Disease Associated Protein-Protein Interaction Network Reconstruction Based on Comprehensive Influence Analysis

Fei Zhu
Soochow University
Feifei Li
Soochow University
Xinghong Ling
Soochow University
Bairong Shen (bairong.shen@scu.edu.cn)
Sichuan University West China Hospital https://orcid.org/0000-0003-2899-1531

Research article

Keywords: Comprehensive influence analysis, Disease associated protein prediction, Network reconstruction, Protein-protein interactions

Posted Date: September 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-60999/v1

RESEARCH

Disease associated protein-protein interaction network reconstruction based on comprehensive influence analysis

Fei Zhu[1,2], Feifei Li[1], Xinghong Ling[1] and Bairong Shen[3*]

* Correspondence:
bairong.shen@scu.edu.cn
[3] Institutes for Systems Genetics, Frontiers Science Center for Disease-related Molecular Network, West China Hospital, Sichuan University, 610041 Chengdu, China
Full list of author information is available at the end of the article

Abstract

Background: Proteins and their interactions are fundamental to biological systems and they are affecting our body. Functional study of protein networks is becoming increasingly essential to get a deep understanding of proteins and their roles in human life and diseases. Although several methods already exist for protein-protein interaction (PPI) network building, the precise reconstruction of disease associated PPI network remains a challenge. In this paper we introduce a novel concept of comprehensive influence of proteins in network, in which direct and indirect connections are adopted for the calculation of influential effects of a protein with different weights. With the optimized weights, we calculate and select the important proteins and their interactions to reconstruct the PPI network for further validation and confirmation.

Results: To evaluate the performance of the method, we compared our model with six existing methods using five standard data sets.

Conclusions: The results indicated that our method outperforms the existed ones. We then applied our model to prostate cancer and Parkinson's disease to predict novel disease associated proteins for the future experimental validation.

Keywords: Comprehensive influence analysis; Disease associated protein prediction; Network reconstruction; Protein-protein interactions

Background

Protein are biological functional units, and protein-protein interaction (PPI) networks are associated with biological signal transduction, gene-expression regulation, energy and substance metabolism, and cell cycle regulation [1]. Network reconstruction and analysis of disease associated PPIs is important to the understanding of protein functions in networks and of the molecular mechanism of diseases [2, 3]. The rapid development of experimental techniques to detect protein interactions has greatly increased the data accumulation for the reconstruction and analysis of PPI network; however, due to the limitations of experimental methods and the complexity of biomedical systems, the experimental results often present high false- positive and false-negative rates [4]. Moreover, experimental verification of the PPI network is time consuming and expensive. Therefore, rapid and accurate computational methods represent useful alternative for predicting protein interactions, with such results providing guidance for experiments, including determination of unknown relationships between disease-causing genes and protein interactions [5].

Such knowledge can be useful to the understanding of protein structure and evolution, as well as the overall function of the network and its associated dynamic processes, because there are many methods based on PPI networks to explore protein structure and other research [6, 7].

Several databases containing protein-protein interactions or association data set are available such as, HPRD [12], IntAct [13], and STRING [14]. However, previous studies suggested that high false-positive rate existed with much of the experimental data used to construct these networks, requiring improvements to yield more accurate PPI networks [15]. There are a number of online resources allowing access to interactome data [16], including those for interactomes associated with Arabidopsis [17] and maize [18], generated using predictive algorithms, the information of interactomes, or a combination of the two. Although high-throughput experiments have allowed generation of large amounts of PPI data, issues concerning false-positive and false-negative results exist. Additionally, small-scale experiments displaying relatively high levels of accuracy require significant resources and man- power. Therefore, many computational methods have been proposed for protein interaction prediction to complement the experimental methods.

The unknown interaction between proteins can be predicted by the theory of link prediction in complex network. PPI network consists of nodes and edges, with the nodes representing proteins and the edges representing associations or interactions between the two proteins [8]. Computational reconstruction of PPI networks are based on different strategies, among which the similarity-based link prediction method is the most widely used. The basic idea is that if two nodes are similar, they are more likely linked. Libennowell et al. considered the influence of common neighbor nodes in the network and introduced similarity indices to include common neighbors [9]. Pujari et al. suggested that each attribute of a pair of connected proteins represents different information [10], and Li et al. proposed the use of a feature set containing user proximity, attributes, and topology for predictions [11]. When analyzing protein interaction pairs, the above-mentioned methods often take the local or path information of the nodes in the network into account, and do not systematically analyze the edge relationships in the network. Based on the network resource allocation, the resource allocation index [19] and Adamix-Adar index [20] are proposed and both of them take the degree information of network nodes into account, and regard the amount of resource transfer information between two nodes as similarity index. Based on the common neighbors index [9], the number of com- mon neighbors is used as a measure of similarity. Considering the influence of node degree, Jaccard index, and other similarity indicators based on local information are derived. Because of its high robustness and easy operation, these algorithms have attracted much attention of the researchers. In addition to extracting structural information, similarity index based on path information is also used in the link pre- diction. The similarity index based on path information includes Katz index and Local Path index [21]. Generally, when two proteins are predicted to be linked, they may share similar functions under certain conditions, suggesting that links are influenced by the interactions and the external neighbors. This concept represents a naive Bayesian model [22]. Additionally, node degree and the clustering coefficient impact on establishing of links by combining information from common neighbor nodes for prediction were also proposed [23, 24].

In the present study, we aim to emphasize the analysis of edges in the PPI network to improve the reconstruction of disease associated PPI network and develop the concept of network influence to evaluate the possibility of interactions between proteins. The model built based on the concept was then compared with the existing methods and was applied in reconstructing several disease associated PPI networks.

Results

Baseline and Evaluation Metrics

To evaluate the performance of our proposed method, six existing methods are used for comparison, they are structural deep-network embedding (SDNE) index [26,27], Common Neighbors (CN) index [9], Jaccard index [28], Sorenson index [29], HPI index [29] and Salton index [29].

To measure the performance of an algorithm, it is necessary to divide the data set into training set and test set. In link prediction, we remove 15% of the edges (retaining the related nodes) as the test set. The remaining 85% of the edges and all the nodes of the network are the training set. If no edges exist between two nodes, we define them as non-existent.

In this work, we used the AUC index to evaluate the performance of the methods [21]. AUC index measures the accuracy as a whole, and it refers to the probability that the score value of a random selected edge is higher than that of a non-existent edge in the test set. In the experiment, it get a similar score between each pair of nodes (edges in train set, test set and non-existent set) in the network by training set with an algorithm. One edge is randomly selected from the test set and another from the non-existent edge. Among the n independent comparisons, if there are n' cases that the edge in the test set has a higher score, then 1 point is added; if there are n" cases that the edge in the test set and that of the non-existent edge shares the same score, 0.5 point is added. Then the AUC index is defined as follows [19].

$$AUC = \frac{n' + 0.5n''}{n} \tag{1}$$

Performance of the proposed model

Based on the perspective of graph theory, proteins and interactions in PPI network can be regarded as nodes and links respectively. The computational challenge about reconstructing PPI network from fixed seed proteins is referred as link prediction problem [30]. In this section, we present our results about the solution of the link prediction problem.

We first divided the data set into training set and test set [35]. The basic format for each record of the data set involved in training and testing is <Protein A, Protein B, Interaction Score>. We take a certain proportion of the whole data set (e.g. 85%) as training set, and then set the record format of the rest (e.g. 15%) to < Protein A, Protein B, whether there is interaction >. After training, we input < Protein A, Protein B > in the test set to determine the interaction between protein A and protein B, and then compare the results with ground truth [27].To evaluate the performance of the algorithms, we use our proposed method to construct interaction

networks with five datasets as Table 1. In this section, we take 85% of each data set as training set and 15% as test set.

Table 1 Interaction networks used for comparative analyses.

Network	*Nodes*	*Edges*	*Description*
PPI[31]	2375	11693	The network is generated by the interaction between proteins extracted from experiments.
Yeast [32]	2361	7182	Data with nodes representing yeast proteins, and links indicating interactions between yeast proteins.
Dolphins [33]	62	159	Data from dolphin studies in New Zealand's Doubtful Sound, with nodes representing dolphins, and links between dolphins indicating dolphins that pair up more often than expected.
USAir	332	2126	A network associated with United States air transportation and available at http://vlado.fmf.uni-lj.si/pub/networks/data/.
GP [34]	259	640	Data obtained from the bibliography of the book Graph Products by Imrich and Klavžar (1999).

Table 2 AUC of comparative analyses.

Dataset	**PPI**	*Yeast*	*Dolphins*	*USAir*	*GP*
Ours	***0.8907***	***0.7148***	0.7748	**0.9263**	**0.7529**
SDNE	0.8868	0.7090	0.7678	0.9163	0.7355
CN	0.8906	0.7147	**0.7778**	0.9261	0.7526
Jaccard	0.8892	0.7076	0.7741	0.8865	0.7451
Sorenson	0.8888	0.7081	0.7740	0.8871	0.7459
HPI	0.8071	0.6002	0.7113	0.7880	0.6394
Salton	0.8899	0.7135	0.7700	0.8989	0.7515

Table 2. shows that area under curve (AUC) of the proposed algorithm was improved relative to previous methods. We can get the following results from five different data sets: (1) In all data sets, the AUC value of our method is greater than 0.71, and the maximum value is more than 0.92, which proves that our method is effective in predicting the relationship between two nodes. (2) The AUC value of proposed method is improved compared with other six methods in five independent data sets. Only in the Dolphin network, the AUC value of our method is 0.003 lower than that of the CN method. The result proves that our method is superior to the other six methods. Fig 1 shows ROC curves of different networks with various methods. It can be seen that our method is effective for various network characterization and analysis.

