Cent. Eur. J. Phys.
DOI: 10.2478/s11534-009-0114-9

VERSITA

Central European Journal of **Physics**

Predicting disease-related genes by topological similarity in human protein-protein interaction network

Research Article

Lei Zhang*, Ke Hu, Yi Tang†

Department of physics, University of Xiangtan, Xiangtan 411105, Hunan, China

Received 7 March 2009; accepted 22 June 2009

Abstract: Predicting genes likely to be involved in human diseases is an important task in bioinformatics field. Nowadays, the accumulation of human protein–protein interactions (PPIs) data provides us an unprecedented opportunity to gain insight into human diseases. In this paper, we adopt the topological similarity in human protein–protein interaction network to predict disease-related genes. As a computational algorithm to speed up the identification of disease-related genes, the topological similarity has substantial advantages over previous topology-based algorithms. First of all, it provides a global measurement of similarity between two vertices. Secondly, quantity which can measure new topological feature has been integrated into the notion of topological similarity. Our method is specially designed for predicting disease-related genes of single disease-gene family. The proposed method is applied to human protein–protein interaction and hepatocellular carcinoma (HCC) data. The results show a significant enrichment of disease-related genes that are characterized by higher topological similarity than other genes.

PACS (2008): 87.10.Vg, 87.15.km

Keywords: predicting disease genes • protein–protein interactions • topological features • topological similarity

1. Introduction

Mining genes associated with human diseases is an important task in bioinformatics field. It can help in understanding the pathogenic mechanism of diseases. There are two traditional approaches for disease gene discovery: the candidate gene approach [1] and positional cloning *via* linkage analysis [2]. In the traditional approaches, researchers would need to analyze a large number of genes. This will cost too much man power and resources. Therefore, an efficient algorithm for predicting disease-related genes is needed, which can help researchers narrow down the search scope and speed up the identification process of disease-related genes.

Large-scale molecular interaction networks in humans such as human protein–protein interaction network have just become available in the past three or four years. The large-scale high-throughput experiments have yielded a large amount of PPIs data, such as yeast two-hybrid (Y2H) system [3] and affinity purification followed by mass spectrometry (AP-MS) [4]. More databases are becoming available in recent years. Gandhi *et al.* [5] have collected

[*] E-mail: twinsheros@126.com

[†] E-mail: tang_yii@126.com (Corresponding author)

over 38,000 human LC (literature-curated) protein interactions in the HPRD database and in OPHID [6] Brown and Jurisica have collected over 90000 human protein interactions. The development of experimental technologies and the availability of more PPIs databases provide an unprecedented opportunity to discover disease-related genes from PPIs network.

Various features and patterns have been exploited to predict disease-related genes, such as sequence features [7], expression patterns [8] and so on. Recently, some researchers have studied methods based on the topological features in PPIs network [9, 10]. The theoretical basis of these topology-based methods is that genes associated with a particular phenotype or function, such as disease, are not randomly positioned in the network. They tend to exhibit high connectivity, cluster together, and reside in central network locations [11]. More topological features have been discovered and exploited in recent years. Tu *et al.* [12] found that the degrees of disease genes are significantly higher than other genes in the PPIs network. Oti *et al.* [9] found that genes neighboring disease related genes were more likely to be also disease related genes. Other methods that measure a variety of topological features have been decribed. Xu *et al.* [10] developed a classifier in which five quantities were employed to measure five different topological features. The same idea of previous topology-based methods is that measurements of similarity are generally determined by local information of topology. The methods based on local topological information are well-suited to predict genes which are neighbors or next neighbors of known disease genes. However when predicting disease-related genes within a single disease-gene family, neighbors and next neighbors of known disease genes just cover a limited scope of network. For example, there are 73 genes which can be treated as high confidence lung cancer genes[1]. In our PPIs network (LC dataset), there are 3558 genes which are neither neighbors or next neighbors of the 73 known disease gene. To solve this problem, we propose a method which is based on topological similarity [13]. The proposed method has substantial advantages over previous topology-based methods. Firstly, it is global. It depends on the whole graph and allows two vertexes to be similar without sharing neighbors. Secondly, quantity which can measure new topological feature has been integrated into topological similarity. A detailed description of topological similarity and our method is presented in the Methods section.

2. Methods

2.1. Hepatocellular carcinoma data and human PPIs datasets

In order to evaluate the results of our method, a certain amount of known disease genes are needed. In this paper, hepatocellular carcinoma (HCC) data originating from OncoDB.HCC [14] was employed. OncoDB.HCC is the first comprehensive oncogenomic database for hepatocellular carcinoma (HCC). In this database, researchers compiled a list of 614 significant genes which were selected under following criteria:

- Criterion 1: genes significantly up- or down- regulated in at least three independent HCC microarray/proteomic reports.
- Criterion 2: genes were selected with consistent expression level changes for at least 2 folds in more than 70% patients after reprocessing Stanford HCC microarray data.
- Criterion 3: genes with wet-lab experimental data from previous reports.

