Physica A 496 (2018) 53–61

Contents lists available at ScienceDirect

Physica A

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

Discovering disease-associated genes in weighted protein–protein interaction networks

Ying Cui [a,b,d,*], Meng Cai [c,d], H. Eugene Stanley [d]

[a] School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China
[b] Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an 710071, China
[c] School of Economics and Management, Xidian University, Xi'an 710071, China
[d] Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA

HIGHLIGHTS

- Weight of links is taken into consideration in the construction of a PPI network.
- Disease genes show distinct topological properties from non-disease genes.
- An improved forest-based model was applied as classifier.
- Weighted networks perform better than unweighted networks.

ARTICLE INFO

Article history:
Received 8 October 2017
Received in revised form 20 November 2017
Available online 26 December 2017

Keywords:
Disease gene discovering
Topological properties
Weighted PPI network
Machine learning

ABSTRACT

Although there have been many network-based attempts to discover disease-associated genes, most of them have not taken edge weight – which quantifies their relative strength – into consideration. We use connection weights in a protein–protein interaction (PPI) network to locate disease-related genes. We analyze the topological properties of both weighted and unweighted PPI networks and design an improved random forest classifier to distinguish disease genes from non-disease genes. We use a cross-validation test to confirm that weighted networks are better able to discover disease-associated genes than unweighted networks, which indicates that including link weight in the analysis of network properties provides a better model of complex genotype–phenotype associations.

1. Introduction

Networks provide a ubiquitous and efficient tool for the analysis of biological systems [1–3]. Researchers found that a disease phenotype rarely results from an aberration in a single gene, but is a consequence of various pathological processes that interact in a complex network [4]. The requirement of discovering disease genes from molecular networks leads to the development of "Network medicine" [4], which recapitulates the molecular complexity of human disease and offers network-based computational methods to unravel how the molecular complexity manipulates human disease.

Researches dedicating to systematically capture the properties of disease-associated genes in molecular networks have demonstrated that genes related to the same or similar diseases, tend to cluster and interact with each other in these networks [5–7]. These findings promote the development of network-based approaches for identifying and prioritizing

[*] Corresponding author at: School of Mechano-Electronic Engineering, Xidian University, Xi'an 710071, China.
E-mail addresses: ycui@xidian.edu.cn (Y. Cui), mcai@xidian.edu.cn (M. Cai).

https://doi.org/10.1016/j.physa.2017.12.080
0378-4371/© 2017 Elsevier B.V. All rights reserved.

candidate disease genes by using biological network data, such as protein–protein interaction (PPI) networks [8–10], disease phenotype networks [11–13], regulatory networks [14–16] and co-expression networks [17–19], etc.

Although has contributed a lot to disease diagnose and therapy [4], discovering disease-associated genes by using biological and biomedical networks, is still a challenging task in human genetics. Current network-based approaches for identifying disease-related genes have important limitations [20]. Plenty of network-based methods depend on sophisticated integrated data source [11–19,21,22], which lead to time consumption and increasing computing complexity. Our response is to propose a novel network-based approach to discover disease-related genes by using only PPI network data. A PPI network consists of physical interactions between proteins and is widely used in discovering disease genes [23]. The connections between proteins and human diseases confirm that proteins that physically interact with each other share a common function [5,24]. Thus, an aberration in one protein tends to replicate similar disease phenotypes. Accordingly, PPI network is a powerful data source for discovering disease-related genes.

Furthermore, in most exiting network-based methods, all the connections between nodes are binary with values being either 1 or 0, which means the networks they used to identify disease genes are unweighted. Unweighted networks can only reflect whether there are any interactions between vertices, but fail to display different interaction weights between nodes. However, as has long been appreciated, many molecular networks, such as PPI networks, are intrinsically weighted, their edges are not merely binary entities, but have different weights that record their strengths relative to one another. The weight of edges in molecular networks plays an important role in deciphering the topological properties of PPI networks. In order to take into account the existing heterogeneity in the capacity and intensity of connections, we employ a weighted PPI network to analyze the distinct topological properties between disease and non-disease genes.

In this paper, we propose a hybrid network-based method for the discovery of disease-related genes. We consider the heterogeneity of interactions between genes and construct a weighted PPI network for the purpose to better capture the network topological properties. We then analyzed topological properties of both weighted and unweighted PPI networks. The analysis results show that disease genes have discriminatory network properties which enable their distinction from non-disease genes in both weighted and unweighted PPI network. We constructed four different classification models based on KNN, SVM, Random Forest and CForest, respectively. The topological properties are combined in tandem to use as inputs of the classifier. We use grid-search and 10-fold cross validation to find the optimal parameters for every classifier and CForest is chosen as the best classification model according the prediction performance. The computational simulation results reported that the weighted and unweighted networks achieve 88.56% and 83.57% classification accuracy, respectively. It demonstrated that, by considering the weight of edges, we successfully improved the discovery of disease genes and contributed a deeper understanding into complex genotype–phenotype relationships.

