Prediction of protein network and functions for yeast using multiple types of genomic data

Yoshihiro Yamanishi[1] yoshi@kuicr.kyoto-u.ac.jp
Tetsuya Sato[1] sato@kuicr.kyoto-u.ac.jp
Jean-Phillipe Vert[2] Jean-Philippe.Vert@mines.org
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] Computational Biology group, Ecole des Mines de Paris, 35 rue Saint-Honoré, 77305 Fontainebleau cedex, France

Keywords: protein network, pathway, multiple genomic data, kernel canonical correlation analysis

1 Introduction and Methods

The problem of predicting a global protein network and protein functions, using all available genomic data about a given organism, is one of the main issues in current computational biology. By protein network, we mean a graph with proteins as vertices and with edges that correspond to various binary relationships between proteins. More precisely we consider below the protein network with edges between two proteins if (i) the proteins interact physically, or (ii) the proteins are enzymes that catalyze two successive chemical reactions in a pathway, or (iii) one of the proteins regulates the expression of the other. This definition of protein network involves various forms of interactions between proteins which should be taken into account for the study of the behavior of biological systems. Unfortunately, the experimental determination of this protein network remains very challenging nowadays. There is therefore an incentive to develop methods in order to predict the protein network from various genomic data generated by high-throughput technologies such as gene expression data [1], physical protein interactions [3], protein localization [2], phylogenetic profiles [5], or regulatory and metabolic pathway knowledge [4].

Recently, we have developed a method to infer protein networks from multiple heterogeneous genomic datasets in a *supervised* context [6]. The systematic experiments highlight the accuracy improvement resulting from the integration of heterogeneous data, and from the supervised learning approach. Because of the sapce limitations we present the detail of the method in a companion paper [6], and we focus in this study on the possible applications of the methods using the result of protein network predicted by our method. We conducted a prediction of biological functions for all hypothetical proteins of the yeast using four datasets: gene expression data, yeast two-hybrid systems, localization data, and phylogenetic profiles. Finally, we confirmed the validity of our prediction with respect to domain.

2 Results and Discussion

We conducted a comprehensive prediction of protein network for all proteins of the yeast. We used four datasets for proteins of *Saccharomyces cerevisiae*: gene expression data obtained from DNA microarrays, protein interaction data obtained by yeast two-hybrid systems, localization data obtained from chromosomally tagged green fluorescent protein fusion proteins, and sequence data encoded into

phylogenetic profiles. The predicted network enabled us to make new biological inferences not only about unknown protein interactions, but also about missing enzymes in biochemical pathways. Next, we predicted the biological functions for hypothetical proteins, based on an assumption that if two proteins are close each other in the predicted network, the proteins are likely to work in similar biochemical pathways or play similar biological functions. We related all the hypothetical proteins to some biological functions. Table 1 shows the examples of candidate proteins and their numbers, which are predicted to be involved in metabolic pathways.

As an example of the function prediction, we focus on the cysteine metabolism pathway. Five proteins known to work in the cysteine metabolism pathway belong to Cys_Met_Meta_PP (Cys/Met metabolism PLP-dependent enzyme) family in the PFAM database. The hypothetical protein YLL058W predicted to work in the cysteine metabolism belong to Cys_Met_Meta_PP family. We guess that the protein YLL058W might work in the cysteine metabolism, because Cys_Met_Meta_PP family is known to be associated with the cysteine metabolism. These results show a possibility of predicting biological functions of hypothetical proteins, and show the usefulness of our method.

Table 1: Examples of the candidate proteins

Metabolism	Number of candidate proteins	Examples
Glutathione	39	YPL225W, YJR119C, etc.
Starch and sucrose	25	YOR059C, YGR149W, etc.
Cysteine	5	YLL058W, YNL191W, etc.
Aminophosponate	4	YDL119C, YNL083W, etc.
Selenoamino acid	1	YJR137C

Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation. The computational resource was provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

[1] Eisen, M.B., Spellman P.T. et al., Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, 95:14863-14868, 1998.

[2] Huh W.K. et al., Global analysis of protein localization in budding yeast, *Nature*, 425:686-691, 2003.

[3] Ito, T, Chiba, T. et al., A comprehensive two-hybrid analysis to explore the yeast protein interacome, *Proc. Natl. Acad. Sci. USA*, 98(8):4569-4574, 2001.

[4] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30:42-46, 2002.

[5] Pellegrini, M., Marcotte, E.M., et al., Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA*, 96:4285-4288, 1999.

[6] Yamanishi, Y., Vert, J.-P., and Kanehisa, M., Protein Network Inference from Multiple Genomic Data: A Supervised Approach, *To appear in Bioinfomatics (ISMB2004)*, 2004.