BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 22 2006, pages 2800–2805
doi:10.1093/bioinformatics/btl467

*Systems biology*

# Discovering disease-genes by topological features in human protein–protein interaction network

Jianzhen Xu* and Yongjin Li

Department of Bioinformatics, Harbin Medical University, Harbin 150086, China

Received on June 6, 2006; revised on August 3, 2006; accepted on August 29, 2006
Advance Access publication September 5, 2006
Associate Editor: Golan Yona

**ABSTRACT**

**Motivation:** Mining the hereditary disease-genes from human genome is one of the most important tasks in bioinformatics research. A variety of sequence features and functional similarities between known human hereditary disease-genes and those not known to be involved in disease have been systematically examined and efficient classifiers have been constructed based on the identified common patterns. The availability of human genome-wide protein–protein interactions (PPIs) provides us with new opportunity for discovering hereditary disease-genes by topological features in PPIs network.

**Results:** This analysis reveals that the hereditary disease-genes ascertained from OMIM in the literature-curated (LC) PPIs network are characterized by a larger degree, tendency to interact with other disease-genes, more common neighbors and quick communication to each other whereas those properties could not be detected from the network identified from high-throughput yeast two-hybrid mapping approach (EXP) and predicted interactions (PDT) PPIs network. KNN classifier based on those features was created and on average gained overall prediction accuracy of 0.76 in cross-validation test. Then the classifier was applied to 5262 genes on human genome and predicted 178 novel disease-genes. Some of the predictions have been validated by biological experiments.

**Contact:** jianzxu@hotmail.com

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 INTRODUCTION

Hunting genes likely to be involved in human genetic disease is of vital importance for both understanding disease pathogenic mechanism and improving clinical practice. Traditional linkage and association studies have been applied to complex disease, but success has been limited due to variable disease penetrance (Botstein and Risch, 2003). Additionally, large number of genes among large family datasets often needed to be analyzed which is a labor-intensive task. Therefore there is an open space for development of timely and relevant computational algorithms to speed the identification of disease-genes.

A long-held and partially proved assumption shared by biologists is that genes related to similar disease phenotypes are likely to be functionally related, for example, participating in a common pathway or signal transduction mechanism (Badano and Katsanis, 2002; Brunner and van Driel, 2004; Gandhi *et al*., 2006). Based on this assumption (observation), some scoring systems that match automatically the possible functional relations of human genes to hereditary diseases are developed which greatly facilitate the disease-gene discovery (Perez-Iratxeta *et al*., 2002; Turner *et al*., 2003). Other groups aim at exploiting sequence-based features and found that for many of them there are significant differences between genes underlying human hereditary disease and those not known to be involved in disease (Adie *et al*., 2005; Lopez-Bigas and Ouzounis, 2004). Since protein plays its function in a modular and interactive fashion and mutations in interacting proteins may lead to similar phenotypes, a more direct and robust manifestations of a functional relationship between genes is through protein–protein interactions (PPIs) network. As found from a recent analysis of the human interaction map, that the genes ascertained, from the Online Mendelian Inheritance in Man (OMIM) database (Hamosh *et al*., 2005), to be associated with a human disease preferentially interacted with other disease-causing genes significantly indicating heritable disease-genes might share some topological features in the PPIs network, whereas the non-disease-genes do not (Gandhi *et al*., 2006).

The rapid identification of genome-wide human PPIs network provided us with new avenues for elucidating the disease-gene directly from PPIs network. In this paper, five topological features, which describe a gene (i.e. a protein) in the PPIs network, were compared between known heritable disease-genes and those not known to be involved in disease were made. By using a machine learning algorithm we created an automatic classifier capable of identifying genes more likely to be involved in hereditary disease based on the topological patterns.

## 2 METHODS

### 2.1 Human interaction datasets

Human PPIs datasets were downloaded from Online Predicted Human Interaction Database (OPID) (Brown and Jurisica, 2005). It collected the identified human PPIs from: (1) literature-curated (LC) interactions from BIND (Bader *et al*., 2003), HPRD (Peri *et al*., 2003) and MINT (Zanzoni *et al*., 2002); (2) interactions identified from high-throughput yeast two-hybrid mapping approach (EXP) (Rual *et al*., 2005; Stelzl *et al*., 2005); (3) predicted interactions(PDT) made from *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *and Mus musculus* by 'interologs' (i.e. potential interactions predicted from interactome data available for model organisms given evolutionary conservation of two known partners). Total number of PPIs and number of unique proteins in

*To whom correspondence should be addressed.
The authors wish it to be known that, in their opinion they should be regarded as joint First Authors.