Discussion

In this section, we applied our method to reconstruct protein interaction networks and to identify potential interactions associated with prostate cancer and Parkinson's disease.

Data Sources

We used the Human Protein Reference Database (HPRD) [12] and IntAct Molecular Interaction Database as the data sources, the STRING database of known and predicted PPIs was used to verify our results [14].

The HPRD integrates information pertaining to domain architecture, posttranslational modifications, interaction networks, and disease associations for each protein in the human proteome. The proteins and interactions in the HPRD

are manually extracted from the literature by experts after interpretation and analysis using do- main knowledge. The IntAct Molecular Interaction Database houses PPI data manually extracted from published literature and includes experimental methods, conditions, and interacting domains, as well as information concerning non-interacting proteins.

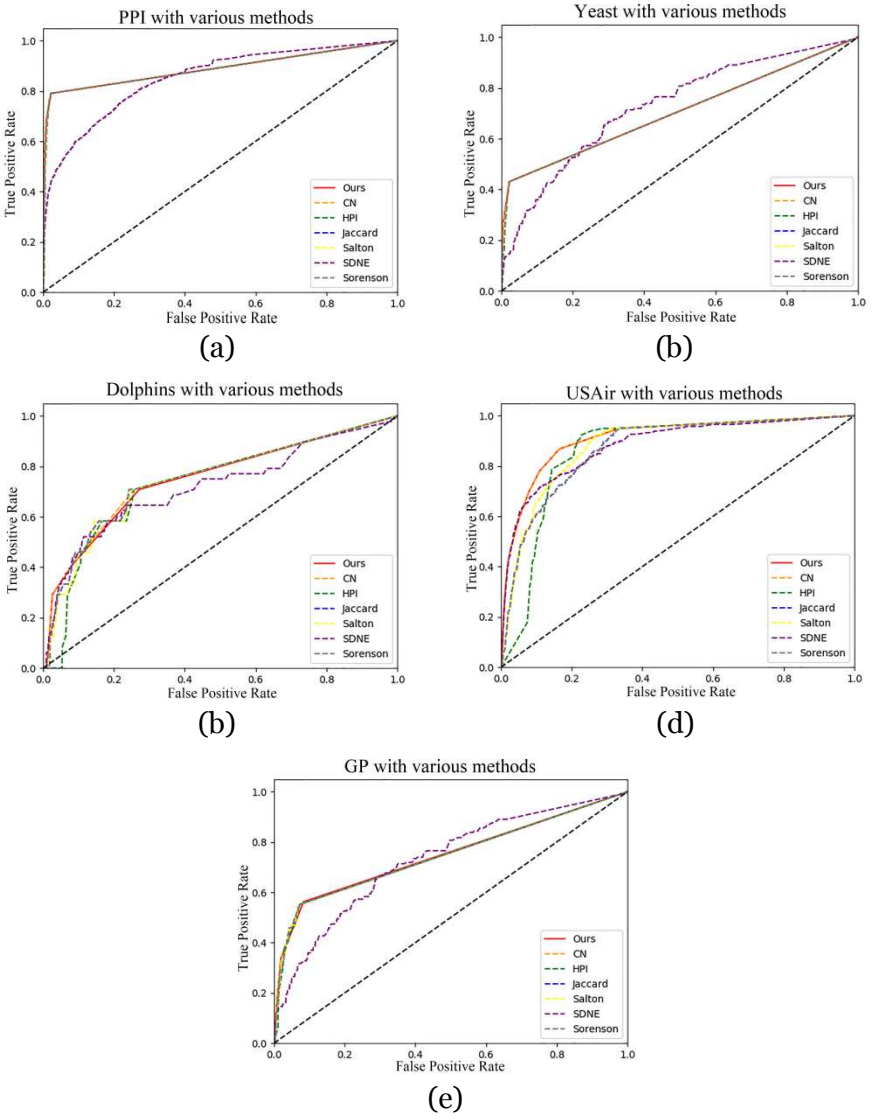


Figure 1 Performance of various methods for different network analysis.

We extracted data from the HPRD and IntAct to gather seed proteins related to prostate cancer and Parkinson's dis-ease, and then unified the various identifiers from multiple databases. After that, we collected the PPI dataset using the seed- protein data with the nearest-neighbor method and initialized the score of each protein-interaction pair.

PPI network of prostate cancer

Prostate cancer is a malignancy in the urinary system that primarily targets middle-aged and elderly men, with an estimated 1,600,000 cases diagnosed and 366,000 deaths annually [36]. To identify genes and proteins related to cancer progression is important to the personalized treatment of this disease [37,38]. To screen prostate cancer associated proteins, we set γ as 0.5 to calculate influence between two nodes, and set β as 0.5 to extract top influence.

Seed-protein identification and unification

Using "prostate" as a retrieval term in the HPRD, we identified 18 proteins for the seed-protein set. We then used UNIPROT to identify prostate-cancer-related genes according to standard gene identifiers, which were subsequently mapped to the seed-protein set according to Swiss-Prot protein identifiers (Table 3.).

Table 3 Prostate cancer seed-protein set.

	HPRDID	*SWISSPROTID*	*PROTEIN*
1	05084	O96017	CHK2 checkpoint homolog
2	01143	P08118	Beta-microseminoprotein
3	02437	P10275	Androgen receptor
4	01885	P12830	Cadherin 1
5	01092	P21757	Macrophage scavenger receptor 1
6	00625	P22455	Fibroblast growth factor receptor 4
7	02997	P29323	EphB2
8	08926	P35680	Hepatocyte nuclear factor 1-beta
9	02486	P50539	Max interacting protein 1
10	02554	P51587	BRCA2
11	03431	P60484	Protein tyrosine phosphatase PTEN
12	01590	Q05823	Ribonuclease L
13	00075	Q15911	Alpha fetoprotein enhancer binding protein
14	10109	Q8NDI1	EH domain-binding protein 1
15	05210	Q92826	Homeobox protein Hox-B13
16	03632	Q99612	Core promoter element binding protein
17	05641	Q9BQ52	ELAC
18	04065	Q9Y6D9	MAD1 mitotic arrest deficient-like 1

Establishment of prostate related PPI network

We collected interacting-protein pairs related to prostate cancer according to our generated seed-protein set. To assess the confidence of each interaction, we used the following scoring rules: 1) protein pairs were assigned a confidence value of 0.8 from the HPRD and 2) other protein pairs were assigned a confidence value from IntAct. This allowed identification of the PPI dataset from the seed-protein set via the nearest-neighbor method, resulting in 488 protein pairs.

The constructed PPI network contained 450 nodes, with an average degree of 2.19 for each node. The network comprised a weakly connected graph that met

the characteristics of a small-world network (Fig 2).

Generation of the scale-free graph resulted in hub nodes associated with the proteins CDH1, EPHB2, BRCA2, PTEN, and CHEK2, with their importance in maintaining network stability suggesting their potentially key roles in the pathogenesis of prostate cancer. Upregulated proteins in tumors likely play a role in tumor invasion and metastasis, with examples including TP53, MDM2, RB1, HDAC1, HDAC2, and SRC. We use cytoscape [25] to visualize the PPI network, and Fig 2 shows the PPI network associated with prostate cancer, and Fig 3

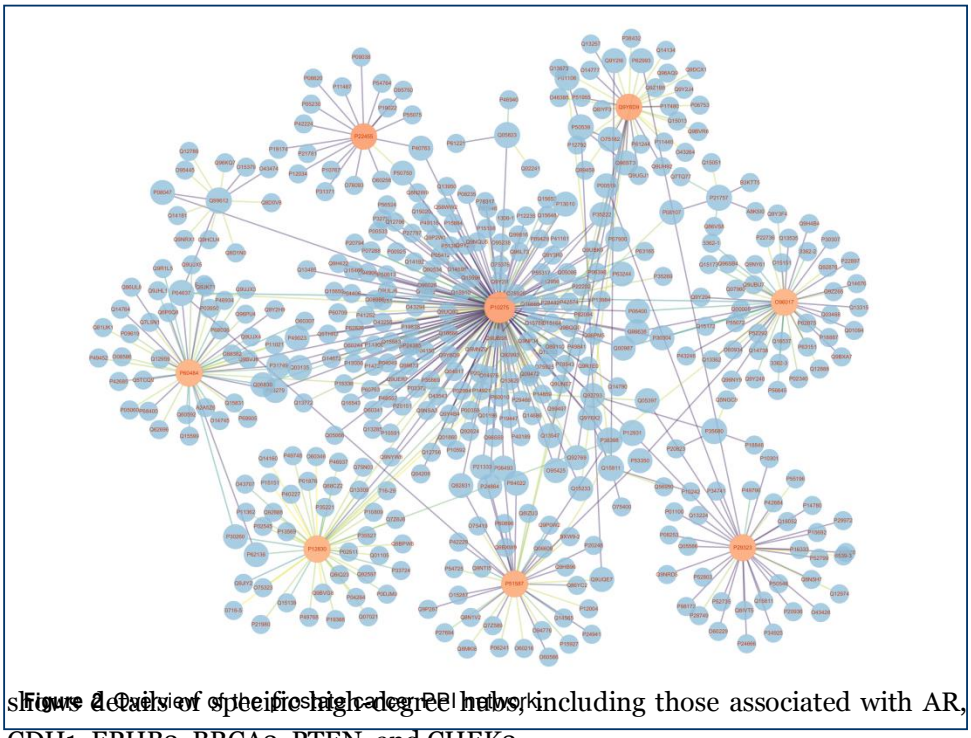
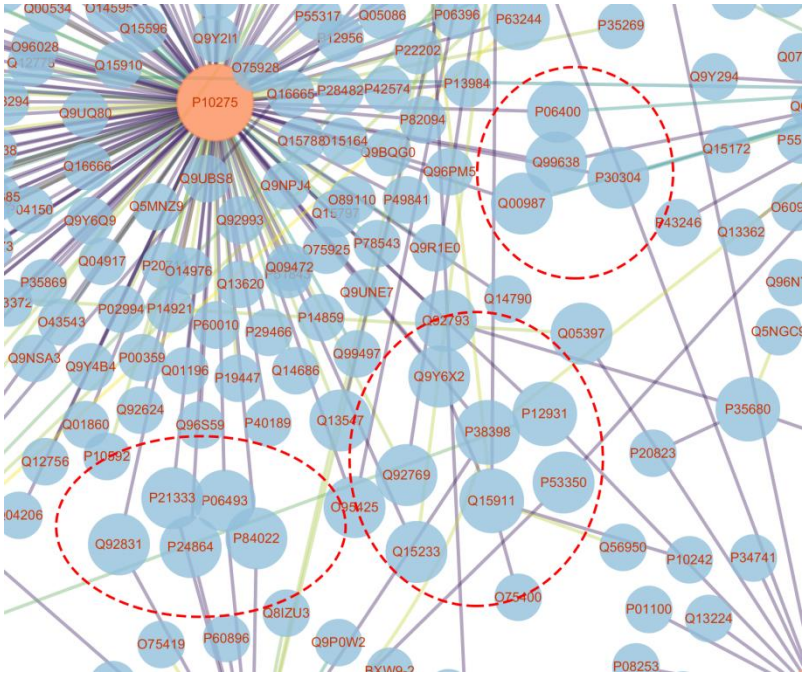