2. Materials and methods

2.1. Data sources

We downloaded high-quality protein–protein interactions from HIPPIE v2.0 (the Human Integrated protein–protein Interaction rEference) [25]. HIPPE database collects human PPIs with experimental annotations from seven major expert-curated databases [26–32]. For each PPI, HIPPIE assigned a stringent confidence score to reflecting its reliability and authenticity. This score is computed by integrating diverse experimental evidence and applying basic network node importance evaluating algorithms. HIPPIE map all source database entries to gene names, Entrez gene ids and UniProt ids or accessions. In this work, we downloaded 71 823 "high confidence" PPIs (confidence score is equal or greater than 0.73) involving 11 813 proteins.

We obtained the list of disease-associated genes and non-disease genes supplied by the Online Mendelian Inheritance in Man (OMIM) [33]. The genemap file of OMIM has 16 161 records with gene symbols, MIM number and related disease phenotypes. We use phenotype mapping key, which appears in parentheses after a disorder, to select gene–disease relations with key (3). This group of gene–disorder associations has well-known molecular basis and a mutation to support. We got 10 980 genes involving 3700 disorder phenotypes, indicating that most diseases are polygenic.

Gene Symbols are used to match the disease and non-disease gene lists from OMIM to protein–protein interactions from HIPPIE. We got 1608 genes that have at least one related disease phenotype and one PPI as disease gene samples, and 3645 genes with interactions in HIPPIE but no disease association record in OMIM as non-disease gene samples. These 1608 disease genes are considered to be positive samples, and 1608 non-disease genes are randomly selected to be negative samples.

2.2. Weighted PPI network construction

We constructed a weighted PPI network by the 71 823 high-quality PPIs we got from HIPPIE. This weighted network is modeled as an undirected graph $G_w = (V, E, w_e)$, where $V = \{v_1, v_2, \ldots, v_M\}$ is the set of nodes, $E = \{e_1, e_2, \ldots, e_N\}$ is the set of edges and w_e is the set of edge weights. Two nodes, referring to proteins in PPI network, are connected by a weighted edge if there is an interaction between them. Based on the confidence score, each edge is assigned a weight

$$w(e_i) = \frac{s_i}{\sqrt{\sum_{i-1}^{N}(s_i - \bar{s})^2/(N-1)}}, \tag{1}$$

Table 1
Topological properties of PPI network.

Features	Function	Description	Ref.
Degree	k_i	The number of edges connected with node *i*.	[35]
ANN	$\frac{\sum_{i=1}^{k} k_i}{k}$	The average nearest neighbor degree of node *i* with degree *k*.	[36]
Authority Centrality	$t(A) * Ax = \lambda x$	The principal eigenvector of $t(A) * A$	[37]
Betweenness Centrality	$g(i) = \sum_{i \neq j \neq k} \frac{\sigma_{jk}(i)}{\sigma_{jk}}$	σ_{jk} is the total number of shortest paths from node *j* to node *k*, $\sigma_{jk}(i)$ is the number of that paths going through node *i*.	[38]
Closeness Centrality	$C(i) = \frac{N}{\sum_{i \neq j} d(i,j)}$	*N* is the network size, $d(i, j)$ refers to the shortest path between node *i* and *j*.	[39]
Eigenvector Centrality	$Ax = \lambda x$	λ is the eigenvalue of adjacency matrix *A*.	[40]
PageRank	$PR(i) = (1 - d) + d \sum_{j \in N(i)} \frac{PR(j)}{k_i}$	$N(i)$ stands for the neighbors of node *i*, *d* s damping factor.	[41]
Burt's constraint	$cons(\upsilon) = \sum localCons(\upsilon, \omega)\,(\omega \in MP(\upsilon), \omega \neq \upsilon)$	$MP(\upsilon)$ is the subset of υ's neighbors that are oth predecessors and successors of υ.	[42]

where $w(e_i)$ is the weight of edge i, $i \in \{1, \cdots, N\}$, s_i refers to the confidence score of edge i and $\bar{s}$ is the mean value of all s. The weight of a protein pair indicates the reliability of the interactions between the two proteins. After discarding self-loops, multiple edges, we have a weighted PPI network with 11 183 nodes and 66 718 edges.

