Science in China Series C: Life Sciences

www.scichina.com
life.scichina.com
www.springerlink.com

Article

Identifying drug-target proteins based on network features

ZHU MingZhu[1], GAO Lei[1], LI Xia[1,2†] & LIU ZhiCheng[1†]

[1] School of Biomedical Engineering, Capital Medical University, Beijing 100069, China;
[2] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, China

Proteins rarely function in isolation inside and outside cells, but operate as part of a highly interconnected cellular network called the interaction network. Therefore, the analysis of the properties of drug-target proteins in the biological network is especially helpful for understanding the mechanism of drug action in terms of informatics. At present, no detailed characterization and description of the topological features of drug-target proteins have been available in the human protein-protein interaction network. In this work, by mapping the drug-targets in DrugBank onto the interaction network of human proteins, five topological indices of drug-targets were analyzed and compared with those of the whole protein interactome set and the non-drug-target set. The experimental results showed that drug-target proteins have higher connectivity and quicker communication with each other in the PPI network. Based on these features, all proteins in the interaction network were ranked. The results showed that, of the top 100 proteins, 48 are covered by DrugBank; of the remaining 52 proteins, 9 are drug-target proteins covered by the TTD, Matador and other databases, while others have been demonstrated to be drug-target proteins in the literature.

drug-target, protein-protein interaction, topological features

With the emergence of high-throughput biological data, such as gene expression profile data, protein-protein interaction (PPI) data, molecular sequence data, RNA structure data, analyzing and mining mass data by informatic methods has become a focus of bioinformatics research. Currently, many studies report machine learning methods employed for predictions in fields such as for the functions of genes or proteins[1–4], RNA structure[5,6], and disease genes.

With the application of high-throughput inspection technologies for protein interaction, such as yeast two-hybrid[7] and tandem affinity purification-mass spectrometry (TAP-MS)[8], the protein interaction patterns for many species have been discovered, and many patterns involving multiple species have been collected in databases (MIPS[9] and BioGRID[10]). Using protein interaction data, and based on network differentiation and the characteristics of a functional knowledge system[1,11–13], a series of functional analyses and calculation methods have been proposed. These studies predicted protein functions by using the topological structure of the network to separate the network into related functional modules. Although disease analysis technologies based on the protein interaction network have already been instituted, recent pattern recognition technologies, such as *k*-nearest neighbors (KNN) and support vector machine (SVM), have also been applied to analyze the topological features of relevant protein in-

Received October 6, 2008; accepted Ocotber 28, 2008
doi: 10.1007/s11427-009-0055-y
†Corresponding author (email: lixia@hrbmu.edu.cn; zcliu@ ccmu.edu. cn)
Supported by National Natural Science Foundation of China (Grant No. 30370798, 30571034 and 30570424), National High-Tech Research and Development Program of China (Grant No. 2007AA02Z329), National Basic Research Program of China (Grant No. 2008CB517302), Natural Science Foundation of Heilongjiang Province, China (Grant No. ZJG0501, 1055HG009, GB03C602-4 and BMFH060044), New Century Hundred-Thousand-Ten Thousand Talents Project of Beijing City, Scientific Research Common Program of Beijing Municipal Commission of Education (KM200610025011).

Citation: ZHU M Z, GAO L, L X et al. Identifying drug-target proteins based on network features. Sci China Ser C-Life Sci, 2009, 52(4): 398-404, doi: 10.1007/s11427-009-0055-y

teraction networks[14].

Screening and identification of drug targets is the foundation of novel drug research and development, and the preliminary step to drug discovery, as well as one of the key factors in rational synthesis. Because research on the genic regulation mechanism of gene knock-out animal models is time-consuming and costly, it is difficult to directly use it for screening targets of a novel drug. However, the advancement of the Human Genome Project provides a new approach for studying drug-targets utilizing bioinformatics methods. The drug target database, DrugBank[15] is the largest and most comprehensive database for pharmaceuticals and drug targets. DrugBank contains abundant detailed information about drug targets and pharmaceuticals regarding chemistry, pharmaceutics, medicine, and molecular biology. In this study, the drug-target information in DrugBank was mapped to the human protein interaction network for analyzing the topological features of target proteins in the interaction network, so that predictions concerning the proteins of the drug targets could be carried out.