Figure 2. Overview of the significant prostate cancer PPI network.

shows details of specific high-degree nodes, including those associated with AR, CDH1, EPHB2, BRCA2, PTEN, and CHEK2.



Based on comprehensive score of each node, we choose top 24 proteins from established PPI network in Fig 3 (as listed in Table 4.) with our proposed method. These proteins are highly connected proteins in the PPI network associated with prostate cancer, which are closely connected with several hub nodes of the network, and it is important to maintain the stability of the network.

Table 4 Highly connected proteins in the PPI network associated with prostate cancer.

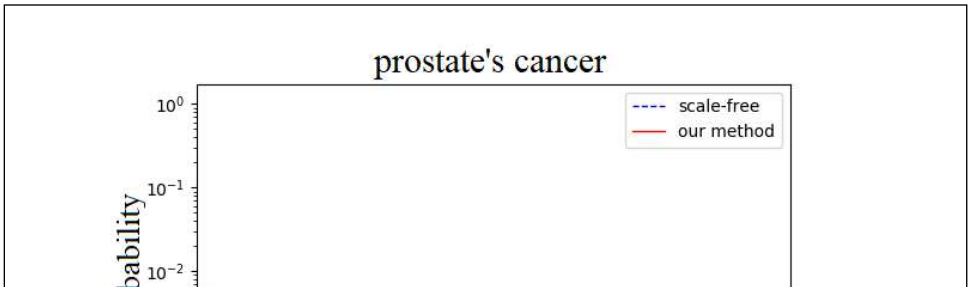
	GENE	*SWISSPROTID*	*PROTEIN*
1	KAT2B	Q92831	PCAF
2	FLNA	P21333	Filaminranks A, alpha
3	CDK1	P06493	CDC2
4	CCNE1	P24864	Cyclin E1
5	SMAD3	P84022	SMAD family member 3
6	GRB2	P62993	Grb2
7	ABL1	P00519	v-abl Abelson murine leukemia viral oncogene homolog 1
8	SVIL	O95425	Supervillin
9	HDAC1	Q13547	Histone deacetylase 1
10	HDAC2	Q92769	Histone deacetylase 2
11	NONO	Q15233	Non pou domain containing octamer binding protein
12	SRC	P12931	c-Src
13	CDC27	P30260	Anaphase promoting complex, subunit 3
14	PPP1CA	P62136	Protein phosphatase 1, catalytic subunit, alpha isoform
15	CAV1	Q03135	Caveolin 1
16	PXN	P49023	Paxillin
17	UBE2I	P63279	Ubiquitin conjugating enzyme E2I
18	PRDX1	Q06830	Peroxiredoxin 1
19	AKT1	P31749	AKT1

Figure 3 Representative central nodes. Nodes included AR(P10275), CDH1(P12830), and CHEK2(O96017) connected with several hub nodes in the network.

20	RB1	P06400	Retinoblastoma 1
21	RAD9A	Q99638	RAD9
22	CDC25A	P30304	CDC 25A
23	MDM2	Q00987	MDM2
24	TP53	P04637	p53

Prostate cancer associated network analysis

The analysis of prostate cancer associated PPI network indicates that only a small portion of the proteins in the network were highly connected, with most having short-path connections which is in agreement with the characterization of a small-world network. Additionally, these results are consistent with the characteristics of PPI networks associated with complex diseases, in that the network extends from a central node to the periphery. In most cases, the node degree of a complex network obeys a power-law distribution rather than a Poisson distribution [39]. For a randomly selected node, the probability of its degree being k is , where r and k are constants, thereby representing a scale-free network. Previous studies suggest that many complex networks are characterized by scale-free properties. Similarly, our analysis indicated that the generated network displayed a node-degree distribution representative of a scale-free network (Fig 4).



We selected proteins with the highest influence scores, resulting in an average of 75% confirmed PPIs with those in STRING, the other 25% verified in the literature. We identified two proteins with novel interactions (AQP1 and MMP2). Studies suggest that AQP1 upregulates MMP2 and MMP9 levels, leading to increased invasion and adhesion of colon cancer cells and migration of bladder cancer cells. Aquaporins (AQP) are widely distributed in tissues and cells, with AQP1 an integrative membrane protein in human erythrocytes responsible for maintaining cell permeability. Urine proteome analysis to identify disease biomarkers for prostate cancer, renal cell carcinoma, bladder cancer, urothelial carcinoma, and renal Fanconi syndrome suggested as a possible marker AQP1 [40 – 44].

MMP2 is a protease capable of degrading collagen and the extracellular matrix and involved in tumor-cell infiltration of connective tissue stroma, small blood vessels, and lymphatic tissue [45]. Additionally, MMP2 is highly expressed in prostate cancer tissues and plays a role similar to that of proto-oncogenes, as attenuated MMP2 levels inhibit tumor-cell proliferation, invasion, and migration and promote apoptosis, thereby alleviating tumor malignancy.

Relationships between AQP1 and prostate cancer invasion and metastasis in connection with MMP2 status remain unknown; however, previous studies implicated both AQP1 and MMP2 are associated with prostate cancer [46, 47]. AQP1 is localized to the kidney, with studies reporting its elevated expression in the prostate gland in testicles of rats. Additionally, Kyoto Encyclopedia of Genes and Genomes (KEGG) analysis indicated that MMP2 is involved in bladder cancer and other cancer pathways. Although investigations have focused on their roles in colon cancer [48,49] and in vivo status in animal studies [50], the relationship between these two proteins and prostate cancer is not reported yet. Our analysis suggested a potential role for these two proteins in prostate cancer; although experimental studies are necessary to confirm this relationship.

PPI network of Parkinson's disease

Parkinson's disease [51] is a common age-related brain disorders defined primarily as a movement disorder accompanied by typical symptoms of resting tremors, rigidity, bradykinesia, and postural instability [52]. For Parkinson's disease, we detected indirect influence have negative effects on results with the same

parameter as that for prostate cancer. So γ was set as 0.45 to reduce the indirect impact, and β was set as 0.05 to improve accuracy in top.

Seed-protein identification and unification

Using Parkinson's disease as retrieval word in the HPRD, we obtained 14 proteins for the seed-protein set and mapped them to gene symbols and Swiss-Prot protein identifiers (Table 5.)

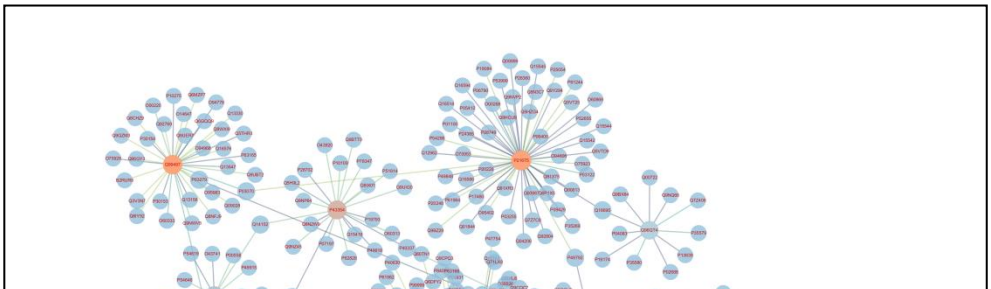
Table 5 Parkinson's disease seed-protein set.

	GENE	*HPRDID*	*SWISSPROTID*	*PROTEIN*
1	MAPT	01142	P10636	Microtubule associated protein tau
2	SNCA	01227	P37840	Synuclein alpha
3	TRPM7	10418	Q96QT4	Transient receptor potential cation channel, subfamily M, member 7
4	PINK1	10514	Q9BXM7	PTEN induced putative kinase 1
5	HTRA2	05919	O43464	HtrA serine peptidase 2
6	PARK7	03961	Q99497	Oncogene DJ1
7	PARK2	03967	O60260	Parkinson disease (autosomal recessive, juvenile) 2, parkin
8	PRKAG2	04119	Q9UGJ0	PRKAG2
9	NCAIP	04804	Q9Y6H5	Synphilin 1
10	TAF1	02436	P21675	Transcription factor IID
11	NDUFV2	02757	P19404	NADH ubiquinone oxidoreductase flavoprotein 2
12	NR4A2	03493	P43354	Nuclear receptor subfamily 4, group A, member 2
13	UCHL1	01877	P09936	Ubiquitin carboxyl terminal esterase L1
14	ADH1C	00066	P00326	Alcohol dehydrogenase 3

Establishment of Parkinson's disease PPI network

Application of the scoring rules described above resulted in identification of 568 pairs of interacting proteins associated with Parkinson' s disease using the nearest neighbor method.

