identify cancer related genes. At earlier stage, most of these methods were based on differential expression analysis. In other words, the aberrantly expressed genes are identified as cancer related lesions. Although partical success has made in identifying cancer related genes, these methods are unable either to infer any details on how a protein's behavior has changed or to reveal what specific mechanisms lead to the pathologic transition [2].

To overcome this drawback, some gene-centric identification methods have been developed to embed themself in the context of cellular network. These methods utilize the known disease relatedness of other nodes in the cellular network to infer some node's disease relatedness. The rationale is that if some neighbors (direct or indirect neighbors) of a gene are disease related, then the gene can also be inferred to be disease related with certain confidence [3]. With such a scheme, Kartik M Mani et al. proposed a novel identification method [2]. Their analysis method works in two steps. That is, this method first identifies dysregulated interactions (interactions showing either a gain of correlation or a loss of correlation pattern) in the phenotype of interest, and then ranks genes according to the statistical significance of dysregulated interaction enrichment among the interactions in which they directly participate [2]. This method's rationale is that if a node or gene's relation with most of their neighbors are changed under the disease state, then it can be inferred with high confidence that the gene itself is arch-criminal and disease related.

Some other gene-centric identification methods aim at the entire pathway or a prior defined gene set[4,5,6,7]. Pathway-based methods use a metric to measure the cohesiveness level of the members of the pathway and represent the tightness of relation between its members. Their rationale is that if the cohesiveness level is descended or elevated under disease state, the pathway can be viewed as a disrupted or newly constructed subsystem under disease state [5]. Gene set-based methods first use some metric to measure the differential expression level of each gene and then a ranked list of differentially expressed gene is obtained. Enrichment analysis of differentially expressed gene in the prior defined gene set is conducted to find which gene set's overall differential expression level is statistically significant [6, 7].

At present, there are also some edge-centric computational identification methods, for which the key is how to define the edges between nodes in the network and how to capture the differential behavior of the edge. Essentially, the definition of edge depends on the data at hand. High-throughput technologies are now producing vast amounts of biological data representing the availability of specific molecular species in a cellular population[2]. These include, among many others, gene expression and genotypic profiles [8], DNA-binding profiles[9], genomic sequences, and protein abundance from mass spectrometry [10]. At the same time, another high-throughput experiments have populated the public databases with thousands of protein-protein interaction (PPI) data and genetic interaction data[11].

Some researchers use the gene co-expression to define the edge between genes [12,13 ,14,15,16]: if two genes's mRNA expression levels are highly correlated under certain condition, then it can say that there is a functional association between two genes, in other words, there exists an edge between two genes. Jung-Kyoon Choi et al.[12] constructed a normal and disease coexpression network respectively based on 10 cancer microarray datasets and 10 their normal counterparts, and then identified the differential coexpression in the network. There are also some other methods based on differential co-expression analysis that was proposed to identify disease related lesions[13,14,15,16].

Other research works use the physical PPI and genetic interaction to identify disease related edges. One weakness of the high-throughput PPI data and genetic interaction data is that it contains no information about the conditions under which the interactions may take place[17]. Under the hypothesis that higher expression correlation of the genes implies genuine interactions of the proteins under the investigated conditions, it is a popular way to use the gene expression information to measure the 'activity 'of an interaction in response to the investigated condition. Zheng Guo et al.[17] scored the edge in PPI network based on the correlation coefficient of two genes's expression levels and the deferential expression of two genes, and then used simulated annealing algorithm to find a statistically significant responsive subnetwork.

On the other hand, protein-protein interaction have recently been recognized as challenging but attractive targets for small chemical drugs[18]. Furthermore, recent research works suggest that PPI inhibition could lead treatments for some human disease[18-23]. Motivated by both the potential pharmaceutic and therapeutic applications of disease related interactions and sparseness of computational methods for identifying disease related PPI or genetic interactions, we propose a new method to identify dysrgulated interactions by exploiting the mechanism of diseases in this paper. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated during disease process.

The remainder paper is organized as follows. Firstly, we describe the details of our method as well as the data set we used. Secondly, the results are presented through numerical tests on prostate cancer case. Finally, the features for the new method of identifying disease related interaction are discussed, and a brief conclusion and directions of further research works are presented in the last section.

2 Methods and materials

2.1 Dataset and data processing

The protein-protein interaction and genetic interaction data was first derived from the BIOGRID database(2008, 2.0.36 version). Then the self-interactions and reduplicate interactions were removed from the dataset. Finally, we have 23791 interactions in the interaction data set, which constitute a protein interaction network.

The prostate microarray data set [24] consists of about 7641 genes measured in 71 prostate tumors as well as 41 normal prostate specimens. In the microarray dataset, if there are multiple probes that correspond to the same gene, we choose the one that contains the least amount of missing values. Then, we only retain genes with missing data smaller than one third of the total sample size. Finally, we convert all values $<= 10$ to 10, and then perform a base 2 log transform. The prostate cancer related genes were obtained from Prostate Gene Database (PGDB)[25].

2.2 Estimation of pairwise gene co-expression

In this paper, the Percentage Bend Correlation [26] with $\beta = 0.1$ is applied to obtain a robust correlation estimate. Percentage Bend Correlation is first adopted to detect outliers in expression values of each gene so as to reduce the effects of those outliers in the correlation calculation[15]. Since the Percentage Bend Correlation may have some bias due

to sample size, Fisher's ż-transform [27] is also performed to reduce sample size effect, which can be formulated as

$$Z = \frac{\sqrt{n-3}}{2} \times \log\sqrt{\frac{1+r}{1-r}} \quad (1)$$

where r and n denote correlation estimate and sample size respectively, while Z corresponds to the Fisher's Z scores. Z score divided by its theoretical standard deviation theoretically has an asymptotically standard normal distribution. However, Min Xu et al. observed that the distributions of the z-score are still different from dataset to dataset [15]. Hence, we further normalize z-scores to enforce the standard normal distribution. After that, standardized correlations r' are obtained by inverting the z-score with a fixed n of 30 as Min Xu did.

2.3 Active interactions under certain condition

We give different definition of active interaction with respect to physical protein-protein interaction and genetic interaction. Suppose a physical protein-protein interaction connects gene A and gene B in cellular interaction network. We define the interaction to be active under normal state if the expression correlation of gene A and gene B in normal data set is higher than some threshold (in this paper, the threshold is set to be 0.20). Otherwise, the physical interaction between A and B are defined as inactive. For genetic interaction, we define it to be active under normal state if the absolute value of its two genes's expression correlation is higher than some threshold. Otherwise, the genetic interaction between A and B are defined as inactive. Similarly, we can define how an interaction is active or inactive under disease state.

2.4 Downregulated and upregulated interactions under disease state

We define an interaction to be upregulated if it is inactive in normal state but active under disease state. We define an interaction to be downregulated if it is active in normal state but inactive under disease state.