 

**Table 1.** PPI datasets

| Datasets | PPIs | Proteins | Disease-genes |
|---|---|---|---|
| LC | 17183 | 5955 | 1269 |
| EXP | 6273 | 3025 | 485 |
| PDT | 25170 | 4313 | 560 |

the largest connected network component (main component) for each of above PPIs network are listed in Table 1.

To be as comprehensive and unbiased as possible we evaluated the classifier on all the three datasets.

### 2.2 Building training samples

List of known hereditary disease-genes were obtained from the 'morbid map' table in the OMIM database. In each of the three PPI datasets, we mapped such genes to the main component networks, the corresponding genes were called the 'disease-gene set' and total number of such genes are listed in Table 1.

Compiling a list of genes that are known not to be involved in hereditary disease is difficult or even impossible currently. A recent study (Tu *et al*., 2006) showed that the human genome may contain thousands of essential genes having features which differ significantly from both disease-genes and the other genes. They proposed to classify them as a unique group for comparisons of disease-genes with non-disease-genes. In the absence of a set of well-defined human essential genes, they compiled a list of ubiquitously expressed human genes (UEHGs) as an approximation of essential genes. Thus in contrast to previous analysis (Adie *et al*., 2005; Lopez-Bigas and Ouzounis, 2004) which regarding that all the gene set were not listed in OMIM as control set, we (1) excluded UEHGs from the gene population that was not listed in OMIM, the remaining genes were called 'control-gene set'; (2) then we randomly selected, from the control-gene set, genes with a size equal to that of the disease-gene that was detected in the corresponding network as negative training samples. Our final training data consisted of the sampled negative representative data and the fixed positive disease data.

### 2.3 Defining features set

For each node $i$ in the PPI network we defined five measures for assessing its topological property (listed in Table 2). Briefly, degree defined as the number of links to node $i$. 1N designates the neighbors of node $i$ and 2N designates the neighbors' neighbor of node $i$. 1N index and 2N index are defined as the proportion of the number of links to other node which listed in 'disease-gene set' among all links for 1N and 2N, respectively. The measure of average distance to disease-genes is used to assess the communication efficiency of node $i$ to overall 'disease-gene set' in PPIs network. The lower this measure it corresponds to a quicker transduction between node $i$ and 'disease-gene set'. The positive topological coefficient is essentially a variant of the classical topological coefficient (Goldberg and Roth, 2003). It is a measure for the extent to which a protein in the network shares interaction partners with other disease-genes.

The above measures were computed for each node in the main component of the network. To assess the statistical significance of the measures between the disease-gene set and control-gene set, Wilcoxon rank sum test for equal medians was applied to the populations. The medians and $P$-values of the features are listed in Table 3.

### 2.4 Classification algorithm and validation

We choose the k-nearest neighbors (KNN) as classification algorithm. The KNN algorithm is a simple yet powerful non-parametric classification algorithm (Duda *et al*., 2000). It is widely used in bioinformatics fields. Despite its simplicity, it can give competitive performance compared with many other methods. The topological measures were computed for each node in the PPIs network and each feature vector was then normalized to have a zero mean and unit variance so that the ranges of the features used in the classifier are comparable. Euclidean distance was used as the distance measure to find the nearest neighbors. The number of nearest neighbors ($K$) is a critical parameter for KNN algorithm. Tests have been done with various values of $K$ ($K = 1, 3, 5, 7$ and 9). For a fixed $K$, each sample is classified based on simple majority of class membership of its KNN in the training data.

We used the 10-fold cross-validation test to get an estimate of how our classifier might perform on unseen data. During the test process a fraction of the data (in this case, 10% of the whole, with the same balance of disease-genes and negative genes) is singled out in turn as a test sample, the remaining genes are used as the training set to calculate the test sample's membership and predict the class. The prediction quality was evaluated by the overall prediction accuracy, prediction sensitivity and precision.

$$\text{Accuracy} = \frac{\text{TP} + TN}{\text{TP} + \text{TN} + \text{FP} + \text{FN}},\ \text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

and

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}},$$

where TP, TN, FP and FN denote true positive, true negative, false positive and false negative, respectively.

