Predicting Types of Protein-Protein Interactions Using a Multiple-Instance Learning Model

Hiroshi Yamakawa[1], Koji Maruhashi[1], and Yoshio Nakao[1]

Fujitsu Laboratories Ltd.
1-1, Kamikodanaka 4-chome, Nakahara-ku, Kawasaki, Kanagawa, 211-8588, Japan
{ymkw, maruhashi.koji, ynakao}@jp.fujitsu.com

Abstract. We propose a method for predicting types of protein-protein interactions using a multiple-instance learning (MIL) model. Given an interaction type to be predicted, the MIL model was trained using interaction data collected from biological pathways, where positive bags were constructed from interactions between protein complexes of that type, and negative bags from those of other types. In an experiment using the KEGG pathways and the Gene Ontology, the method successfully predicted an interaction type (phosphorylation) to an accuracy rate of 86.1%.

1 Introduction

In recent molecular biology and its application fields including drug discovery, analysis of protein-protein interactions (PPIs) is an emerging issue to elucidate the mechanism of biological processes. Since PPIs play a central role in numerous cellular processes, understanding PPIs provides us with clues to determining potential drug targets in cases where drug targets are identified from a known pathway related to a disease. Although a large volume of PPI data has been collected as described later, many other unknown PPIs are believed to exist (Rhodes et al., 2005). Moreover, only a few PPIs have been elucidated at the functional level. In order to resolve this situation, an approach that combines wet and dry technologies is promising, so that one might construct hypotheses on a biological mechanism based on confirmed PPI data produced by a wet technology with plausible PPI data predicted by a dry technology.

As for PPI prediction, Rhodes et al. (2005) proposed a probabilistic method that integrates model organism interactome data, protein domain data, genome-wide gene expression data and functional annotation data; and Lee et al. (2005) presented an assessment scheme for the reliability of PPI candidates based on a neural network algorithm. However, the aim of these studies is discovering novel PPIs or filtering correct PPIs, not identifying the interaction types between proteins (e.g., activation, inhibition, phosphorylation, or the like).

Although PPI types are essential in describing the mechanism of a biological process, only a few existing PPI databases provide interaction types (Ekins, Nikolsky, and Nikolskaya, 2005). For example, the Human Protein Reference

Database (HPRD) (Peri et al., 2003), which is one of the most famous databases providing a large volume of PPI data, does not provide interaction types although it stores 33,710 entries of PPI data at the time of September 2005 and its size continues to grow. Although the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa and Goto, 2000) provides pathway data in terms of protein-protein interaction networks with interaction types, the number of its PPI entries is only several thousands.

PPI type prediction is, therefore, an important issue to understand biological processes. In this paper, we propose a method for predicting PPI types based on a machine learning model using known PPI types provided by the KEGG database (Kanehisa and Goto, 2000) as training data.

2 A Model of Interaction between a Protein Complex Pair

A PPI described in existing pathways, as provided by the KEGG database (Kanehisa and Goto, 2000), often corresponds to a pair of protein complexes (see rounded rectangles in Fig. 1), each of which is composed of several subunits (simple proteins). On the other hand, functional annotations, as provided by the Gene Ontology (Ashburner et al., 2000), have been accumulated mainly for simple proteins.

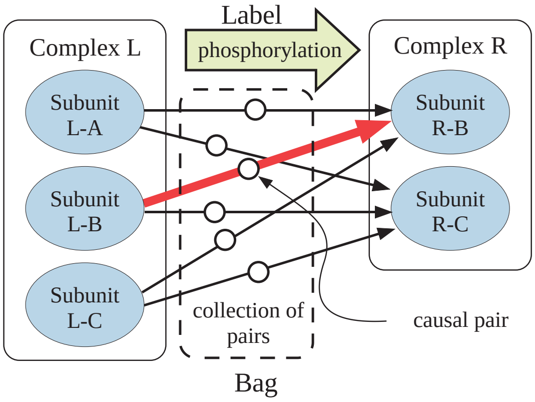

Fig. 1. A MIL Bag Model for PPI Type Prediction: each rounded rectangle depicts a protein complex; the upper block arrow depicts an interaction type (i.e., a label to be learned) between complexes; each simple arrow depicts a possible subunit pairing (i.e., an instance of the MIL model) and the broad arrow indicates the subunit pair that is the only cause of the process that Complex L phosphorylates Complex R.