2.5 Enrichment analysis

The GO term enrichment analysis is done by the hypergeometric test on genes involved in downregulated interactions and upregulated interactions respectively through submitting them to DAVID online webserver(http://david.abcc.ncifcrf.gov /home.jsp). The prostate cancer and cancer related gene enrichment analysis are also done by the hypergeometric test.

Finally, the whole procedure of the method is summarized as Figure 1.

3 Results and discussion

Under the different thresholds, there are different numbers of interactions being active under normal state or disease state. In this paper, we present the result obtained when setting threshold being 0.20.

Under the threshold of 0.20, there are 1289 interactions that are active under normal state, while there are 1310 interactions that are active under disease state. Accordingly,

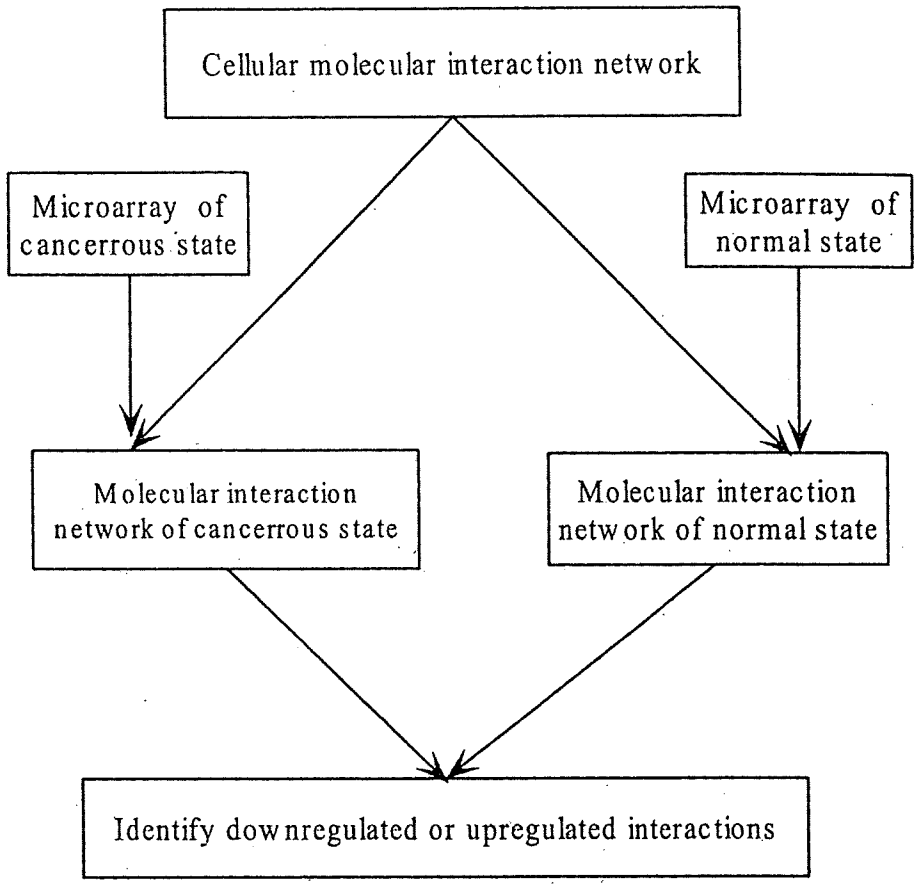

Figure 1: Flowchat of the proposed method

there are 213 interactions that are upregulated and 228 interactions that are downregulated. To evaluate the biological relevance of this identified dysregulated interactions, we perform some enrichment analysis. Firstly, the identified dysregulated interactions involve many genes. If these genes are cancer related, then we can infer that these interactions are also cancer related to some extent. There are 327 genes involved in upregulated interactions, of which 17 genes are known cancer related. There are 337 genes involved in downregulated interactions, of which 18 genes are known cancer related. Furthermore, there are 8042 genes involved in the interaction network. The known 118 cancer related genes that are included in these 8042 genes are used as background. We performed enrichment analysis on genes involved in downregulated and upregulated interactions respectively. The p-value of enrichment analysis is $4.9685e-006$ and $1.7217e-006$ respectively. The small p value shows that the enrichment of cancer related genes on the identified dysregulated interactions is statistically significant, and the identified dysregulated interactions are biological relevant and cancer related.

To further verify its biological relevance and cancer relatedness, we also performed the enrichment analysis of GO terms on the identified dysregulated interactions. There are many GO terms that are enriched. In this paper, we only present GO terms belonging to biological process category for the sake of simplicity. Some representative GO terms are listed in Tables 1 and 2 respectively. Enriched GO terms on downregulated inter-

Table 1: Representative enriched GO terms on downregulated interactions

GO term	P-value
Regulation of transcription	1.34E-10
Cell differentiation	2.77E-10
Programmed cell death	1.11E-10
Apoptotic program	0.0017
Cell proliferation	3.04E-7
Cell death	2.005E-10
Intracellular receptor-mediated signaling pathway	1.707E-6
RNA biosynthetic process	1.45E-10

actions include regulation of transcription, cell differentiation, programmed cell death, apoptotic program, cell proliferation, cell death, intracellular receptor-mediated signaling pathway, RNA biosynthetic process, which are all well known cancer related GO terms. Enriched GO terms on upregulated interactions include intracellular signaling cascade, negative regulation of metabolic process, regulation of transcription, cell differentiation, programmed cell death, apoptotic program, cell proliferation, cell death, intracellular receptor-mediated signaling pathway, and RNA biosynthetic process. It can be seen that most of these enriched terms were identified with small p-value. In summary, the enrichment of cancer related GO terms further verifies the biological relevance and cancer relatedness of our identified dysregulated interactions.

However, enrichment of cancer related genes and GO terms are just indirect evidence for the cancer relatedness of the identified dysregulated interactions. Finding direct evidence supporting cancer relatedness of dysregulated interactions is a challenging but important work.

Note that the method proposed in this paper is similar to the method presented in [15]. Next, we outline the main differences between the proposed method and the existing methods below.

(1). We use Percentage Bend Correlation to measure correlation, while mutual information was applied in [15].

(2). Method in [15] needs large background population to measure activity of interactions, while only counterpart samples of tissue samples are needed in our method.

(3). Difference between correlation of two genes in background population and tissue samples are used to define gain or loss of interactions in the existing methods. On the other hand, in our method, each interaction is classified as active or inactive under some condition, and thereby the downregulated or upregulated interactions are defined.

(4). The most important difference is that the goal of the research in [15] is to find disease related genes or perturbed target. However, our work aims to directly at dysregulated interactions. In other words, our goal is to exploit the impact of dysregulated

Table 2: Representative enriched GO terms on upregulated interactions

GO term	P-value
Intracellular signaling cascade	1.89E-13
Negative regulation of metabolic process	3.91E-7
Positive regulation of transcription	8.84E-10
Cell differentiation	3.30E-9
Programmed cell death	1.71E-8
Apoptotic program	2.85E-4
Regulation of cell proliferation	5.11E-6
Cell death	7.58E-8
Intracellular receptor-mediated signaling pathway	0.005
RNA biosynthetic process	0.0035

interactions on cellular function and their relation to disease.

