: NWE: Node-weighted expansion for protein complex prediction using random walk distances, *Proteome Science*, Vol. 9, No. Suppl 1, p. S14 (2011).

[6] Bader, G. D. and Hogue, C. W.: An automated method for finding molecular complexes in large protein interaction networks, *BMC Bioinformatics*, Vol. 4, p. 2 (2003).

[7] King, A., Prulj, N. and Jurisica, I.: Protein complex prediction via cost-based clustering, *Bioinformatics*, Vol. 20, pp. 3013–3020 (2004).

[8] Altaf-Ul-Amin, M., Shinbo, Y., Mihara, K., Kurokawa, K. and Kanaya, S.: Development and implementation of an algorithm for detection of protein complexes in large interaction networks, *BMC Bioinformatics*, Vol. 7, p. 207 (2006).

[9] Chua, H., Ning, K., Sung, W. K., Leong, H. and Wong, L.: Using indirect protein-protein interactions for protein complex prediction, *Journal of Bioinformatics and Computational Biology*, Vol. 6, pp. 435–466 (2008).

[10] Liu, G., Wong, L. and Chua, H. N.: Complex discovery from weighted PPI networks, *Bioinformatics*, Vol. 25, pp. 1891–1897 (2009).

[11] Wu, M., Li, X., Kwoh, C. and Ng, S.: A core-attachment based method to detect protein complexes in PPI networks, *BMC Bioinformatics*, Vol. 10, p. 169 (2009).

[12] Pu, S., Wong, J., Turner, B., Cho, E. and Wodak, S.: Up-to-date catalogues of yeast protein complexes, *Nucleic Acids Research*, Vol. 37, pp. 825–831 (2009).

[13] Tatsuke, D. and Maruyama, O.: Sampling strategy for protein complex prediction using cluster size frequency, *Gene*, Vol. 518, pp. 152–158 (2013).

[14] Ruan, P., Hayashida, M., Maruyama, O. and Akutsu, T.: Prediction of heterodimeric protein complexes from weighted protein-protein interaction networks using novel features and kernel functions, *PLoS ONE*, Vol. 8, No. 6, p. e65265 (online), DOI: 10.1371/journal.pone.0065265 (2013).

[15] Maruyama, O.: Heterodimeric protein complex identification, *ACM Conference on Bioinformatics, Computational Biology and Biomedicine 2011*, pp. 499–501 (2011).

[16] Ruan, P., Hayashida, M., Maruyama, O. and Akutsu, T.: Prediction of heterotrimeric protein complexes by two-phase learning using neighboring kernels, *BMC Bioinformatics*, Vol. 15, No. suppl 2, p. S6 (2014).

[17] Vapnik, V.: *Statistical Learning Theory*, Wiley-Interscience (1998).

[18] Tipping, M. E.: The relevance vector machine, *Advances in Neural Information Processing Systems*, pp. 652–658 (2000).

[19] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M. and Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome, *Proc Natl Acad Sci USA*, Vol. 98, pp. 4569–4574 (2001).

[20] 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 analysis of protein-protein interactions in Saccharomyces cerevisiae, *Nature*, Vol. 403, pp. 623–627 (2000).

[21] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M. and et al.: The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic Acids Research*, Vol. 33, pp. D418–D424 (2005).

[22] Gavin, A.-C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L. J., Bastuck, S., Dümpelfeld, B., Edelmann, A., Heurtier, M.-A., Hoffman, V., Hoefert, C., Klein, K., Hudak, M., Michon, A.-M., Schelder, M., Schirle, M., Remor, M., Rudi, T., Hooper, S., Bauer, A., Bouwmeester, T., Casari, G., Drewes, G., Neubauer, G., Rick, J. M., Kuster, B., Bork, P., Russell, R. B. and Superti-Furga, G.: Proteome survey reveals modularity of the yeast cell machinery, *Nature*, Vol. 440, pp. 631–636 (2006).

[23] Chang, C.-C. and Lin, C.-J.: LIBSVM: A library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, Vol. 2, pp. 27:1–27:27 (2011).

[24] Tipping, M. E. and Faul, A. C.: Fast marginal likelihood maximisation for sparse Bayesian models, *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics* (2003).