The average node degree in the PPI network was 2.12, which fit the characteristics of a small-world network (Fig 5). Fig 5 shows the PPI network associated with Parkinson's disease, and Fig 6 shows details of representative central nodes including SNCA, PARK2, MAPT connected with several hub nodes in the network.



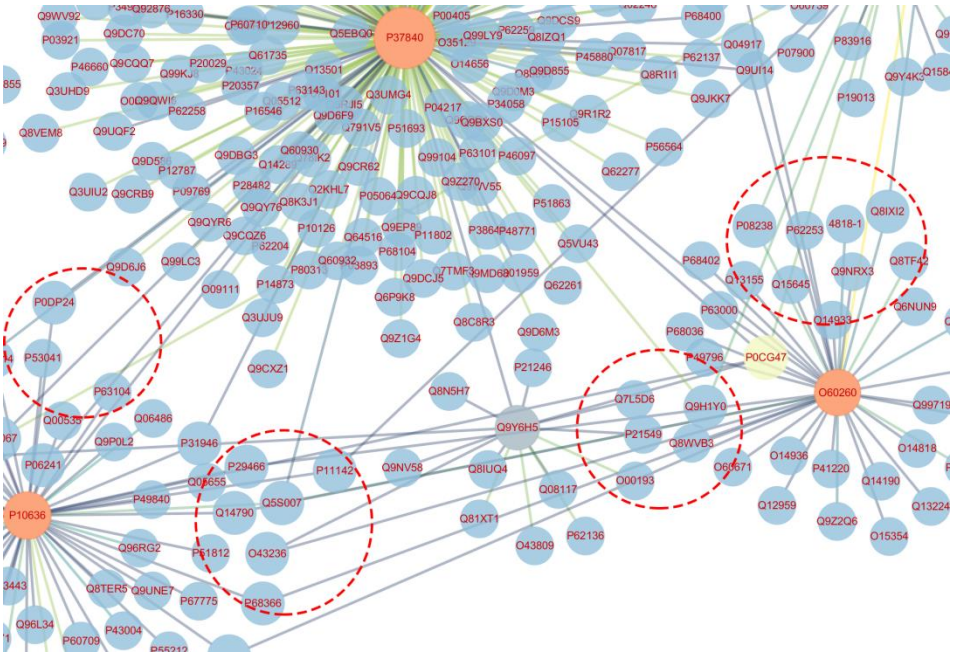


Figure 6 Representative central nodes in the Parkinson's disease PPI network. Nodes includeding SNCA(P37840). PARK2(O60260). MAPT(P10636) connected with several hub nodes in the network.

We choose top 21 highly connected proteins from the PPI network associated with Parkinson's disease. We can see that these proteins are almost identical to the blue proteins in Fig 6. These proteins maintain the stability of the network and are connected with several hub nodes of the network. Proteins such as UBB, HSPA8, ABL1 and CASP8 are involved in biological processes of Parkinson's disease, for example, UBB shows relationship with positive regulation of protein monoubiquitination [53] and HSPA8 is related to regulation of protein import and regulation of protein complex assembly [54]. LRRK2 may play a role in the phosphorylation of proteins central to Parkinson's disease, which has certain effects on biological processes and molecular functions [55]. Table 6 shows highly

connected proteins in the PPI network associated with Parkinson’ s disease.

Table 6 Highly connected proteins in the PPI network associated with Parkinson’s disease.

	Gene	*SwissprotID*	*Protein*
1	UBB	P0CG47	Ubiquitin B
2	LRRK2	Q5S007	Leucine rich repeat kinase 2
3	APP	P05067	Amyloid beta A4 protein
4	GSK3B	P49841	Glycogen synthase kinase 3 beta
5	FYN	P06241	Fyn
6	YWHAB	P31946	14-3-3 Beta
7	SEPT4	O43236	Septin 4
8	CASP8	Q14790	Caspase 8
9	CASP1	P29466	Caspase 1
10	TUBA4A	P68366	Tubulin alpha 1
11	TUBB	P07437	Tubulin, beta
12	HSPA8	P11142	Heat shock 70 kDa protein 8
13	PSEN1	P49768	Presenilin 1
14	YWHAH	Q04917	14-3-3 Eta
15	ATG5	Q9H1Y0	Autophagy protein 5-like
16	MAP1LC3B	Q9GZQ8	Microtubule-associated protein 1 light chain 3 beta
17	HSP90AB1	P08238	HSP90B
18	RHOT1	Q8IXI2	Mitochondrial Rho 1
19	ABL1	P00519	v-abl Abelson murine leukemia viral oncogene homolog 1
20	UBE2I	P63279	Ubiquitin conjugating enzyme E2I
21	RANBP2	P49792	Ran binding protein 2

Parkinson’s disease associated network Analysis

As shown in Fig 7, the node-degree distribution in the Parkinson’s disease associated network also met the characteristics of a scale-free network, which conforms to a power-law distribution. For the Parkinson’s disease data, 69% of the predicted PPI can be validated by STRING. The others are the novel Parkinson’s disease associated PPI and are listed in Table 7.

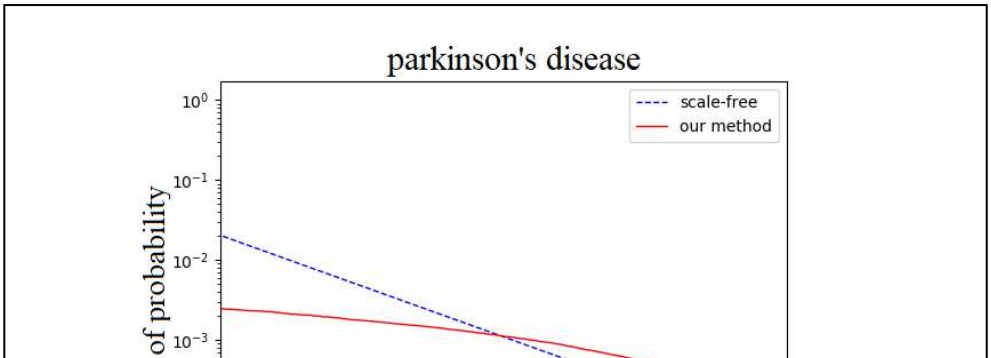


Table 7 Highly connected proteins in the PPI network associated with Parkinson's disease.

PROTEIN A	*PROTEIN B*	*DESCRIPTION*
APOE	CASP6	APOE is up-regulated in brain tumors in Parkinson's patients [56]. CASP6 modulates apoptosis in Parkinson's disease (https: //www.genome.jp /kegg/pathway/hsa/hsa05012.html).
GSK3B	RNF19A	GSK3β is possibly associated with neuron apoptosis Parkinson's disease [57], Alzheimer's disease, and schizophrenia. GSK3β possibly regulates RNF19A stability [58].
MAPT	CSNK2A2	MAPT is associated with several neurodegenerative disorders, such as Parkinson's disease [59]. The CSNK2A2 paralog CSNK2A1 is associated with MAPT related activities [60].

We investigated the functions of the identified interacting proteins with the highest influence scores (APOE, CASP6, GSK3B, RNF19A, MAPT, and CSNK2A2). The PPI network included several central nodes (TAF1, PARK7, SNCA, PARK2, UCHL1, MAPT, HTPR2, and NR4A2) involved in Parkinson's disease [61–63], as well as others (UBB, LRRK2, APP, GSK3B, FYN, SEPT4, RANBP2, ABL1, and MAP1LC3B).

Understanding complex biological processes involves high levels of data beyond the scope of experimental methods; therefore, computational approaches for reconstruction of disease associated PPI network have become increasingly important. Here, we compare our proposed methods with six existing methods and the results indicate that our method is effective to identify disease associated proteins and their interactions. We then generated two novel PPI networks associated with prostate cancer and Parkinson's disease using our method. A novel interaction between AQP1 and MMP2 in prostate cancer and three novel potential interaction pairs for APOE and CASP6, GSK3B and RNF19A, MAPT and CSNK2A2 in Parkinson's disease were predicted for further experimental validation. Our results suggest that this method could be applied to different diseases for identifying disease associated PPIs.

Conclusions

The diverse protein-protein interaction networks have dramatic effects on

biological system. The disease associated PPI networks are generally reconstructed from experimental data with computational models but with limited accuracy. We developed a novel concept of comprehensive influence of proteins in network for reconstructing the PPI network. Our model outperforms the state-of-the-art ones and we then applied our model to identify novel interactions for further validation.

Methods

A PPI network can be described as an undirected graph, G = (V, E), where V denotes the set of nodes, and E represents the set of edges. The importance of the interaction or node in a network can be described by different weights. Fig 8 shows an example of PPI network, where the left is a part of PPI network extracted from STRING database with TP53 as the keyword and then visualized in Cytoscape [25] and the right is the weighted PPI network abstracted from the left. The initial weights in Figure are obtained from protein interaction database and they are transformed accordingly to understand easy.

In a concrete PPI network(such as network in section 5), the function of a protein is influenced by both its direct and indirect interactions, therefore it is necessary to consider the effect of its surrounding sub-network. To evaluate the influence of a protein in a network, we introduce the following definitions, taking the network shown in Fig 1 as an example.