4 Conclusion and future work

The computational identification of disease related lesions is still a key open problem in biomedicine and systems biology. In this paper, we proposed a new method to exploit the mechanism of disease by identifying dysregulated interactions. Specially, we present a method to detect which protein-protein interactions or genetic interactions are downregulated or upregulated. Experiment on a prostate cancer case shows that the identified dysregulated interactions are disease related, which confirms the effectiveness of our method.

However, our method indirectly verifies disease relatedness of the dysregulated interactions. This is still far away from our ultimate goal that elucidates the role that the dysregulated interactions play in disease. To reach this goal, some further work should be made in the future:

(1). Find direct evidence that can demonstrate cancer relatedness of dysregulated interactions.

(2). Identify the relation between downregulated interactions and upregulated interactions. For instance, we want to know if or not there exists some switch-like behavior from this study. We also want to know which cellular function or process is disturbed and which is newly emerged with the deletion of some old interactions and the addition of inclusion of new interactions.

(3). Integrate the methodology of pathway detection with our method. Now the computational identification of protein-protein target is mainly based on structural properties. We can exploit those techniques to provide a primary candidate list of protein-protein targets.

Acknowledgment

The authors would like to thank Dr Yong Wang, Xingming Zhao and Ruisheng Wang, Prof.Zengrong Liu and Prof. Luonan Chen for helpful discussions and suggestions. The authors would like to thank Dr Min Xu (Program in Molecular and Computational Biology, University of Southern California, Los Angeles, CA, USA) for providing the microarray data sets.

References

[1] Bert Volgelstein, Kenneth W Kinzler: Cancer genes and the pathways they control. NATURE MEDICINE 10(8) (2004) 789–799.

[2] Katik M Mani, Celine Lefebvre, Kai wang, Wei Keat Lim, Katia Basso, Riccardo Dallafavera, Andrea Califano: A systems biology approach to prediction of oncogenes and molecular perturbation tragets in B-cell lymphomas. Molecular Systems Biology 4:169 (2008) doi:10.1038/msb.2008.2.

[3] Ramon aragues, Chris Sander, Baldo Oliva: Predicting cancer involvement of genes from hetrogeneous data. BMC Bioinformatics 9:172 (2008) doi:10.1186/1471-2105-9-172.

[4] Igor Ulitsky, Richard M Karp, Ron Shamir: Detecting Disease-Specific Dysregulated Pathways Via Analysis of Clinical Expression Profiles. Lecture Notes in Computer Science(RECOMB2008) (2008) 347-359.

[5] Ruili Huang, Anders Wallqvist, David G Covell: Targeting changes in cancer: assessing pathway stability by comparing pathway gene expression coherence levels in tumor and normal tissues. Molecular Cancer Therapeutics, 5(9) (2006) 2417–2427.

[6] Aravind Subramaniana, Pablo Tamayoa, Vamsi K Moothaa, Sayan Mukherjeed, Benjamin L Eberta, Michael A Gillettea, Amanda Paulovichg, Scott L Pomeroyh, Todd R Goluba, Eric S Landera, Jill P Mesirova: Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy Science(USA), 102(43) (2005) 15545–15550.

[7] Trey Ideker, Owen Ozier, Benno Schwikowski, Andrew F Siegel: Discovering regulatory and signaling circuits in molecular interaction networks. Bioinformatics, 18(Suppl.1) (2002) S233–S240.

[8] M schena, D Shalon, R W Davis, P O Brown: Quatitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 270 (2005) 467–470.

[9] B Ren, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T L Volkert, C J wilson, S P Bell, R A Yong: Genome-wide location and function of DNA binding proteins. Science, 290 (2000) 2306–2309.

[10] O D Perez, G P Nolan: Simutaneous measurement of multiple active kinase states using polychromatic flow cytometry. Nature Biotechnology, 20 (2002) 155–162.

[11] P Uetz: A comprehensive analysis of protein-protein interactions in Sacharomyces cerevisiae. Nature, 403 (2000) 623–627.

[12] Jung Kyoon Choi, Ungsik Yu, Ook Joon Yoo, Sangsoo Kim: Differential coexpression analysis using microarray data and its application to human cancer. Bioinformatics, 21(24) (2005) 4348-4355.

[13] Kerby Shedden, Jeremy Taylor: Differential Correlation Detects Complex Associations Between Gene Expression and Clinical Outcomes in Lung Adenocarcinomas. Methods of Microarray Data Analysis, Springer-Verlag, Heidelberg (2005) 121–131.

[14] Yinglei Lai, Baolin Wu, Liang Chen, Hongyu Zhao: A statistical method for identifying differential gene-gene co-expression patterns. Bioinformatics, 20(17) (2004) 3146–3155.

[15] Min Xu, Ming-Chih J Kao, Juan Nunez-Iglesias, Joseph R Nevins, Mike West, Xianghong Jasmine Zhou: An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. BMC Genomics, 9(Suppl 1):S12 (2008) doi:10.1186/1471-2164-9-S1-S12.

[16] Ker Chau Li: Genome-wide coexpression dynamics: Theory and application. Proceedings of the National Academy of Sciences(USA), 99(26) (2002) 16875–16880.

[17] Zheng Guo, Yongjin Li, Xue Gong, Chen Yao, Wencai Ma, Dong Wang, Yanhui Li, Jing Zhu, Min Zhang, Da Yang, Jing Wang: . Bioinformatics, 23(16) (2007) 2121–2128.

[18] M R Arkin, J A Wells: Samll-molecular inhibtors of protein-protein interactions: progressing towards the dream. Nature Reviews Drug Discovery, 3 (2004) 301–317.

[19] P L Toogood: Inhibtion of protein-protein association by small molecules: approaches and progress . Journal of Medicinal Chemistry, 45 (2002) 1543–1558.

[20] A I Archakov, V M Govorun, A V Dubanov, Y D Ivanov, A V Veselovsky, P Lewis, P Jassen: Protein-protein interactions as a target for drugs in proteomics. Proteomics, 3 (2003) 380–391.

[21] L Pagliaro, J Felding, K Audouze, S J Nielsen, R B Terry, K J Christian, S Butcher: Emerging classes of protein-protein interactions inhitors and new tools for their development. Current Opinion in Chemical Biology, 8 (2004) 442–449.

[22] S Fletcher, A D Hamilton: Targeting protein-protein interactions by rational design: mimicry of protein surfaces. Journal of the Royal Society Interface, 3 (2006) 215–233.

[23] Nobuyoshi Sugaya, Kazuyoshi Ikeda, Toshiyuki Tashiro, Shiru Takeda, Jun Otomo, Yoshiko Ishida, Akiko Shiratori, Atushi Toyoda, Hideki Noguchi, Tadayuki Takeda, Satoru Kuhara, Yoshiyuki Sakaki, Takao Iwayanagi: An integrative in silico approach for discovering candidates for drug-targetable protein-protein interactions in interactome data. BMC Pharmacology, 7:10 (2007) doi:10.1186/1471-2210-7-10.

