Journal of Theoretical Biology 349 (2014) 82–91

Contents lists available at ScienceDirect

Journal of Theoretical Biology

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

Analysis and identification of toxin targets by topological properties in protein–protein interaction network

Lei Yang [a], Jizhe Wang [a], Huiping Wang [a], Yingli Lv [a], Yongchun Zuo [b], Wei Jiang [a,*]

[a] *College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China*
[b] *The National Research Center for Animal Transgenic Biotechnology, Inner Mongolia University, Hohhot, 010021, PR China*

HIGHLIGHTS

- Protein–protein interaction networks are used to study toxin targets.
- The toxin targets are analyzed by 12 topological properties in the PPI network.
- The SVM is presented to predict the toxin targets based on topological properties and sequence information.

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form
20 January 2014
Accepted 1 February 2014
Available online 7 February 2014

Keywords:
Support vector machine
Sequence information
Statistical test
Topological features

ABSTRACT

Proteins do not exert their function in isolation of one another, but interact together in protein–protein interaction (PPI) networks. Analysis of topological properties of proteins in the PPI network is very helpful to understand the function of proteins. However, until recently, no one has ever undertaken to investigate toxin targets by topological properties. In this study, for the first time, 12 topological properties are used to investigate the characteristics of toxin targets in the PPI network. Most of the topological properties are found to be statistically discriminative between toxin targets and other proteins, and toxin targets tend to play more important roles in the PPI network. In addition, based on the topological properties and the sequence information, support vector machine (SVM) is used to predict toxin targets. The results obtained by the jackknife test and 10-fold cross validation are encouraging, indicating that SVM is a useful tool for predicting toxin targets, or at least can play complementary roles in relevant areas.

1. Introduction

Toxins are important classes of poisonous compounds include pollutants, pesticides, preservatives, drugs and venoms. With more and more new toxins are found during the past decades, the applications of toxins as tools in drug discovery and cellular biology become an important part of toxin research. The fundamental research on toxins and their biological targets is one of the most attractive topics in descriptions of the mechanism of action, their metabolism in the human body, their lethal or toxic dose levels, their potential carcinogenicity, exposure sources and recommended treatments. Therefore, identification and analysis of toxin targets are important for both medicine and biology. The ability to rapidly identify toxin targets has been described as the most important task of toxicogenomics. Experimental approaches for identification of toxin targets are time consuming and expensive. So, developing a fast and effective way to identify toxin targets by computational methods would be very necessary for toxicology research (Kavlock et al., 2008). In 2010, Toxin and Toxin-Target Database (T3DB) (Lim et al., 2010) collected the existing information for toxins and their targets, which provided an opportunity for characterizing the common properties of toxin targets by computational methods (Zhou et al., 2013).

Because the majority of proteins interact with each other for proper function in a cell, the knowledge about interactions between proteins is essential for the understanding of molecular and cellular functions (Chaurasia et al., 2007; Rual et al., 2005; Stelzl et al., 2005). Therefore, the study of protein–protein interaction (PPI) networks provides many new insights into protein function in the context of a network. With the development in the high-throughput protein interaction detection technology Yeast Two-Hybrid (Y2H) technology (Uetz and Hughes, 2000) and the tandem affinity purification-mass spectrometry technique (TAP-MS) (Gavin et al., 2002), large-scale data sets of protein–protein

* Corresponding author. Tel.: +86 451 8666 9617; fax: +86 451 8661 5922.
E-mail address: bioccjw@yahoo.com (W. Jiang).

http://dx.doi.org/10.1016/j.jtbi.2014.02.001

interactions are created (Brown and Jurisica, 2005; Hermjakob et al., 2004; Stark et al., 2006; Von Mering et al., 2003; Xenarios et al., 2000). This provides great opportunities for researchers to elucidate the process of life activities from the system-level of the PPI networks. However, most of the PPI networks are too complex to easily understand. By using graph theoretic concepts to investigate the topological properties of the PPI networks, this problem can be overcome. The topological properties have been applied to study social networks in social sciences (Wasserman and Faust, 1994). Furthermore, the topological properties are used to evaluate the properties of the PPI networks. Xu and Li (2006) used five topological properties to describe disease genes in the PPI networks, the topological properties were found to be statistically discriminative between disease genes and non-disease genes. Furthermore, based on the work of Xu et al., a new topological property was proposed by Zhang et al. (2010) to calculate the statistical significance. The dynamic method was used to investigate the topology characteristics of regulatory network (Ding et al., 2013a). The work of Zhu et al. showed that the topological properties of drug targets were significantly different from those of non-drug-targets in the human PPI networks (Zhu et al., 2009), and Wang et al. (2011) found these differences were mainly caused by mir-drug-targets and there was no difference in topological properties between non-mir-drug-targets and non-drug-targets. In 2009, 10 topological properties and 4 sequence properties were used by Hwang et al. to describe essential genes in the PPI networks (Hwang et al., 2009). There were significant differences in these properties between essential genes and non-essential genes. Kotlyar et al. found that there were significant differences in degree, betweenness and clustering coefficients between drug targets, drug-regulated genes and unaffected genes (Kotlyar et al., 2012). Network-based methods were also used by other researchers in different networks (Coulomb et al., 2005; Florez et al., 2010; Han et al., 2013a; Han et al., 2013b; Hwang et al., 2008; Sualp and Can, 2011; Wachi et al., 2005). However, until recently, no network-based method is applied in the dataset of toxin targets.

