Comparison of Prediction Methods for Protein-Protein Interactions Using Co-Evolutionary Information

Tetsuya Sato 1
sato@kuicr.kyoto-u.ac.jp
Yoshihiro Yamanishi 2
yoshi@cg.ensmp.fr
Hisako Ichihara 1
ichihara@kuicr.kyoto-u.ac.jp

Minoru Kanehisa 1
kanehisa@kuicr.kyoto-u.ac.jp
Hiroyuki Toh 3
toh@bioreg.kyushu-u.ac.jp

1 Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
2 Centre de Geostatistique, Ecole des Mines de Paris, 35 rue Saint-Honore, 77305 Fontainebleau cedex, France
3 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 development of computational prediction method for protein-protein interaction (PPI) using the genomic information is an important issue in bioinformatics. Mirror tree was recently proposed as a method to predict PPIs from the similarity of the phylogenetic tree 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. To solve this problem, we have developed the two different methods to improve the mirror tree prediction by using partial correlation coefficients and projection operators. In our previous work, we improved the mirror tree method by excluding the information about the phylogenetic relationships from the co-evolutionary information, using the projection operator [3]. We also showed that the partial correlation coefficient is a useful statistical measure for predicting PPIs with high accuracy [2].

In this paper, we compared the prediction accuracy among the three methods, original mirror tree method, the method with projection operators and that with partial correlation coefficients. The ability and characteristics of these methods to predict PPIs were demonstrated with the dataset of pairs of physically contacting proteins.

2 Method

2.1 Dataset

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 13 pairs of interacting proteins (26 proteins) of Escherichia coli from the DIP database, which is a repository of experimentally identified interactions between proteins. Then, the putative orthologues corresponding to the 26 proteins derived from E. coli were collected from 40 different bacterial species, according to the description in the KEGG/KO database [5]. We also extracted the 16S ribosomal RNA sequences derived from the same source species as the proteins from the KEGG/GENES database [5] to infer the information about the phylogenetic relationship among the source species.

2.2 Procedure

Our methods were developed in the framework of the mirror tree method. The procedure 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 the 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 protein set.
4. Calculate Pearson’s correlation coefficient between phylogenetic vectors as the result of mirror tree.
5. Calculate Pearson’s correlation coefficient between the residual vectors after applying the projection operator constructed from the distance matrix of the 16S ribosomal RNA genes to the phylogenetic vectors.
6. Calculate the partial correlation coefficient matrix from the correlation coefficient matrix, which contains Pearson’s correlation coefficient of every pair of protein.

3 Results and Discussions

To investigate how different threshold values affect the prediction accuracy, we introduced three thresholds for the three correlation coefficients, 0.8, 0.7 and 0.6. The performance of the mirror tree method and our methods were evaluated with regard to sensitivity and specificity (Table 1). When two proteins have a correlation coefficient or partial correlation coefficient greater than the threshold value, the pair was predicted to interact with each other. The advantage of the partial correlation was high specificity for any threshold. In contrast, mirror tree showed high sensitivity in any case. The projection operator indicated unique results that the advantage of this method was high specificity under the threshold of 0.8 and high sensitivity under the threshold of 0.6. These results, therefore, suggest that we should switch the prediction method according to what is required.

<table>
<thead>
<tr>
<th>Method</th>
<th>0.8</th>
<th>0.7</th>
<th>0.6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>Sensitivity</td>
</tr>
<tr>
<td>mirror tree</td>
<td>84.62</td>
<td>6.21</td>
<td>100.00</td>
</tr>
<tr>
<td>projection operator</td>
<td>23.08</td>
<td>75.00</td>
<td>46.15</td>
</tr>
<tr>
<td>partial correlation</td>
<td>7.69</td>
<td>100.00</td>
<td>15.38</td>
</tr>
</tbody>
</table>

Sensitivity = true positive / (true positive + false negative) × 100%
Specificity = true positive / (true positive + false positive) × 100%

Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, 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