1 Materials and methods

1.1 The dataset of human protein interaction

Protein interactive data, including both physical and genetic interactions, are applicable to predict biological functions[16]. Human PPI data from the BioGRID database, based on the literature, are processed beforehand by (i) eliminating self-interactions, and (ii) counting one time for identical interactions observed by different detection methods (e.g. yeast two hybrid, co-immunoprecipitation). The processed dataset contains 24063 interactions among 7832 different human proteins. This data is utilized to construct a protein interaction network, in which the protein is represented as its node and the interactive relationship as its edge (Figure 1).

1.2 The dataset of drug-target proteins

In the DrugBank database, there are 1103 listed drug-target proteins that show an interactive relationship in BioGRID (including FDA-approved and those under investigation, which are also tagged 'experimental drugs'). Among them, 760 proteins belong to FDA-approved drug-targets and 343 proteins belong to experimental drug-targets. In the following analysis, the 1103 drug-targets are used as the positive samples. The data were downloaded from the DrugBank database in January, 2008.

1.3 The definition of the topological feature set

For each node (protein) in the PPI network, five measures for assessing its topological properties were defined (Table 1). The most elementary characteristic of a node is its degree (or connectivity), which indicates how many links the node *i* holds with other nodes. The 1N index designates the neighbors of node *i*, which is defined as the proportion of the number of links to other nodes listed in the 'drug-target set' among all links. The measurement of the average distance to drug-target proteins is used to assess the communication efficiency of node *i* with the overall 'drug-target set' in the PPI network. A shorter average distance corresponds to a quicker transduction between node *i* and the 'drug-target set'. In many conditions, if node A is connected to B, and B is connected to C, then it is highly probable that A also has a direct link to C. This phenomenon is quantified using a clustering coefficient. The shortest distance to a drug-target protein measures the path of a protein to its nearest drug-target protein.

1.4 The construction of a negative set

In contrast to building the positive set, compiling a set of the negative samples is indefinite, because the DrugBank database is in a state of continual renewal. Some proteins belonging to drug-targets are definitely known, but those excluded as drug-targets cannot be clearly recognized. A recent study[17] showed that the human proteome may contain thousands of essential proteins which are clearly distinguishable from other proteins in topological features. Yildirim et al.[18] took essential proteins as a separate group to make a contrastive study on drug-target and non-drug-target proteins. In this paper, we consider the essential proteins which were selected by Tu et al.[17]. They expounded that the gene products expressed ubiquitously in the human genome are fundamental for cellular physiological processes. Thus, they proposed to use the 1789 ubiquitously expressed gene products as substitutes for the essential proteins. In this study, it was confirmed that: (i) If a protein is not in the DrugBank database (including FDA-approved proteins and those at the experimental stage); (ii) essential proteins are excluded; and (iii) the protein with interactive information may then be used as a negative candidate set. The same number of proteins as in the positive set are randomly selected as the negative set (control set).