In this study, the literature-curated (LC) human PPI network was obtained from BIND (Bader et al., 2003), HPRD (Peri et al., 2003) and MINT (Ceol et al., 2010). 12 topological properties are calculated for each node in the PPI network. Significant differences are found between the topological properties of toxin targets and the topological properties of other nodes. In order to avoid statistical bias, 751 non-toxin targets are randomly selected from the non-toxin target dataset and 751 nodes are randomly selected from the whole network, this process was repeated 1000 times. The results show that the toxin targets always have the most extreme index in all properties. In addition, by using 12 topological properties and 400 dipeptides (Lin et al., 2010), a machine learning approach is proposed to predict toxin targets in both a redundant dataset and a non-redundant dataset. Good performances are obtained by the jackknife test and 10-fold cross validations in both datasets. The performance indicates that our model could be a powerful tool for predicting toxin targets. The workflow of our study is shown in Fig. 1.

2. Material and method

2.1. Dataset of human protein–protein interactions

Human protein–protein interaction (PPI) datasets were downloaded from Online Predicted Human Interaction Database (version 1.95) (Brown and Jurisica, 2005) on May 22, 2012. Three data sources (1) literature-curated (LC) human PPI from BIND (Bader et al., 2003), HPRD (Peri et al., 2003) and MINT (Ceol et al., 2010); (2) interactions identified from high-throughput yeast two hybrid mapping approach (EXP); (3) interactions predicted from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* are included in Online Predicted Human Interaction Database. In order to obtain the high quality of human PPIs, only the literature-curated human PPIs were used. The entire LC network comprises 12,265 nodes and 83,818 interactions. After removing self loops and duplicate edges, the final network comprises 12,265 nodes and 61,170 interactions (the nodes represent proteins and the edges represent interactions). This network contains 228 connected components, and the main component comprises 11,952 nodes and 61,081 interactions. Because the topological properties are incalculable for proteins which do not belong to the main component, so only the main component is considered in this study.

2.2. Dataset of toxin targets

Toxin targets were downloaded from Toxin and Toxin-Target Database (T3DB) (Lim et al., 2010) on May 22, 2012. T3DB is a resource that compiles information about toxins and their targets. The dataset currently contains over 2900 small molecule and peptide toxins, 1300 toxin targets and more than 33,000 toxin target associations. 993 human targets were used in this study. There were 751 human toxin targets in the PPI network.

2.3. Dataset of non-toxin targets

Because there is still no protein which has been identified as non-toxin targets now, thus in contrast to previous work (Huang et al., 2010; Li and Lai, 2007) of establishing non-drug target dataset, the non-toxin target dataset was established as follows. First, 20,250 human proteins were downloaded from Swiss-Prot (Bairoch and Boeckmann, 1991) on May 22, 2012. 993 toxin targets covered 540 protein families in Pfam database (Bateman et al., 2004) which contained 6931 human proteins in Swiss-Prot. 6931 human proteins were eliminated from 20,250 human proteins. Finally, 13,319 proteins remained in the putative non-toxin target dataset, and there were 6758 proteins in the PPI network. Although novel toxin targets might exist in this dataset, the chance was pretty low. Then, 751 non-toxin targets were randomly selected from 6758 non-toxin targets. This dataset was defined as control dataset one.

2.4. Randomly selected nodes

In order to compare the topological properties of toxin targets with those of other nodes in the PPI network, 751 nodes were randomly selected from the PPI network and this dataset was defined as control dataset two. So, our final training datasets consisted of toxin target dataset and two control datasets.

2.5. Topological properties

In this study, the following topological properties are calculated for illustrating the behavior of the proteins in the PPI network (Table 1). Degree is the most elementary and simplest topological index, which is defined as the number of nodes directly connected to a given node i. Average shortest path (ASP) is defined as the average shortest path between a node and all the nodes in the PPI network. The average distance to toxin targets (ADT) is defined as the average shortest distance between a protein and all the toxin targets in the PPI network. The shortest distance to toxin targets (SDT) is defined as the shortest path between a protein and its nearest toxin target. 1N index (Xu and Li, 2006; Zhu et al., 2009) of node i is defined as the proportion of toxin targets among all its neighbors (Fig. S1). In undirected networks, clustering coefficient