[24] Jacques Lapointe, Chunde Li, John P Higgins, Matt van de Rijna, Eric Bair, Kelli Montgomery, Michelle Ferrari, Lars Egevad, Walter Rayford, Ulf Bergerheim, Peter Ekman, Angelo M DeMarzo, Robert Tibshirani, David Botstein, Patrick O Brown, James D Brooks, Jonathan R Pollacka: Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proceedings of the Academy Science(USA), 101(3) (2004) 811–816.

[25] Longcheng Li, Hong Zhao, Hiroaki Shiina, Christopher J Kane, Rajvir Dahiya: PGDB: a curated and integrated database of genes related to the prostate. Nucleic Acids Research, 31 (2003) 291–293.

[26] R R Wilcox: Introduction to robust estimation and hypothesis testing, Academic Press, San Diego (1997).

[27] T W Anderson: An introduction to multivariate statistical analysis. Wiley-Interscience, Hoboken N.J. (2003).

Inference of Protein-Protein Interactions by Using Co-evolutionary Information

Tetsuya Sato[1], Yoshihiro Yamanishi[2], Katsuhisa Horimoto[3], Minoru Kanehisa[2], and Hiroyuki Toh[1]

[1] Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka 812-8582, Japan
sato@bioreg.kyushu-u.ac.jp, toh@bioreg.kyushu-u.ac.jp
[2] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
yoshi@kuicr.kyoto-u.ac.jp, kanehisa@kuicr.kyoto-u.ac.jp
[3] Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, 2-42 Aomi, Koto-ku, Tokyo, 135-0064, Japan
k.horimoto@aist.go.jp

Abstract. The mirror tree is a method to predict protein-protein interaction by evaluating the similarity between distance matrices of proteins. It is known, however, that predictions by the mirror tree method include many false positives. We suspected that the information about the evolutionary relationship of source organisms may be the cause of the false positives, because the information is shared by the distance matrices. Therefore, we excluded the information from the distance matrices and evaluated the similarity of the residuals as the intensity of co-evolution. We developed two different methods with a projection operation and partial correlation coefficient. The number of false positives were drastically reduced by our methods.

Keywords: protein-protein, co-evolution, projection operation, partial correlation coefficient.

1 Introduction

Information about protein–protein interactions in living cells provides deep insight into the biological functions of proteins at the cellular level. The development of large-scale experimental analyses, such as the yeast 2-hybrid system [7,21] and pull-down method [3,6], has facilitated understanding the protein–protein interaction network in cells. However, such experimental approaches have problems in coverage and accuracy [20,22]. Following the trend, the prediction of protein–protein interactions has become one of the major issues in bioinformatics. The predicted protein–protein interactions can provide complementary or supporting evidence to the large-scale experimental studies on protein–protein interactions although computational analyses also have the same drawbacks as experimental studies, that is, low coverage and low accuracy.

H. Anai, K. Horimoto, and T. Kutsia (Eds.): AB 2007, LNCS 4545, pp. 322–333, 2007.

Various computational methods to predict protein–protein interactions have been developed until today. Co-evolutionary behavior between interacting proteins provides useful information for the prediction of protein-protein interaction. The mirror tree method [15] and the *in silico* 2-hybrid system method [14] are two representative methods to predict protein-protein interaction with co-evolutionary information. In this paper, we explain our studies [18,19] aiming at improvement of the mirror tree method. The mirror tree method was developed by Pazos and Valencia [15], although there are several preceding works, such as Goh *et al.* [5]. The mirror tree method predicts protein–protein interactions under the assumption that the interacting proteins show similarity in molecular phylogenetic tree because of the co-evolution through the interaction. To avoid the difficulty to evaluate the similarity between a pair of phylogenetic trees, however, the mirror tree method compares a pair of distance matrices. Consider two proteins, proteins A and B. The orthologous amino acid sequences of protein A are collected from n species. The n sequences of protein A are aligned and the distance matrix, D_{A}, is calculated. The size of D_{A} is $n \times n$, and each row or column of the matrix corresponds to a species under consideration. An element of the matrix, $D_{\mathrm{A}}(i,j)$, represents the genetic distance between species i and j, which is calculated by comparing the amino acid sequences of protein A between the two species. A distance matrix is symmetric, and only the upper or lower half of the matrix includes sufficient information for tree construction. Likewise, the orthologous amino acid sequences of protein B are collected from the same n species, and the distance matrix, D_{B}, is calculated. The intensity of co-evolution between proteins A and B is evaluated as Pearson's correlation coefficient, $\rho_{\mathrm{AB}}^{\mathrm{MIRROR}}$, between the distance matrices D_{A} and D_{B}, which is calculated as follows:

$$\rho_{\mathrm{AB}}^{\mathrm{MIRROR}} = \frac{\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (D_{\mathrm{A}}(i,j) - \mathrm{Ave}(D_{\mathrm{A}}))(D_{\mathrm{B}}(i,j) - \mathrm{Ave}(D_{\mathrm{B}}))}{\sqrt{\mathrm{Var}(D_{\mathrm{A}})\mathrm{Var}(D_{\mathrm{B}})}}, \quad (1)$$

where Ave and Var represent the average and the variance of the upper (or lower) half elements of a distance matrix. High correlation between the distance matrices indicates the resemblance of the corresponding phylogenetic trees. Therefore, a pair of proteins are predicted to interact with each other, when the distance matrices of the proteins show high correlation. Because of the simplicity, modification and improvement have been introduced into the mirror tree method by several groups [4,9,16]. On the other hand, it has been recognized that the mirror tree predictions include many false positives. That is, even protein pairs that are known not to interact often show high correlation coefficients. Then, such pairs are predicted to interact in error. The abundance of false positives in the mirror tree prediction reduces the reliability of the method in actual applications. We suspected that the cause of the false positives is the information about the evolutionary relationship among the source organisms of the collected orthologous sequences. The distance matrices of orthologous proteins from the same set of n source organisms are compared in the mirror tree method. Therefore, all of the

distance matrices of the proteins are considered to include the same information about the evolutionary relationships among the same n sources. The information shared by the distance matrices would generate high correlation even between the matrices of non-interacting proteins. If our hypothesis is correct, the number of false positives in the predictions could be reduced by excluding such information from the distance matrices. We developed two different methods to exclude the information from the distance matrices for the prediction of protein-protein interaction. One of them uses a projection operator, whereas the other is based on multiple regression. The two methods were applied to physically contacting proteins, to evaluate their performances. Then, it was found that our methods drastically reduced the number of false positives in the predicted protein–protein interactions as expected.

2 Material and Method

2.1 Data Preparation

13 pairs of *Escherichia coli* proteins that are physically in contact were selected from the Database of Interacting Proteins (DIP) [17]. The pairs are listed in the legend for Table 1. Each pair was selected so that neither of the interacting proteins participated in the remaining 12 pairs of interacting proteins. Then, (putative) orthologues corresponding to the 26 proteins were collected from 40 different bacterial species, according to the KEGG KO database [10]. We assumed that a pair of proteins, which are orthologous to the interacting proteins of *E. coli*, are also physically in contact.

