Comprehensive Analysis and Prediction of Synthetic Lethality Using Subcellular Locations

Takuji Yamada[1] takuji@kuicr.kyoto-u.ac.jp
Shuichi Kawashima[2] shuichi@hgc.jp
Hiroshi Mamitsuka[1] mami@kuicr.kyoto-u.ac.jp
Susumu Goto[1] goto@kuicr.kyoto-u.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
[2] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

Abstract

The lethality of a gene is a fundamental and representative measure for understanding the function of a gene and its associated bio-systems. Recently, many research groups have started focusing on the concept of synthetic lethality. The synthetic lethality between genes is defined by the combination of mutations in two genes causing cell death. Here, we confirm that synthetic lethality and cellular location have close relationships among the *Saccharomyces cerevisiae* genes. Furthermore, we attempt the prediction of candidate gene pairs with synthetic lethality. The prediction is based on the hierarchical aspect model (HAM) which learns from a data set of cellular location to estimate a likelihood value indicating the synthetic lethality between genes.

Keywords: synthetic lethality, subcellular location, prediction

1 Introduction

In the post-genomic era, analyses about relationships among genes or proteins have become popular topics. There are qualitatively different types of relationships for the same genes and/or proteins (e.g. sequence similarity, protein-protein physical interaction, *etc*). The whole of these relationships form a complex network, which represents the underlying structure of bio-systems. Some research groups have developed high through-put experimental tests to obtain comprehensive data of these relationships [4, 7, 13]. Recently, there have been integrative analyses of these multiple relationships. These approaches are considered effective methods to solve unknown functions of genes or proteins [2, 11]. Furthermore, some integrative methods have been extended for the prediction of the relationships of elements using statistical or mathematical methods [15]. Here, we applied such an integrative method with protein co-localization, synthetic lethality of genes, and protein relationships in the metabolic and regulatory pathways.

Protein co-localization is one of the representative protein features with a long history. Recently, Huh *et al.* [3] analyzed the localization of proteins comprehensively, and they also investigated the co-localization of protein pairs. One of the most important protein roles is catalyzing chemical reactions. A whole set of reactions in a particular species is defined as a metabolic pathway. On the other hand, a regulatory pathway consists of a chain of physical protein-protein interactions, which play a role of information processing. These pathways are comprehensive representations of the relationships between proteins or compounds [5, 6]. The synthetic lethality between genes is defined by the combination of mutations in two genes causing cell death. This has also been studied for a long

time, and lately one group developed high-throughput analysis to investigate the synthetic lethality between genes, so large scale data sets could be obtained [10, 12].

In this analysis, we address the synthetic lethality between genes in *Saccharomyces cerevisiae* and investigate the effects of other relationships (protein co-localization, relationships in the metabolic pathway) to synthetic lethality. We propose that there exists a high correlation between synthetic lethality and protein co-localization. Using high correlation as a measure, we attempted the prediction of synthetic lethality for gene pairs. The prediction method is based on the Hierarchical Aspect Model (HAM) developed by Mamitsuka [8], which estimates the likelihood of query gene pairs. This method takes as input a dataset for learning, which are categorical data of entries and entry pairs with particular features. In this paper, our method takes gene sets with their subcellular locations and gene pairs with synthetic lethality as the learning data sets. We successfully obtained some candidate gene pairs with synthetic lethality. We confirmed this by comparing our results with random data using jack-knife cross-validation. The direct relevance of this prediction is to reduce the time and cost of comprehensive experimental tests for synthetically lethal gene pairs.

We did not utilize the relationships between synthetic lethality and the pathway maps for the prediction because this data is currently not available over all pathways. However, a few of these relationships could be observed in some of the regulatory pathways, such as the MAPK signaling pathway. Some metabolic pathways are also related to the regulatory pathways through synthetically lethal relationships. Our results suggest that the gene relationships affect synthetic lethality between regulatory pathways more than between metabolic pathways. However, the number of regulatory pathways is few compared with metabolic pathways. In the future, our method may take pathways as input for prediction with the improvement of the data set used.