We choose 310 genes evidenced by wet–lab experimental results and PPI data and were therefore treated as a high confidence disease-gene set (we call it disease-gene set for convenience). Herein, we propose to provide a detailed explanation as follows. Cancer, similar to common disease, is a disease due to malfunction involing multiple factors. It is very difficult to say with hundred percent certainty that some genes are related to cancer but other genes are not. Although many mutated genes been identified in HCC or other cancers, it is very difficult to conclude that specifically mutated genes are cancer genes. Therefore compiling a list of genes that are are positively cancer genes is a difficult or even impossible task. In OncoDB.HCC, genes which both have additional references from wet-lab animal experiments and human tissue data might be recognized as high confidence cancer genes [14]. Human protein–protein interactions datasets were downloaded from database OPHID [6]. PPIs data of OPHID were collected from three sources: (1) Literature-curated (LC) interactions; (2) High-throughput experiments (EXP);(3) Interactions predicted from model organisms (PDT), such as Drosophila, Saccharomyces cerevisiae *etc.* We will focus on proteins that are located in the largest connected network component (main component) because the topological similarity is incalculable for proteins which do not belong to the main component. The total number of unique proteins and disease genes in the main component are listed in Tab. 1. It should be noted

[1] *Lung Cancer Gene Database. Obtained through the internet: www.bioinformatics.org/LuGenD/genelist.htm*

that there are 1332 genes which are neither neighbors nor next neighbors of known HCC genes in LC dataset.

2.2. Building training samples

In recent studies, researchers have found that there are thousands of essential genes in the human genome, which are different from both disease and non-disease genes. They proposed to divide the gene population into three parts namely essential genes, disease genes, and non-disease genes. Researchers compiled a list of ubiquitously expressed human genes (UEHGs) as an approximation of human essential genes [12]. The number of these essential genes in the main component is also listed in Tab. 1.

In the main component of LC network, we exclude disease genes and essential genes from the 9894 proteins and the remaining 8210 genes are called ‘control-gene set’. Some genes are randomly selected from the control-gene set. These selected genes are regarded as negative training samples. The “control-gene set” of PDT and EXP datasets can be obtained in the same way.

Table 1. The PPI datasets.

			Disease	Essential	Control
Datasets	PPIs	Proteins	genes	genes	genes
LC	45099	9894	310	1445	8210
PDT	33833	5048	155	975	3957
EXP	7904	3464	114	589	2792

2.3. The notion of topological similarity

Firstly, we want to introduce the notion of topological similarity which was first proposed by Leicht *et al.* [13]. In this definition of similarity, vertex i is said to be similar to vertex j if i has any network neighbor v that is itself similar to j. This definition is apparently recursive, because vertex v could also be similar to vertex j through any neighbor of v. In order to make the results converge to a useful limit, a starting point for the recursion should be provided. The starting point we have selected is to make each vertex similar to itself. Thus the definition of topological similarity has two components: the neighbor term and the self-similarity term. The definition can be written as follows

$$S_{ij} = \phi \sum_{v} A_{iv} S_{vj} + \psi \delta_{ij}. \tag{1}$$

Here, S_{ij} is the similarity between gene i and gene j. A_{iv} is the iv element of adjacency matrix A. δ_{ij} is the Kronecker's function

$$\delta_{ij} = \begin{cases} 1 \text{ if } i = j, \\ 0 \text{ if } i \neq j. \end{cases} \tag{2}$$

The first term of Eq. (1) is the neighbor term, which is determines whether i has any network neighbor v that is itself similar to j. Parameter ϕ can be treated as the weight of neighbor term. The second term of Eq. (1) says that a vertex is similar to itself. Parameter ψ can be treated as the weight of the self-similarity term. ϕ and ψ control the balance between these two components of the similarity.

We can write Eq. (1) in matrix form as

$$S = \phi AS + \psi I, \tag{3}$$

where S is the similarity matrix, and S_{ij} is the ij element of S. Matrix **A** is the adjacency matrix of the network. **I** is the identity matrix. Eq. (3) can also be written as $S = \psi[I - \phi A]^{-1}$. Because we only consider the relative similarity of different pairs of vertices, Parameter ψ can be safely set to 1. Now, Eq. (3) can be written as

$$S = [I - \phi A]^{-1}. \tag{4}$$

Then we can expand Eq. (4) as a power series

$$S = I + \phi A + \phi^2 A^2 + \phi^3 A^3 + \cdots. \tag{5}$$

It should be noted that $[A^l]_{ij}$ is equal to the number of network paths of length l from i to j. Eq. (5) gives us a term-by-term interpretation of the topological similarity. The first term denotes that a vertex is similar to itself. The second term denotes that vertices that are immediate neighbors of one another have similarity ϕ. The third term denotes that vertices that are next neighbors of one another have similarity ϕ^2, and so forth. ϕ^l represents the weight of paths of length l and short paths have higher weight than long paths. In previous research, measurements of similarity are generally determined by local information of topology. From Eq. (5), we can see that, topological similarity is a global measurement of similarity. It is based on the global topology. Not only paths length of 1 or 2, but also paths of any length can contribute to the similarity. Although genes which are functionally related usually locate in one or more modules, the number of known disease genes is so few when predicting disease-related genes of single disease-gene family. Neighbors and next neighbors of known disease genes just cover a limited scope of the network. By applying the definition of topology similarity to predict disease genes, some important disease genes that are distant from known disease genes may be found. Our thoughts are tested by hepatocellular carcinoma (HCC) data from OncoDB.HCC. We will discuss it in the Results section.