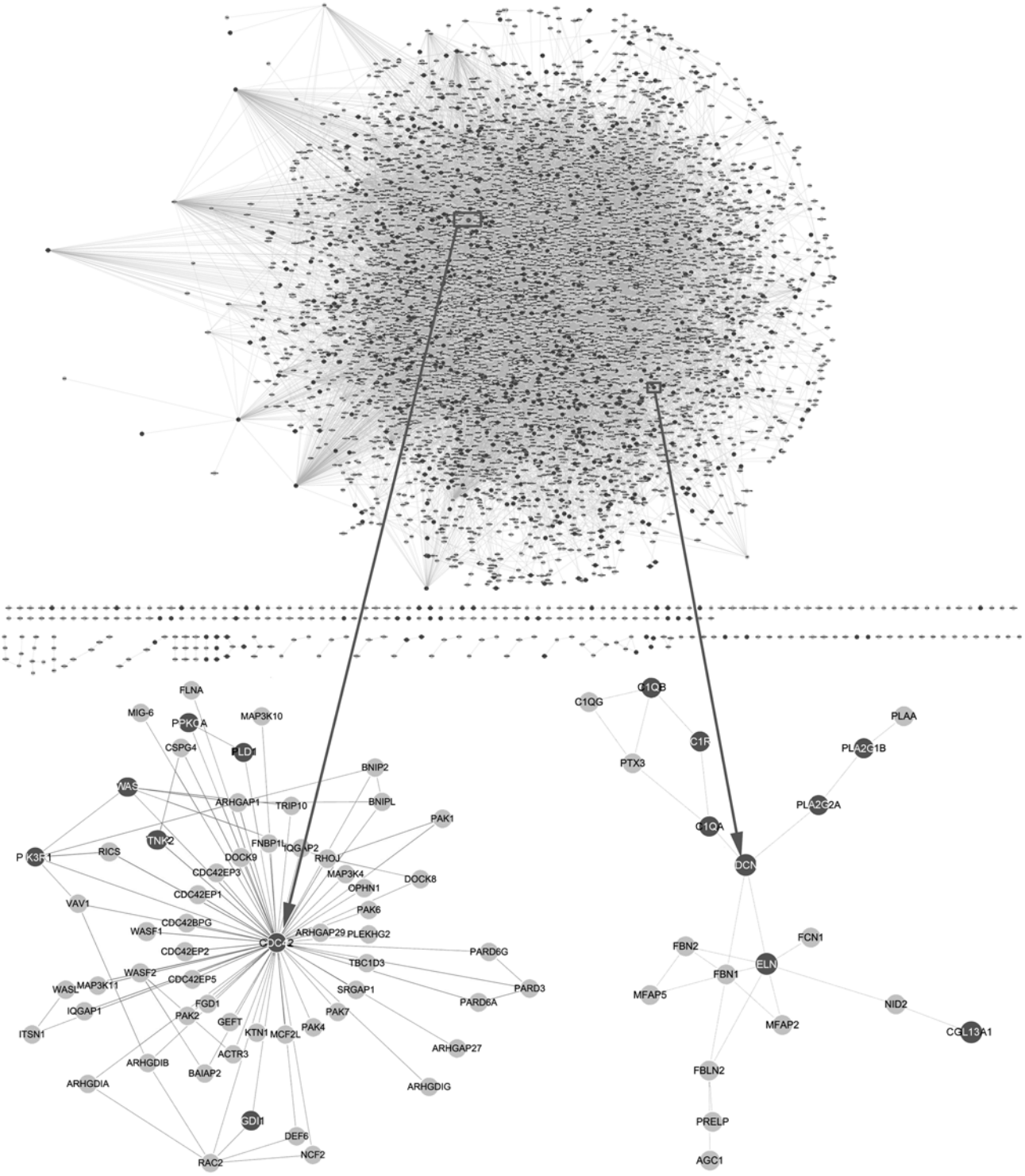

Figure 1 The distribution of drug-target proteins in the human protein-protein interaction network. The deep-colored nodes represent drug-target proteins and the light-colored nodes represent other proteins in the BioGRID interaction database. The graph in the top portion of this figure shows the entire set of interaction relationships, while the graph below shows two local interaction relationships.

Table 1 Topological feature set.

Feature	Function	Description
Degree	k_i	the number of links to node i;
1N index	k_i^P / k_i	k_i^P is the number of links between node i and drug-targets
The average distance to drug-targets	$\sum_{j \in M} d_{ij} / \lvert M \rvert$	M denotes the set of indices corresponding to drug-target; d_{ij} denotes length of the shortest path between node i and node j;
Clustering coefficient	$2n_i / k_i(k_i - 1)$	n_i is the number of links connecting the k_i neighbors of node i to each other.
The shortest distance to drug-targets	d_i	the shortest distance from node i to drug-target protein

1.5 The *k*-core analysis of the PPI network

An important method in network analysis, the *k*-core algorithm, has been applied in the study of yeast and other species[19,20], and is an iterative process. For the iterative parameter *k*, the nodes with a degree smaller than *k* are gradually deleted until the degree of the remaining nodes is larger than or equal to *k*.