2 Data Set

2.1 Synthetically Lethal Gene Pairs

We used the 6234 *S. cerevisiae* proteins obtained from SGD (Saccharomyces Genome Database) [1], and 2374 gene pairs with synthetic lethality (1140 genes) were extracted from the Tong and Ozier's data [10, 12].

2.2 Subcellular Location of Proteins

The sub cellular locations of the proteins in the *S. cerevisiae* genome were extracted from the MIPS database [9]. Although this database stores information of the subcellular locations of proteins from the literature, we extracted the data published by a single group, Huh *et al.* [3], to obtain data of consistent quality. There are 22 subcellular locations, and the locations for 6077 proteins are defined. Furthermore, for each protein pair, subcellular location pairs are also defined. We constructed the localization pairs of all against all protein pairs.

2.3 Pathway Map from KEGG/PATHWAY

Some proteins catalyze chemical reactions as enzymes, and also play a role in a part of the signal transduction cascade. The KEGG/PATHWAY database contains these reactions and proteins from fully sequenced genomes [5]. Over 100 pathway maps are contained in the database, and we extracted the *S. cerevisiae* proteins from each pathway map. Totally, 76 pathway maps contained 891 *S. cerevisiae* proteins. Each protein is included in particular pathway maps, so pathway map pairs are defined in each protein pair, as was done for subcellular location.

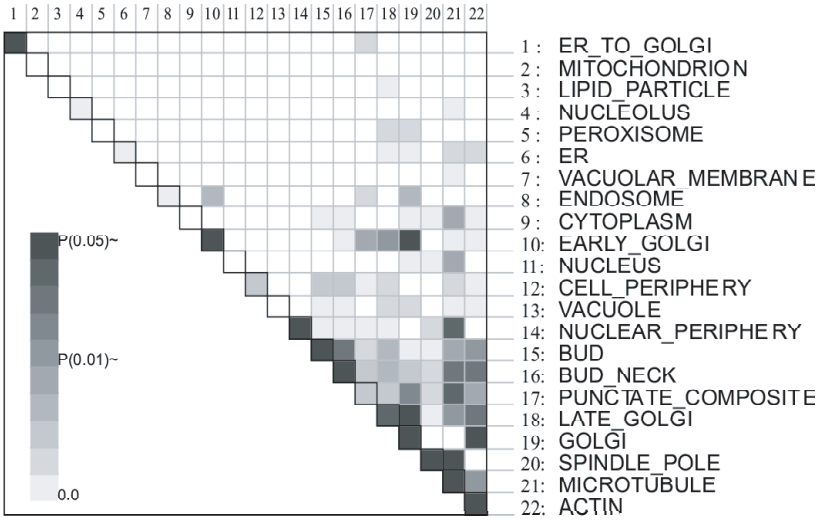

Figure 1: The similarity matrix of subcellular locations. Each cell represents the p-value indicating the synthetic lethality of the location pair. The p-value is calculated from the distribution of the SL-score. The order of the subcellular location obeys the result of the hierarchical clustering using the p-value.

3 Method

3.1 Synthetic Lethality and Subcellular Localization

Each protein pair corresponds to a number of sub cellular location pairs, so each subcellular location pair also has information of protein pairs. We constructed a similarity matrix of the subcellular locations. Each cell in this matrix corresponds to a score indicating the relationship of subcellular location pairs to synthetic lethality. This score, which we call the SL-score, is defined by Sn/Tn (Sn: the number of synthetically lethal gene pairs in a particular location pair, Tn: the number of gene pairs in the subcellular location pair). The number of proteins depends on each subcellular location, so the score is normalized for the bias. We inferred the p-value from the distribution of this score.

