Relationship between Data Size and Accuracy of Prediction of Protein-Protein Interactions by Co-Evolutionary Information

Tetsuya Sato\textsuperscript{1} \texttt{sato@kuicr.kyoto-u.ac.jp} \quad Yoshihiro Yamanishi\textsuperscript{1} \texttt{yoshi@kuicr.kyoto-u.ac.jp} \quad Hisako Ichihara\textsuperscript{1} \texttt{ichihara@kuicr.kyoto-u.ac.jp}

Kei-ichi Kuma\textsuperscript{1} \texttt{kuma@kuicr.kyoto-u.ac.jp} \quad Minoru Kanehisa\textsuperscript{1} \texttt{kanehisa@kuicr.kyoto-u.ac.jp} \quad Hiroyuki Toh\textsuperscript{2} \texttt{toh@bioreg.kyushu-u.ac.jp}

\textsuperscript{1} Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
\textsuperscript{2} Division of Bioinformatics, Medical Institute of Bioregulation, Kyushu University, 3-1-1, Maidashi, Higashi-ku, Fukuoka, Fukuoka 812-8582, Japan

Keywords: protein-protein interactions, projection operator, partial correlation coefficient, co-evolution

1 Introduction

The prediction of protein-protein interaction (PPI) with genomic information is an important issue of bioinformatics. Mirror tree is a method to predict PPIs by evaluating the similarity of the phylogenetic trees or distance matrices [1]. In this method, the intensity of the co-evolution between a pair of proteins is evaluated by Pearson's correlation coefficient between a pair of distance matrices of the proteins. However, it has been recognized that predictions by the mirror tree method include many false positives. We have developed two different methods to reduce the number of false positives. One of them uses a projection operator [3], and the other method calculates a partial correlation coefficient [2], to exclude the information about the evolutionary relationship among the source organisms from the distance matrices. Both methods succeeded to drastically reduce the number of false positives, as compared with the mirror tree method. In this paper, we examined the relationship between the prediction accuracy and the data size by using the three methods, original mirror tree method, the method with projection operators and that with partial correlation coefficients.

2 Materials and Method

The dataset used in this study was re-constructed from the entries of the Database of Interacting Proteins (DIP) database [4] and the KEGG GENES database [5]. At first, we selected 821 pairs of interacting proteins (179 proteins) of \textit{Escherichia coli} from the DIP database, which is a repository of experimentally identified interactions between proteins. Then, the putative orthologues corresponding to the 179 proteins derived from \textit{E. coli} were collected from 65 different bacterial species, according to the description in the KEGG KO database [5].

The procedure used in this study is summarized as follows:

1. Construct a distance matrix representing the phylogenetic tree from a multiple sequence alignment of orthologous proteins.
2. Transform the off-diagonal elements in a distance matrix into a vector, which we refer to as a phylogenetic vector.
3. Repeat 1-2 for all the target proteins and obtain a set of the phylogenetic vectors for the proteins under consideration.
4. Calculate four types of correlation coefficients: a Pearson's correlation coefficient by the original mirror tree approach, two Pearson's correlation coefficients by the projection method, where projection operators were obtained by averaging the phylogenetic vector and by applying the
principal component analysis to the phylogenetic vector, and a partial correlation coefficient. For more
detail, see our papers [2, 3].
5. Apply the procedures 1-4 to four datasets of different sizes and evaluate the accuracy of the
predictions with the four types of correlation coefficients.

3 Results and Discussion

The relationship between the prediction accuracy and four different sizes of datasets was examined, changing the
data size from 10, 20, 40 to 80. For each dataset, the accuracies of the predictions with the four types of
correlation coefficients were calculated under four threshold values, 0.9, 0.8, 0.7 and 0.6. The performances
of the four methods were evaluated with sensitivity and specificity (Figure 1). When two proteins have a
correlation coefficient or partial correlation coefficient greater than a given threshold value, the pair was
predicted to interact each other. The prediction with partial correlation coefficient showed highest specificity under
any threshold among the four methods. The specificity of the prediction with partial correlation coefficient
increases and those by projection method decreases, as the size of data increases. In contrast, mirror tree showed
the highest sensitivity in any case, which was independent from the size of data. These results suggest that the
prediction with partial correlation coefficients is more efficient to reduce the number of false positives, particularly
when a large size of dataset is available.

Figure 1: The relationship between the prediction accuracy and the size of PPI datasets.

Acknowledgments

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

References

distance matrices as a new indicator of protein-protein interactions, Bioinformatics , in press.
coevolutionary analysis is improved by excluding the information about phylogenetic relationships,