The *k*-core analysis will result in a series of sub-networks which gradually reveal the framework of the central area in the original network. Larger values of the index *k* will generate nodes with a larger degree in sub-networks and those closer to the central area in the network structure. In order to measure the centrality of the set of drug-targets in the network, for each *k*, the ratio of the number of drug-targets to the number of all proteins in the sub-network was calculated. This ratio was used to describe the proportion of drug-targets in the sub-network. If the ratio gradually increases with the increasing value of *k*, it means that the drug-targets tend to be closer to the central area in the network, i.e. they have centrality.

2 Results and analysis

2.1 The analysis of topological features

In order to compare the difference of topological features between the set of drug-target proteins and the set of proteins in the whole network, drug-target proteins are used to form a positive sample set, and then the same number of proteins as the positive samples from the 7832 human proteins in the BioGRID database are randomly selected to form a control set. The samplings are repeated 1000 times and the average value of indices in the control set are calculated. The experimental results show that the average degree of the drug-target set is 9, significantly higher than the average degree of all the proteins in the network. This is consistent with the previous finding by Tu et al., that drug-target proteins have a larger degree[17]. This study also finds that, among all the neighbors of drug-target proteins, the proportion of target proteins is significantly larger than that of target proteins in the proteome (the level of significance is $P<10^{-16}$). The average distance between target proteins and target proteins is also significantly smaller than that between the proteins in the proteome and the target proteins (the level of significance is $P<10^{-7}$), indicating that drug-target proteins communicate with each other more rapidly. The clustering coefficient reflects the compactness of the module formed by the protein and its interactive neighbors. The results in Table 2 show that the clustering coefficient of the drug-target protein set is lower than that of the proteome set, but it does not mean that the modularity of drug-target proteins is inferior. The reason for this result is that proteins with a higher degree in the interaction network tend to have a lower clustering coefficient and proteins with a lower degree tend to have a higher clustering coefficient. Yildirim et al. have discussed the relationship between the degree and its clustering coefficient in a study of net-biology[18]. Therefore, it is not the drug-target proteins which have a looser distribution in the PPI network, for they in fact have a higher degree within the network. The shortest distance to the drug-target protein reflects the length of the shortest path from a protein to its closest drug-target protein. The index is lower in the set of drug-target proteins.

In further analysis, these topological features of the drug-target set were compared with those of the non-drug-target set. The results are shown in Table 3. The results indicate that the topological features of the drug-target protein set are significantly distinguished from those of the non-drug-target protein set in the interaction network.

2.2 The rank of proteins in the PPI network.

The five indices, including the degree, the 1N index, the average distance to drug-targets, the clustering coefficient, and the shortest distance to drug-targets, describe

Table 2 The average value and significance level of topological features of the BioGRID's protein set and the drug-target set (*K-S* test).

Feature[a)]	Drug-target set	The proteins set of BioGRID		*P*-value
		Mean	SD	
Degree	9.3019	6.1440	0.2803	3.9724e-4
1N index	0.3230	0.2082	0.0081	1.7990e-17
ADT	4.2742	4.5581	0.0215	7.2322e-8
CC	0.1048	0.1539	0.0083	5.8445e-3
SDT	1.3330	1.5940	0.0189	3.8613e-14

a) ADT denotes average distance to drug-targets; CC denotes clustering coefficient; SDT denotes the shortest distance to drug-targets.

Table 3 The average value and significance level of topological features of the control set and the drug-target set (*K-S* test).