Table 1: The list of cellular location pairs which have high SL-score described in the Method section. The first and second columns represent specific subcellular locations (L1, L2), and the third and fourth columns represent Sn and Tn. The fifth and the sixth column represent SL-score and SL-ratio (see Section 3.1). This table was sorted by SL-score, and only the top 20 are shown. Shaded rows represent pairs of the same subcellular location.

L1	L2	Sn	Tn	SL-score	SL-ratio
ER_TO_GOLGI	ER_TO_GOLGI	2	15	0.13333	1169.9309
MICROTUBULE	MICROTUBULE	19	190	0.1	877.4701
ACTIN	ACTIN	16	496	0.03226	283.0718
MICROTUBULE	SPINDLE_POLE	39	1256	0.03105	272.4545
SPINDLE_POLE	SPINDLE_POLE	31	2145	0.01445	126.7944
ACTIN	GOLGI	14	1376	0.01017	89.2387
GOLGI	EARLY_GOLGI	18	1789	0.01006	88.2735
EARLY_GOLGI	EARLY_GOLGI	14	1485	0.00943	82.7454
GOLGI	GOLGI	8	903	0.00886	77.7438
NUCLEAR_PERIPHERY	NUCLEAR_PERIPHERY	16	1830	0.00874	76.6909
BUD_NECK	BUD_NECK	40	4753	0.00842	73.883
BUD	BUD	20	2628	0.00761	66.7755
GOLGI	LATE_GOLGI	14	1978	0.00708	62.1249
MICROTUBULE	PUNCTATE_COMPOSITE	14	2819	0.00497	43.6103
MICROTUBULE	NUCLEAR_PERIPHERY	6	1220	0.00492	43.1715
LATE_GOLGI	LATE_GOLGI	5	1035	0.00483	42.3818
BUD_NECK	BUD	23	5129	0.00448	39.3107
ACTIN	BUD_NECK	13	3136	0.00415	36.415
MICROTUBULE	BUD_NECK	8	1960	0.00408	35.8008

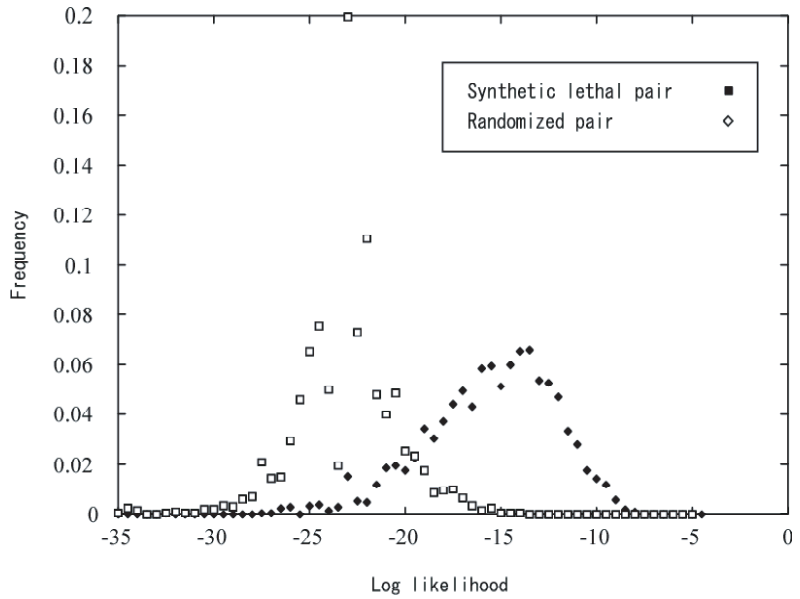

Figure 2: The distribution of the log-likelihood.

3.2 Prediction of Synthetically Lethal Protein Pairs Using HAM

Hierarchical Aspect Model (HAM) is a prediction model for co-occurrence data [8]. It is designed to estimate the likelihood of a certain relationship using categorical data. In this model, entries with categorical data and pairs of these entries are required as the learning dataset. Given a test set (a particular entry pair) to HAM after the learning procedure is completed, the likelihood will be returned according to the categories of the entry pair.