Definition 1: <A, B> represents a direct influence and is used to evaluate direct interactions between two proteins, A and B.

For example, in Fig 1, the influence of <A, B> for nodes A and B is 15 (direct influence), and that of <B, C> is 20.

Definition 2: Indirect influence is used to evaluate an indirect interaction between two proteins through other proteins.

For example, in Fig 1, nodes A and B exhibit only a direct influence (i.e., <A, B>), whereas we define node A as having an interaction with node C through node B (i.e., an indirect influence; <A, B, C>). Because indirect influence is exerted by other proteins, it is relevant to more than two primary nodes. In Fig 1, nodes A and C might indirectly influence one another through node B.

Definition 3: An indirect influence involves the shortest distance between two nodes.

For example, in Fig 1, two pathways from D to F include <D, E, F> and <D, C, E, F>, respectively. When there are additional pathways between two proteins, for the purpose of simplification, we only consider the case with the shortest distance.

Definition 4: A hop between nodes i and j is the minimum number of nodes required to traverse from node i to node j in the network.

We use the following approach to generate a score for each protein by evaluating its value and influence to other proteins in the PPI network:

$$\underset{j\in N(\mathrm{i})\cap \mathrm{P}}{Influence_{ij}} = \underset{j\in N(\mathrm{i})\cap \mathrm{P}}{DI_{ij}} + \gamma \underset{j\in N(\mathrm{i})\cap \mathrm{P}}{IDI_{ij}} \qquad (2)$$

$$IDI = \gamma^{hop} \sum DI \tag{3}$$

where i and j represent two solid nodes, γ is a discount($0 < \gamma < 1$), P is the set of nodes in protein network, and N(i) represents the set of proteins directly interacting with node i. In Equation (1), Influenceij denotes the influence between nodes i and j and is the sum of all direct and indirect influence scores. DIij represents the set of direct links for nodes without considering the possible influence of shared nodes. In the case of indirect links, we assume that node i can possibly influence node j through node k, which is termed a "one-hop link":

$$IDI_{ikj} = \gamma DI_{ik} + \gamma DI_{kj} \tag{4}$$

Because there are likely many one-hop links, there will consequently be many one-hop influence scores. In Equation (3), k is a variable node that links nodes i and j.

IDIiklj is used to evaluate the two-hop influence between nodes i and j and assumes that node i can get to node j through nodes k and l, and vice versa. For example, node i is linked to variable node k, which is linked to node l, and node l is linked to node j. We then assume that node i has possible influence on node j as part of a two-hop link:

$$IDI_{iklj} = \gamma^2 DI_{ik} + \gamma^2 DI_{kl} + \gamma^2 DI_{lj} \tag{5}$$

Link prediction-based methods for predicting protein-protein interactions have their own merits, and data preprocessing has different effects on the performance of the algorithm. In this work, in order to reduce the problem caused by data bias, multiple data sources were used to optimize the parameters and maintain the data diversity for the generality of model. We proposed a link-based prediction method to extract protein interaction information from multiple data sources, and then evaluate candidate proteins. For the reconstruction of PPI network, proteins and links with high scores are selected based on Equation 1, as explained in Algorithm 1 and Algorithm 2.

Algorithm 1 describes the calculation of comprehensive influence between two nodes (i.e. proteins) in a network. Algorithms 2 describes how to build protein interaction networks based on the selected proteins from Algorithm 1.

Algorithm 1: Computing comprehensive influence in the network

Input： graph, *threshold*, discount factor γ

Return: comprehensive influence between two nodes

1: get all paths between node i and node j
2: **for** each path in paths:
3: **if** i and j are directly connected:
4: calculate direct influence by
DI = weight of i and j as directing influence score between i and j
5: **if** i and j are indirectly connected:
6: get all paths from i to j with $hop < threshold$
7: **for** each path in paths
8: get hop number hop between i and j
9: calculate indirect influence score by influence

$$IDI = \gamma^{hop} \sum_{all\ in\ path} \text{weight of interaction}$$

10: calculate influence score between i and j by $influence=DI+IDI$

Algorithm 2: Constructing PPI network via comprehensive influence

Input：interaction data, threshold β
Return: G
1: extract interactive data (node a, node b, weight)
2: initialize the score between interaction pairs
3: divide Train list and Test list
4: create the network G of relationships by retrieving direct partners from Train list
5: **for** each node in G
6: calculate influence score between a and b with Algorithm 1
7: sort candidate proteins of a in descending order
8: acquire the top β% of the candidate proteins
9: **end for**

Abbreviations

PPI:protein-protein interactions; AUC: area under curve; SDNE:structural deep-network embedding; CN:Common Neighbors

Ethics approval and consent to participate

Not applicable.

Consent to publish

Not applicable.

Availability of data and materials

The data of USAir and Yeast were downloaded from http://vlado.fmf.uni-lj.si/pub/networks/data/. Dolphins data was downloaded from http://www-personal.umich.edu/ mejn/netdata/. GP data was obtained from the bibliography of the book Graph Products in https://www.routledge.com/Handbook-of-Product-Graphs/Hammack-Imrich-Klavzar/p/book/9781138199088. And PPI data was found from https://www.nature.com/articles/nature750.pdf

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by National Natural Science Foundation of China (61303108); The National Key Research and Development Program of China (Grant No. 2016YFC1306605), The Natural Science Foundation of Jiangsu Higher Education Institutions of China (17KJA520004); Suzhou Key Industries Technological Innovation-Prospective Applied Research Project (SYG201804); Program of the Provincial Key Laboratory for Computer Information Processing Technology (Soochow University) (KJS1524); A Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

Authors's contributions

FZ, FL, XL and BS designed the study. FZ created the algorithm with WC. Then FL made the implementation and analyzed the data. The paper was jointly written by FZ, FL, and BS. All authors have read and approved the final manuscript.

Acknowledgements

Not applicable.

Author details

[1] School of Computer Science and Technology, Soochow University, 215006 Suzhou, China. Key Laboratory for Computer Information Processing Technology, Soochow University, 215006 Suzhou, China. Institutes for Systems Genetics, West China Hospital, Sichuan University, 610041 Chengdu, China.

References

1. Keskin, O., Tuncbag, N., Gursoy, A.: Predicting protein¨cprotein interactions from the molecular to the proteome level. Chemical Reviews **116**(8), 4884–4909 (2016)
2. Alvarez, M.J., Shen, Y., Giorgi, F.M., Lachmann, A., Ding, B.B., Ye, B.H., Califano, A.: Functional characterization of somatic mutations in cancer using network-based inference of protein activity. Nature Genetics **48**(8), 838–847 (2016)
3. Azevedo, H., Khaled, N.A., Santos, P., Bertonha, F.B., Moreirafilho, C.A.: Temporal analysis of hippocampal ca3 gene coexpression networks in a rat model of febrile seizures. Disease Models Mechanisms **11**(1) (2018) 4. Cheng, F., Liu, C., Shen, B., Zhao, Z.: Investigating cellular network heterogeneity and modularity in cancer: a network entropy and unbalanced motif approach. BMC Systems Biology **10**(3), 65–65 (2016)
5. Zhu, F., Liu, Q., Zhang, X., Shen, B.: Protein interaction network constructing based on text mining and reinforcement learning with application to prostate cancer. Iet Systems Biology **9**(4), 106–112 (2015)
6. Cowen, L.J., Ideker, T., Raphael, B.J., Sharan, R.: Network propagation: a universal amplifier of genetic associations. Nature Reviews Genetics **18**(9), 551–562 (2017)
7. Hoeng, J., Talikka, M., Martin, F., Sewer, A., Yang, X., Iskandar, A.R., Schlage, W.K., Peitsch, M.C.: Case study: the role of mechanistic network models in systems toxicology. Drug Discovery Today **19**(2), 183–192 (2014)
8. Gani, O.A.B.S.M., Thakkar, B., Narayanan, D., Alam, K.A., Kyomuhendo, P., Rothweiler, U., Tellofranco, V., Engh, R.A.: Assessing protein kinase target similarity: Comparing sequence, structure, and cheminformatics approaches. Biochimica et Biophysica Acta **1854**(10), 1605–1616 (2015)
9. Libennowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the Association for Information Science and Technology **58**(7), 1019–1031 (2007)
10. Pujari, M., Kanawati, R.: Link prediction in complex networks by supervised rank aggregation **1**, 782–789 (2012)
11. Li, Y., Niu, K., Tian, B.: Link prediction in sina microblog using comprehensive features and improved svm algorithm, 18–22 (2014)
12. Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., Jonnalagadda, C.K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T.K., Gronborg, M., *et al.*: Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Research **13**(10), 2363–2371 (2003)
13. Kerrien, S., Alamfaruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R.P., *et al.*: Intact–open source resource for molecular interaction data. Nucleic Acids Research **35**, 561–565 (2007)
14. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C.J., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., *et al.*: String 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research **37**(37), 412–416 (2009)
15. Szeto, A., Fung, S., Lee, E.A., Wong, A.K.C.: Prediction of protein-protein interaction via co-occurring aligned pattern clusters. Methods **110**, 26–34 (2016)
16. Szklarczyk, D., Jensen, L.J.: Protein-protein interaction databases. Methods of Molecular Biology **1278**, 39–56 (2015)

17. Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., Shim, J.E., Shim, H., Kim, H., Kim, C., *et al.*: Aranet v2: an improved database of co-functional gene networks for the study of arabidopsis thaliana and 27 other nonmodel plant species. Nucleic Acids Research **43**, 996–1002 (2015)
18. Zhu, G., Wu, A., Xu, X., Xiao, P., Lu, L., Liu, J., Cao, Y., Chen, L., Wu, J., Zhao, X.: Ppim: A protein-protein interaction database for maize. Plant Physiology **170**(2), 618–626 (2016)
19. Zhou, T., Lu, L., Zhang, Y.: Predicting missing links via local information. European Physical Journal B **71**(4), 623–630 (2009)
20. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social Networks **25**(3), 211–230 (2003)
21. Lu, L., Jin, C., Zhou, T.: Similarity index based on local paths for link prediction of complex networks. Physical Review E **80**(4), 046122 (2009)
22. Liu, Z., Zhang, Q., Lu, L., Zhou, T.: Link prediction in complex networks: A local naïve bayes model. EPL **96**(4), 48007 (2011)
23. Lingling, Y., Zengqiang, C., Qing, Z.: Analysis of key nodes in china's aviation network based on the degree centrality indicator and clustering coefficient. CAAI Transactions on Intelligent Systems (2016)
24. Yang, G., Zhang, Y.P., Qian, F.L., Zhao, S.: Combined with node degree and node clustering coefficient of link prediction algorithm. Journal of Chinese Computer Systems (2017)
25. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., Ideker, T.: Cytoscape: A software environment for integrated models of biomolecular interaction networks. Genome Research **13**(11), 2498–2504 (2003)
26. Crichton, G., Guo, Y., Pyysalo, S., Korhonen, A.: Neural networks for link prediction in realistic biomedical graphs: a multi-dimensional evaluation of graph embedding-based approaches. BMC Bioinformatics **19**(1), 176
27. Wang, D., Peng, C., Zhu, W.: Structural deep network embedding. In: the 22nd ACM SIGKDD International Conference (2016)
28. Güne?, S., Gündüz-??üdücü, u., ?ataltepe, Z.: Link prediction using time series of neighborhood-based node similarity scores. Data Mining Knowledge Discovery **30**(1), 147–180
29. Martinez, V., Berzal, F., Cubero, J.: A survey of link prediction in complex networks. ACM Computing Surveys **49**(4), 69 (2017)
30. Lei, C., Ruan, J.: A novel link prediction algorithm for reconstructing protein?protein interaction networks by topological similarity. Bioinformatics **29**(3), 355–364 (2013)
31. Lu, L., Zhou, T.: Link prediction in complex networks: A survey. Physica A-statistical Mechanics and Its Applications **390**(6), 1150–1170 (2011)
32. Bu, D., Zhao, Y., Cai, L., Xue, H., Zhu, X., Lu, H., Zhang, J., Sun, S., Ling, L., Zhang, N., *et al.*: Topological structure analysis of the protein–protein interaction network in budding yeast. Nucleic Acids Research **31**(9), 2443–2450 (2003)
33. Newman, M.E.J.: Finding community structure in networks using the eigenvectors of matrices. Physical Review E **74**(3), 036104 (2006)
34. Hasan, M.A., Zaki, M.J.: A Survey of Link Prediction in Social Networks. Springer, ??? (2011)
35. Hashemifar, S., Neyshabur, B., Khan, A.A., Xu, J.: Predicting protein?protein interactions through sequence-based deep learning. Bioinformatics **34**(17) (2018)
36. Torre, L.A., Bray, F.I., Siegel, R.L., Ferlay, J., Lortettieulent, J., Jemal, A.: Global cancer statistics, 2012. CA: A Cancer Journal for Clinicians **65**(2), 87–108 (2015)
37. Zhang, D., Park, D., Zhong, Y., Lu, Y., Rycaj, K., Gong, S., Chen, X.M., Liu, X., Chao, H., Whitney, P., *et al.*: Stem cell and neurogenic gene-expression profiles link prostate basal cells to aggressive prostate cancer. Nature Communications **7**(1), 10798–10798 (2016)
38. Zhao, S.G., Chang, S.L., Erho, N., Yu, M., Lehrer, J., Alshalalfa, M., Speers, C., Cooperberg, M.R., Kim, W., Ryan, C.J., *et al.*: Associations of luminal and basal subtyping of prostate cancer with prognosis and response to androgen deprivation therapy. JAMA Oncology **3**(12), 1663–1672 (2017)
39. Virkar, Y.S.: Power-law distributions and binned empirical data. The Annals of Applied Statistics **8**(1), 89–119 (2014)
40. Morrissey, J.J., Mobley, J., Song, J., Vetter, J., Luo, J., Bhayani, S.B., Figenshau, R.S., Kharasch, E.D.: Urinary concentrations of aquaporin-1 and perilipin-2 in patients with renal cell carcinoma correlate with tumor size and stage but not grade. Urology **83**(1), 2–6 (2014)
41. Morrissey, J.J., Mellnick, V.M., Luo, J., Siegel, M.J., Figenshau, R.S., Bhayani, S.B., Kharasch, E.D.: Evaluation of urine aquaporin-1 and perilipin-2 concentrations as biomarkers to screen for renal cell carcinoma: A prospective cohort study. JAMA Oncology **1**(2), 204–212 (2015)
42. Rodrigues, D., Monteiro, M., Jeronimo, C., Henrique, R., Belo, L., Bastos, M.D.L., De Pinho, P.G., Carvalho, M.: Renal cell carcinoma: a critical analysis of metabolomic biomarkers emerging from current model systems. Translational Research **180**, 1–11 (2017)
43. Rubenwolf, P., Georgopoulos, N.T., Clements, L.A., Feather, S., Holland, P., Thomas, D.F.M., Southgate, J.: Expression and localisation of aquaporin water channels in human urothelium in situ and in vitro. European Urology **56**(6), 1013–1024 (2009)
44. Tomita, Y., Dorward, H., Yool, A.J., Smith, E., Townsend, A., Price, T.J., Hardingham, J.E.: Role of aquaporin 1 signalling in cancer development and progression. International Journal of Molecular Sciences **18**(2), 299 (2017)
45. Wu, W., Gao, H., Li, X., Peng, S., Yu, J., Liu, N., Zhan, G., Zhu, Y., Wang, K., Guo, X.: ?-hcg promotes epithelial ovarian cancer metastasis through erk/mmp2 signaling pathway. Cell Cycle **18**(1), 46–59 (2019)
46. Lacorte, L.M., Rinaldi, J.C., Justulin, A.L., Delella, F.K., Moroz, A., Felisbino, S.L.: Cadmium exposure inhibits mmp2 and mmp9 activities in the prostate and testis. Biochemical and Biophysical Research Communications **457**(4), 538–541 (2015)
47. Brundl, J., Wallinger, S., Breyer, J., Weber, F., Evert, M., Georgopoulos, N.T., Rosenhammer, B., Burger, M., Otto, W., Rubenwolf, P.: Expression, localisation and potential significance of aquaporins in benign and malignant human prostate tissue. BMC Urology **18**(1), 75 (2018)
48. Knight, J., Kim, E., Ivanov, I., Davidson, L.A., Goldsby, J.S., Hullar, M.A.J., Randolph, T.W., Kaz, A.M., Levy, L., Lampe, J.W., *et al.*: Comprehensive site-specific whole genome profiling of stromal and epithelial colonic gene signatures in human sigmoid colon and rectal tissue. Physiological Genomics **48**(9), 651–659 (2016)
49. Yde, J., Keely, S.J., Wu, Q., Borg, J.F., Lajczak, N., Odwyer, A.M., Dalsgaard, P., Fenton, R.A., Moeller, H.B.: Characterization of aqps in mouse, rat, and human colon and their selective regulation by bile acids. Frontiers in Nutrition 3, 46 (2016)
50. Jiang, Y., Zhang, C., Ma, J., Wang, L., Gao, J., Ren, J., He, W., Wang, S., Sheng, S., Huang, X.: Expression of

