Prediction of Protein-Protein Interactions from Phylogenetic Trees Using Partial Correlation Coefficient

Tetsuya Sato[1] sato@kuicr.kyoto-u.ac.jp
Yoshihiro Yamanishi[1] yoshi@kuicr.kyoto-u.ac.jp
Katsuhisa Horimoto[2] khorimoto@ims.u-tokyo.ac.jp
Hiroyuki Toh[1] toh@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] Laboratory of Biostastistics, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan

Keywords: protein-protein interactions, phylogenetic tree, co-evolution, partial correlation coefficient

1 Introduction

Computational prediction of protein-protein interactions from the sequence information is an important issue in bioinformatics. In recent years, the phylogenetic profile method [3] has been developed for predicting protein functions and discovering specific protein interactions, and the mirror tree method [2] has been proposed as a generalization of the idea of the phylogenetic profile in order to measure the evolutionary distance between proteins more precisely. Both methods basically stem from the assumption that functionally correlated proteins evolve in a correlated manner. In the mirror tree method, the intensity of the correlation between proteins is evaluated by Peason's correlation coefficient based on phylogenetic trees, but it has been pointed out that a number of false positives tend to be introduced in the prediction.

This paper presents a new method to predict protein-protein interactions from the evolutionary information by using partial correlation coefficient in order to extract direct interactions rather than indirect interactions between proteins. Our method is successfully tested on its ability to predict physical protein interactions, from the comparison of phylogenetic trees of proteins. Using the dataset containing known interactions, we show that our approach removes false positives to a large extent and improves the specificity for predicting physical protein-protein interactions.

2 Method

2.1 Datasets

The dataset in this study is constructed from the DIP database [4] and the KEGG database [1]. For the protein interaction data, we used 13 interacting protein pairs of *Escherichia coli* in the DIP database, which is a repository of experimentally determined interactions between proteins. For the protein sequence data, we selected the orthologous proteins across 24 different species (e.g., proteobacteria and bacillales) from the KEGG/GENES and the KEGG/SSDB databases. We used ClustalW for the multiple sequence alignment, and the JTT matrix as a similarity score matrix whose elements represent similarity scores between amino acids.

2.2 Methodology

Our method is based on a combination of the ordinary mirror tree method [2] and the use of partial correlation coefficient. Proposed procedure in this study is summarized as follows:

Step 1. Construct the distance matrix representing the phylogenetic tree for a protein based on multiple sequence alignment across different species.

Step 2. Transform the off-diagonal elements in the distance matrix into a vector, which we refer to as phylogenetic vector.

Step 3. Repeat Step 1-2 for all the proteins, and obtain a set of phylogenetic vectors for the protein set.

Step 4. Compute partial correlation coefficients based on the phylogenetic vectors for all possible combinations of the proteins.

Step 5. Select high scoring protein-protein pairs as candidates of protein interactions.

3 Results and Discussion

Figure 1 shows the distribution of the scores for both Peason's correlation coefficients (left figure) and the partial correlation coefficients (right figure). In the figures, vertical axes indicate the correlation scores, and horizontal axes indicate 325 possible pairs of proteins. We selected the high scoring protein-protein pairs and compared them with the known protein-protein interactions stored in the DIP database. True positive hits are highlighted by asterisks in the figures. It is found that our method has effects of increasing true positives and reducing false positives in terms of the number of correctly detected protein-protein interactions in comparison with Peason's correlation method. For instance, when top ranking five predictions are examined, the prediction accuracy is 80% (4/5) in the case of partial correlation, but it is 20% (1/5) in the case of Peason's correlation. So these results suggest that the use of partial correlation coefficient can improve the specificity for predicting physical protein interactions.

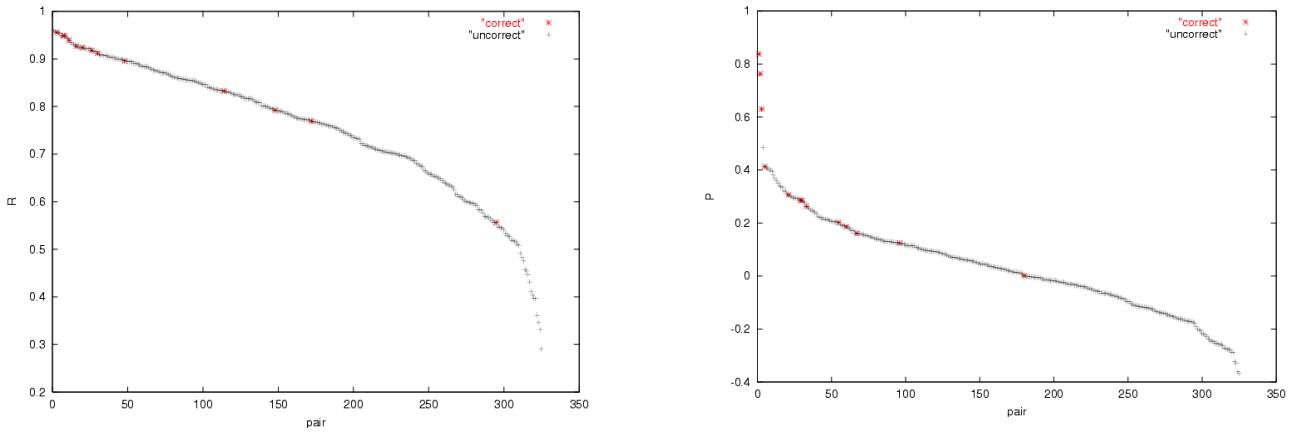

Figure 1: The distributions of the correlation coefficients (left) and partial correlation coefficients (right).

Acknowledgments

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] Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A., The KEGG databases at GenomeNet, *Nucleic Acids Res.*, 30(1):42–46, 2002.

[2] Pazos, F. and Valencia, A., Similarity of phylogenetic trees as indicator of protein-protein interaction, *Protein Eng.*, 14(9):609–614, 2001.

[3] Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D., and Yeates, T.O., Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc. Natl. Acad. Sci. USA.*, 96(8):4285–4288, 1999.

[4] Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., and Eisenberg, D., DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.*, 30(1):303–305, 2002.