Feature[a)]	Drug-target set	Control set		*P*-value
		Mean	SD	
Degree	9.3019	5.3617	0.2524	3.6287e-6
1N index	0.3230	0.1931	0.0085	1.1396e-21
ADT	4.2742	4.6083	0.0205	3.6143e-10
CC	0.1048	0.1671	0.0093	7.0466e-4
SDT	1.333	1.6461	0.0203	1.1802e-17

a) ADT denotes average distance to drug-targets; CC denotes clustering coefficient; SDT denotes the shortest distance to drug-targets.

the modularity of proteins in the network from different facets. $S(i)$ is calculated based on the five indices and the proteins are ranked in the PPI network from 1 to 7832. The $S(i)$ is calculated as follows:

$$S(i) = \sum_j (rank_j(i))$$

First, each protein i is ranked according to a different property j, and $S(i)$ is the sequence of the five topological features after they are integrated. It is found that, among the 1103 drug-targets, 285 targets rank in the top 1000. In the top 100 proteins, 48 proteins are recorded as drug-targets within the DrugBank database. The detailed results are illustrated in Figure 2. The horizontal axis denotes that the proteins are ranked from 1 to 7832 according to $S(i)$, and the ratio in the vertical axis denotes the ratio between the observed number of drug-targets and the expected number in each interval. The larger ratio shows the higher probability of the occurrence of drug-targets. The figure shows that the drug-targets have a higher probability of occurrence in the top intervals within the sequence. Among 52 proteins in the top 100 proteins, but excluded from the DrugBank database, 9 proteins are recorded in other drug-target databases (TTD, Matador), while other proteins may be validated to be drug-targets in related literatures[21–27] (Table 4). Along with the deepening of research on BRCA1, it is regarded that BRCA1 has antagonized

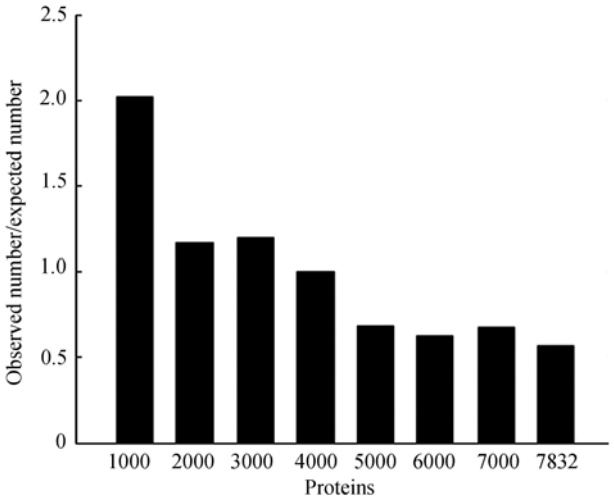

Figure 2 The distribution of ratios. The horizontal axis shows the intervals of the sequence based on $S(i)$, and the vertical axis shows the ratio between the observed number of drug-targets and the expected number in each interval.

Table 4 Database and literature validation information

Validation	Protein
TTD	ACLY, CD36, CDC25A
Pgeni	ARRB1, ARRB2
Matador	BCR, BGN, BRCA1, CD36, GNAI2
Literature	FHL2, HSPCA, ADAM15, AKAP13, BAT2, BDKRB2

multiple functions of estrogen. BRCA1 mutation directly performs function deficiency on the tumor inhibition effect. The BRCA1 phosphorylated interactor as a small molecule inhibitor is a possible biomarker for predicting the cure effect of breast cancer and directing the treatment of breast cancer[25].

In order to test the validity of the five topological features in discriminating between drug-targets and non-drug-targets, the support vector machine (SVM) classifier was applied to classify the samples. In this paper, the cubic polynomial function $K(x, x_i)=[(x, x_i)+1]^d$ (d= 3)is introduced as the kernel function of the classifier, and the five indices (i.e. the degree, the 1N index, the average distance to drug-target proteins, the clustering coefficient, and the shortest distance to drug-target proteins) are regarded as the features of the SVM classifier. The fivefold cross-validation test is also used, and the ROC curve (Figure 3) is applied to evaluate the results. The area below the ROC curve accounts for 70.60% of the total area.