The calculation of the similarity matrix S is achieved by this equation:

$$DSD = \frac{\alpha}{\lambda_1} A(DSD) + I. \tag{6}$$

Eq. (6) is the final form of deduction and it can be directly used in the algorithm. The deducing process from Eq. (5) to Eq. (6) is beyond the scope of this paper. For more details of deduction, we refer the reader to Leicht *et al.* [13]. In Eq. (6), D denotes the diagonal matrix having the degrees of the vertices in its diagonal elements: $D_{ij} = K_i \delta_{ij}$. K_i is the degree of vertex $i.\lambda_1$ is the largest eigen-value of matrix **A.** λ_1 can be calculated after we know the adjacency matrix A. α is a tunable parameter. The effect of the parameter α is to reduce the contribution of long paths relative to short ones. Leicht *et al.* developed a model and used it to test the performance of topological similarity. Their experience suggested that the measurement results would be closest to the underlying model if α was set to 0.97 [13]. Detailed meaning of parameter α can be found in Leicht *et al.* [13]. Because the diagonal matrix D includes the information of the degrees of vertices, we will use DSD as the similarity matrix so that our algorithm can cover the feature of degrees. The elements of matrix DSD are initially set to constant 0. If Eq. (6) is iterated a approx. a 100 times, and then a good convergence will be found.

2.4. K-nearest neighbors classification algorithm (KNN)

After we get the similarity matrix S, we then seek for the K-highest values of every column, which represent the K-nearest neighbors of every gene. In our method, a gene will be predicted to be disease-related genes only if all of the K-nearest neighbors are known disease genes. For example, if the K value is set to 2, it means that a gene will be predicted to be a disease-related gene only if the 2 nearest neighbors are known disease genes.

3. Results

3.1. The validity of topological similarity

In order to prove that topological similarity is an effective measurement which can differentiate disease genes and control genes, we have performed two different statistical analyses. One is to test whether the average level of topological similarity between two disease genes is different from two control genes. Here, we take LC dataset as an example where there are 310 known disease genes in the dataset. After obtaining the similarity matrix S of LC network, the submatrix S_D is extracted from S. There are 310 rows and 310 colomns in S_D where every row and column of S_D corresponds to a disease gene. The ij element of S_D represents the topological similarity between disease gene i and disease gene j. The average topological similarity between two disease genes is calculated by

$$\overline{S_{ij}} = \frac{\sum\limits_{i\in D, i\neq j}\ \sum\limits_{j\in D, j\neq i} S_{ij}}{n_D^2 - n_D}. \tag{7}$$

D denotes the disease-gene set which has 310 disease genes. n_D is the number of genes in D. $\overline{S_{ij}}$ is the average value of elements of matrix S_D. We don't consider the similarity between a gene and itself, therefore the diagonal elements are excluded from the sum of matrix elements. We gradually decrease the percentage of disease genes in the submatrix S_D. For example, in Fig. 1 (LC dataset), the x-coordinate "90%" denotes that the percentage of disease genes in submatrix S_D is reduced to 90%. That means 279 disease genes (90% of D) are randomly sampled from the disease-gene set D. The remaining 31 genes are randomly sampled from the control-gene set which has 8210 control genes. In order to avoid sampling bias, the random sampling is repeated 1000 times and the average values of $\overline{S_{ij}}$ are shown in Fig. 1 (LC dataset). The same procedures are also performed on PDT and EXP datasets. The results are shown in Fig. 1. Because disease genes of single disease-gene family are functionally related, Fig. 1 actually suggests that gene pairs which are functionally related will have higher topological similarity. The decrease of $\overline{S_{ij}}$ is a consequence of the decrease of such functionally related genes. In Fig. 1 (LC dataset), the x-coordinate "0%" means that all the 310 disease genes has been replaced by control genes. The values here represent the average topological similarity between two control genes. From Fig. 1 (LC dataset), we can see that the average topological similarity between two disease genes (100%) is more than 10-fold higher than that between two control genes (0%).

Another issue is to test whether the average level of topological similarity between a disease gene and the whole disease-gene set D is different from that between a control gene and D. The average topological similarity between single gene i and the whole disease-gene set D is calculated by

$$\overline{S_{ij}}(j \in D) = \frac{\sum\limits_{j\in D} S_{ij}}{n_D}. \tag{8}$$

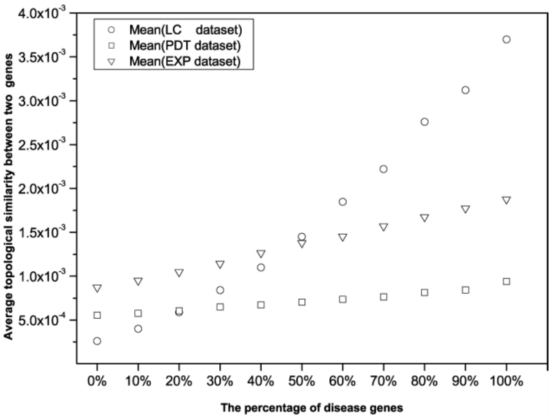

Figure 1. The mean of $\overline{S_{ij}}$ with the decrease of disease genes. The x-coordinate denotes that the percentage of disease genes in submatrix S_D. The y-coordinate denotes the $\overline{S_{ij}}$. "100%" denotes that all genes in submatrix S_D are disease genes. We gradually decrease the percentage of disease genes in the submatrix S_D until all disease genes are replaced by control genes (0%). At each percentage, the random sampling is repeated 1000 times so that we can get the mean of $\overline{S_{ij}}$.

