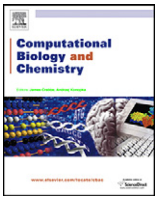
Contents lists available at SciVerse ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem



Semantically predicting protein functions based on protein functional connectivity

Wei Zhu[a], Jingyu Hou[b,*], Yi-Ping Phoebe Chen[a,*]

[a] *Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Australia*
[b] *School of Information Technology, Deakin University, Melbourne, Australia*

ARTICLE INFO

Article history:
Received 12 September 2012
Received in revised form 3 January 2013
Accepted 22 January 2013

Keywords:
Protein function
Function prediction
Functional connectivity

ABSTRACT

Background: The current availability of public protein–protein interaction (PPI) databases which are usually modelled as PPI networks has led to the rapid development of protein function prediction approaches. The existing network-based prediction approaches mainly focus on the topological similarities between immediately interacting proteins, neglecting the protein functional connectivity which is the functional tightness between proteins. In this paper, we attempt to predict the functions of unannotated proteins based on PPI networks by incorporating the protein functional connectivity, as well as the similarity of protein functions, into the prediction procedure.
Results: An approach named *S*emantic protein function *P*rediction based on protein *F*unctional *C*onnectivity (SPFC) is proposed to achieve a higher accuracy in predicting functions of unannotated protein. We define the functional connectivity and function addition for each protein, and incorporate them into the prediction. We evaluated the SPFC on real PPI datasets and the experiment results show that the SPFC method is more effective in function prediction than other network-based approaches.
Conclusion: Incorporating the functional connectivity of each protein into the function prediction can significantly improve the accuracy of protein prediction.

1. Introduction

Protein–protein interaction (PPI) is one of the most important tasks required for a living cell to carry out its biological functions such as DNA replication, transcription, translation, and signal transduction (Hartwell et al., 1999). The classic PPI laboratory detecting technique is yeast two-hybrid (Y2H) system developed by Fields and Song (1989). The principle of the system is to split a transcription factor into two separate domains (which is called binding domain (BD) and activating domain (AD)), and then bind two proteins (the binding proteins are named “bait” and “prey” protein) with the two domains, there is another gene named “report gene” to test whether “bait” and “prey” protein interacts or not. If the two proteins interact with each other when they physically get close, the transcript factor will start to regulate the expression of “report” gene. Therefore, detect the expression product of “report” gene can effectively determine whether two proteins interact or not. Most reliable available PPI resources are originated from laboratory technique. It also can be seen from the description of Y2H that protein functions cannot be determined in the laboratory detecting process. In this work, we refer to the proteins that have already-known functions as the *annotated proteins*, and the proteins whose functions have not been documented as the *unannotated proteins*.

In order to predict the functions of unannotated proteins, some effective approaches in gene sequence level have been developed. One way to do this, named sequence alignment approach is to search similar sequences based on the alignment of nucleotide or amino acid. Another way is the sequence motif searching, which aims to find similar patterns in proteins for function predictions. The most representative methods are the BLAST (Altschul et al., 1990) and PROSITE (Pearson, 1990). Research indicates that more than 30% of unannotated protein functions can be identified by searching for homologues proteins (Ofran et al., 2005). However, it is difficult to determine the sequence similarity of a protein with other proteins. Therefore, the sequence alignment search and motif search approaches cannot effectively predict functions in most cases. Structure-based approaches, the most popular of which are FATCAT (Ye and Godzik, 2004) and ProCAT (Zhu et al., 2006), have been developed structure-based approaches (Skolnick and Fetrow, 2000; Baker and Sali, 2001; O’Donoghue et al., 2001) can predict protein functions based on the exhibited function when a protein folds. Although structure-based approaches are effective in predicting protein functions for some cases, due to a lack of enough protein structure data, only 20–50% of function prediction accuracy can be achieved by the structure-based approaches (Sleator and

* Corresponding authors.
E-mail addresses: w6zhu@students.latrobe.edu.au (W. Zhu), jingyu@deakin.edu.au (J. Hou), phoebe.chen@latrobe.edu.au (Y.-P.P. Chen).