In this paper, the entries and their categories correspond to proteins and their subcellular locations respectively. Furthermore, entry pairs for the learning data are protein pairs which are encoded in gene pairs with synthetic lethality. As a test set to estimate the likelihood, we prepared two kinds of protein pairs. One is synthetically lethal gene pairs. In fact, this data is utilized for learning data set, so we applied jack knife cross validation for each pair. The other is the all against all *S. cerevisiae* gene pairs.

3.3 Synthetic Lethality and Pathway Map Category

We constructed a similarity matrix of the pathway map categories similarly to the previous subcellular locations matrix. However, only a few number of proteins (265 proteins) on the pathway maps related to synthetic lethality, so the scores for each pathway map pair are the number of protein pairs which is encoded by the gene pairs with synthetic lethality.

4 Results

4.1 Synthetically Lethal and Subcellular Localization

Figure 1 illustrates the similarity matrix of subcellular locations described in Section 3.1. The darkness of the color in each cell corresponds to the p-value inferred from the distribution of the scores. The darker the color, the higher the intensity of the synthetic lethality between the subcellular locations.

Generally, the same subcellular location pairs tend to have relatively high scores, which means that the protein pairs with synthetic lethality tend to be located in the same compartment in the cell. On the other hand, some proteins in the ACTIN and the MICROTUBLE compartments have

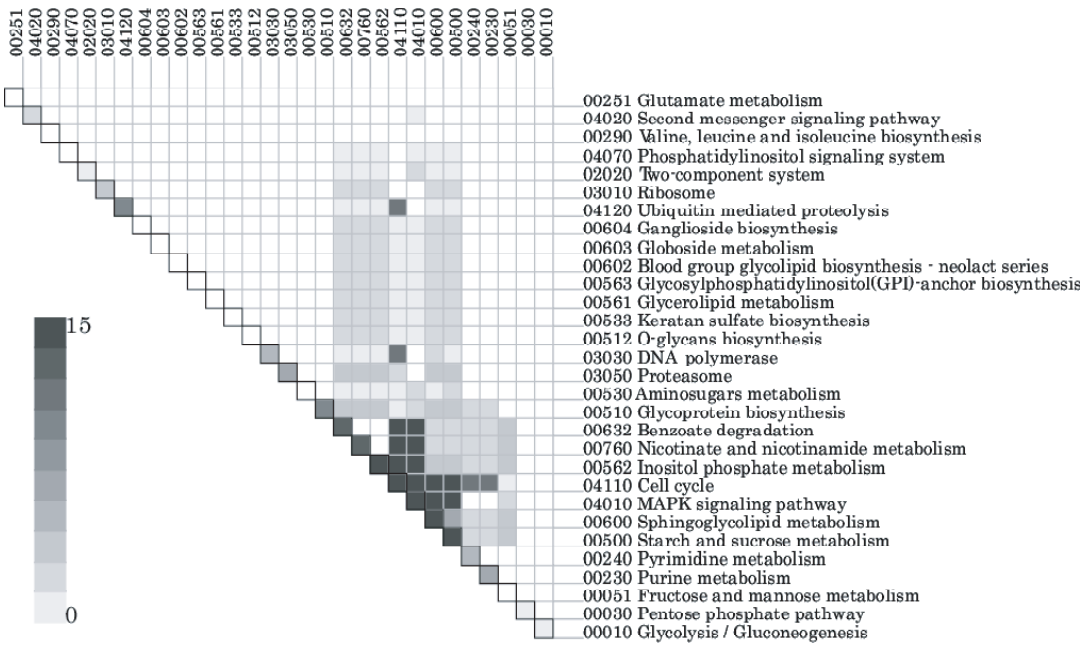

Figure 3: The similarity matrix of pathway maps. Each cell represents to the number of synthetic lethal gene pairs.

strong synthetically lethal relationships with proteins located in other cellular components. Subcellular location pairs with particularly high scores are indicated in Table 1. Rows representing pairs of the same subcellular location are shaded. The second to the last column is the SL-score, as described in Section 3.1. This score also clearly shows that pairs in the same location have higher score; pairs within the ER_TO_GOLGI, MICROTUBULE and ACTIN compartments are especially high.