KNN classification and validation procedures were implemented using SPIDER package (http://www.kyb.tuebingen.mpg.de/bs/people/spider/main.html), all the other computations were carried out by custom written MATLAB R13 (Mathworks, Natick, MA) scripts (available on request).

## 3 RESULTS

### 3.1 Disease-genes are clustered in LC PPIs network but not in EXP and PDT PPIs networks

In the LC PPIs network, our results demonstrate, that the degree of disease-genes is significantly higher than that of genes in the control dataset. This confirms previous findings that disease-genes have larger degree (Tu *et al*., 2006). We also found among all the neighbors of a disease node (gene) the proportion of being another disease node is significantly larger than that of a node in control set ($P < 2.22 \times 10^{-16}$) which is consistent with recent observation that human disease-gene preferentially interacted with other disease-causing genes significantly (Gandhi *et al*., 2006). Similarly, we observed that the neighbors of a disease-gene, more likely to be another disease-gene, preferentially interacted with other disease-genes too ($P < 2.22 \times 10^{-16}$). Compared with the control-gene set, we found that the measure of average shortest path to disease-gene set is significantly lower in the known disease-genes set ($P < 0.004$), indicating that the disease-genes communicated with each other quickly in LC PPIs network. Intuitively, if a gene shares more neighbors with known hereditary disease-genes in the PPIs network, it should more likely be a disease-gene too. So we adopted positive topological coefficient to assess the level of a gene that shares neighbors with known hereditary disease-genes. Indeed our empirical results show this measure is significantly larger for known hereditary disease-genes ($P < 0.003$). Proteins seldom function individually, but rather in a modular fashion. All our measures, except the degree measure, assess such modularity of proteins from different but related facets. The above results suggested that the disease-genes are heavily clustered into modules in the LC dataset.

For EXP datasets, we found the differences between the control-gene set and disease-genes set measured by 1N index, 2N index and

**Table 2.** Topological feature set

| Feature | Function | Description |
|---|---|---|
| Degree | $k_i$ | the number of links to node $i$ |
| 1N index | $k_i^p/k_i$ | $k_i^p$ is the number of links between node $i$ and disease-genes |
| 2N index | $\sum_{j\in N_i} k_j^p / \sum_{j\in N_i} k_j$ | $N_i$ denotes the set of indices of the neighbors of node $i$ |
| Average distance to disease-genes | $\sum_{j\in M} d_{ij}/\lvert M\rvert$ | $M$ denotes the set of indices corresponding to disease-genes; $d_{ij}$ denotes length of the shortest path between node $i$ and node $j$ |
| Positive topology coefficient | $\sum_{j\in M_i} \frac{C_{ij}}{\min(k_i,k_j)} \Big/ \lvert M_i\rvert$ | $C_{ij}$ denotes the number of nodes to which both $i$ and $j$ are linked; $M_i$ is the set of indices of the disease-genes sharing neighbors with node $i$ |

**Table 3.** Medians of the topological features between the control-genes set and disease-genes set

| | LC | | | EXP | | | PDT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Disease | Control | *P*-value | Disease | Control | *P*-value | Disease | Control | *P*-value |
| Degree | 3.000 | 2.000 | 2.22E−16 | 3.000 | 3.000 | 7.36E−01 | 2.000 | 2.000 | 5.17E−01 |
| 1N index | 0.313 | 0.077 | 2.22E−16 | 0.049 | 0.000 | 8.74E−03 | 0.000 | 0.000 | 3.86E−01 |
| 2N index | 0.333 | 0.211 | 2.22E−16 | 0.181 | 0.111 | 2.22E−16 | 0.217 | 0.136 | 2.22E−16 |
| Average distance to disease-genes | 4.476 | 4.515 | 3.70E−03 | 4.484 | 4.446 | 2.56E−01 | 4.684 | 4.635 | 3.18E−01 |
| Positive topology coefficient | 0.095 | 0.094 | 2.87E−03 | 0.037 | 0.033 | 3.48E−02 | 0.088 | 0.088 | 6.51E−01 |

Significances between the two gene populations were also listed (calculated by the Ranksum test).