matrix metalloproteinases-2 and aquaporin-1 in corneoscleral junction after angle-closure in rabbits. BMC Ophthalmology **19**(1), 43 (2019)
51. Berg, D., Adler, C.H., Bloem, B.R., Chan, P., Gasser, T., Goetz, C.G., Halliday, G.M., Lang, A.E., Lewis, S.J.G., Li, Y., *et al.*: Movement disorder society criteria for clinically established early parkinson's disease. Movement Disorders **33**(10), 1643–1646 (2018)
52. Aarsland, D., Creese, B., Politis, M., Chaudhuri, K.R., Ffytche, D.H., Weintraub, D., Ballard, C.: Cognitive decline in parkinson disease. Nature Reviews Neurology **13**(4), 217–231 (2017)
53. Kazlauskaite, A., Kondapalli, C., Gourlay, R., Campbell, D.G., Ritorto, M.S., Hofmann, K., Alessi, D.R., Knebel, A., Trost, M., Muqit, M.M.K.: Parkin is activated by pink1-dependent phosphorylation of ubiquitin at ser65. Biochemical Journal **460**(1), 127–141 (2014)
54. Orenstein, S.J., Cuervo, A.M.: Chaperone-mediated autophagy: molecular mechanisms and physiological relevance. Seminars in Cell and Developmental Biology **21**(7), 719–726 (2010)
55. Angeles, D.C., Gan, B.H., Onstead, L., Zhao, Y., Lim, K., Dachsel, J.C., Melrose, H.L., Farrer, M.J., Wszolek, Z.K., Dickson, D.W., *et al.*: Mutations in lrrk2 increase phosphorylation of peroxiredoxin 3 exacerbating oxidative stress?induced neuronal death. Human Mutation **32**(12), 1390–1397 (2011)
56. Tropea, T.F., Xie, S.X., Rick, J., Chahine, L.M., Dahodwala, N., Doshi, J., Davatzikos, C., Shaw, L.M., Van Deerlin, V.M., Trojanowski, J.Q., *et al.*: Apoe, thought disorder, and spare?ad predict cognitive decline in established parkinson's disease. Movement Disorders **33**(2), 289–297 (2018)
57. Yue, P., Gao, L., Wang, X., Ding, X., Teng, J.: Intranasal administration of gdnf protects against neural apoptosis in a rat model of parkinson's disease through pi3k/akt/gsk3? pathway. Neurochemical Research 42(5), 1366–1374 (2017)
58. Ma, P., Yang, X., Kong, Q., Li, C., Yang, S., Li, Y., Mao, B.: The ubiquitin ligase rnf220 enhances canonical wnt signaling through usp7-mediated deubiquitination of ?-catenin. Molecular and Cellular Biology **34**(23), 4355–4366 (2014)
59. Davis, A.A., Andruska, K.M., Benitez, B.A., Racette, B.A., Perlmutter, J.S., Cruchaga, C.: Variants in gba, snca, and mapt influence parkinson disease risk, age at onset, and progression. Neurobiology of Aging **37**, 2–9 (2016)
60. Raftery, M.J., Campbell, R., Glaros, E.N., Rye, K., Halliday, G.M., Jessup, W., Garner, B.: Phosphorylation of apolipoprotein-e at an atypical protein kinase ck2 psd/e site in vitro? Biochemistry **44**(19), 7346–7353 (2005)
61. Campelo, C.L.D.C., Silva, R.H.: Genetic variants in snca and the risk of sporadic parkinson's disease and clinical outcomes: A review. Parkinson's Disease **2017**, 4318416–4318416 (2017)
62. Domingo, A., Amar, D., Grutz, K., Lee, L.V., Rosales, R.L., Bruggemann, N., Jamora, R.D.G., La Paz, E.M.C., Rolfs, A., Dressler, D., *et al.*: Evidence of taf1 dysfunction in peripheral models of x-linked dystonia-parkinsonism. Cellular and Molecular Life Sciences **73**(16), 3205–3215 (2016)
63. Rocha, E.M., De Miranda, B.R., Sanders, L.H.: Alpha-synuclein: Pathology, mitochondrial dysfunction and neuroinflammation in parkinson?s disease. Neurobiology of Disease **109**, 249–257 (2018)

Figures

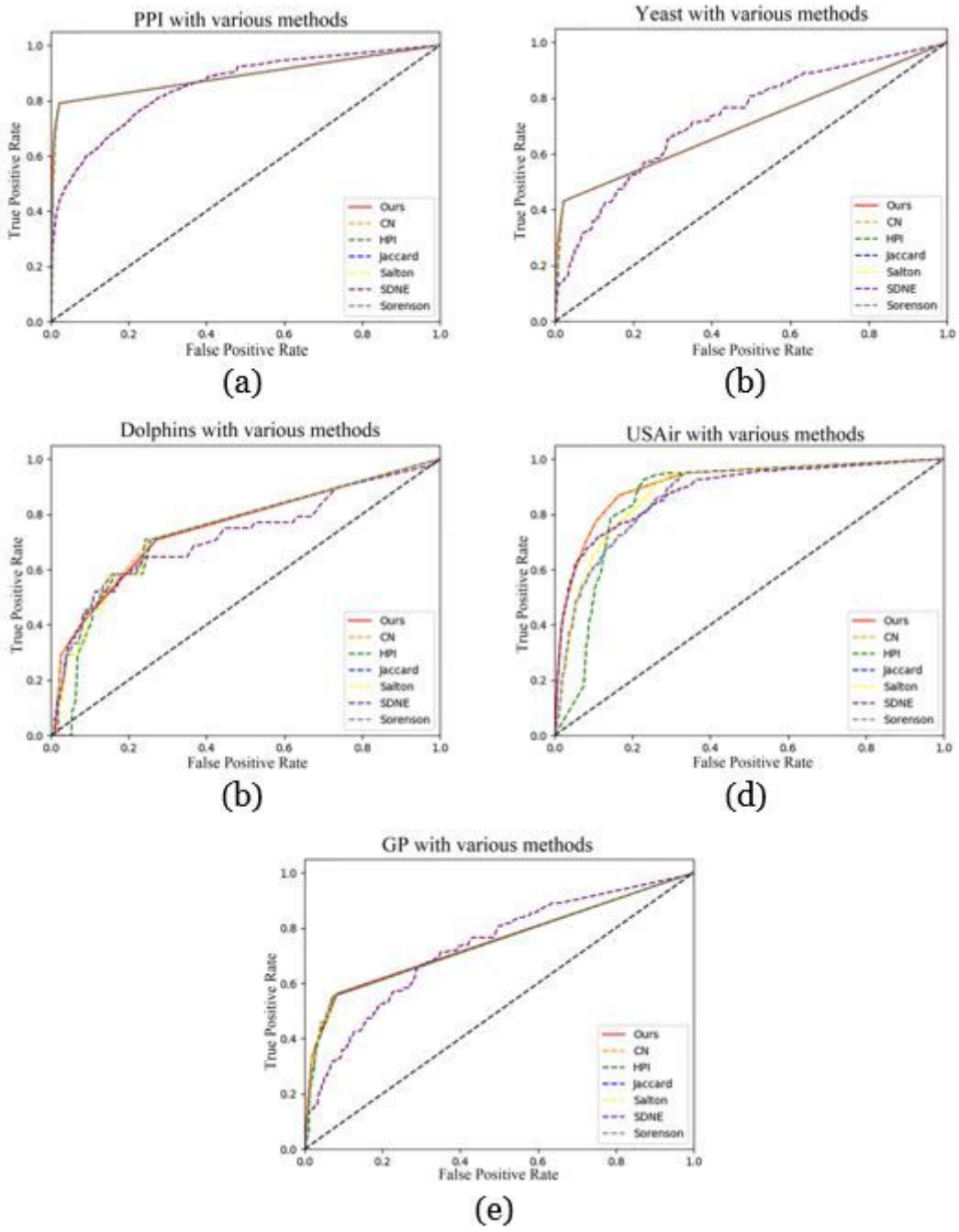


Figure 1

Performance of various methods for different network analysis.

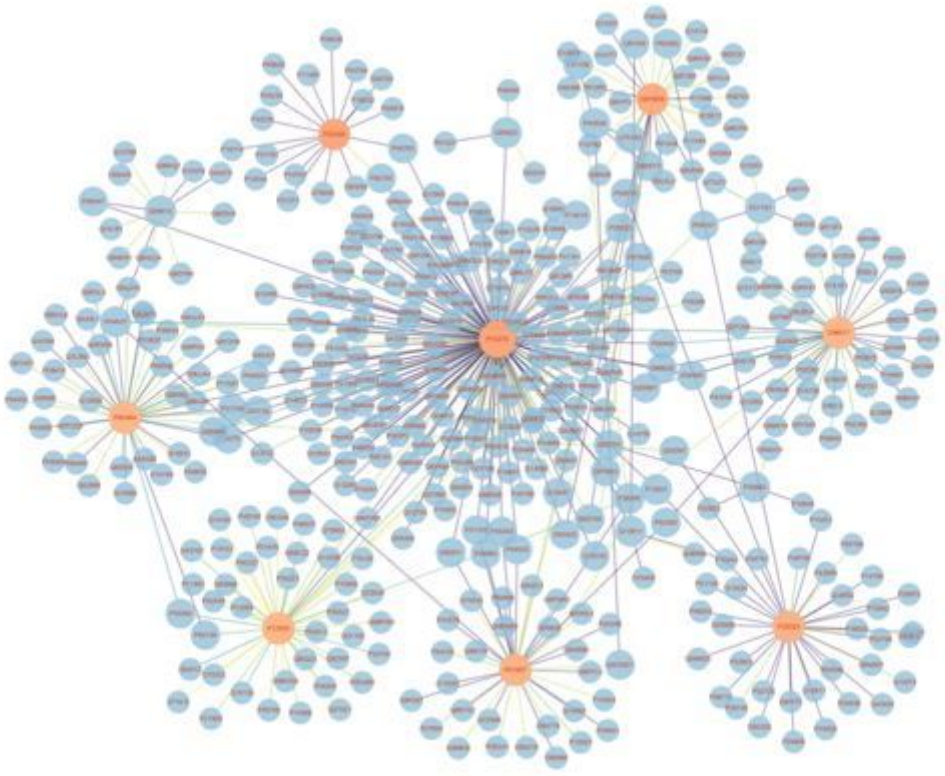


Figure 2

Overview of the prostate cancer PPI network.