2.2 Multiple Sequence Alignment and Distance Matrix

A multiple alignment of each set of orthologous amino acid sequences was made with the alignment software MAFFT [11]. A distance matrix for the orthologous sequences was calculated from the multiple alignment. A genetic distance between every pair of aligned sequences was calculated as a maximum likelihood estimate using the PROTDIST in the PHYLIP package [2]. JTT model [8] was used as a model for the amino acid substitution for the estimation.

2.3 Transformation from Distance Matrix to Phylogenetic Vector [19]

The distance matrix was transformed into a vector. The upper or lower half of the non-diagonal elements of the distance matrix was arranged as a one-dimensional array of the numerical values in a certain order. All of the matrices were transformed into vectors with the same arrangement of the elements. When the matrix has a size of $n \times n$ the dimension of the vector is $n(n-1)/2$. The vector is hereafter referred to as a 'phylogenetic vector'. The dimension of the phylogenetic vector is 820, because n is 41. Consider a pair of phylogenetic vectors, which are transformed from distance matrices D_i and D_j. The subscripts

i and j indicate different sets of orthologues, that is, different proteins. Then, the elements of each vector are normalized with the average and the standard deviation of the elements as follows:

$$\left|\nu_i^{\#}\right\rangle = \frac{|\nu_i\rangle - |\mu\rangle}{\sqrt{\mathrm{Var}(\nu_i)}}, \tag{2}$$

where $|\mu\rangle$ is a vector with the same dimension as $|\nu_i\rangle$. All the elements of $|\mu\rangle$ are constant, and are equal to the arithmetic average over the elements of $|\nu_i\rangle$. $\mathrm{Var}(\nu_i)$ is the variance over all the elements of $|\nu_i\rangle$. The superscript # in $\left|v_i^{\#}\right\rangle$ indicates that the vector is normalized. Then, the inner product between a pair of normalized vectors is the Pearson's correlation coefficient used for the mirror tree method, which is defined by formula (1). Hereafter, the correlation coefficient by the mirror tree method is denoted as $\rho_{ij}^{\mathrm{MIRROR}}$.

$$\rho_{ij}^{\mathrm{MIRROR}} = \langle \nu_i^{\#} | \nu_j^{\#} \rangle. \tag{3}$$

2.4 First Method with Projection Operator [19]

Consider an $n(n-1)/2$-dimensional unit vector $|u\rangle$, which represents the evolutionary relationship of the source species under consideration. Given such a vector, following projection operator P can be defined:

$$P = I - |u\rangle \langle u|. \tag{4}$$

The projection operator is a matrix with the size of $n(n-1)/2 \times n(n-1)/2$. The method to obtain $|u\rangle$ is explained below. I represents an identity matrix with the size of $n(n-1)/2 \times n(n-1)/2$. By applying the projection operator (4) to a phylogenetic vector, say, $|\nu_i\rangle$, the component within $|\nu_i\rangle$, which is orthogonal to $|u\rangle$, is generated:

$$|\varepsilon_i\rangle = P|\nu_i\rangle = |\nu_i\rangle - |u\rangle \langle u|\nu_i\rangle. \tag{5}$$

$|\varepsilon_i\rangle$ is a residual vector obtained by excluding the information about the evolutionary relationship from the phylogenetic vector . The same projection operator was applied to all of the phylogenetic vectors under consideration. Each of the residual vectors was then normalized with the average and the standard deviation of the elements. The inner product between the two residual vectors $\left|\varepsilon_i^{\#}\right\rangle$ and $\left|\varepsilon_j^{\#}\right\rangle$ represents the Pearson's correlation coefficient between the residual vectors:

$$\rho_{ij}^{\mathrm{PROJECTION}} = \langle \varepsilon_i^{\#} | \varepsilon_j^{\#} \rangle \tag{6}$$

was used as a new measure to evaluate the intensity of co-evolution between proteins i and j.

In order to obtain the unit vector representing the phylogenetic relationship of the source organisms, three different methods were considered. In the first method, 16S rRNA was used for the calculation. Basically, at least one copy of the 16S rRNA gene is encoded by each genome. Therefore, the distance matrix or the phylogenetic vector of the 16S rRNAs is considered to represent the evolutionary relationship among the source organisms. The nucleotide sequences of rRNA were collected from the same sources as the proteins under consideration according to the KEGG GENES database [10] and the Ribosomal Database Project-II Release 9 [1]. The nucleotide sequences of the 16S rRNA were aligned, and the distance between every pair of the aligned nucleotide sequences was calculated by using the F84 model [12] with the DNADIST in the PHYLIP package [2]. The distance matrix was then transformed into a phylogenetic vector $|\nu_{16S}\rangle$. Then, a unit vector $|u_{16S}\rangle$ was obtained as $|\nu_{16S}\rangle/\||\nu_{16S}\|$.

In the second method, all of the phylogenetic vectors of proteins under consideration were normalized so that the size of the elements in each protein was '1' at first. Then, they were averaged as

$$|\nu_{AVE}\rangle = \frac{1}{m}\sum_{i=1}^{m}\frac{|\nu_i\rangle}{\||\nu_i\|}, \tag{7}$$

where m is the number of proteins. So, m was 26 here. The second unit vector $|u_{AVE}\rangle$, was obtained as $|\nu_{AVE}\rangle/\||\nu_{AVE}\|$.

In the third method, the phylogenetic vectors were used again. Let X be a matrix of $n(n-1)/2 \times m$ in which the i-th column corresponds to a normalized phylogenetic vector of protein i. Then, a correlation coefficient matrix Y of $m \times m$ was calculated as $X^T X$. The superscript T indicates the transpose of a matrix. The principal component analysis for the data corresponding to X is equivalent to solving the eigenvalue problem of Y. Then, $|\nu_{PC1}\rangle$ was obtained as $|\nu_{PC1}\rangle = X|z_1\rangle$, where $|z_1\rangle$ is a vector corresponding to the first principal component axis. Then, $|\nu_{PC1}\rangle/\||\nu_{PC1}\|$ generated the third unit vector, $|u_{PC1}\rangle$.

The Pearson's correlation coefficients between the residual vectors for a pair of proteins i and j, which were generated by the projection operations constructed with $|u_{16S}\rangle$, $|u_{AVE}\rangle$ and $|u_{PC1}\rangle$, were represented by ρ_{ij}^{16S}, ρ_{ij}^{AVE} and ρ_{ij}^{PC1}. The type of correlation coefficient is collectively represented by ρ^* without the subscripts, i and j where the superscript indicates the type of correlation coefficient.