**Table 4.** More detailed classifier performance statistics

| | LC | | | EXP | | | PDT | | |
|---|---|---|---|---|---|---|---|---|---|
| | Accuracy | Sensitivity | Precision | Accuracy | Sensitivity | Precision | Accuracy | Sensitivity | Precision |
| $K = 1$ | 0.76 | 0.75 | 0.76 | 0.82 | 0.85 | 0.81 | 0.77 | 0.79 | 0.76 |
| $K = 3$ | 0.76 | 0.75 | 0.77 | 0.80 | 0.82 | 0.79 | 0.78 | 0.80 | 0.77 |
| $K = 5$ | 0.75 | 0.74 | 0.76 | 0.78 | 0.79 | 0.77 | 0.78 | 0.77 | 0.78 |
| $K = 7$ | 0.74 | 0.74 | 0.75 | 0.77 | 0.78 | 0.76 | 0.78 | 0.78 | 0.78 |
| $K = 9$ | 0.74 | 0.74 | 0.75 | 0.76 | 0.77 | 0.75 | 0.78 | 0.79 | 0.77 |

positive topological coefficient are statistically significant. However we could not find a difference on the degree measure between the two gene populations (Table 3). Surprisingly, compared with that of the control-gene set, we found the measure of average shortest path to disease-gene set is even slightly higher for the known disease-genes set. In PDT datasets, we only found statistically significant difference for 2N index between the control-gene set and disease-genes set. These results indicate the structures of PPI network in EXP and PDT datasets are different from that of LC datasets and currently there is not enough evidences for suggesting that the disease-genes are clustered in EXP or PDT datasets.

We think the disparity of PPIs network topology could be explained by the fact that due to the priority interests of disease-genes in current literature LC dataset is expected to be biased toward known disease-genes. Thus lots of heavily connected disease-genes (as they are the center of study) lay in the LC dataset. On the other hand, genes tested in the EXP/PDT datasets were not intentionally selected thus strong bias should not exist. Considering the estimated 25 000 human genes and 15 interactions per protein, there would be <375 000 interactions in the complete human protein interaction network (Ramani *et al*., 2005), the known human PPIs currently is rather small, a similar but more comprehensive investigation on topological features is warranted when more PPIs are available.

## 3.2 Satisfactory performance of KNN classifier on all the three PPIs networks

Despite the distinct topologies of the three PPIs networks, KNN classifiers work equally well. On the LC dataset, KNN classifier ($K = 3$) correctly recovered 75% of known hereditary disease-genes with a precision of 77% during 10-fold cross-validation. On average, overall accuracy is 0.76 which outperforms previous predictions significantly (Adie *et al*., 2005; Lopez-Bigas and Ouzounis, 2004). Comparable classifier performances could be obtained on both EXP and PDT datasets (Table 4). In the practice of machine learning, it is common to see two individually irrelevant features for classification as judged by univariate methods, which may become relevant when used in combination because of the underlying feature of independent assumptions made by univariate

methods (Guyon *et al*., 2006). In our case, most measures were designed to assess modularity of proteins from different but related facets. It is not surprising that a complex web of relatively weak correlations could be found among those measures. For example, 1N and 2N index is positively correlated with Spearman correlation coefficient of 0.34. The feature of independent assumption is obviously violated, so those features appear to be useful in the prediction of disease-genes collectively. Indeed, none of the five topological features could accurately predict disease-genes individually as evaluated by 10-fold cross-validation (Supplementry Table S1).

It can also be seen that the prediction accuracy does not change significantly when $K$ is set differently. Thus we demonstrated that the KNN based classifier is robust to the setting of a number of nearest neighbor under current scenario. To further avoid sampling bias of negative training data, we repeated the sampling procedure 1000 times and in each time a classifier was trained and evaluated. The prediction performance of classifiers were essentially invariant indicating that the result was largely free of sampling bias (Supplementry Table S2–4). The results reported in the following sections were obtained using $K = 3$ on LC datasets.

### 3.3 KNN classifier is robust to the changing of disease training sample

As new disease-genes are continually being discovered, it would therefore be interesting to see how the KNN classifier would perform companying with the novel disease-genes discovery process. On the LC network we imitated the circumstance as follows: first only a proportion of disease-genes were random sampled from the original disease-gene set (from 60 to 90%) and used as positive training samples for KNN classifier learning. The remaining true disease-genes were regarded as 'unknown novel disease-genes' and pooled with the 'control-gene set' from which the negative samples were randomly picked out (thus some true disease-genes could be sampled and deemed as negative non-disease-genes in training). Then the trained classifier was used to predict the whole unknown genes except for the genes used for training. At each percentage, the sampling was repeated 1000 times and the average performance are shown in Figure 1. From the result we can see that the KNN classifier works considerably well even as the positive training sample was reduced to 60% of the original one and negative sample was mixed with some positive disease samples. As the size of the training samples increased there is no doubt that the performance increases towards the highest, where all the disease-genes are used for training. Thus we can expect the performance to improve even further when new disease-genes are continually discovered and added into the training samples in the future.