In other words, training data for interaction type predictions are given in terms of a collection of complex pairs labeled with an interaction type (i.e., the target variable); while available features (i.e., the input variables) for machine learning are given for each subunit (simple protein) composing complexes. Figure 1 depicts a case where the target variable is whether the interaction type is phosphorylation and the input variables are given for each subunit of L-A, L-B, ..., and R-B.

It is difficult for a standard supervised learning method to solve a problem of this kind because the relationship between the input variables and the target variable is ambiguous. With regard to this point, we assume that the interaction type between complexes can be determined by a subunit pair across those complexes. In other words, we ignore those cases, for example, where two or more subunits of a complex work cooperatively, or where the active site of a subunit of a complex is hidden by other subunits of that complex. Hereafter, this assumption is referred to as the *subunit reaction assumption*.

This assumption can be validated based on the following discussion. From the protein structure viewpoint, an interaction between a protein pair is often explained by one or a small number of domain[1]-level interactions across those proteins. This fact suggests that an interaction between a protein complex pair can also be reduced to just a few domain-level interactions even though it cannot always be reduced to those of a protein pair across those complexes. In molecular biology, target validation (e.g., validation of a hypothesis that a particular protein causes a particular disease) is often performed by knocking-out the gene that encodes a target protein. The effectiveness of this approach suggests that many individual biological functions originate from a specific protein. Accordingly, the subunit reaction assumption is almost as valid as the tacit assumption based on which biologists may take the above-mentioned approach to target validation.

The assumption also has another advantage in that prediction methods based on it are expected to predict both the interaction types of complex pairs and those of subunit (protein) pairs, the former of which may describe the behavior of a biological system while the latter may indicate the molecular function that could be controlled with chemical substances (i.e., potential drugs).

3 A Method for Predicting Interaction Types Based on Multiple Instance Learning

In this section, we propose a method for predicting PPI types based on a multiple-instance learning (MIL) scheme. As shown in Fig. 1, the PPI-type prediction task based on the subunit reaction assumption can be formalized as a problem of MIL as follows: a complex pair (two rounded rectangles) with an interaction type (the upper block arrow) is formulated as a labeled bag, and a possible subunit pair across a complex pair (the six simple arrows) as an instance.

[1] A portion of a protein thought to have specific molecular functions and often corresponding to a characteristic sequence of amino acids.

3.1 Multiple-Instance Learning Scheme

Multiple-instance learning (MIL) is a scheme of semi-supervised learning for problems with incomplete knowledge concerning the labels of the training data. In the MIL training data, labels to be learned are only assigned to bags of instances but not to individual instances, while every training instance is labeled for supervised learning. For example, in a binary classification problem, a bag is labeled positive if at least one instance in that bag is positive, while the bag is labeled negative only if all the instances in it are negative. The goal of MIL is to predict labels of unseen bags and/or instances based on those labeled bags.

In the pioneering work of (Dietterich, Lathrop, and Lozano-Perez, 1997), MIL was applied to drug activity estimation. Following this, many MIL methods have been proposed and applied to various fields including image classification (Maron and Lozano-Pérez, 1998), stock selection (Maron and Lozano-Pérez, 1998), text classification, face labeling in broadcasting news video (Yang, Yan, and Hauptmann, 2005), Web mining (Zhou, Jiang, and Li, 2005), etc.

3.2 Diverse Density

In this research, we solve the problem using a modified version of the diverse density (DD) framework (Maron and Lozano-Pérez, 1998), which is one of the well-known MIL solutions. The main idea of the DD framework is to find a desired concept point (i.e., the point desired for positive identification) in the feature space that is close to at least one instance from every positive bag and far away from any instances in negative bags. Desired concept points are found according to a score called diverse density (dd), which is a measure of how many different positive bags have instances near the point, and how far the negative instances are away from that point. The optimal concept point is defined as the one with the maximum dd.