The ratio of the total number of synthetic lethal gene pairs against the total number of gene pairs is about $1/e^{-5}$. Comparing this ratio with the SL-score in column 5, most of the location pairs are very specialized. This is indicated by values in column 6. For example, ER_TO_GOLGI - ER_TO_GOLGI is over 1000 times more specialized. This result obviously illustrates the relationship between subcellular locations and synthetic lethality.

4.2 Prediction of Synthetically Lethal Pairs

Figure 2 shows the likelihood distribution. Black squares correspond to the likelihoods from the results of the jack knife test of each synthetically lethal gene pair, and white squares correspond to the likelihood from the same number of gene pairs extracted at random from all against all gene pairs.

The likelihoods of the synthetically lethal gene pairs are relatively high compared with randomized pairs. This obvious difference between the two distributions indicates clearly that subcellular location is related to the synthetic lethality between genes.

From the likelihoods of all against all gene pairs, we extracted the top 500 with the highest likelihoods. In this list of gene pairs, some particular genes were observed very frequently. Table 2 shows these frequently appearing genes. These genes are candidates for causing synthetic lethality.

Although high though-put methods for deciding synthetic lethal gene pair has been developed, much cost and time must be spent for every gene pair. Thus, our candidate prediction method is very helpful for inferring the most interesting genes with synthetic lethality.

Table 2: The list of the pathway map pairs which includes much synthetic lethal pairs. Prefix number in the first, and second column is the index of the pathway map in the KEGG/PATHWAY database. Third column is the number of synthetic lethal pairs between these maps. This is sorted by the third column, and only the top 30 was shown. Shaded rows represent pairs of the same subcellular location.

pathway map 1	pathway map 2	#
04110:Cell cycle	04110:Cell cycle	37
00600:Sphingoglycolipid metabolism	04010:MAPK signaling pathway	28
04010:MAPK signaling pathway	00562:Inositol phosphate metabolism	27
04010:MAPK signaling pathway	00760:Nicotinate and nicotinamide metabolism	26
04010:MAPK signaling pathway	00632:Benzoate degradation	26
00500:Starch and sucrose metabolism	04110:Cell cycle	26
00500:Starch and sucrose metabolism	04010:MAPK signaling pathway	26
00600:Sphingoglycolipid metabolism	04110:Cell cycle	25
04110:Cell cycle	00562:Inositol phosphate metabolism	24
04110:Cell cycle	00760:Nicotinate and nicotinamide metabolism	23
04110:Cell cycle	00632:Benzoate degradation	23
04010:MAPK signaling pathway	04010:MAPK signaling pathway	20
00500:Starch and sucrose metabolism	00500:Starch and sucrose metabolism	19
00600:Sphingoglycolipid metabolism	00600:Sphingoglycolipid metabolism	18
04010:MAPK signaling pathway	04110:Cell cycle	16
00562:Inositol phosphate metabolism	00562:Inositol phosphate metabolism	15
00760:Nicotinate and nicotinamide metabolism	00760:Nicotinate and nicotinamide metabolism	14
00632:Benzoate degradation	00632:Benzoate degradation	14
04110:Cell cycle	04120:Ubiquitin mediated proteolysis	13
04110:Cell cycle	03030:DNA polymerase	13
00240:Pyrimidine metabolism	04110:Cell cycle	12
00230:Purine metabolism	04110:Cell cycle	12
04120:Ubiquitin mediated proteolysis	04120:Ubiquitin mediated proteolysis	11
00510:Glycoprotein biosynthesis	00510:Glycoprotein biosynthesis	11
03050:Proteasome	03050:Proteasome	8
00500:Starch and sucrose metabolism	00600:Sphingoglycolipid metabolism	8
00230:Purine metabolism	00230:Purine metabolism	8
03030:DNA polymerase	03030:DNA polymerase	6
00240:Pyrimidine metabolism	00240:Pyrimidine metabolism	6
03010:Ribosome	03010:Ribosome	5

