Prediction of Protein-Protein Interactions Based on Real-Valued Phylogenetic Profiles Using Partial Correlation Coefficient

Tetsuya Sato[1]
sato@kuicr.kyoto-u.ac.jp

Yoshihiro Yamanishi[1]
yoshi@kuicr.kyoto-u.ac.jp

Minoru Kanehisa[1]
kanehisa@kuicr.kyoto-u.ac.jp

Hiroyuki Toh[1]
toh@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

Keywords: protein-protein interactions, phylogenetic profile, partial correlation coefficient

1 Introduction

The improvement of computational prediction method for protein-protein interactions (PPI) through genome comparisons is an important issue in bioinformatics. The phylogenetic profile method [2] has previously been developed for predicting protein functions and discovering specific PPI. This method basically stems from an assumption that interacting protein pairs are likely to have similar patterns of gene inheritance under the evolutionary restriction depending on PPI. Consider (n+1) different organisms whose genomes are completely sequenced. Suppose that one of the organisms has m genes in the genome, presence or absence of the orthologous gene for each of the m genes is examined for the remaining n organisms. If we assign an integer 1 to the presence and in integer 0 to the absence, we can construct an n-dimensional bit vector for a gene, each element of which corresponds to one of the n organisms. The bit vector is called a phylogenetic profile of the gene. When a pair of genes shares the same or similar bit patterns in their phylogenetic profiles, the gene products are predicted to interact each other. One of the recent progresses for phylogenetic profile method is the development of real-valued profiles [1]. The profiles are represented by continuous numerical values such as sequence alignment scores or its p-values instead of binary values. This method enables us to evaluate the similarity between profiles as Pearson's correlation coefficient. When a correlation coefficient matrix is given, the detailed information about the interactions between the variables is obtained by calculating the partial correlation coefficient matrix. In our previous work, we have shown that partial correlation coefficient is a useful statistical measure for predicting PPI with high accuracy [3].

In this paper, we developed a new method to predict PPI by using the partial correlation coefficients in order to improve the accuracy of the phylogenetic profile method based on the continuous numerical values. The ability of our method to predict PPI was tested by using real-valued phylogenetic profile constructed from the comparison of completely sequenced genomes. The result suggested that our method could improve the accuracy of the prediction of PPI.

2 Method

2.1 Dataset

The dataset used in this study was re-constructed from the entries of the DIP database [4] and the KEGG/GENES database [5]. At first, 13 pairs of interacting proteins (26 proteins) of *Escherichia coli* were selected from the DIP database, which is a repository of experimentally identified interactions between proteins. Then, we selected the putative orthologues of the 26 proteins of *E. coli* from 205 different species in the KEGG/GENES. Finally, a set of real-valued phylogenetic profiles was constructed as described below.

2.2 Construction of real-valued phylogenetic profiles

We used the SSEARCH program for computing the pairwaise sequence alignment scores between proteins of *E. coli* and the putative orthologues. The real-valued phylogenetic profile was constructed with the alignment

scores. Each element in the phylogenetic profile was normalized as S/S_{max}, where S is the Smith-Waterman (SW) score between a protein of *E. coli* and a putative orthologue from a different organism, and S_{max} is the SW score for alignment of the *E. coli* protein with itself.

2.3 Procedure

Our method is based on a combination of the phylogenetic profile method [1,2] and partial correlation coefficients. The procedure developed in this study is summarized as follows:

1. Construct a set of real-valued phylogenetic profiles for the genes of *E. coli*.
2. Compute the Peason's correlation coefficient matrix in the real-valued phylogenetic profiles among the genes from *E. coli*.
3. Compute the partial correlation coefficient matrix on the basis of the Peason's correlation coefficient matrix.
4. Select gene pairs with high partial correlation coefficients as the candidates for interacting pairs of proteins.

3 Results and Discussions

To test the performance of our method, we compared the result of prediction using the ordinary correlation coefficients with those using the partial correlation coefficients. The prediction accuracy of both methods is assessed on the gold standard dataset in this study, by their capacity to detect PPI. The Figure 1 shows the ROC curves representing the number of true positives (predicted interactions that are indeed present in the gold standard) as a function of the number of false positives (predicted interactions that are absent from the gold standard). The measure of ROC curve shows that, the upper area of the figure the curve lies in, the higher the accuracy is. It seems that the partial correlation has effects of increasing the ratio of true positives in the bottom-left area in a large extent, although the ordinary correlation slightly works better in the upper-right area. In the application to the real-world problem, we focus on only high scoring protein-pairs, which are predicted with high reliability, shown in bottom-left area in Figure 1. These results, therefore, suggest that the use of partial correlation coefficients is a useful tool for predicting PPI computationally.

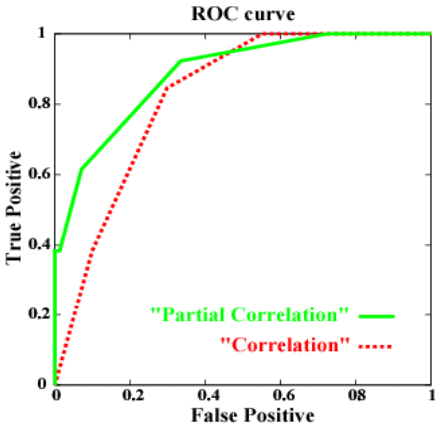

Figure 1: The ROC curves. The solid line and dot line indicate the partial correlation and ordinary correlation coefficients, respectively.

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] Marcotte, E. M., Xenarios, I., van Der Bliek, A. M., Eisenberg, D., Localizing proteins in the cell from their phylogenetic profiles, *Proc. Natl. Acad. Sci. U S A.*, 97(22): 12115-12120, 2000.

[2] 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. U S A.*, 96(8): 4285-4288, 1999.

[3] Sato, T., Yamanishi, Y., Horimoto, K., Toh, H., and Kanehisa, M., Prediction of Protein – Protein interactions from Phylogenetic Trees Using Partial Correlation Coefficient, *Genome Informatics*, 14: 496-497, 2003.

[4] http://dip.doe-mbi.ucla.edu/

[5] http://www.genome.jp/kegg/