A probabilistic version of dd at a point x is calculated using positive bags B_i^+ and negative bags B_i^- by the following formula:

$$dd\left(x\right) \propto \prod\nolimits_i \Pr\left(x|B_i^{\pm}\right) \tag{1}$$

where $B_{ij}^{\pm}$ denotes the j-th instance in the bag $B_i^{\pm}$; $Pr(x|B_i^{\pm})$ denotes a contribution score of an instance in the bag $B_i^{\pm}$, which is calculated as a likelihood score whether an instance in the bag $B_i^{\pm}$ is near the location x (described later).

Each bag B_x (i.e., a case whose interaction type is predicted) is evaluated by a score at the optimal concept point in that bag, which can be calculated by the following formula:

$$dd(B_x) = \max_m dd(x^{(m)})$$

where $x^{(m)}$ is the point at which instance m is located. It is judged positive if the score is larger than a threshold value(≥ 0).

Every contribution score is calculated using a *noisy-or* model as follows:

$$\begin{aligned}\Pr\left(x|B_i^+\right) &= 1 - \prod_j \left(1 - \Pr\left(x|B_{ij}^+\right)\right) \\ \Pr\left(x|B_i^-\right) &= \prod_j \left(1 - \Pr\left(x|B_{ij}^-\right)\right) \qquad (2)\end{aligned}$$

where $\Pr\left(x|B_{ij}^\pm\right)$ denotes the contribution score of an instance $B_{ij}^\pm$, which is calculated using a Gaussian-like distribution as follows:

$$\Pr\left(x|B_{ij}^\pm\right) = \exp\left[-\sum\nolimits_k s\left(B_{ijk} - x_k\right)^2\right] \qquad (3)$$

where k denotes the index of the axes of the feature space, and s denotes the scale factor.

3.3 Modified Diverse Density as a Weighted Voting System

In a preliminary experiment, we found that the diverse density score (*dd*) did not work well for our problem. We found many false negative cases related to a few instances from negative bags. This was because the *dd* score calculated by (1) is very sensitive to the contribution score of a negative bag: i.e., the *dd* score is calculated as very low even in the case where only one negative instance is near the point x.

Relating this to our particular problem, this could represent a case where a subunit pair with an interacting potential is inhibited from interaction in such a condition that the active domain of a subunit of a complex is hidden by other subunits of that complex. Accordingly, a point around which many positive instances are near should be given a higher score even though there are a few negative instances near to that point. In this respect, we introduced another diverse density score, *vdd*, based on the following formula:

$$vdd(x) \propto \sum_i \mathrm{sign}_i \left[1 - \prod_j \left(1 - \Pr\left(x|B_{ij}^\pm\right)\right)\right] \qquad (4)$$

where sign_i is $+1$ for positive bags and -1 for negative bags. The *vdd* score can be interpreted as a weighted voting system where the absolute value of a voting weight is a likelihood score of whether any of the instances in a bag are at the point. Hereafter, this score is referred to as the *voting diverse density* (*vdd*).

Each bag B_x is evaluated by a score $vdd(B_x) = \max_m vdd(x^{(m)})$ ($x^{(m)}$ is the point at which instance m locates). It is judged positive if the score is larger than a threshold value (≥ 0). The experiments described later used a threshold value of 0.

4 PPI Dataset and Feature Space

In this study, we use the KEGG pathways (Kanehisa and Goto, 2000) for dataset construction, and the Gene Ontology (Ashburner et al., 2000) for feature space construction.

4.1 PPI Dataset Obtained from the KEGG Pathways

The PPI dataset for the experiment was constructed from the KEGG pathways as of March 2006. Firstly, we obtained human pathway data in XML format from the KEGG site[2], and then extracted PPI data, i.e., the *relation* elements whose value for the *type* attribute was *PPrel*. Each record in the PPI data comprises two groups of proteins, either or both of which may correspond to a protein complex, a protein family, or a simple protein, and one or more interaction types, which is described in the *subtype* element in the XML file. As a result, 1,279 different PPI records were obtained.

Table 1 summarizes the distribution of interaction types. The rows from *state* to *compound* individually correspond to an interaction type; the columns from 1 to 17 show the assignment patterns of the interaction types where the cells with a value of 1 in the same column indicate that all the interaction types corresponding to those rows were assigned to one or more identical PPI records. For example, the 6th column indicates that there were six records labeled with both interaction types of *ubiquination* and *inhibition*.