2.3. Network topological properties analysis

It has been noticed that disease genes have distinct network topological properties with non-disease genes [34]. We therefore analyze how the topological properties of disease genes in the network differ from those of non-disease genes for the purpose to discover disease genes from non-disease genes. We use eight network topological measurements to capture the topological properties of the PPI network, which can be broadly categorized into (i) neighbor-based measurements, including degree and average nearest-neighbor degree, (ii) path-based measurements, including betweenness centrality, closeness centrality and Burt's constraint, (iii) eigenvector-based measurements, including authority centrality, eigenvector centrality, and PageRank. In Table 1, we display an introduction of these topological features with function, description, and reference.

These features evaluate the topological importance of a gene in the PPI network from various perspectives. For instance, the degree of a node shows how many edges that incident to this node, it believes that the more neighbors a node owns, the more influential it is. While *K*-Nearest Neighbor displays the average nearest neighbor degree of a node, which is obviously complementary with degree, cause different neighbors' degree of different nodes will unavoidably lead to different topological importance.

Betweenness, closeness and Burt's constraint are path-based centrality measurements. Betweenness captures the control of nodes on the network flow along the shortest path in the network. The definition of betweenness suggests that if a node is the only way to communicate between other node pairs in the network, it will have an essential position in the network. Closeness judges the node importance by the average path length of information propagation in the network. High closeness score means a node is close to the center of the network, therefore more important. Burt's constraint measures the network closure and structural holes. A structural hole is understood as a gap between two individuals who have complementary sources to information. The smaller the constraint value of a node is, the larger the structural hole is, and the more important is the node in the network.

The other three are all eigenvector-based measurements, whose basic idea is to assign relative scores to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Eigenvector centrality uses the adjacency and eigenvector matrices to score the relative importance of all nodes in the network by weighting connections to highly important nodes more than connections to nodes of low importance. As graph G is undirected and loop-free, the adjacency matrix *A* is symmetric, and all diagonal entries are 0. Eigenvector centrality can be computed by finding the principal eigenvector of the adjacency matrix *A*. Eigenvector centrality is suited to measure nodes' power to influence other nodes in the network both directly and indirectly through its neighbors. Connections to neighbors that are in turn well connected themselves are rated higher than connections to neighbors that are weakly connected. Authority centrality and PageRank are variants of the eigenvector centrality measure. Authority centrality estimates the value of the content of the page. A node is important if it contains valuable content and hence receives many links from other important sources. A good authority represents a page that contain reliable information on the topic of interest and is linked by many different hubs. PageRank works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. It assigns a numerical weighting to each element of a hyperlinked set of documents, such as the World Wide Web, with the purpose of "measuring" its relative importance within the set. Although both authority centrality and PageRank are initially proposed as link analysis algorithms to rate Web pages, they also widely applied to any collection of entities with interactions such as molecular network.

The theoretical analysis of the above topological importance measurements indicates that using a solo indicator to evaluate the importance of a node in the network has a great one-sidedness. Therefore, we synthesize eight indicators to describe the complex network topology from different perspectives in this paper.

2.4. Classifier

We employ four binary classifiers with different principles to distinguish disease-associated genes from non-disease genes based on the network topological properties.

- KNN classifier.

The *k*-nearest neighbors (KNN) algorithm is a linear supervised pattern recognition method used for classification and regression [43]. KNN performs classification by comparing a specific test tuple to a collection of labeled examples in a training set. Each new sample in prediction set is classified according to the class of the majority of its k-nearest neighbors in the training set. Parameter "*K*" indicates the number of neighbors taken into account in determining the class. It has a great influence on the identification rate of KNN model.

- SVM classifier.

The support vector machine (SVM) is a supervised machine learning algorithm based on the statistical learning theory [44]. SVM is usually used to solve classification and regression problems and has been successfully applied to bioinformatics investigations, such as the identification of alternative splice sites [45]. The basic thought of SVM is to map the original data into a high-dimensional feature space through a nonlinear mapping function and then to construct a hyperplane as the discriminative surface between the positive and negative data.

- RF classifier.

Random Forest (RF) algorithm is an ensemble machine-learning method developed by Breiman [46], and has been widely applied to classification problems in bioinformatics area [47]. RF is consisted of many base tree-structured classifiers such as CART (Classification and Regression Tree) and is proven to be robust to noise, no over-fitting and computationally feasible. By applying CART as a base classifier, RF collects the outputs of all decision trees to vote for the final result, then a sample is classified according the voting result.

- CF classifier.