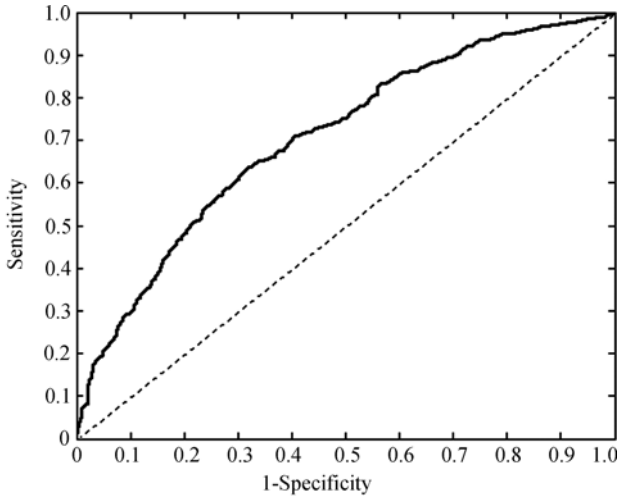

Figure 3 The ROC curve of the SVM classifier. Specificity is TN/(TN+FP), and sensitivity is TP/(TP+FN).

2.3 The *k*-core analysis of the PPI network.

Using *k*-core analysis, drug-target distances to the network center are measured (Figure 4). Figure 4 shows that along with the increase of *k* the proportion of drug-targets in the sub-network is higher. The larger value of *k* denotes that the obtained sub-graph is closer to the center of the original network (*k*=1). The ratio here represents the proportion of the number of drug-targets in the network to the number of all proteins (nodes) in the sub-network. A larger value of *k* results in a higher ratio, indicating that drug-targets tend to be closer to the topological center of the network.

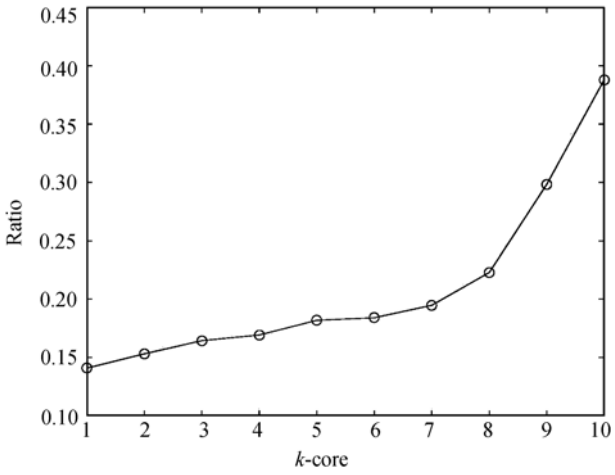

Figure 4 The *k*-core analysis of drug-target proteins in the PPI network.

3 Discussion

In this paper, the drug-targets in the DrugBank have been mapped onto the human PPI network, and the topological features of drug-targets analyzed. It has been confirmed, as reported in previous studies, that drug-targets have a higher degree in the interaction network, and it has been discovered that drug-targets have more interactions between them and a shorter average distance, indicating that the signals between drug-targets are more rapidly transmitted. Simultaneously, the k-core analysis of the PPI network showed that drug-targets tend to be closer to the topological center of the network.

Furthermore, according to the topological features of proteins in the PPI network, the proteome was scored and ranked. The results show that the proteins ranked in the top are more likely to be drug-targets. These results have been confirmed in the literature and in various databases.

In this paper, we investigated the approach to studying and predicting drug-targets according to the topological features of the proteins in the PPI network, and found that the topological features of drug-targets in the PPI network provide helpful information in differentiating drug-targets and non-drug-targets. We made valuable preparation and theoretical discussion for predicting and discovering new drug-targets through network features. With the increasing quantity of human PPI data and additional new drug-targets, this approach has significance for future investigation.