Table 1. Distribution of Interaction Types in PPI Records from the KEGG Pathways

	Assignment Pattern [Dataset Type]																	
Interaction Type	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	#
	[−]	[−]	[−]					[+]	[+]	[+]	[−]		[+]			[+]	[−]	Recs.
state	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	4
ubiquination	.	1	.	.	.	1	.	.	.	.	.	.	.	.	.	.	.	13
dephosphorylation	.	.	1	.	.	.	1	.	.	.	.	.	.	.	1	.	.	25
dissociation	.	.	.	1	.	.	.	.	1	.	.	.	.	.	.	.	.	14
inhibition	.	.	.	.	1	1	1	.	.	1	.	.	.	.	.	.	.	198
phosphorylation	.	.	.	.	.	.	.	1	1	1	.	.	1	.	.	1	.	249
binding association	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	.	181
indirect	.	.	.	.	.	.	.	.	.	.	.	1	1	.	.	.	.	102
activation	.	.	.	.	.	.	.	.	.	.	.	.	.	1	1	1	.	588
compound	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	27
#Records	4	7	8	13	145	6	6	150	1	41	181	97	5	525	11	52	27	1,279

4.2 Feature Space Encoding the Gene Ontology

Feature Vector: A feature vector of each instance, which comprises two proteins and an interaction type, was constructed using the Gene Ontology (GO) (Ashburner et al., 2000). The GO is a vocabulary that describes the attributes of genes. Each term in the vocabulary, called a GO term, represents a possible attribute value possessed by a gene or a protein encoded by genes. The GO has a hierarchical structure, i.e., GO terms are connected by *is-a* relations and

[2] ftp://ftp.genome.jp/pub/kegg/xml/KGML_v0.5/hsa

construct a directed acyclic graph. GO currently consists of three standard gene ontologies that describe biological processes, cellular components, and molecular functions. Many biological resources use the GO terms to annotate the properties of genes or their products (i.e., proteins or RNAs encoded by genes).

An initial feature vector of each instance consists of two components, each of which represents a list of GO terms annotated to one of the proteins composing that instance. The list of GO terms were obtained from the *gene2go* file[3] provided by the National Center for Biotechnology Information (NCBI). The list of GO terms for each protein was extended by adding all the ancestors in the GO hierarchy for each term in the initial list.

Feature Space Construction by Singular Value Decomposition: It is difficult for MIL to use the initial feature vectors. This is because a GO term in the higher layers of the GO hierarchy often relates to too many instances to discriminate different ones, and ones in the lower layers often relate to too few instances to generalize similar instances. In addition, the large size of the GO term set[4] requires expensive computational cost for MIL.

To overcome these problems, the feature space was compressed by using Singular Value Decomposition (SVD).

Logarithmic Probability Weighting: The SVD process tends to emphasize those properties that appear frequently. On the other hand, a GO term in the higher layers of the GO hierarchy often relates to many instances. If all the GO terms are equally weighted, an ineffective feature space will be obtained that emphasize those GO terms in the higher layers of the GO hierarchy.

In this respect, each GO term, j, is weighted with a logarithmic probability weight, W_j, which highlights a moderate abstraction level (depth) in the GO hierarchy. W_j is calculated using the following formula:

$$W_j = -\log(p_j) = -\log(\frac{m_j}{m}) \tag{5}$$

where m is the total number of genes appearing in the *gene2go* file, and m_j is the number of genes with term j in that file. For instance, the topmost term in the hierarchy, which relates to all the genes, is weighted with zero ($W_j = 0$), and is then ignored. On the other hand, a term in the middle layers, which is expected to have an appropriate specificity, is emphasized.

Singular Value Decomposition (SVD): In preparation for the SVD process, a matrix G that represents the relationship between the instances and the GO terms was constructed. A cell G_{ij} is set as W_j if the first protein of instance i relates to term j, or is set as $W_{j'}$ if the second protein of instance i relates to

[3] ftp://ftp.ncbi.nih.gov/gene/DATA/

[4] The molecular function ontology contains 622 different terms.

term j' where $j' = j - n_1$ (n_1 is the number of elements for the first proteins of instances), otherwise it is set as 0.

