Evaluating Protein Sequence Signatures Inferred from Protein-Protein Interaction Data Using Gene Ontology Annotations

Osamu Maruyama\textsuperscript{1} om@math.kyushu-u.ac.jp
Hideki Hirakawa\textsuperscript{2} hh@kazusa.or.jp
Takao Iwayanagi\textsuperscript{3} tiwayana@lab.nig.ac.jp
Yoshiko Ishida\textsuperscript{4} yoshiko.ishida.aq@hitachi.com
Shizu Takeda\textsuperscript{4} shizu.takeda.me@hitachi.com
Jun Otomo\textsuperscript{4} jun.otomo.aq@hitachi.com
Satoru Kuhara\textsuperscript{5} kuhara@grt.kyushu-u.ac.jp

\textsuperscript{1} Faculty of Mathematics, Kyushu University, 744 motooka, nishi-ku, Fukuoka-city 819-0395, Japan
\textsuperscript{2} Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan
\textsuperscript{3} R&D Solution Center, Research & Development Group, Hitachi, Ltd., 1-6-6 Marunouchi, Chiyoda-ku, Tokyo 100-8280, Japan (Present address: Center for Information Biology and DNA Data Bank of Japan, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, JAPAN)
\textsuperscript{4} Central Research Laboratory, Hitachi, Ltd., 1-280 Higashi-koigakubo, Kokubunji-shi, Tokyo 185-8601, Japan
\textsuperscript{5} Graduate School of Genetic Resources Technology, Kyushu University, Hakozaki, Higashi-ku, Fukuoka-shi, Fukuoka 812-8581, Japan

Keywords: protein-protein interaction, interacting site, gene ontology

1 Introduction

Protein-protein interactions (PPIs) are involved in all cellular processes, e.g. signal transduction, gene expression, and molecular transmission. A protein can express its function through interacting with other proteins or molecules. Therefore, the detection of interaction sites would be crucial to understanding the mechanism of interaction of protein and to further applications of proteins like drug design.

2 Methods

In this paper, we propose a novel systematic method of predicting protein sequence signatures from PPIs of a target species (see Fig. 1). For a particular protein, called a host protein, we collect all the proteins that are known in some PPI database to be interacting partners with the host protein. From the sequences of the guest proteins, statistically significant signatures are found. Then, for each of the found signatures, two kinds of sets of proteins can be identified: (i) the set of the guest proteins whose sequences match the signature and (ii) the set of all the non-guest proteins whose sequences match the signature, found from the proteome of a target species. Notice that these two sets are mutually exclusive. Next, the functional similarity between these two groups is calculated. The similarity measure we use here is based on gene ontology, proposed by [3], which has shown to be effective in predicting whether given two proteins interact with each other. The Gene Ontology project, or
GO, provides a controlled vocabulary to describe gene and gene product attributes in any organism [1]. Although the assignments of annotations are incomplete and are ongoing due to the dynamic nature of the GO project, these annotations are already applicable to evaluating and predicting many interactions between proteins. For example, it has been reported that the gene ontology annotations assigned to proteins interacting with each other are often identical or similar to each other [4, 2] because interacting proteins have the same function or closely related functions. Finally, statistical significances of signatures are calculated. By this way, we try to find interesting signatures from a PPI data set. Furthermore, this method can provide good candidates for new interacting partners of the host protein. They are members of the protein group of (ii), i.e., those having occurrences of the signatures but being unknown in the PPI database we extract the guest proteins whether they interact with the host protein.

3 Results

The set of PPIs we use here is our original experimental data determined by high-throughput screening yeast two-hybrid (HTS-Y2H) assays, whose previous version has already been used to find drug-targetable PPIs [5]. One of the features of the data is that it includes relatively less false positives due to the rigorous procedures to determine interactions of proteins. In our computational experiments on the PPI data set, we have succeeded in detecting interesting signatures, some of which correspond to interacting sites. Thus, our method is empirically shown to be useful.

References