Gene i could be any control gene or disease gene. Medians of $\overline{S_{ij}}(j \in D)$ are shown in Tab. 2. There are two different gene populations in Tab. 2: control-gene set and disease-gene set. Statistical significance between the two gene populations (P-values) is calculated by rank sum test. In LC dataset, we can see that there are significant differences between control-gene set and disease-gene set. In PDT and EXP datasets, the differences are weaker than those in LC dataset. Tab. 2 tells us that the average topological similarity between single disease gene and the whole disease-gene set D is significantly higher than that between a control gene and D(except EXP dataset).

3.2. Performance of the classification algorithm

We use the 5-fold cross validation to evaluate the prediction quality of our algorithm. The whole gene population is divided into 5 subsets. Each time, one of the 5 subsets is used as the test set and the other 4 subsets are put together to form a training set. After that, we use the training set to classify the test set. Three quantities are employed to evaluate the prediction quality. These quantities are accuracy, sensitivity and precision.

Table 2. Medians of $\overline{S_{ij}}$ $(j \in D)$ between the disease-gene set and control-gene set. The statistical significance between these two gene populations is calculated by rank sum test.

LC dataset			PDT dataset			EXP dataset		
Disease	Control	P-value	Disease	Control	P-value	Disease	Control	P-value
6.81E-04	1.26E-04	2.82E-56	1.63E-04	8.34E-05	6.46E-04	4.80E-04	4.01E-04	7.52E-02

$$\begin{aligned} \text{Accuracy} &= \frac{TP+TN}{TP+TN+FP+FN}, \\ \text{Sensitivity} &= \frac{TP}{TP+FN}, \\ \text{Precision} &= \frac{TP}{TP+FP}. \end{aligned} \tag{9}$$

In Eq. (9), TP, FP, TN, FN represents true positive, false positive, true negative and false negative. The detailed information of prediction quality is listed in Tab. 3. When the K value is set to 2, in LC dataset, the classification algorithm can correctly recover 72% of known disease genes with a precision of 70%. The corresponding accuracy is 73%. The performance of classification algorithm on PDT and EXP datasets is not as good as that on LC dataset. There are two possible reasons for this. First, our method is based on the topological structure of a network, but the topologies of these three PPIs networks are different. Each network has its own features. For example, in LC network, disease genes communicate with each other more quickly. Researchers also found that there exist a lot of heavily connected disease genes in the LC network [10]. The differences of topologies may be a potential reason for the disparity of performance. Second, in PDT and EXP datasets, the differences between disease-gene set and control-gene set are weaker than those in LC dataset. This may be another important factor which can influence the performance of algorithm.

We adopt the method mentioned in Xu *et al.* [10] to evaluate the performance of our method. 50%-80% of disease genes are randomly sampled from the original disease-gene set and used as positive training samples. The remaining true disease genes are called “leave-out genes”. These leave-out genes are treated as “unknown novel disease genes” and mixed with the negative training samples which are randomly sampled from the control-gene set (discussed in Sec. 2.2). In order to avoid sampling bias, at each percentage, the random sampling is repeated 1000 times and the average values of accuracy, sensitivity, precision are listed in Tab. 4. From Tab. 4, we can learn that the classification algorithm has certain robustness to the changing of positive training samples during 5-fold cross validation and the performance becomes better with the increase of positive training samples.

We use the leave-out genes to test the ability of disease gene predicting of our algorithm. In LC network, 50% of disease genes (155 genes) are randomly sampled from the original disease-gene set and used as positive training samples. The remaining 155 disease genes are treated as “unknown novel disease genes” and mixed with negative training samples which are randomly sampled from the control-gene set. It is not appropriate to perform statistical analysis on negative training sample which size is either too large or too small. Therefore, we randomly select 2945 control genes as negative training samples. The classification algorithm is used to predict disease genes from these 3100(155+2945) unknown genes. In Fig. 2b, when the K value is set to 2, on average, among the predicted novel disease genes, 13.5% genes are already listed in the 155 leave-out genes. Before using the method, among the 3100 candidate genes, 5 %(155/3100) genes are leave-out genes. Therefore we can see about 2.7-fold (13.5/5) enrichment relative to the random prediction. The fluctuation of the dots becomes larger if we adopt more strict restrictions (K=2). This is due to so few known disease genes when dealing with the case of single disease-gene family. Because K=3 is too strict, many disease genes will be excluded. In practical use, we set safely the K value to 2.

Table 3. Statistical performance of the classification algorithm.

	LC dataset			PDT dataset			EXP dataset		
	Accuracy	Sensitivity	Precision	Accuracy	Sensitivity	Precision	Accuracy	Sensitivity	Precision
K=1	0.73	0.74	0.74	0.71	0.70	0.73	0.67	0.63	0.61
K=2	0.73	0.72	0.70	0.70	0.72	0.70	0.65	0.62	0.60

Table 4. The average performance of classification algorithm with the increase of positive training samples (50%-80%). Results were obtained using K=2.