SVD decomposes matrix G as follows:

$$G = USD^T \tag{6}$$

where U and D are a unitary matrix that satisfies $U^T U = I_m$ and $D^T D = I_n$ respectively. The column vector of U is called the left singular vector. The column vector of D is called the right singular vector. The diagonal entries of S are called singular values.

A compressed feature matrix $\tilde{G}$ can be composed using selected right singular vectors d_x, d_y, ... as follows:

$$\tilde{G} = G\tilde{D} \text{ where } \tilde{D} = [d_x, d_y, \ldots] \tag{7}$$

5 Experiments: Binary Classification Task concerning Phosphorylation

This section reports on an experiment to evaluate the proposed method using a PPI prediction task as binary classification concerning an interaction type of phosphorylation. In the experiment, the proposed method was implemented on MATLAB. This is because MATLAB provides an efficient implementation of SVD for a sparse matrix.

5.1 Binary Classification Task concerning Phosphorylation

Since the GO provides three kinds of ontologies as described in Sect. 4.2, the input variables (i.e., feature vectors) can be constructed by several different approaches concerning which ontologies should be encoded into the feature vectors. In addition, since there are two proteins comprising an instance, the relation across the properties of those proteins can be encoded into the feature vectors. On the other hand, since some bags (i.e., PPI records) have two or more interaction types (Table 1), we can construct complicated tasks where certain combinations of interaction types are formalized as the target variable.

Among those possible tasks, we chose one of the simplest ones: the task of classifying PPIs into binary classes based on whether the PPI type is phosphorylation or not by using the feature vectors constructed from the molecular function ontology. This is because preliminary experiments suggested that this task is most promising among the simpler ones.

By taking account of the distribution of PPI type assignment patterns shown in Table 1, positive and negative bags were constructed as follows.

A set of positive bags was constructed by selecting those PPI records of which one of the interaction types was phosphorylation. As a result, 249 PPI records, those columns with a '[+]' mark in Table 1, were selected as positive bags.

A set of negative bags was constructed by selecting those PPI records of which neither of the interaction types appeared in any records in the set of positive bags. As a result, 227 PPI records, those columns with a '[-]' mark in Table 1, were selected as negative bags.

The remaining 803 PPI records were those whose interaction types included a type appearing in some positive bags but not phosphorylation. For example, PPI records that had only one interaction type of *activation* or *inhibition*.

5.2 Results

There are two major parameters of the proposed method: the dimension of the feature space and the scale factor of the Gaussian-like distribution (3) concerning the voting diverse density. The optimal values of these two parameters were explored in the range described below.

The compressed feature space based on the first few singular vectors by the SVD process is expected to represent major components of the initial feature space. In the experiment, twelve kinds of feature spaces of different dimensions were constructed by (7) with topmost n right singular vectors where $n \in \{[2..10], 12, 15, 20\}$.

The experiments also used the following twelve values in exploring the scale factor s: {500, 1000, 2500, 5000, 10000, 20000, 40000, 80000, 160000, 320000, 640000, 1280000}.

Figure 2 summarizes the results by the accuracy score in leave-on-out cross-validation for each parameter setting. The best accuracy of 86.1% was obtained at the point of $n = 4$ and $s = 160,000$. A comparable accuracy of 85.6% was obtained at the point of $n = 10$ and $s = 10,000$, and local peaks of the accuracy were found along a line connecting those two points. These observation suggest that a smaller scale factor is better for the feature space of a higher dimension. This could be explained by a tendency that the higher the dimension of the feature space, the longer the distance between instances.

Figure 3 is a scatter graph that plots every instance in the condition when the best accuracy was obtained ($n = 4$ and $s = 160,000$). The x-axis of this graph indicates the contribution of negative bags by a value of the 0.25th power of the negative components obtained by (4); and the y-axis of this graph indicates the contribution of positive bags by a value of the 0.25th power of the positive components obtained by (4).

Since the voting diverse density at a point equals to the difference of the value indicated by the y-axis and that indicated by the x-axis at that point, an instance located in the area above the dotted line in Fig. 3 is judged positive; while that located in the area below the dotted line is judged negative.