At each percentage, the leave-out disease-genes mixed with other unknown genes were used to assess the ability of novel disease prediction for classifiers. For instance when 90% of the disease-genes were sampled as training data and the other 127 disease-genes (10%) used as test data, we noted that, on average, 754 genes were predicted to be disease-genes based on simple majority of the class of the $K = 3$ nearest neighbors in the training data. Of these predicted disease-genes, 66 genes (8.8%) are already listed in the 10% leave-out genes. Used as baseline, in the 2978 genes for prediction, 127 genes (4.3%) were known disease-genes. Thus this represents in percentage a 8.8/4.3 i.e. ~2.0-fold enrichment relative to random prediction directly. If we require a gene was predicted to be disease-gene only if all of the $K = 3$ nearest neighbors are disease-genes, the disease-gene pool could be further enriched. Of the 184 disease-gene predictions, 35 (~19%) are already listed in 10% leave-out disease-genes. This represents in percentage a 19/4.3 i.e.~4.4-fold enrichment. Similar results at other percentages are shown in Supplementry Table S5.

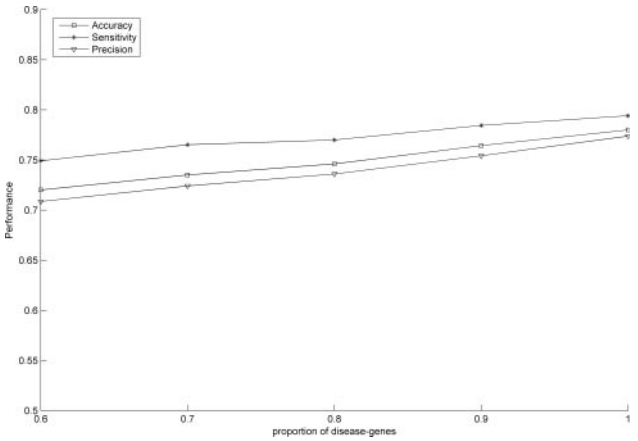


**Fig. 1.** Performance of KNN classifier with adding new disease-genes.

It should be noted that although the positive disease-genes obtained from OMIM can generally be trusted, the randomly selected negative examples were from the gene population that were not listed in OMIM and presumably were not known to be involved in disease. However, some of these genes may well be involved in disease, although they have not been identified yet. In the above case, as the positive training sample was reduced to 60% of the original one, there should be more disease-genes sampled as negative non-disease-genes. The stable performance suggested KNN classifier, at least to some degree, is insensitive to negative sample impurity.

### 3.4 Prediction novel disease-genes from human PPIs network

Finally, the trained classifier from LC when $K = 3$ was applied to all the other 5262 genes except the UEHGs in the LC network. Genes (1872) were predicted to be hereditary disease-genes based on simple majority of the class of the $K = 3$ nearest neighbors in the training data of which 1087 genes are already known disease-genes. If we require a gene was predicted to be disease-gene only if all of the $K = 3$ nearest neighbors are disease-genes, 970 disease-gene were identified of which 792 genes were already listed in 'morbid map' table in the OMIM. In the novel 178 putative disease-genes (listed in Supplementry Table S6), some have shown experimental evidences. For example, TGF-beta-activated kinase 1 (TAK1, Swissprot ID O43318), a member of the serine/threonine protein kinase, mediates the signaling transduction induced by TGF beta and morphogenetic protein (BMP), and controls a variety of cell functions including transcription regulation and apoptosis. It has been shown hypohidrotic/anhidrotic ectodermal dysplasia (ED), a disorder characterized by sparse hair, lack of sweat glands and malformation of teeth, was caused by dysfunction of Edar (Ectodysplasin receptor) signaling. The Edar signaling pathway

stimulates NF-kappa B transcription factors via an activation of the IkappaB kinase (IKK). Dominant negative forms of TAK1, TAB2 (TAK1-binding protein 2) and TRAF6 (TNF-receptor-associated factor 6) blocked the NF-kappa B activation induced by Edaradd (Edar-associated death domain). These results support the involvement of the TAK1/TAB2/TRAF6 signaling complex in the Edar signal transduction pathway and implied TAK1 as a new disease-gene candidate for ED (Morlon *et al*., 2005)