1476-9271/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2013.01.002

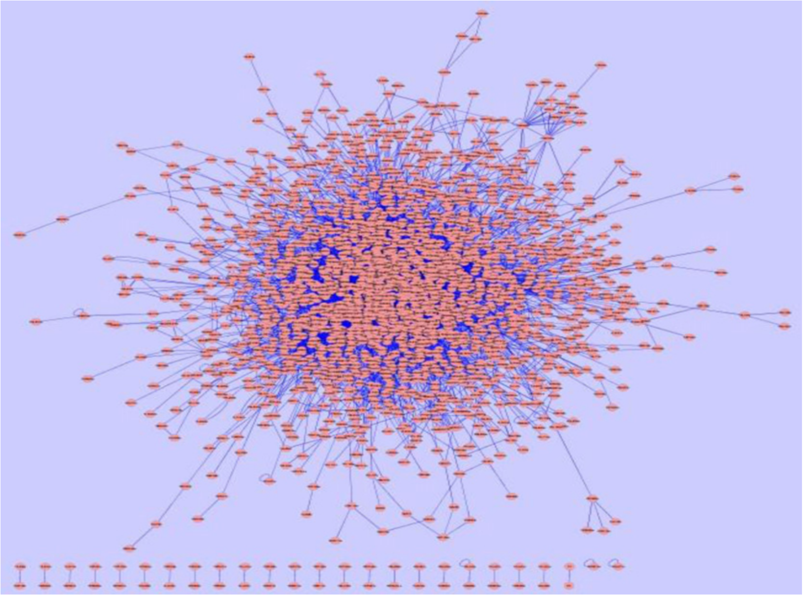


Fig. 1. Yeast PPI network sample (drawn by cytoscape: www.cytoscape.org/).

Walsh, 2010). To overcome the limitations of the above approaches, researchers tried to find other more effective approaches to predict protein functions. To this end, an indirect approach to predict protein functions by discovering the knowledge of interactions between proteins was developed.

Protein–protein interactions can be modelled as an undirected network named PPI network (a PPI network sample is presented in Fig. 1). A PPI network G consists of vertices V and edges E, e.g., $G=(V, E)$, where each vertex denotes a protein and each edge denotes the interaction or correlation between two proteins. Each protein in a PPI network has its functions. For proteins P, P_0, P_1 and P_2, if protein P has an interaction with protein P_0 in a PPI network, then P_0 is the *neighbour* of protein P. The number of neighbours of protein P is named connectivity of protein P. If protein P_1 interacts with P_0 but does not directly interact with P, then P_1 is a *layer 2 neighbour* of P. Use the same way, we can define *layer n neighbour* of protein P. Protein similarity is defined to measure the extent of tightness between two interacting proteins. The connectivity and similarity are two important properties in protein function prediction based on PPI network. The purpose of protein function prediction based on the PPI network is to pursue the functions of unannotated proteins by exploiting the PPI network properties.

Based on the PPI network scale used by the prediction algorithms, we divide the previous approaches into global network-based approaches and local network-based approaches. Global network-based approaches focus on the whole PPI network when predicting functions. These approaches usually use hierarchical or graph clustering methods to infer unannotated protein functions (Maciag et al., 2006; Adamcsek et al., 2006; Pandey et al., 2009; Bork et al., 2004). Local network-based approaches, however, usually focus on the direct (Samanta and Liang, 2003; Deng et al., 2003; Schwikowski et al., 2000) or indirect (Chua et al., 2006) neighbours of the unannotated proteins to predict protein functions.

No matter whether the approaches are global network-based or local network-based, how to determine the similarity between proteins in the PPI network is a key to improve the prediction accuracy. The protein similarity definition varies in different predicting algorithms. In the early stage, the protein similarity is defined according to the existence of an association between two proteins. If an association between two proteins exists, the similarity is “1”; otherwise the similarity is “0”. Some researchers use the number of shared neighbour proteins of two proteins to measure the similarity of two proteins, which is within a range between “0” and “1” (Samanta and Liang, 2003; Brun et al., 2003). With the development of biological vocabularies such as the Gene Ontology (Ashburner et al., 2000) (http://www.geneontology.org) and function annotation scheme such as the MIPS FunCat (Ruepp et al., 2004) (http://mips.helmholtz-muenchen.de/proj/funcatDB/), researchers have developed various methods to semantically measure protein similarities (Guzzi et al., 2012; Wang et al., 2012; Chen et al., 2009; Lubovac et al., 2005) based on gene similarity measures. These approaches are reasonable because they not only calculate the similarity between proteins but also extend the similarity calculation to gene level.

In addition to the protein similarity, the protein connectivity is also paid more attention when predicting protein functions. It is widely accepted that a protein with a high connectivity plays more significant roles than the one with a low connectivity. When a protein has a connectivity of more than 8 (this number is defined differently, for example, in Ref. (Maslov and Sneppen, 2002) the number is set to 20), it can be called a *hub protein*. A hub protein can be classified into the *party hub* or the *date hub* according to its partner’s correlating state (Han et al., 2004). A party hub interacts with the partner at the same time in a static network and is therefore a static hub, while a date hub interacts with partners at different times and locations in a dynamic network and is therefore a dynamic hub. Hub proteins have proved to be high conserved and play a pivotal role in the whole PPI network (Han et al., 2004). Therefore, in protein function prediction, it is necessary to take the protein connectivity into consideration.

Differently, we use a functional connectivity feature to represent the strength of a protein’s impact on its neighbours, function addition, however, is to measure the strength of a function correlating with other functions in a protein. The approach to predict protein functions the functional connectivity feature and function addition based on the PPI network is proposed in this paper. In Section 2, we present the details of our approach, including the definitions of the basic functional similarity, function addition, protein functional connectivity, predicting score and the prediction algorithm. The results of algorithm evaluation are given in Section 2, and a discussion about our method and the results are presented in Section 3. Finally, we summarize our approach in Section 4.