Note the instances in a positive bag (plotted with '*') and that in a negative bag (plotted with '·'). A significant amount (20.5%) of positive instances appear below the dotted line while most of instances at the optimal concept point (plotted with a small circle) in a positive bag appear above the dotted line. This is because the vdd score at the optimal point in a bag, according to its definition, equals to the maximum score among instances in that bag. For the same reason,

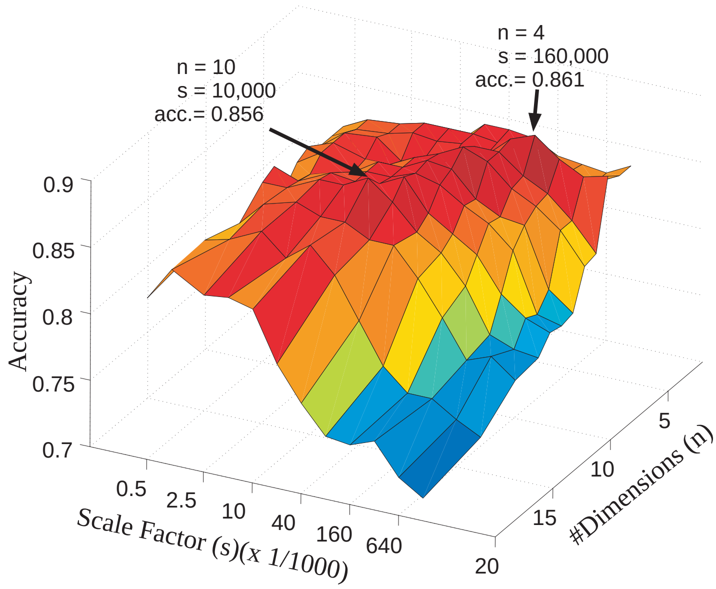

Fig. 2. Prediction Accuracy according to Two Parameters

a relatively large amount of instances at the optimal concept point (plotted with small rectangles) in a negative bag appear above the dotted line. This is the cause of false positive bags.

6 Conclusion

In this paper, we addressed the issue of PPI type prediction, and pointed out the problem that it is difficult for a standard supervised learning method to handle this issue because the relationship between the input variables (annotations for subunits) and the target variable (PPI type) is ambiguous.

We introduced the subunit reaction assumption and proposed that the PPI type prediction task based on this assumption can be formalized as a problem of MIL as follows: a complex pair with an interaction type is formulated as a labeled bag, and a possible subunit pair across a complex pair as an instance.

To solve that MIL problem, we proposed a method, a type of weighted voting system, based on Maron's Diverse Density (Maron and Lozano-Pérez, 1998), and evaluated it based on a binary classification version of that problem.

In the experimental evaluation, we have constructed a dataset, consisting of 1,279 different PPI records, from the KEGG pathways (Kanehisa and Goto, 2000), and a feature space for instances using the Gene Ontology (GO) (Ashburner et al., 2000). We then applied the method to a binary classification task concerning phosphorylation and achieved a highest accuracy of 86.1% in leave-

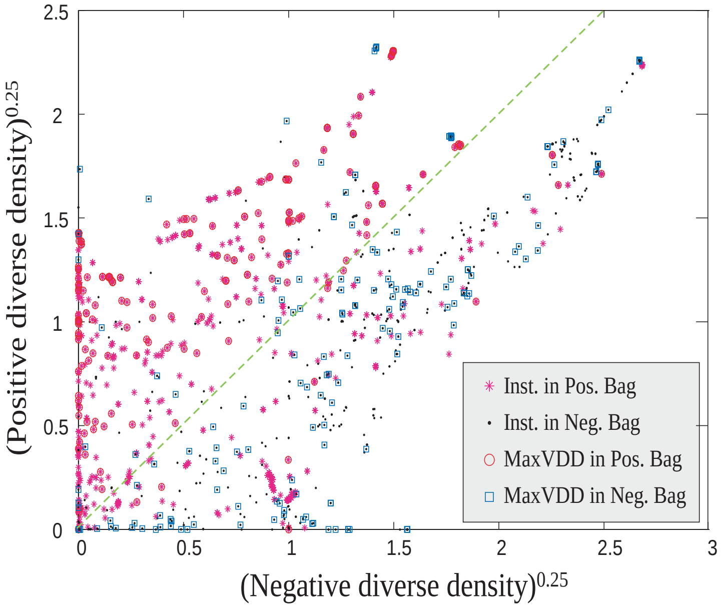

Fig. 3. Contribution of Positive and Negative Bags to Voting Diverse Density Score

one-out cross-validation when using the four major topmost components of the feature space obtained by SVD.