	LC dataset			PDT dataset			EXP dataset		
	Accuracy	Sensitivity	Precision	Accuracy	Sensitivity	Precision	Accuracy	Sensitivity	Precision
50%	0.71	0.73	0.75	0.67	0.65	0.67	0.61	0.59	0.57
60%	0.74	0.72	0.77	0.69	0.67	0.68	0.61	0.61	0.58
70%	0.75	0.75	0.76	0.69	0.68	0.71	0.63	0.61	0.60
80%	0.78	0.76	0.79	0.72	0.70	0.72	0.64	0.62	0.60

3.3. Predicting HCC genes from human PPIs network

Hepatocellular carcinoma (HCC) is one of the major cancers in the world. Every year more than 250 thousand people die of HCC. Mining genes associated with HCC is an important step towards understanding the detailed mechanisms of hepatocarcinogenesis and discovering new target molecules for drugs. In this section we take HCC data from OncoDB.HCC as a biological example and analyze it with our method. We set the K value to 2 and apply the classification algorithm to LC network. 482 genes are predicted to be HCC-related genes. These 482 genes are

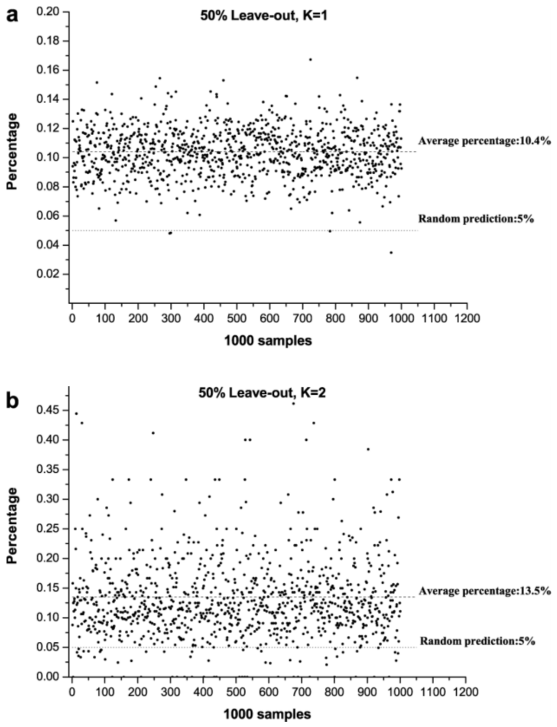

Figure 2. The ability of disease gene predicting of our algorithm. (a) The parameter K is set to 1. (b) The parameter K is set to 2. 50% of the disease genes are randomly sampled from the original disease-gene set and used as positive training samples. The leave-out disease genes are mixed with other unknown genes to test the ability of disease genes predicting for our algorithm. The Horizontal axes represents the random sampling process. The vertical axes represents the percentage of leave-out genes in the predicted novel disease genes. The sampling is repeated 1000 times and the average percentage is obtained. The dot line represents the random prediction.

listed in supplementary materials. Among the predicted HCC genes, some genes have been shown to be evident from experiments or literatures. These genes are listed in Tab. 5. There are three types of evidence: abnormal expression level in HCC, literatures and mammalian experiments (such as murine model of HCC). If a gene has evidential support from one of these sources, the corresponding term will be “YES”. From Tab. 5, we can see that, when predicting disease-related genes of single disease-gene family, the proposed method can find out not only genes which are neighbors or next neighbors of known disease genes, but also genes which are neither neighbors nor next neighbors. The detecting scope is the whole network. There are 110 genes listed in Tab. 5 and we hope that information compiled in Tab. 5 can help researchers narrow down the search scope and speed up the identification process of HCC related genes.

In Tab. 5b, gene CCR7 is the receptor for the chemokine. It can be found in various lymphoid tissues and activates B and T lymphocytes. CCR7 is reported to be involved in various cancers, such as breast cancer, colon cancer, lung cancer, gastric carcinoma, and thyroid cancer [15]. In gastric carcinoma and lung cancer, CCR7 was found to be associated with lymph node metastasis [16, 17]. In HCC, researchers also found that CCR7 is significantly associated with locally progressed tumors and lymph node metastases [15]. These evidences imply CCR7 as a new disease gene candidate for HCC. CCR7 may be closely related to the process of lymph node metastases existing in various cancers.

In Tab. 5b gene BAAT encodes a liver enzyme that catalyzes the conjugation of bile acid with glycine or taurine. It is mainly expressed in fully differentiated and quiescent liver cells. Researchers have found that low expression of BAAT is significantly associated with a decreased probability of survival in HCC patients [18]. Massive and uncontrollable recurrence of HCC tends to occur in the group of patients who have low BAAT expression. BAAT was also found to be a reliable indicator for both survival and recurrence in HCC patients [18]. These evidences suggest that gene BAAT may play a role in the recurrence of HCC.