2.5 Second Method with Multiple Regression [18]

Suppose that m proteins are given and we want to predict interacting pairs from them. Consider multiple regressions of $|\nu_i\rangle$ and $|\nu_j\rangle$ with $(m-2)$ phylogenetic vectors:

$$|\nu_i\rangle = \alpha_0 + \sum_{k\neq i,j}^{m}\alpha_k|\nu_k\rangle + |\delta_i\rangle, \tag{8}$$

$$|\nu_j\rangle = \beta_0 + \sum_{l\neq i,j}^{m}\beta_l|\nu_l\rangle + |\delta_j\rangle, \tag{9}$$

where α_i and β_j are parameters. The residual vectors, $|\delta_i\rangle$ and $|\delta_j\rangle$, are expected to lack the evolutionary information of the source organisms. Note that $|\nu_j\rangle$ is excluded from the summation on the right side of the equation (8). Likewise, $|\nu_i\rangle$ is excluded from the summation on the right side of the equation (9). The similarity between the two residual vectors is considered to indicate the intensity of co-evolution between proteins i and j. To evaluate the similarity between the residual vectors, the Pearson's correlation coefficient between $|\delta_i\rangle$ and $|\delta_j\rangle$ was calculated. As described above, the inner product between the normalized residual vectors is equivalent to the Pearson's correlation coefficient between them:

$$\rho_{ij}^{\mathrm{PARTIAL}} = \langle \delta_i^{\#} | \delta_j^{\#} \rangle. \tag{10}$$

The correlation coefficient is called the partial correlation coefficient between $|\nu_i\rangle$ and $|\nu_j\rangle$. In actual practice, the following formula was used to obtain the partial correlation coefficient, instead of performing multiple regression.

$$\rho_{ij}^{\mathrm{PARTIAL}} = \frac{-(R^{-1})_{ij}}{\sqrt{(R^{-1})_{ii}}\sqrt{(R^{-1})_{jj}}}, \tag{11}$$

where R is the correlation coefficient matrix whose (i,j)-th element is $\rho_{ij}^{\mathrm{MIRROR}}$, and the superscript -1 indicates inverse. ρ^{PARTIAL} without subscripts, i and j, collectively represents that the type is partial correlation coefficient.

3 Results and Discussions

We calculated five types of correlation coefficients, ρ^{MIRROR}, $\rho^{16\mathrm{S}}$, ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL}, for all of the possible pairs of 26 proteins, that is, 325 pairs of proteins. The performance of each correlation coefficient was evaluated with specificity and sensitivity. Out of the 325 pairs, the interactions of 13 pairs have been experimentally identified. Only top 20 of the five types of correlation coefficients are shown in Table 1, where the actually interacting pairs are highlighted with circles. As shown in the table, the top ranks of $\rho^{16\mathrm{S}}$, ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL} were occupied by pairs of actually interacting proteins. In contrast, non-interacting proteins were present within the top ranks of ρ^{MIRROR}. The decreasing patterns of the five correlation coefficients are seen in this table. The decrease of ρ^{MIRROR} was quite slow, whereas ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL} decreased rapidly. The rate of the $\rho^{16\mathrm{S}}$ decrease was rather moderate. The decreasing patterns shown in Table 1 clearly demonstrates the problem of the original mirror tree method. Even if a high value, e.g. 0.9, is used as a threshold for the correlation coefficient to predict a protein–protein interaction, ρ^{MIRROR} produces many pairs with high correlation, including non-interacting pairs, which likely lead to the generation of many false positives. However, the occupation of the top ranks by interacting proteins and the rapid decreases of $\rho^{16\mathrm{S}}$, ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL} guarantee the specificity of prediction, if the threshold is set at a sufficiently high value.

Table 1. Comparison of top 20 protein pairs sorted in decreasing order of the correlation coefficients

rank	ρ^{MIRROR}		ρ^{16S}		ρ^{AVE}		ρ^{PC1}		ρ^{PARTIAL}	
1	dnaN-rpoB	0.977	sucD-sucC	0.924 O	sucD-sucC	0.910 O	sucD-sucC	0.915 O	sucD-sucC	0.885 O
2	dnaK-secY	0.963	atpA-atpD	0.803 O	trpA-trpB	0.792 O	trpA-trpB	0.754 O	trpA-trpB	0.744 O
3	dnaK-rpoB	0.963	carA-carB	0.803 O	rpoA-rpoB	0.654 O	carA-carB	0.649 O	tufB-tsf	0.612 O
4	sucD-sucC	0.962 O	dnaK-secY	0.801	carA-carB	0.642 O	atpA-atpD	0.640 O	carA-carB	0.597 O
5	dnaN-dnaK	0.960	trpA-trpB	0.797 O	dnaN-rpoB	0.634	dnaN-rpoB	0.587	rpoA-carA	0.519
6	atpA-atpD	0.959 O	dnaE-secA	0.783	atpA-atpD	0.615 O	dnaK-atpD	0.560	dnaN-iscS	0.505
7	rpoA-rpoB	0.958 O	dnaK-atpD	0.774	iscS-iscU	0.607 O	dnaK-secY	0.560	grpE-rpoB	0.462
8	rpoB-secY	0.955	dnaN-rpoB	0.772	grpE-clpP	0.553	iscS-iscU	0.555 O	dnaK-secY	0.457
9	secY-secA	0.954 O	rpoA-rpoB	0.768 O	dnaK-carB	0.541	grpE-clpP	0.545	rpoA-rpoB	0.450 O
10	dnaK-atpD	0.953	dnaN-carA	0.761	grpE-tsf	0.541	dnaK-carB	0.542	atpA-atpD	0.443 O
11	dnaN-secY	0.953	dnaN-dnaK	0.760	dnaK-secY	0.516	secY-carB	0.531	ruvA-ruvB	0.439 O
12	dnaK-atpA	0.952	dnaK-carB	0.758	dnaK-atpD	0.514	dnaK-rpoB	0.500	dnaA-dnaB	0.358 O
13	dnaE-secA	0.945	dnaN-secY	0.758	ruvA-ruvB	0.511 O	rpoA-rpoB	0.498 O	grpE-dnaN	0.355
14	dnaN-rpoA	0.945	dnaE-secY	0.755	secY-carB	0.501	grpE-tsf	0.497	clpP-atpD	0.345
15	dnaK-secA	0.944	rpoB-secY	0.753	rpoB-secY	0.498	dnaK-atpA	0.496	dnaK-tufB	0.339
16	dnaE-secY	0.944	dnaK-atpA	0.747	tsf-trpB	0.476	ruvA-ruvB	0.489 O	dnaB-sucD	0.336
17	dnaN-clpX	0.943	dnaE-dnaK	0.743	dnaA-ruvB	0.469	dnaE-secA	0.480	secA-carB	0.333
18	dnaN-secA	0.940	secY-carB	0.727	secA-trpB	0.453	dnaE-secY	0.477	dnaK-sucC	0.329
19	clpX-rpoB	0.937	iscS-iscU	0.717 O	tufB-tsf	0.445 O	dnaA-ruvB	0.455	atpA-sucC	0.306
20	dnaN-carA	0.937	dnaK-carA	0.716	dnaK-atpA	0.438	secY-secA	0.431 O	clpX-dnaE	0.302

The abbreviated names of the interacting proteins are as follows: sucC-sucD, succinyl-CoA synthetases alpha - beta; atpA-atpD, ATP synthases alpha - beta; rpoA-rpoB, DNA - directed RNA polymerases alpha - beta; secA-secY, preprotein translocase secA - secY; carA-carB, carbamoyl-phosphate synthases small - large; ruvA-ruvB, Holliday junction DNA helicases ruvA - ruvB; iscS-iscU, putative aminotransferase - NifU-like protein; dnaE-dnaN, DNA polymerases III alpha - beta; trpA-trpB, tryptophan synthases alpha - beta; tufB-tsf, elongation factors EF-Tu - EF-Ts; dnaA-dnaB, DNA helicase - dnaA; grpE-dnaK, heat shock protein grpE - dnaK protein; and clpX-clpP, ATP-dependent clp proteases ATP-binding subunit - protease proteolytic subunit.

