Journal of Theoretical Biology 396 (2016) 144–153

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Prediction of human protein–protein interaction by a domain-based approach

Xiaopan Zhang [a], Xiong Jiao [a,*], Jie Song [a], Shan Chang [b]

[a] Institute of Applied Mechanics and Biomedical Engineering, College of Mechanics, Taiyuan University of Technology, Taiyuan 030024, China
[b] Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, Changzhou 213001, China

HIGHLIGHTS

- A decent model for predicting PPIs was built via the information of PPIs and DDIs.
- The confidence probability was used for measuring the likelihood of the PPI.
- With using our predicting model we obtained 113 new probable PPIs.
- We constructed PPI network graphs to identify intriguing signal pathways.

ARTICLE INFO

Article history:
Received 7 October 2015
Received in revised form
29 January 2016
Accepted 20 February 2016
Available online 27 February 2016

Keywords:
Optimizing model
Predicting
Domain–domain interaction
Method
Network graph

ABSTRACT

Protein–protein interactions (PPIs) are vital to a number of biological processes. With computational methods, plenty of domain information can help us to predict and assess PPIs. In this study, we proposed a domain-based approach for the prediction of human PPIs based on the interactions between the proteins and the domains. In this method, an optimizing model was built with the information from InterDom, 3did, DOMINE and Pfam databases. With this model, for 147 proteins in the integrin adhesome PPI network, 736 probable PPIs have been predicted, and the corresponding confidence probabilities of these PPIs were also calculated. It provides an opportunity to visualize the PPIs by using network graphs, which were constructed with Cytoscape, so that we can indicate underlying pathways possible.

1. Introduction

In biological systems, the knowledge about the interaction situation of some proteins with the others is very important for the understanding of some biological processes (Goehler et al., 2004). Protein–protein interactions (PPIs) have been used to explain many important biological problems, such as predicting protein function (Chou, 2011b), speculating signal regulatory pathways (Bi-Qing et al., 2012; Hu et al., 2012), and identifying colorectal cancer related genes (Zhang et al., 2014). However, the identification of a protein–protein interaction (PPI) is a hard task. Even though many computational and experimental methods have been developed to observe or predict the PPIs in biological systems (Chou and Cai, 2006; Diao et al., 2015), there still exist serious technical difficulties for the prediction of PPI (Singhal and Resat, 2007). As a structural and functional subunit of protein, protein domain plays a dominant role in studying the function and structure of proteins (Cai et al., 2003; Chou and Shen, 2008). If two domains can physically interact, proteins containing these two domains are also likely to interact (Chou and Maggiora, 1998; Riley et al., 2005). Therefore, a better understanding of domain–domain interactions (DDIs) is an important step to identify PPIs (Chou and Cai, 2004; Kuo-Chen and Yu-Dong, 2002).

So far, several groups have developed methods to deduce PPI based on domain–domain interaction (DDI) information. Wan Kyu Kim has used a method, which is based on the co-occurrence of domains in interacting protein pairs to predict the PPIs (Kim et al., 2002). Wojcik and Schachter have devised a domain pair profile method. They characterized proteins using InterPro structural domains which had been known, and then they identified PPIs correlated with structural domain pairings (Wojcik and Schächter, 2001). Chou has developed a PseAAC model, which can be expanded to a functional domain mode, and this expanded model can be used to predict the protein subcellular localization, the

* Corresponding author.
E-mail address: jiaoxiong@tyut.edu.cn (X. Jiao).

http://dx.doi.org/10.1016/j.jtbi.2016.02.026
0022-5193/

membrane protein types and other core features of the protein (Chou, 2011a). Mudita Singhal and Haluk Resat have taken a multi-parameter optimization method in which they used the available PPI information to derive a quantitative scoring scheme about the domain–domain pairs, and then, they used the obtained domain interaction scores to predict whether two proteins interact (Singhal and Resat, 2007). Morihiro Hayashida has used mutual information of DDIs to predict PPIs via the method of conditional random fields (Hayashida et al., 2011). Binny Priya has developed a matrix based method for predicting PPIs with using the "all against all" datasets (Priya et al., 2013). These methods effectively used all kinds of information of DDIs, and they had obtained rational results at that time. Simultaneously, for a large proportion of these methods, the first step was to get all the possible protein interacting pairs. Then, they calculated the possibility of this protein pair with the information of DDI, which belong to the protein pair. However, it made some false negative predictions in these ways. At present, the false negative is unavoidable in a large number of computational methods for predicting PPI. In addition, the possibility of false predictions may limit the computational methods to being as a useful supplement to the experimental observations (Singhal and Resat, 2007).

In order to decrease the false negative ratio of prediction (Nemade and Pardasani, 2015), we propose a domain-based method for the prediction of PPIs based on the interactions between the proteins and the domains. First of all, let the proteins to construct PPI pairs randomly, and then build the DDI matrix for every protein pair. After that, we can obtain the possibly interactional protein pairs at the domain level. With the information from some DDI databases, we get a set of DDI confidence probabilities, and these scores are used to calculate the confidence probability of the predicted PPIs. In this process, in order to get an accurate prediction, we take a set of certified PPI data as the benchmark to train our predicted model.