In the future, we would like to apply this method to the tasks of predicting other interaction types. In addition, we could improve the method by introducing a process to discriminate whether the protein group comprising a PPI record corresponds to a protein complex or a protein family, because the current version of the algorithm ignored those different types of protein groups, which the KEGG pathways may include in the same format. It will also be a future issue to improve the feature space, for example, by constructing an appropriate combination of features, or by encoding the domain or structural information of a protein, or relational information across a protein pair.

Acknowledgement

The authors would like to thank Kentaro Doi of Fujitsu Limited for the very useful discussions on potential applications of this research.

References

Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler1, J. Michael Cherry, Allan P. Davis, Kara Dolinski, Selina S. Dwight, Janan T. Eppig, Midori A. Harris, David P. Hill, Laurie Issel-Tarver, Andrew Kasarskis, Suzanna Lewis, John C. Matese, Joel E. Richardson, Martin Ringwald, Gerald M. Rubin, and Gavin Sherlock. 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29. [http://www.geneontology.org/].

Dietterich, Thomas G., Richard H. Lathrop, and Tomas Lozano-Perez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1–2):31–71.

Ekins, Sean, Yuri Nikolsky, and Tatiana Nikolskaya. 2005. Techniques: Application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in Pharmacological Sciences*, 26(4):202–209.

Kanehisa, Minoru and Susumu Goto. 2000. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30. [http://www.genome.jp/kegg/].

Lee, Min Su, Seung-Soo Park, and Min Kyung Kim. 2005. A protein interaction verification system based on a neural network algorithm. In *CSB Workshops*, pages 151–154. IEEE Computer Society.

Maron, Oded and Tomás Lozano-Pérez. 1998. A framework for multiple-instance learning. In Michael I. Jordan, Michael J. Kearns, and Sara A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10. MIT Press.

Peri, Suraj, J. Daniel Navarro, Ramars Amanchy, Troels Z. Kristiansen, Chandra Kiran Jonnalagadda, Vineeth Surendranath, Vidya Niranjan, Babylakshmi Muthusamy, T.K.B. Gandhi, Mads Gronborg, Nieves Ibarrola, Nandan Deshpande, K. Shanker, H.N. Shivashankar, B.P. Rashmi, M.A. Ramya, Zhixing Zhao, K.N. Chandrika, N. Padma, H.C. Harsha, A.J. Yatish, M.P. Kavitha, Minal Menezes, Dipanwita Roy Choudhury, Shubha Suresh, Neelanjana Ghosh, R. Saravana, Sreenath Chandran, Subhalakshmi Krishna, Mary Joy, Sanjeev K. Anand, V. Madavan, Ansamma Joseph, Guang W. Wong, William P. Schiemann, Stefan N. Constantinescu, Lily Huang, Roya Khosravi-Far, Hanno Steen, Muneesh Tewari, Saghi Ghaffari, Gerard C. Blobe, Chi V. Dang, Joe G.N. Garcia, Jonathan Pevsner, Ole N. Jensen, Peter Roepstorff, Krishna S. Deshpande, Arul M. Chinnaiyan, Ada Hamosh, Aravinda Chakravarti, and Akhilesh Pandey. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13(10):2363–2371.

Rhodes, Daniel R, Scott A Tomlins, Sooryanarayana Varambally, Vasudeva Mahavisno, Terrence Barrette, Shanker Kalyana-Sundaram, Debashis Ghosh, Akhilesh Pandey, and Arul M Chinnaiyan1. 2005. Probabilistic model of the human protein-protein interaction network. *Nature Biotechnology*, 23(8):951–959.

Yang, Jun, Rong Yan, and Alexander G. Hauptmann. 2005. Multiple instance learning for labeling faces in broadcasting news video. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 31–40.

Zhou, Zhi-Hua, Kai Jiang, and Ming Li. 2005. Multi-instance learning based web mining. *Applied Intelligence*, 22(2):135–147.