The unit vector $|u\rangle$ seems to be a crucial factor for the prediction of a protein–protein interaction when a projection operator is used. Therefore, we examined the relationships among $|u_{16S}\rangle$, $|u_{AVE}\rangle$ and $|u_{PC1}\rangle$ by calculating absolute value of Pearson's correlation coefficients $|r|$ among them. $|r|$ between $|u_{16S}\rangle$ and $|u_{AVE}\rangle$ was 0.947, whereas $|r|$ between $|u_{16S}\rangle$ and $|u_{PC1}\rangle$ was 0.946. The highest correlation, $|r| = 0.998$, was observed between $|u_{AVE}\rangle$ and $|u_{PC1}\rangle$. The high correlation between $|u_{16S}\rangle$ and the other unit vectors suggests that the information except for the evolutionary relationship of source organisms can be approximately canceled out by the average operation or principal component analysis.

The ρ^{16S}, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$ seem to outperform the ρ^{MIRROR}. That is, the exclusion of the information about the evolutionary relationships among the source organisms from the distance matrices is effective to reduce the number of the false positives from the mirror tree predictions. The specificities and the sensitivities of the five types of correlation coefficients under four different threshold values, 0.9, 0.8, 0.7 and 0.6, are shown in Table 2. When a pair of proteins had a correlation coefficient greater than the threshold the proteins were predicted to interact with each other. Three types of correlation coefficients, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$, showed high specificity under any threshold value, whereas ρ^{16S} showed high specificity only when threshold was 0.9 or 0.8. The high specificities of ρ^{16S}, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$ mean the drastic reduction of false positives, compared with ρ^{MIRROR} [18,19]. Recently, Pazos *et al.* [13] have independently developed a method to exclude the information of evolutionary relationship among the source organisms by using 16S rRNA. They adjust the scale of the distance matrix of rRNA to that of the distance matrix of a protein, and simply subtract the former from the latter. Then, correlation coefficient is calculated between the sets of residual elements. Improvement in specificity is also observed by their operation, although the mathematical framework of their method is different from those of ours.

Table 2. Specificity and Sensitivity of the prediction

Method	Specificity				Sensitivity			
	0.9	0.8	0.7	0.6	0.9	0.8	0.7	0.6
ρ^{MIRROR}	13.79	6.21	4.96	4.17	61.54	84.62	100.00	100.00
ρ^{16S}	100.00	75.00	28.57	24.32	7.14	21.43	42.86	64.29
ρ^{AVE}	100.00	100.00	100.00	85.71	7.14	7.14	14.29	42.86
ρ^{PC1}	100.00	100.00	100.00	100.00	7.14	7.14	14.29	28.57
$\rho^{PARTIAL}$	–	100.00	100.00	100.00	0.00	7.14	14.29	21.43

$$\text{Specificity} = \frac{\text{true positive}}{(\text{true positive} + \text{false positive})} \times 100\%,$$

$$\text{Sensitivity} = \frac{\text{true positive}}{(\text{true positive} + \text{false negative})} \times 100\%.$$

When threshold was set to 0.9, no interacting pair was predicted with $\rho^{PARTIAL}$, and specificity was not calculated in the case.

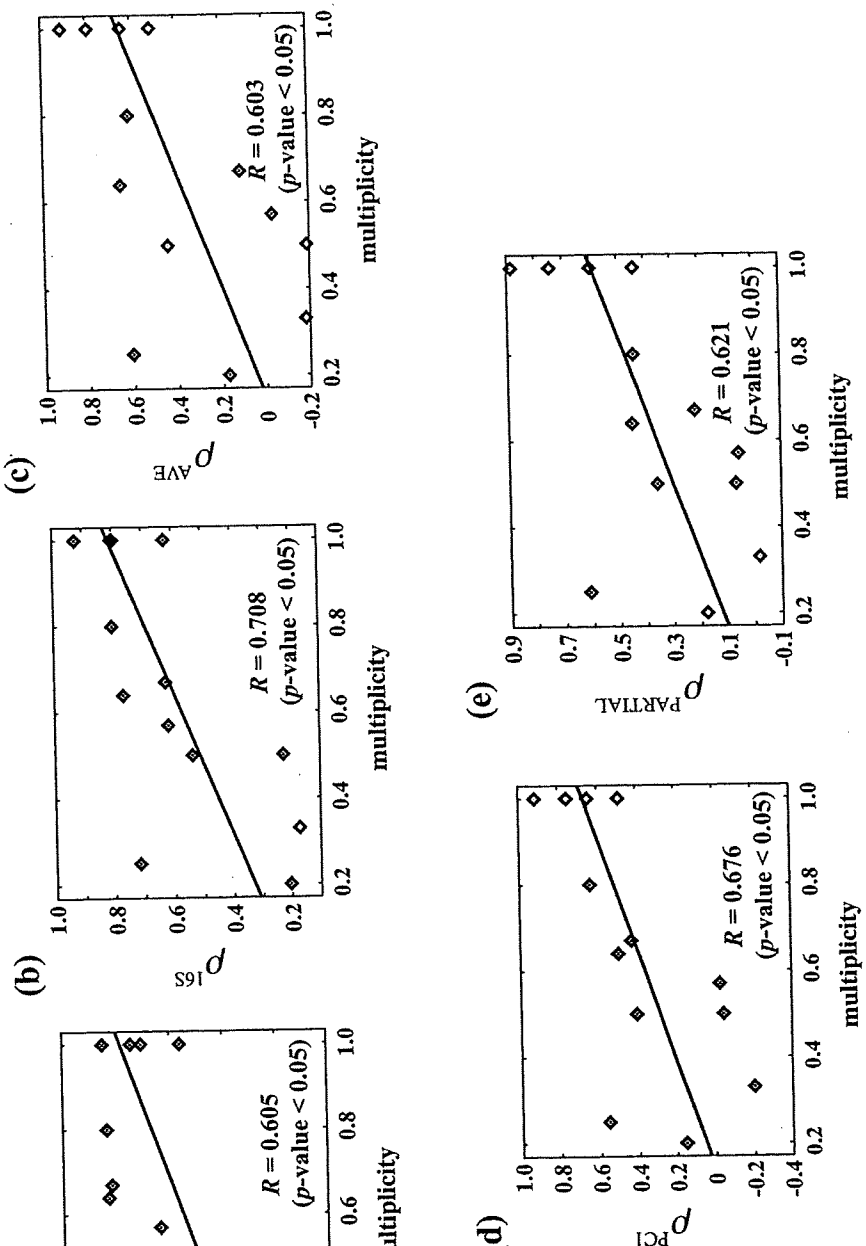

Fig. 1. The relationship between multiplicity and five types of correlation coefficients, ρ^{MIRROR}, ρ^{16S}, ρ^{AVE}, ρ^{PC1} and ρ^{PARTIAL}