2. Materials and methods

2.1. Data collection and data set construction

2.1.1. Benchmark data set preparation

The PPI data was collected from the Human Protein References Database (HPRD): version Release 9 (http://www.hprd.org/) (Keshava Prasad et al., 2009) and the MatrixDB: version matrixdb_141216 (http://matrixdb.ibcp.fr) (Launay et al., 2015). More than 95% of the interactions in the HPRD have been derived from individual in vitro or in vivo experiments. A large proportion of the interactions in the MatrixDB is based on experimental data. The data quality of these two databases is good enough to construct the PPI prediction models. After the processing of removing the repeat interactions, we got 18,656 human PPIs from these two databases. All of these human PPIs were used in preparing the positive data set. It is essential to select a negative data to the reliability of the prediction model. Nevertheless, the non-interacting protein pairs are not readily available, and it is difficult to generate such a data set (Yanzhi et al., 2008). We constructed the negative data as follows. First strategy, the non-interacting pairs were composed of the proteins that appeared in the positive data set. For example, if AB and CD were positive interaction pairs, the protein pairs AC, AD, BC, or BD could be selected as candidates for negative pairs (Shen et al., 2007). The second strategy has been described by Guo and colleagues in detail (Yanzhi et al., 2008), and it is based on such an assumption that human proteins situated in different subcellular localizations do not interact. The non-interacting pairs were constructed via pairing proteins from one subcellular location site with those from the other one. In this study, the subcellular location information of proteins was extracted from Swiss-Prot (http://www.expasy.org/sprot/) (Bairoch and Apweiler, 2000). Additionally, these two strategies must meet the following requirements: (1) the non-interacting pairs cannot appear in the whole HPRD and MatrixDB protein interacting pairs and (2) the contribution of proteins composing non-interacting pairs should be as harmonious as possible (Pitre et al., 2006; Shen et al., 2007). Since the number of negative pairs should equal to the total number of the positive pairs, 18,656 non-interacting pairs for constructing negative data set were composed via using the two strategies above.

In this study, we constructed a benchmark data set, which consisted of 37,312 human protein pairs; they were half the interacting protein pairs, half the non-interacting protein pairs. For the details of these pairs and their Swiss-Prot IDs, see Supplementary material Table S.1 and Table S.2.

2.1.2. Optimization of the data of DDIs

The InterDom database contains a set of confidence scores of DDIs and focuses on providing evidence from the detected PPIs (Ng et al., 2003). The InterDom database uses an integrated approach to predict potential DDIs and its confidence scores come from four different sources, domain fusions, protein complexes, scientific literature and PPIs (Björkholm and Sonnhammer, 2009). Therefore, the confidence score of the InterDom is very trustworthy. However, the confidence scores of the InterDom have a wide range and an extremely keen-edged effect (Singhal and Resat, 2007), so that they cannot reflect the relationship of scores well, as shown in Fig. 1A. In addition, Domain interactions with low confidence scores can be identified as likely false positives and the InterDom currently uses 1.5 as the cut-off score (Ng et al., 2003). Therefore, the score that is not lower than the cut-off score was selected and then we handled it by the following equation.

$$P_{D_{mn}} = \frac{\log S_{D_{mn}}}{\log S_{max}} \tag{1}$$

In Eq. (1), $S_{D_{mn}}$ represents the InterDom confidence score for the pair of domain m and domain n. S_{max} represents the maximum confidence score through the InterDom database. Parameter P indicates the confidence of DDI of the InterDom, which range from 0 to 1 (except 0). The D_{mn} represents domain pair of domains m and n in the whole paper.

The database of the 3did is a collection of DDIs in proteins for which high-resolution three-dimensional structures have been known (Mosca et al., 2014; Stein et al., 2005). The 3did exploits domain information to provide structural details necessary for understanding how DDIs occur. The database also provides Z-scores for the interaction, which indicates the confidence of DDIs (Stein et al., 2009). In other word, the higher the Z-score the higher the confidence of DDIs is (Stein et al., 2011). Although, the Z-scores do not have a wide range as vast as the confidence scores of the InterDom, in order to make them between 0 and 1, we also have encoded them with the equation as in the following.

$$Q_{D_{mn}} = \frac{Z_{D_{mn}}}{Z_{max}} \tag{2}$$

In Eq. (2), $Z_{D_{mn}}$ represents the Z-score for the domain pair of domain m and domain n. Z_{max} represents the maximum Z-score of the 3did database. Parameter Q indicates the confidence degree of DDI of the 3did, which ranges from 0 to 1 (except 0). The details of the treated result of the Z-score are shown in Fig. 1C.

DOMINE is a database of known and predicted DDIs compiled from 15 different sources (Raghavachari et al., 2008). It contains the interaction information about the domain pairs, which are predicted by 13 different computational approaches from PDB entries. There are HCP, MCP and LCP three grades represent high,