The JAK/STAT signaling pathway plays a central role in principal cell fate decisions, regulating the processes of cell proliferation, differentiation and apoptosis. It has been reported in association with cancer by many researchers [19–21]. The activation of this pathway may promote the occurance of cancer. In Tab. 5a, nine genes are involved in this pathway. They are JAK2, STAT1, STAT3, STAT5, PIAS1, PIAS4, PTPN11, PIK3R1, and CCND2. We can see that, most central member of JAK/STAT signaling pathway have been predicted by our method. Among these genes, a structurally abnormal form of JAK2 has been reported in human cancer, and the inhibition of JAK2 may play an important role in the treatment of HCC [21]. STAT1, STAT3, and STAT5 are signal transducer and activator of transcription. Constitutive activation of STAT1, STAT3, and STAT5 has been discovered in various cancer types [22–24]. Each STAT gene has different functions. STAT1 functions as a tumor suppressor gene [22], whereas STAT3 and STAT5 have shown to play a role in development and tumor progression [23, 24]. Recently, researchers found that activation of STAT5 was associated with HCC aggressive behavior, it induced HCC invasiveness through Epithelial-mesenchymal transition (EMT) [25]. PIAS1 and PIAS4 are a E3 SUMO-protein ligase and protein inhibitor of activated STAT respectively, and has been reported that these two genes will influence the ability of tumor suppressor p53 PIAS1 may involve in the Sumoylation of p53 [26], PIAS4 inhibited the DNA-binding activity of p53 in nuclear extracts and blocked the ability of p53 to induce expression of two of its target genes [27].

Table 5. The predicted genes which have evidences from databases and literatures,$l$denotes the length of shortest path to the nearest disease neighbor.

a. neighbors and next neighbors of known disease genes ($l \leq 2$)

	Swiss-Prot ID	Protein name	Gene	Abnormal	Literatures	mammalian
				Expression		experiments
Cell cycle						
	Q6FG59	CDC37 protein	CDC37		YES	YES
	P49454	Centromere protein F	CENPF	YES		
	P06493	Cell division control protein 2 homolog	CDK1	YES		
	P24941	Cell division protein kinase 2	CDK2		YES	
	P30279	G1/S-specific cyclin-D2	CCND2	YES		
	P98082	Disabled homolog 2	DAB2		YES	
	Q00535	Cell division protein kinase 5	CDK5	YES		
	O60729	Dual specificity protein phosphatase	CDC14B		YES	
Transcription						
	Q14186	Transcription factor Dp-1	TFDP1		YES	
	P08047	Transcription factor Sp1	SP1		YES	
	P40763	Signal transducer, activator of transcription 3	STAT3		YES	
	Q96T58	Msx2-interacting protein	SPEN		YES	
	P51532	Probable global transcription activator	SMARCA4		YES	
	P42224	Activator of transcription 1-alpha/beta	STAT1	YES		
	O15164	Transcription intermediary factor 1-alpha	TRIM24		YES	YES
	P42229	Signal transducer,activator of transcription	STAT5	YES	YES	
	Q06945	Transcription factor SOX-4	SOX4		YES	
Tumor-associated						
	O14763	Tumor necrosis factor receptor 10B	TNFRSF10B		YES	
	Q9H3D4	Tumor protein 63	TP63		YES	
	O00220	Tumor necrosis factor receptor 10A	TNFRSF10A		YES	
	P46108	Proto-oncogene C-crk (p38)	CRK		YES	
	P51587	Breast cancer type 2 susceptibility protein	BRCA2		YES	
MAPK pathway						
	Q92918	Mitogen-activated protein kinase 1	MAP4K1	YES		
Apoptosis						
	Q14790	Caspase-8	CASP8		YES	
	Q9UET8	Apoptosis signaling receptor FAS	FAS	YES	YES	
Miscellaneous						
	Q5T186	SHC-transforming protein 1	SHC1	YES		
	O00230	Cortistatin	CORT		YES	
	P84022	Mothers against decapentaplegic homolog 3	SMAD3		YES	
	Q05639	Elongation factor 1-alpha 2	EEF1A2	YES	YES	
	P12757	Ski-like protein	SKIL		YES	
	P08069	Insulin-like growth factor 1 receptor	IGF1R	YES	YES	
	P09619	Platelet-derived growth factor receptor	PDGFRB		YES	
	P22681	E3 ubiquitin-protein ligase CBL	CBL		YES	
	P29317	Ephrin type-A receptor 2	EPHA2	YES		
	P11388	DNA topoisomerase 2-alpha	TOP2A		YES	
	P63244	Guanine nucleotide-binding protein subunit	GNB2L1		YES	
	Q86UL8	Membrane-associated guanylate kinase	MAGI2		YES	
	O60674	Tyrosine-protein kinase JAK2	JAK2		YES	
	P27986	Phosphatidylinositol kinase regulatory subunit	PIK3R1		YES	
	Q06124	Tyrosine-protein phosphatase non-receptor	PTPN11		YES	
	P35968	Vascular endothelial growth factor receptor 2	KDR		YES	
	O14828	Secretory carrier- membrane protein 3	SCAMP3	YES		
	P05106	Integrin beta-3,Platelet membrane glycoprotein	ITGB3		YES	
	Q7Z419	E3 ubiquitin-protein ligase RNF144B	RNF144B		YES	
	P05783	Keratin, type I cytoskeletal 18	KRT18	YES		
	P06748	Nucleophosmin,Nucleolar phosphoprotein	NPM1	YES		