Despite the improvement described above, the sensitivities of ρ^{16S}, ρ^{AVE}, ρ^{PC1} and $\rho^{PARTIAL}$ were lower than that of ρ^{MIRROR}. This means that a pair of proteins i and j did not always show high ρ_{ij}^{16S}, ρ_{ij}^{AVE}, ρ_{ij}^{PC1} and $\rho_{ij}^{PARTIAL}$ even when proteins, i and j, interact with each other. In other words, the number of false negatives increased when our methods were used, compared with the original mirror tree method. Here, we calculated the intensity of co-evolution between a pair of proteins as the correlation coefficient after excluding the information about the evolutionary relationship among the source organisms from the phylogenetic vectors. However, the pairs may also interact with other proteins. If such proteins exist, it would be difficult to detect the interaction with the pair, because the co-evolution with the other partners may function as noise for the prediction of interaction of a pair. To examine this hypothesis, we investigated the relationship between the multiplicity of the interaction [19] and the correlation coefficient (Fig 1). The multiplicity, or a modified Jaccard coefficient, is a measure defined between a pair of interacting proteins. Consider an interacting pair of proteins A and B. Let $\mathbf{M}$ and $\mathbf{N}$ be the sets of interaction partners of proteins A and B. The information about the interaction partners were obtained from the DIP database [17]. Protein B belongs to $\mathbf{M}$, whereas $\mathbf{N}$ includes protein A. The multiplicity between proteins, A and B, is defined as follows:

$$\text{Multiplicity (modified Jaccard coefficient)} = \frac{|\mathbf{M} \cap \mathbf{N}| + 1}{|\mathbf{M} \cup \mathbf{N}| - 1}. \tag{12}$$

When proteins A and B interact each other without other interaction partners, multiplicity takes a value 1. When proteins A and B have other interaction partners, the multiplicity decreases. However, when proteins A and B share the other interaction partners, the multiplicity takes a value close to 1. In contrast, when proteins A and B have their own interaction partners respectively, the multiplicity is close to 0. As shown in Fig. 1, the intensities of co-evolution calculated by any method show positive correlation with the multiplicity. That is, the intensities of co-evolution were high when proteins A and B formed a complex without other interaction partners or share the other interaction partners. When proteins A and B had their own interaction partners, that is, the multiplicity was low, the intensities of co-evolution were low. The observation suggests that the false negatives are generated by the presence of unshared interaction partners. Further accumulation of experimental knowledge is required to ascertain this hypothesis.

4 Conclusion

The mirror tree method is a simple approach for the prediction of protein–protein interactions. Here, we reviewed our methods to improve the performance of the original mirror tree method. In the experiment, we confirmed that our methods could drastically reduce the number of false positives in the prediction. Our method, however, generated more false negatives than the original mirror tree method. Our analysis suggested that the presence of unshared interaction

partners may be the cause of the false negatives. However, if we select protein pairs with a high correlation coefficient, e.g. > 0.8, by any one of our methods, we can predict interacting protein pairs with high reliability.

Acknowledgments. This work was supported by Grants-in-Aid for Scientific Research on Priority Areas 'Systems Genomics' (K.H.), 'Comprehensive Genomics' (M.K.) and 'Membrane Interface' (H.T.) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

1. Cole, J.R., Chai, B., Farris, R.J., Wang, Q., Kulam-Syed-Mohideen, A.S., McGarrell, D.M., Bandela, A.M., Cardenas, E., Garrity, G.M., Tiedje, J.M.: The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. Nucleic Acids Res. 35, D169–D172 (2007)
2. Felsenstein, J.: PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle (2004)
3. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147 (2002)
4. Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S., Rothschild, B.: Inferring protein interactions from phylogenetic distance matrices. Bioinformatics 19, 2039–2045 (2003)
5. Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., Cohen, F.E.: Co-evolution of proteins with their interaction partners. J. Mol. Biol. 299, 283–293 (2000)
6. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M.: Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. Nature 415, 180–183 (2002)
7. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. In: Proc. Natl. Acad. Sci. USA 98, pp. 4569–4574 (2001)
8. Jones, D.T., Taylor, W.R., Thornton, J.M.: The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. 8, 275–282 (1992)

9. Jothi, R., Cherukuri, P.F., Tasneem, A., Przytycka, T.M.: Co-evolutionary analysis of domains in interacting proteins reveals insights into domain-domain interactions mediating protein-protein interactions. J. Mol. Biol. 362, 861–875 (2006)
10. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M., Hirakawa, M.: From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res. 34, D354–D357 (2006)
11. Katoh, K., Kuma, K., Toh, H., Miyata, T.: MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 33, 511–518 (2005)
12. Kishino, H., Hasegawa, M.: Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. 29, 170–179 (1989)
13. Pazos, F., Ranea, J.A., Juan, D., Sternberg, M.J.: Assessing protein co-evolution in the context of the tree of life assists in the prediction of the interactome. J. Mol. Biol. 352, 1002–1015 (2005)
14. Pazos, F., Valencia, A.: In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47, 219–227 (2002)
15. Pazos, F., Valencia, A.: Similarity of phylogenetic trees as indicator of protein–protein interaction. Protein Eng. 14, 609–614 (2001)
16. Ramani, A., Marcotte, E.M.: Exploiting the co-evolution of interacting proteins to discover interaction specificity. J. Mol. Biol. 327, 273–284 (2003)
17. Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., Eisenberg, D.: The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 32, D449–D451 (2004)
18. Sato, T., Yamanishi, Y., Horimoto, K., Kanehisa, M., Toh, H.: Partial correlation coefficient between distance matrices as a new indicator of protein-protein interactions. Bioinformatics 22, 2488–2492 (2006)
19. Sato, T., Yamanishi, Y., Kanehisa, M., Toh, H.: The inference of protein-protein interactions by co-evolutionary analysis is improved by excluding the information about the phylogenetic relationships. Bioinformatics 21, 3482–3489 (2005)
20. Sprinzak, E., Sattath, S., Margalit, H.: How reliable are experimental protein–protein interaction data? J. Mol. Biol. 327, 919–923 (2003)
21. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., 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., Rothberg, J.M.: A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000)
22. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399–403 (2002)