1 Nabieva E, Jim K, Agarwal A, et al. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 2005, 21 (Suppl 1): i302—310
2 Vazquez A, Flammini A, Maritan A, et al. Global protein function prediction from protein-protein interaction networks. Nat Biotechnol, 2003, 21(6): 697—700
3 Gao L, Li X, Guo Z, et al. Widely predicting specific protein functions based on protein-protein interaction data and gene expression profile. Sci China C-Life Sci, 2007, 50(1): 125—134
4 Zhu M, Gao L, Guo Z, et al. Globally predicting protein functions based on co-expressed protein-protein interaction networks and ontology taxonomy similarities. Gene, 2007, 391(1—2): 113—119
5 Rivas E, Eddy S R. A dynamic programming algorithm for RNA structure prediction including pseudoknots. J Mol Biol, 1999, 285(5): 2053—2068
6 George R A, Liu J Y, Feng L L, et al. Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucleic Acids Res, 2006, 34(19): e130
7 Uetz P, Hughes R E. Systematic and large-scale two-hybrid screens. Curr Opin Microbiol, 2000, 3(3): 303—308
8 Gavin A C, Bosche M, Krause R, et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002, 415(6868): 141—147
9 Mewes H W, Dietmann S, Frishman D, et al. MIPS: analysis and annotation of genome information in 2007. Nucleic Acids Res, 2008, 36(Database issue): D196—201
10 Stark C, Breitkreutz B J, Reguly T, et al. BioGRID: a general repository for interaction datasets. Nucleic Acids Res, 2006, 34(Database issue): D535—539
11 Chen Y, Xu D. Global protein function annotation through mining genome-scale data in yeast *Saccharomyces cerevisiae*. Nucleic Acids Res, 2004, 32(21): 6414—6424
12 Karaoz U, Murali T M, Letovsky S, et al. Whole-genome annotation by using evidence integration in functional-linkage networks. Proc Natl Acad Sci USA, 2004, 101(9): 2888—2893
13 Jiang T, Keating A E. AVID: an integrative framework for discovering functional relationships among proteins. BMC Bioinformatics, 2005, 6: 136
14 Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics, 2006, 22(22): 2800—2805
15 Wishart D S, Knox C, Guo A C, et al. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res, 2008, 36(Database issue): D901—906
16 Reguly T, Breitkreutz A, Boucher L, et al. Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. J Biol, 2006, 5(4): 11
17 Tu Z, Wang L, Xu M, et al. Further understanding human disease genes by comparing with housekeeping genes and other genes. BMC Genomics, 2006, 7: 31
18 Yildirim M A, Goh K I, Cusick M E, et al. Drug-target network. Nat Biotechnol, 2007, 25(10): 1119—1126
19 Wuchty S, Almaas E. Peeling the yeast protein network. Proteomics, 2005, 5(2): 444—449
20 Wachi S, Yoneda K, Wu R. Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. Bioinformatics, 2005, 21(23): 4205—4208
21 Wang J, Yang Y, Xia H H, et al. Suppression of FHL2 expression induces cell differentiation and inhibits gastric and colon carcinogenesis. Gastroenterology, 2007, 132(3): 1066—1076
22 Maloney A, Workman P. HSP90 as a new therapeutic target for cancer therapy: the story unfolds. Expert Opin Biol Ther, 2002, 2(1): 3—24
23 James C R, Quinn J E, Mullan P B, et al. BRCA1, a potential predictive biomarker in the treatment of breast cancer. Oncologist, 2007, 12(2): 142—150
24 Horiuchi K, Weskamp G, Lum L, et al. Potential role for ADAM15 in pathological neovascularization in mice. Mol Cell Biol, 2003, 23(16): 5614—5624
25 Raponi M, Harousseau J L, Lancet J E, et al. Identification of molecular predictors of response in a study of tipifarnib treatment in relapsed and refractory acute myelogenous leukemia. Clin Cancer Res, 2007, 13(7): 2254—2260
26 Hutson S M, Lieth E, LaNoue K F. Function of leucine in excitatory neurotransmitter metabolism in the central nervous system. J Nutr, 2001, 131(3): 846S—850S
27 Doggrell S A. Bradykinin B2 receptors as a target in diabetic nephropathy. Curr Opin Investig Drugs, 2006, 7(3): 251—255