4.3 Synthetic Lethal and Pathway Map Category

Figure 3 illustrates the similarity matrix of the pathway maps described in Section 3.3. The scores in each cell of the matrix are the number of synthetic lethal gene pairs in each pathway map pair. The majority of the pathway map pairs did not include synthetic lethal gene pairs, so this matrix is only a subset of all possible pairs. Table 2 shows the list of pathway map pairs including many synthetic lethal gene pairs. Shaded lines correspond to the same pathway map pairs.

Gene pairs with synthetic lethality tend to belong to the same pathway map. However, several maps, such as Cell cycle and the MAPK signaling pathway, had high scores against many different maps. There are some maps which have particularly high scores in general. This tendency is similar to subcellular location. These higher-scoring maps tend to be those in the regulatory system.

We did not utilize this pathway data as the learning data set for prediction due to the small number of *S. cerevisiae* genes assigned to the pathways (265 entries). However, we claim that there exist relationships between synthetic lethality and the pathway maps based on the bias found in this matrix.

5 Discussion

Figures 1 and 2 showed that there are many synthetically lethal relationships in the same subcellular locations. The synthetic lethality of the genes seems to be attributable to the co-localizations of the proteins. However, we claim that there is an underlying function between these relationships which is related to the relative positions of the proteins in the cell. The construction of a complex is one possible case because proteins constructing a complex are necessarily located in the same subcellular location. YLR208W and YHR098C are one example of a synthetically lethal gene pair which encodes proteins located in ER_TO_GOLGI. These proteins are part of the COPII complex. Likewise, some synthetically lethal protein pairs in ACTIN are often a part of either the ARP3 or the RVS complex.

There were many gene pairs that construct complexes with synthetic lethality. This result indicates that construction of the complex is one of the dominant factors that cause synthetic lethality. However, to understand the details of this phenomenon, more investigation of such things as conformational changes is required. Furthermore, some binding proteins and their targets were also observed in synthetically lethal protein pairs. In particular, proteins containing the SH3-domain and purine-binding domain appeared frequently. Comprehensive analysis of domain distribution in the proteins related to synthetic lethality may be one research topic to gain further understanding of synthetic lethality. Based on the above, protein-protein physical interaction data may be useful as a learning data set for prediction.

Some regulatory pathways have strong relationships with synthetic lethality. The regulatory pathways are composed of the signaling and information processing pathways, which contain more direct interactions compared with metabolic pathways. As mentioned earlier, protein-protein physical interactions may contribute to synthetic lethality. Furthermore, the synthetic lethality of some gene pairs may connect the regulatory and metabolic pathways. These gene relationships have the possibility to clarify the involvement between regulatory and metabolic pathways. Unfortunately, the synthetic lethality of the genes in the pathway maps was in the minority. As described in Section 4.3, one reason may be the number of genes assigned in the pathway. The total number of *S. cerevisiae* genes in the pathway is 891, of which 265 are related by synthetic lethality. On the contrary, we utilized over 6000 entries for protein subcellular locations. 189 of these 265 genes are related to the genes assigned in the pathways. The other 76 genes related to these 265 genes are outside of the pathways. This means that these genes have critical unknown relationships with genes outside of the pathways. These types of synthetically lethal relationships are one of the cross-links from the pathways to other systems. Therefore, this may be a clue for uncovering more complex systems that include the pathways.

The above indicates that there are various factors that induce synthetic lethality. The identification of these other factors is the next step. The information of protein physical interactions and of domains are dominant features. Our prediction method, HAM, can utilize multiple categorical data as an integrated learning data set. In fact, although we used only the information of subcellular locations in the current work, the accuracy of HAM can be improved by adding other factors of synthetic lethality.