a. continued

O75925	E3 SUMO-protein ligase PIAS1	PIAS1		YES	
Q8N2W9	E3 SUMO-protein ligase PIAS4	PIAS4		YES	
Q8TEW0	Partitioning defective 3 homolog	PARD3		YES	
Q13153	Serine/threonine-protein kinase PAK 1	PAK1		YES	
Q15172	PP2A,B subunit,PR61 alpha isoform	PPP2R5A	YES	YES	
Q92963	GTP-binding protein Rit1	RIT1		YES	
Q9H2X6	Homeodomain-interacting protein kinase 2	HIPK2		YES	
P41240	Tyrosine-protein kinase CSK	CSK	YES	YES	
Q14289	Protein tyrosine kinase 2 beta	PTK2B		YES	
Q9H4B4	Serine/threonine-protein kinase PLK3	PLK3		YES	
O60496	Docking protein 2, tyrosine kinase 2	DOK2		YES	
P19174	Phosphoinositide phospholipase C	PLCG1		YES	
P04083	Phospholipase A2 inhibitory protein	ANXA1		YES	
P17706	Tyrosine-protein phosphatase non-receptor	PTPN2	YES		
P51813	Cytoplasmic tyrosine-protein kinase BMX	BMX		YES	
Q14451	Growth factor receptor-bound protein 7	GRB7		YES	
Q8IZW8	Tensin-4,C-terminal tensin-like protein	CTEN		YES	
Q03135	Caveolin-1	CAV1		YES	
P10599	Thioredoxin,ATL-derived factor	TRX		YES	
Q12778	Forkhead box protein O1	FOXO1		YES	YES
P51398	28S ribosomal protein S29, mitochondrial	DAP3		YES	
Q9UPS7	Tensin-like C1 domain-containing phosphatase	TENC1		YES	
Q9UM63	Zinc finger protein PLAGL1	PLAGL1		YES	
P30291	Wee1-like protein kinase	WEE1		YES	
P19793	Retinoic acid receptor RXR-alpha	RXRA	YES		
P41161	ETS translocation variant 5	ETV5		YES	
P10826	Retinoic acid receptor beta	RARB		YES	
Q04756	Hepatocyte growth factor activator	HGFAC	YES		
P42167	Lamina-associated polypeptide 2	TMPO	YES		
Q99466	Neurogenic locus notch homolog protein 4	NOTCH4		YES	
Q14116	Interleukin-18 (IL-18)	IL18		YES	
Q9UMN6	Histone-lysine N-methyltransferase MLL4	MLL2		YES	
O00497	DNA mismatch repair protein	hMLH1		YES	
O43392	Aryl hydrocarbon receptor nuclear translocator	ARNT		YES	
Q7L311	Armadillo repeat-containing X-linked protein 2	ARMCX2		YES	YES
Q9H2L5	Ras association domain-containing protein 4	RASSF4		YES	
P48029	Sodium, chloride-creatine transporter	SLC6A8		YES	
P58658	Uncharacterized protein C21orf63 (SUE21)	C21orf63		YES	
Q14432	cGMP-inhibited 3',5'-cyclic phosphodiesterase	PDE3A		YES	
Q96KS0	Egl nine homolog 2	EGLN2		YES	
Q13547	Histone deacetylase 1	HDAC1	YES	YES	
Q02297	Pro-neuregulin-1,membrane-bound isoform	NRG1	YES		

b. neither neighbors nor next neighbors of known disease genes ($l > 2$)

Swiss-Prot ID	Protein name	Gene	Abnormal Expression	Literatures	mammalian experiments
Q9NTJ3	Chromosome-associated polypeptide C	SMC4	YES		
Q15493	Regucalcin (RC)	RGN	YES		
P32248	C-C chemokine receptor type 7	CCR7		YES	
P27707	Deoxycytidine kinase	DCK		YES	
Q96FF9	Cell division cycle-associated protein 5	CDCA5	YES		
O15444	C-C motif chemokine 25	SCYA25		YES	
O75936	Gamma-butyrobetaine dioxygenase	BBOX1	YES		
P50120	Retinol-binding protein 2	RBP2		YES	
P16422	Tumor-associated calcium signal transducer 1	TACSTD1		YES	
Q9NZN3	EH domain-containing protein 3	EHD3	YES		
Q5VZM2	Ras-related GTP-binding protein B	RAGB		YES	
P22033	Methylmalonyl-CoA mutase, mitochondrial	MUT	YES		
O14657	Torsin-1B (Torsin family 1 member B)	TOR1B		YES	

b. continued

Q8WWA0	Intelectin-1 (ITLN-1)	ITLN1	YES	
P04798	Cytochrome P450 1A1	CYP1A1		YES
P22003	Bone morphogenetic protein 5	BMP5		YES
P32754	4-hydroxyphenylpyruvate dioxygenase	HPD	YES	
Q9UMW8	Ubl carboxyl-terminal hydrolase 18	USP18		YES
P51684	C-C chemokine receptor type 6	CCR6		YES
Q16563	Synaptophysin-like protein 1	SYPL1	YES	
Q02985	Complement factor H-related protein 3	CFHR3	YES	
Q14032	Bile acid-CoA:amino acid N-acyltransferase	BAAT		YES

It is well known that p53 is an important tumor suppressor, the dysfunction of p53 will cause various cancers [28]. So PIAS1 and PIAS4 may have influence on cancer *via* gene p53. PTPN11 is widely expressed in most tissues and plays a regulatory role in various cell signaling events. Researchers have found a class of PTPN11 mutants which shown oncogenic activity in Hepatocellular carcinoma [29]. Gene PIK3R1 is a subunit of PI3K (Phosphatidylinositol-3-kinase). Evidences from literatures have shown that proper liver function and development depend on intact PI3K signal transduction. When dysregulated, the PI3K pathway is linked to the development of hepatocellular carcinoma [30]. Protein encoded by CCND2 will form a complex and function as a regulatory subunit of gene CDK4 or CDK6, whose activity is required for cell cycle G1/S transition. CCND2 is discovered to be involved in various cancers [31], abnormal expression level of CCND2 has also been observed in Hepatocellular carcinoma [32]. Evidences described above suggest that these nine genes may be potential HCC candidate genes and the JAK/STAT signaling pathway may play a more important role in the progression of HCC than previously thought.