Another example comes from adenylate cyclase 5 (*ADCY*5, Swissprot ID: O95622). Okumura *et al*. (2003) examined the effects of pressure overload, induced by thoracic aortic banding in type 5 adenylyl cyclase disrupted mice. They showed disruption of type 5 adenylyl cyclase gene preserves cardiac function against pressure overload and seemed to involve Bcl-2, which was up-regulated significantly more in mutated mice with pressure overload. Other research also demonstrated expression of ADCY5 mRNA and protein is up-regulated in cyanotic infant human myocardium (Zhao *et al*., 2002). These lines of evidences agree with our prediction that ADCY5 as a potential disease candidate gene.

Finally we looked for experimental knowledge support for our prediction of basic charge, Y-linked 2 (*BPY*2, Swissprot ID:O14599). Interestingly, this gene is located in the non-recombining portion of the Y-chromosome and expressed specifically in testis. To determine the possible relationship of BPY2 with the pathogenesis of male infertility, Tse *et al*. (2003) examined a group of infertile men with and without Y-chromosome microdeletions and with known testicular pathology using BPY2 antibody. The impaired expression of BPY2 in infertile men suggests its involvement in the pathogenesis of male infertility. Other study that analyzed the deletion of six Y-chromosome-specific genes in prostate cancer samples also shows *BPY*2 gene was lost in 42% of tested cases. Especially the loss of *BPY*2 genes was more frequent in higher stages and grades of prostate cancer, suggesting its role in pathogenesis of this disease (Perinchery *et al*., 2000).

## 4 DISCUSSION

We report here that the disease-genes from OMIM (most are Mendelian disease-genes) follow a specific topological property pattern in human LC PPIs network. In addition to the previous observed features such as a disease node has more links and preferentially interacted with other disease node, we found the neighbor of a disease node, more likely to be another disease protein, which also preferentially interacted with other disease nodes. Similarly, we demonstrated disease nodes which share more neighbors with other disease nodes and communicate with each other quickly. However those patterns could not be detected from high-throughput yeast EXP and PDT PPIs networks, which perhaps suggested that the LC PPIs network is biased towards known disease-genes due to the priority interests of disease-genes in current literatures. We have trained an efficient KNN classifier based on these topological features in PPI network to predict which of the genes in the human genome are more likely to be involved in disease.

Our classifier was trained using all disease-genes from OMIM and provide the insights into the general nature of human Mendelian diseases genes in the PPIs network. A further exploration is to study whether the classifier could work equally efficiently when users are interested in a particular disease, for example sudden cardiac death (SCD; Arking *et al*., 2004). Alternatively, one may use our prediction result as an evidence when searching for the candidate gene in a predefined disease loci for later bench identification. If in a disease loci there are several genes picked out by our classifier to be disease-genes from many others, those genes are highly possible to play a role in the pathogenic mechanism of disease and have a high priority for further evaluation. Recently Oti *et al*. (2006) took a heuristic approach to search for candidate disease-genes that interact with a known disease-gene and were located within one of these predefined disease loci. Some promising results were reported.

In previous studies a variety of sequence features such as cDNA and protein sizes, the evolutionary conservation rates, number of exons (Adie *et al*., 2005; Lopez-Bigas and Ouzounis, 2004) were extensively investigated between the sets of genes known to be involved in human hereditary disease and those not known. Decision tree classifiers were constructed based on common sequence patterns. Computational algorithms for prediction of disease-genes based on shared functional annotation to known disease-genes have also been reported (Perez-Iratxeta *et al*., 2002; Turner *et al*., 2003). We proposed a KNN based classifier to explicitly and directly exploit the recent PPIs data. Each method has its pros and cons. For example, due to the incompleteness of functional knowledge, although successful cases has been reported, algorithms relying on functional annotation system are inherently biased towards a particular well-studied subset of genes. Our PPIs based prediction achieved higher accuracy while the genome coverage needed further improvement when more human PPIs are available. A promising approach for disease-gene discovery is to integrate all source of genomic evidences such as PPIs, transcriptional expression, sequence, function annotation and linkage data to make the final prediction through, for instance, Bayesian network learning framework as demonstrated recently for mitochondrial disease-genes mining (Calvo *et al*., 2006) and others for more general applications (Aerts *et al*., 2006; Franke *et al*., 2006).

