Computational Biology and Chemistry 56 (2015) 41–48

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Exploring the relationship between hub proteins and drug targets based on GO and intrinsic disorder

Yuanyuan Fu, Yanzhi Guo*, Yuelong Wang, Jiesi Luo, Xuemei Pu, Menglong Li*, Zhihang Zhang

College of Chemistry, Sichuan University, Chengdu 610064, PR China

ARTICLE INFO

Article history:
Received 9 September 2014
Received in revised form 13 March 2015
Accepted 21 March 2015
Available online 23 March 2015

Keywords:
Protein–protein interactions (PPIs)
Gene ontology (GO) terms
Intrinsic disorder
Drug target

ABSTRACT

Protein–protein interactions (PPIs) play essential roles in many biological processes. In protein–protein interaction networks, hubs involve in numbers of PPIs and may constitute an important source of drug targets. The intrinsic disorder proteins (IDPs) with unstable structures can promote the promiscuity of hubs and also involve in many disease pathways, so they also could serve as potential drug targets. Moreover, proteins with similar functions measured by semantic similarity of gene ontology (GO) terms tend to interact with each other. Here, the relationship between hub proteins and drug targets based on GO terms and intrinsic disorder was explored. The semantic similarities of GO terms and genes between two proteins, and the rate of intrinsic disorder residues of each protein were extracted as features to characterize the functional similarity between two interacting proteins. Only using 8 feature variables, prediction models by support vector machine (SVM) were constructed to predict PPIs. The accuracy of the model on the PPI data from human hub proteins is as high as 83.72%, which is very promising compared with other PPI prediction models with hundreds or even thousands of features. Then, 118 of 142 PPIs between hubs are correctly predicted that the two interacting proteins are targets of the same drugs. The results indicate that only 8 functional features are fully efficient for representing PPIs. In order to identify new targets from IDP dataset, the PPIs between hubs and IDPs are predicted by the SVM model and the model yields a prediction accuracy of 75.84%. Further research proves that 3 of 5 PPIs between hubs and IDPs are correctly predicted that the two interacting proteins are targets of the same drugs. All results demonstrate that the model with only 8-dimensional features from GO terms and intrinsic disorder still gives a good performance in predicting PPIs and further identifying drug targets.

1. Introduction

Proteins are significantly essential for almost all functions in the cells and they work together with others by forming a huge network of protein–protein interactions (PPIs) (Gavin et al., 2002). Nowadays PPIs have become a hotspot in the field of biochemistry and bioinformatics, so computational prediction of PPIs is of great practical significance. In the past decades, there were many descriptors and machine learning methods for prediction of PPIs. For example, domain-based random forest of decision trees (Chen and Liu, 2005), sequence similarity and the conservation of the proteins' functions (Kotelnikova et al., 2007), sequences correlation coefficient transformation based support vector machine method (Shi et al., 2010), ensemble extreme learning machine model using the information of protein sequences (You et al., 2013). With the help of the effective descriptors and machine learning methods, researchers could identify new PPIs. Biochemical experiments are informative but it is time-consuming and cannot provide a global perspective to further research on interactions between proteins (Patil et al., 2010a). So many computational methods were developed to predict PPIs based on large numbers of PPIs verified by experiments. From the aspect of feature extraction, almost all the existing methods rely on the sequence and structure information of the proteins in PPIs (Sprinzak et al., 2006; Guo et al., 2008; Roy et al., 2009; Valente et al., 2013). They analyzed PPIs by measuring various sequence attributes of each interacting pair and the accuracy can reach 88.09% (Guo et al., 2008). Other researchers provided methods to predict PPIs by using structural information (Tuncbag et al., 2011; Zhang et al., 2012; Baspinar et al., 2014). In addition, some researchers also used the functional information to find new PPIs, e.g., Zhang et al. (2012) provided an algorithm, PrePPI to predict

* Corresponding authors. Tel.: +86 28 85413330; fax: +86 28 85412356.
E-mail addresses: yzguo@scu.edu.cn (Y. Guo), liml@scu.edu.cn (M. Li).

http://dx.doi.org/10.1016/j.compbiolchem.2015.03.003
1476-9271/© 2015 Elsevier Ltd. All rights reserved.

PPIs with functional clues. Although these methods have achieved promising results, the number of features proposed by them is always as high as several hundreds even thousands. With so high dimensional feature vector, models by the machine learning methods, such as random forest (RF), Bayes and support vector machine (SVM) might give over-fitting results and so many features make the explanation of results very difficult on the functions of two interacting proteins, especially there are diverse PPIs, such as physical interaction, co-complex relationship and pathway co-membership (Qi et al., 2006). Identifying PPIs correctly is helpful to explore molecular mechanism and biological function of organizations. According to the previous studies, the PPIs could also be potential drug targets (Wang et al., 2011).

In PPI networks, hub proteins, defined as those that interact with more than 10 partners are clearly more disordered than end proteins that interact with just one partner (Haynes et al., 2006). It has been proven that intrinsic disorder is a general characteristic of hub proteins and the disorder might determine proteins' interactivities (Haynes et al., 2006). Shimizu and Toh (2009) have classified PPIs into three types: interactions between disorder proteins, interactions between structure proteins and interactions between disorder proteins and structure proteins. Compared to interactions between structure proteins, they proved that the occurrence of interactions between IDPs was significantly frequent. Obviously, intrinsic disorder could promote hubs to interact with partners as many as possible. Besides, it has been shown that intrinsic disorder regions are abundant in proteins associated with signaling, regulation, cancer and other diseases, so they are desirable targets for inhibition (Metallo, 2010). Since hubs were central to the functions of PPI network, hubs are involved in many diseases and they may also constitute an important source of potential drug targets (Wang et al., 2011; Hsing et al., 2008).