4. Discussion

The availability of large-scale molecular interaction networks in humans such as PPIs network provides an opportunity to understand the basis of human diseases. In this paper, we proposed a topological similarity-based method which is designed to predict disease-related genes from single disease-gene family. Topological similarity has substantial advantages over previous topology-based methods. First, it has a transparent theoretical rationale. Topological similarity can gather information from the whole graph, and gives a global measurement of similarity between two vertices. It allows candidate genes and known disease genes to be similar without sharing neighbors. Secondly, quantity which can measure a new topological feature has been integrated into topological similarity, such as the number of paths of any given length. In the notion of topological similarity, vertices that have many paths of a given length are considered more similar than those that have few. From the results, we can see that topological similarity is an effective measurement in differentiating disease genes and control genes (Sec. 3.1), and there is a 2.7-fold higher likelihood of disease gene prediction than random prediction when the K value was set to 2 (Fig. 2).

There is still ample room to develop our topology-based methods. First, there is practical limitation to method based on single pattern or feature. A promising way to improve the prediction is to integrate various patterns, such as sequence features and expression patterns. We believe that method based on various patterns can enhance the difference between disease genes and non-disease genes. Secondly, methods which can integrate various large-scale biological datasets are needed in the future. However researchers need to develop methods which can extract useful information from different kinds of large-scale data sets and properly handle the difference among them. With the development of large-scale experimental technologies, we believe that the accuracy of disease genes prediction will become better.

Acknowledgements

This work was supported by Project of National Natural Science Funds under Grant no.30570432, the Key Project of Hunan Provincial Educational Department of China under Grant no. 04A058, and the General Project of Hunan Provincial Educational Department of China under Grant no. 07C754.

References

[1] T. P. Dryja et al., Nature 343, 364 (1990)
[2] M. L. Drumm et al., Cell 62, 1227 (1990)
[3] S. Fields, O. Song, Nature 340, 245 (1989)

[4] A. Kumar, M. Snyder, Nature 415, 123 (2002)
[5] T. K. Gandhi et al., Nat. Genet. 38, 285 (2006)
[6] K. R. Brown, I. Jurisica, Bioinformatics 21, 2076 (2005)
[7] N. Lopez-Bigas, C. A. Ouzounis, Nucleic Acids Res. 32, 3108 (2004)
[8] M. A. Van-Driel, K. Cuelenaere, J. A. Leunissen, H. G. Brunner, Eur. J. Hum. Genet. 11, 57 (2003)
[9] M. Oti, B. Snel, M. A. Huynen, H. G. Brunner, J. Med. Genet. 43, 691 (2006)
[10] J. Z. Xu, Y. J. Li, Bioinformatics 22, 2800 (2006)
[11] T. Ideker, R. Sharan, Genome. Res. 18, 644 (2008)
[12] Z. Tu et al., BMC Genomics 7, 1471 (2006)
[13] E. A. Leicht, P. Holme, M. E. J. Newman, Phys. Rev. E 73, 026120 (2006)
[14] W. H. Su et al., Nucleic Acids Res. 35, D727 (2007)
[15] C. C. Schimanski et al., Oncol. Rep. 16, 109 (2006)
[16] B. Schmaußer et al., Clin. Exp. Immunol. 139, 323 (2005)
[17] S. Maekawa et al., Oncol. Rep. 19, 1461 (2008)
[18] M. Furutani et al., Hepatology 24, 1441 (1996)
[19] J. F. Bromberg et al., Cell 98, 295 (1999)
[20] T. Yoshida et al., J. Exp. Med. 196, 641 (2002)
[21] H. Yoshikawa et al., Nat. Genet. 28, 29 (2001)
[22] S. Xi et al., J. Natl. Cancer I. 98, 181 (2006)
[23] N Diaz et al., Clin. Cancer Res. 12, 20 (2006)
[24] H Li et al., Clin. Cancer Res. 16, 5863 (2005)
[25] T. K. Lee et al., Cancer Res. 66, 9948 (2006)
[26] K. Tomoaki, N. Tamotsu, Y. Hideyo, Mol. Cell 8, 713 (2001)
[27] V. Nelson, G. E. Davis, S. A. Maxwell, Apoptosis 6, 221 (2001)
[28] W. T. David, J. Pathol. 180, 118 (1999)
[29] D. Miyamoto et al., Oncogene 27, 3508 (2008)
[30] X. He et al., Cancer Res. 68, 5591 (2008)
[31] D. M. Euhus et al., Cancer Epidem. Biomar. 17, 1051 (2008)
[32] C. F. Lee et al., World J. Gastroentero. 15, 356 (2009)