6 Conclusion

We investigated the relationship between synthetic lethality and subcellular locations of proteins. We confirm their close relationship because gene pairs with synthetic lethality were located in particular subcellular locations. One factor for the close relationship may be direct protein-protein interactions. Furthermore, we attempted computational prediction, which is based on HAM, of synthetically lethal gene pairs. Taking synthetically lethal gene pair and their subcellular locations as learning data sets, HAM estimates likelihood of synthetic lethality of the gene pair. We can obtain some gene pairs with high likelihood for synthetic lethality. This prediction of candidates enables the reduction of time and costs of experimental tests. In the future, we will improve the prediction method by investigating some additional characteristic features which may be related to synthetic lethality.

Acknowledgments

Authors are grateful to Dr. K. F. Aoki Kinoshita for helpful comments on an earlier draft of the manuscript. This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for Promotion of Science and the Japan Science and Technology Agency. Computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University and the Super Computer system, Human Genome Center, Institute of Medical Science, The University of Tokyo.

References

[1] Christie, K.R. *et al.*, *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms, *Nucleic Acids Res.*, 32:D311–314, 2004.

[2] DeRisi, J.L., Iyer, V.R., and Brown, P.O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[3] Huh, W.K., Falvo, J.V., Gerke, L.C., Carroll, A.S., Howson, R.W., Weissman, J.S., and O'Shea, E.K., Global analysis of protein localization in budding yeast, *Nature*, 425:686–691, 2003.

[4] 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*, 98:4569–4574, 2001.

[5] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M., The KEGG resource for deciphering the genome, *Nucleic Acids Res.*, 32:D277–280, 2004.

[6] Keseler, I.M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I.T., Peralta-Gil, M., and Karp, P.D., EcoCyc: A comprehensive database resource for *Escherichia coli*, *Nucleic Acids Res.*, 33:D334–337, 2005.

[7] Lee, T.I. *et al.*, Transcriptional regulatory networks in *Saccharomyces cerevisiae*, *Science*, 298:799–804, 2002.

[8] Mamitsuka, H., Hierarchical latent knowledge analysis for co-occurrence data, *Proc. the Third SIAM International Conference on Data Mining (SDM 2003)*, 504–511, 2003.

[9] Mewes, H.W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., Munsterkotter, M., Pagel, P., Strack, N., Stumpflen, V., Warfsmann, J., and Ruepp, A., MIPS: Analysis and annotation of proteins from whole genomes, *Nucleic Acids Res.*, 32:D41–44, 2004.

[10] Ozier, O., Amin, N., and Ideker, T., Global architecture of genetic interactions on the protein network, *Nat. Biotechnol.*, 21:490–491, 2003.

[11] Rison, S.C., Teichmann, S.A., and Thornton, J.M., Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in *Escherichia coli*, *J. Mol. Biol.*, 318:911–932, 2002.

[12] Tong, A.H. *et al.*, Global mapping of the yeast genetic interaction network, *Science*, 303:808–813, 2004.

[13] Velculescu, V.E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M.A., Bassett, D.E., Jr., Hieter, P., Vogelstein, B., and Kinzler, K.W., Characterization of the yeast transcriptome, *Cell*, 88:243–251, 1997.

[14] Volchuk, A., Ravazzola, M., Perrelet, A., Eng, W.S., Di Liberto, M., Varlamov, O., Fukasawa, M., Engel, T., Sollner, T.H., Rothman, J.E., and Orci, L., Countercurrent distribution of two distinct SNARE complexes mediating transport within the Golgi stack, *Mol. Biol. Cell*, 15:1506–1518, 2004.

[15] Yamanishi, Y., Vert, J.P., and Kanehisa, M., Protein network inference from multiple genomic data: A supervised approach, *Bioinformatics*, 20:i363–i370, 2004.