In our work, we selected hubs from human PPI networks and IDPs as basic datasets. In order to predict PPIs and further identify drug targets, only the 8 variables were computed to characterize the functions of proteins, including 6 semantic similarities of gene ontology (GO) terms and genes, and two that are the rates of disorder residues the two interacting proteins. As a result, the support vector machine (SVM) model yielded accuracy, sensitivity and specificity of 83.72%, 75.73% and 89.05% respectively in predicting interactions between target hubs, and 75.84%, 70.32% and 75.87% respectively in predicting interactions between target hubs and IDPs. Moreover, 118 of 142 PPIs between hubs and 3 of 5 PPIs between target hubs and IDPs were successfully predicted that both the interacting proteins were targets of the same drugs. Different from the general methods that usually use physical and chemical properties to produce hundreds even thousands of features to train the model, our method just used 8 features based on the functional similarity and also yielded a good performance in predicting PPIs and identifying drug targets.

2. Materials and methods

2.1. Data collection

Protein interaction data used in our work were obtained from the Human Protein Reference Database (HPRD) (Release9_062910), containing 39,240 experimentally determined PPIs. In human PPI network, nodes represent proteins and edges represent PPIs. We calculated the degrees of all nodes (Pavlopoulos et al., 2011) and only considered proteins with their degrees $\geq$10 as hubs (Haynes et al., 2006). Then a part of hubs have been proved to be drug targets through DrugBank (Version 3.0), KEGG DRUG (Release 63.0) and Therapeutic Targets Database (TTD, Version 4.3.02) (Zhu et al., 2012). They were selected as target hubs, so 661 target hubs were finally extracted. For model building, the training data of PPIs were 2953 PPIs among the target hubs from HPRD. However, the non-PPIs of training data are not currently available, so they were generated according to the subcellular location based on the hypothesis that two proteins in two different subcellular locations usually not interact with each other (Guo et al., 2008). The subcellular localization information of the 661 target hubs was extracted from UniProtKB (http://www.uniprot.org/uniprot/). After removing proteins without subcellular location information, the remaining proteins were grouped into eight subsets based on the eight main types of subcellular localizations, including cytoplasm, nucleus, mitochondrion, endoplasmic reticulum, golgi apparatus, peroxisome, vacuole, cytoplasm and nucleus (Guo et al., 2008). The non-interacting pairs were generated by pairing proteins with different subcellular locations and thus 4430 non-interacting pairs were used as the negative samples in the training set according to the ratio of 1:1.5 for the positives VS negatives.

For external evaluation, from DisProt (Release 6.01), we chose 165 human IDPs that were not included from target hubs and 41 of them were drug targets. Then 438 PPIs between 249 target hubs and 68 IDPs were obtained from HPRD. In order to test the specificity of the model, the non-PPIs between target hubs and human IDPs were also extracted. Considering all possible protein pairs of the target hubs and IDPs, after eliminating the positive samples, the remaining 71,886 protein pairs were considered as the negative samples. Table 1 shows the number of samples in training and testing sets respectively.

2.2. Features extraction

2.2.1. Intrinsic disorder

Most of PPI networks are proved to be scale-free (Goh et al., 2002; Jeong et al., 2000; Barabási and Albert, 1999), which means that few hubs possess tens even hundreds of interactions and most of proteins only has one link. Meanwhile, the intrinsic disorder is a distinctive feature of hubs and promotes hubs to interact with partners as many as possible in PPI networks. Firstly, intrinsic disorder can serve as the structural basis for hubs interacting with other proteins. Then intrinsic disorder can provide flexible linkers between partners and enables to enhance the ability of hub proteins to participate in multiple signaling pathways. At last, many IDPs can bind to a structured hub (Uversky, 2011; Dunker et al., 2005; Dosztanyi et al., 2005). In addition, the high specificity and low affinity (Zhou, 2012) in the IDPs are controlled by the rate of association and dissociation (Prakash, 2011) and further allow their PPIs to be reversible (Yura and Hayward, 2009; Bertolazzi et al., 2012), which has an important effect on the functions of proteins. So, IDPs and part of hubs are common in eukaryotes, existing as dynamic ensembles and resembling "protein clouds" (Uversky, 2011; Uversky and Dunker, 2010). IDPs were significantly enriched in disorder-promoting amino acids of Ala-Arg-Gly-Gln-Ser-Glu and Lys but substantially depleted in Pro and in so-called order-promoting amino acids such as Ile-Leu-Val-Trp-Tyr-Phe-Cys and Asn. The current researches mostly focus on the number of intrinsic disorder distribution in the PPIs (Haynes et al., 2006; Shimizu and Toh, 2009; Kim et al., 2008; Patil et al., 2010b). The functions of intrinsic disorder in the PPIs play essential roles in regulation of signal and path ways (Metallo, 2010; Dunker et al., 2008; Singh et al., 2007). Now several tools have been developed

Table 1
The number of the samples in training and testing sets respectively.

Data sets	Positive samples	Negative samples
Training data set	2953	4430
Testing data set	438	71886