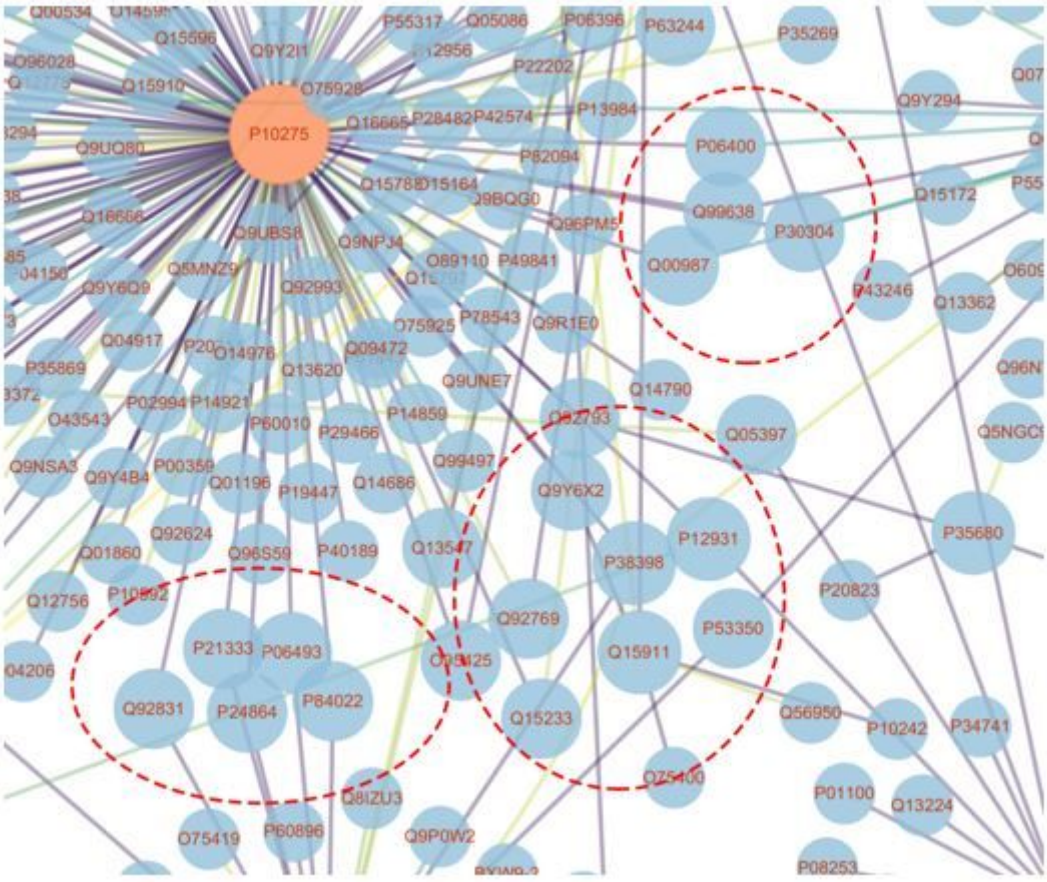


Figure 3

Representative central nodes. Nodes included AR(P10275), CDH1(P12830), and CHEK2(O96017) connected with several hub nodes in the network.

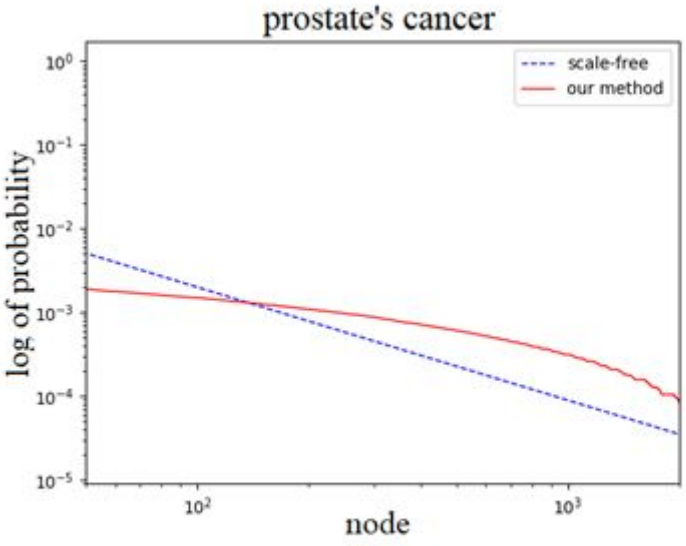


Figure 4

Probability of the node-degree distribution of the PPI network associated with prostate cancer and representative of a scale-free network.

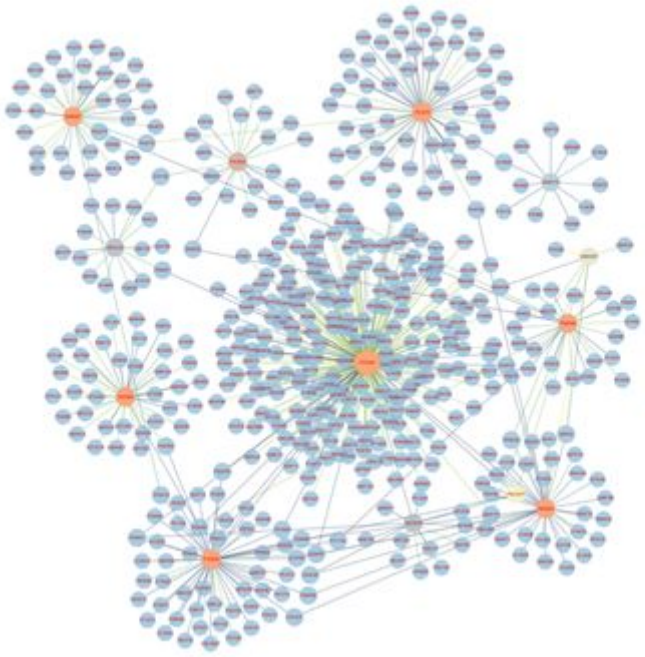


Figure 5

Overview of the Parkinson's disease PPI network.

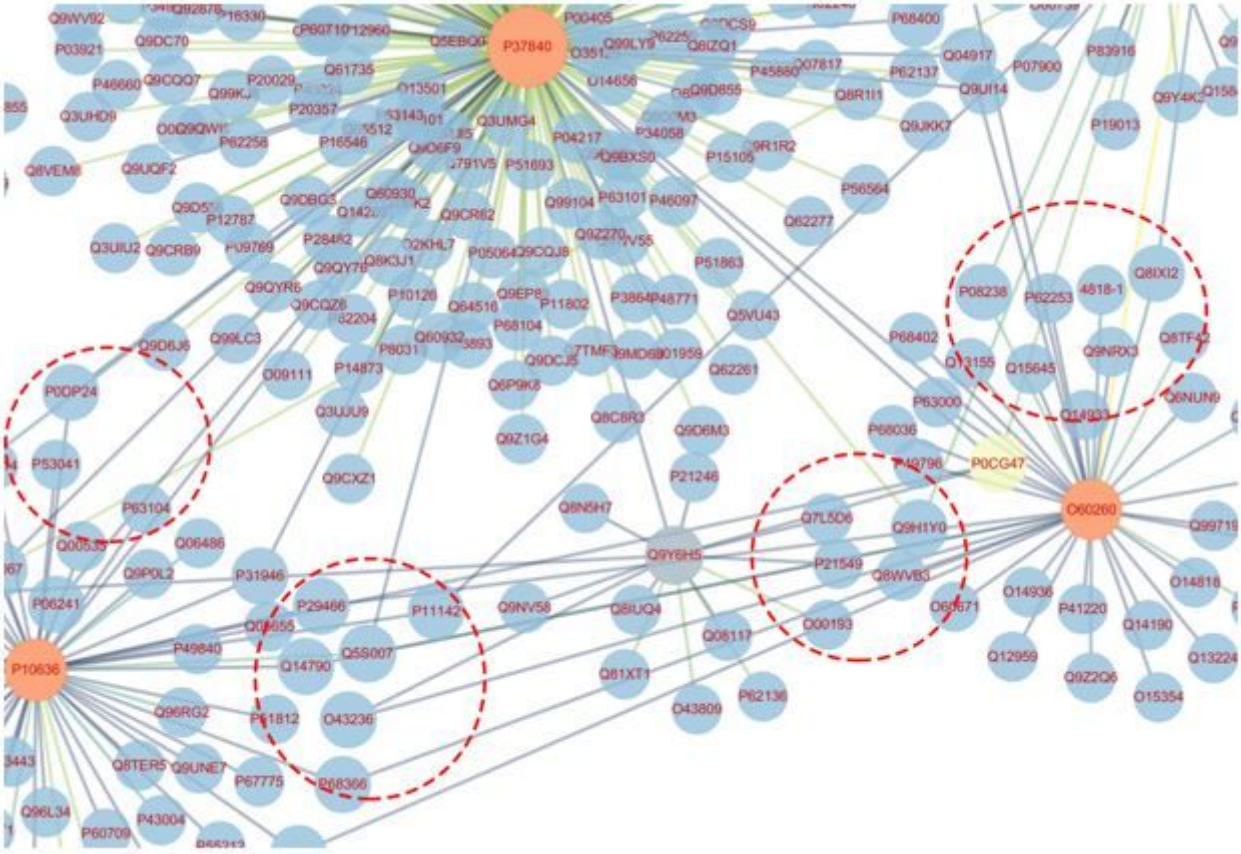


Figure 6

Representative central nodes in the Parkinson's disease PPI network. Nodes includeding SNCA(P37840), PARK2(O60260), MAPT(P10636) connected with several hub nodes in the network.

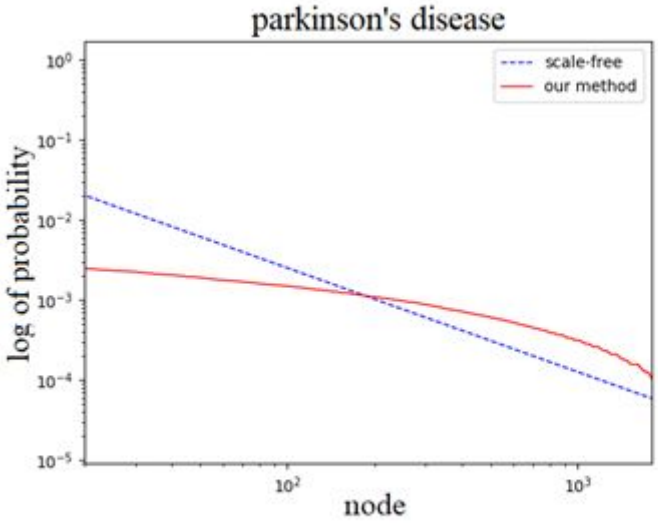


Figure 7

Probability of the node-degree distribution of the PPI network associated with Parkinson’s disease and representative of a scale-free network.