To our knowledge, this is the first time topological properties in three different human PPI networks were systematically compared between the sets of genes known to be involved in human hereditary disease and those not known. Based on those features, KNN classifier was constructed and applied to 5262 genes and predicted 178 putative disease-genes. With increasing quantity and quality of human interaction and phenotypic data, the performance and utility of this approach in facilitating biologists to detect novel disease-genes should improve even further.

## ACKNOWLEDGEMENTS

We are grateful for the comments and suggestions from the three anonymous reviewers.

*Conflict of Interest*: none declared.

## REFERENCES

Adie,E.A. *et al*. (2005) Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, **6**, 55.

Aerts,S. *et al*. (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol*., **24**, 537–544.

Arking,D.E. *et al*. (2004) Genomics in sudden cardiac death. *Circ. Res*., **94**, 712–723.

Badano,J.L. and Katsanis,N. (2002) Beyond Mendel: an evolving view of human genetic disease transmission. *Nat. Rev. Genet*., **3**, 779–789.

Bader,G.D. *et al*. (2003) BIND: the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250.

Botstein,D. and Risch,N. (2003) Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat. Genet*., **33** (Suppl.), 228–237.

Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.

Brunner,H.G. and van Driel,M.A. (2004) From syndrome families to functional genomics. *Nat. Rev. Genet*., **5**, 545–551.

Calvo,S. *et al*. (2006) Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat. Genet*., **38**, 576–582.

Duda,R.O., Hart,P.E. and Stork,D.G. (2000) *Pattern Classification*. Wiley, NY.

Franke,L. *et al*. (2006) Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet*., **78**, 1011–1025.

Gandhi,T.K. *et al*. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet*., **38**, 285–293.

Goldberg,D.S. and Roth,F.P. (2003) Assessing experimentally derived interactions in a small world. *Proc. Natl Acad. Sci. USA*, **100**, 4372–4376.

Guyon,I.M., Gum,S.R., Nikravesh,M. and Zulch,I. (2006) *Feature Extraction, Foundations and Applications*. Springer.

Hamosh,A. *et al*. (2005) Online Mendelian Inheritance in Man (OMIM), a knowledge-base of human genes and genetic disorders. *Nucleic Acids Res*., **33**, D514–D517.

Lopez-Bigas,N. and Ouzounis,C.A. (2004) Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res*., **32**, 3108–3114.

Morlon,A. *et al*. (2005) TAB2, TRAF6 and TAK1 are involved in NF-kappaB activation induced by the TNF-receptor, Edar and its adaptator Edaradd. *Hum. Mol. Genet*., **14**, 3751–3757.

Okumura,S. *et al*. (2003) Disruption of type 5 adenylyl cyclase gene preserves cardiac function against pressure overload. *Proc. Natl Acad. Sci. USA*, **100**, 9986–9990.

Oti,M. *et al*. (2006) Predicting disease genes using protein–protein interactions. *J. Med. Genet*.

Perez-Iratxeta,C. *et al*. (2002) Association of genes to genetically inherited diseases using data mining. *Nat. Genet*., **31**, 316–319.

Peri,S. *et al*. (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res*., **13**, 2363–2371.

Perinchery,G. *et al*. (2000) Deletion of Y-chromosome specific genes in human prostate cancer. *J. Urol*., **163**, 1339–1342.

Ramani,A.K. *et al*. (2005) Consolidating the set of known human protein–protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol*., **6**, R40.

Rual,J.F. *et al*. (2005) Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, **437**, 1173–1178.

Stelzl,U. *et al*. (2005) A human protein–protein interaction network: a resource for annotating the proteome. *Cell*, **122**, 957–968.

Tse,J.Y. *et al*. (2003) Specific expression of VCY2 in human male germ cells and its involvement in the pathogenesis of male infertility. *Biol. Reprod*., **69**, 746–751.

Tu,Z. *et al*. (2006) Further understanding human disease genes by comparing with housekeeping genes and other genes. *BMC Genomics*, **7**, 31.

Turner,F.S. *et al*. (2003) POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol*., **4**, R75.

Zanzoni,A. *et al*. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.

Zhao,Y. *et al*. (2002) Expression of adenylyl cyclase V/VI mRNA and protein is upregulated in cyanotic infant human myocardium. *Pediatr Cardiol*, **23**, 536–541.