The majority-voting rule in traditional RF algorithm renders the minority categories more likely to be misclassified. In response to this limitation of traditional RF, an improved RF function called CForest (CF) is proposed [48]. In contrast to standard RF based on CART with unfair splitting criterion, CForest is established by unbiased base classification trees based on a conditional inference framework.

CForest provides an unbiased variable importance measure based on a conditional permutation scheme for feature selection [49]. The new permutation importance scheme is based on a partition of the entire feature space that is determined directly by the fitted forest model. Accordingly, it is practicable for demonstrating the influence of a variable and computing its permutation importance conditional on correlated covariates of any type.

2.5. Performance assessment

In this work, *precision*, *recall*, *accuracy* and AUC (area under curve) of ROC (receiver operating characteristic) curve are used to evaluate the prediction performance,

$$precision = \frac{TP}{TP + FP}, \tag{2}$$

$$recall = \frac{TP}{TP + FN}, \tag{3}$$

$$accuracy = \frac{TP + TN}{TP + TN + FN + FP}. \tag{4}$$

We calculate these performance measurements using the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) at a score cut-off that distinguishes predicted from non-predicted. Positives are disease genes and negatives are non-disease genes.

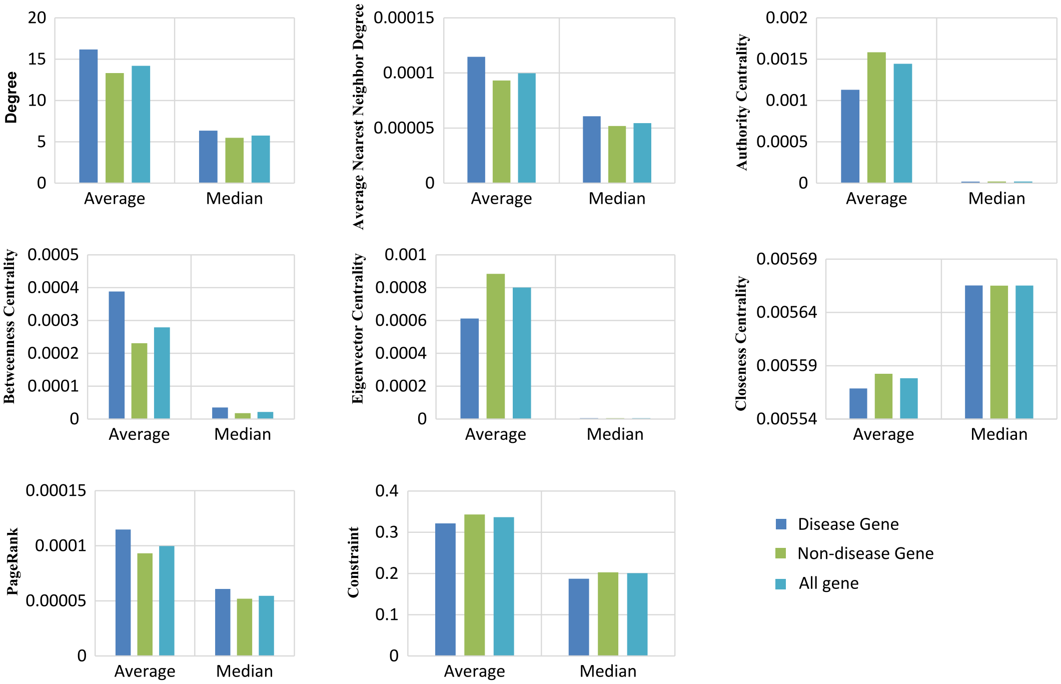

Fig. 1. Comparatively analysis of network topological properties between different gene groups in weighted PPI network.

3. Results

3.1. Results of network topological properties analysis

We calculated all the eight network topological importance indicators of both weighted and unweighted networks in R 3.2.3 and comparatively analyzed the mean value and median of these features in three different gene sets, including disease genes, non-diseases genes, and all genes, in order to explore the topological discrimination of disease and non-disease gene. The analysis results of weighted and unweighted networks are displayed in Fig. 1 and supplementary file S1, respectively. We use t test to measure the statistically significant differences between disease and non-disease genes and the p values show that the difference is significant in both weighted and unweighted networks.

It can be observed from Fig. 1 that, comparing to non-disease genes, disease genes tend to get higher scores of degree ($p = 0.0005156$), average nearest neighbor degree ($p = 0.006554$), betweenness centrality ($p = 0.000002979$), and PageRank ($p = 0.00005328$), but lower scores of authority centrality ($p = 0.0003492$), eigenvector centrality ($p = 0.0364$), closeness centrality ($p = 0.04986$) and constraint ($p = 0.0005156$). These analysis results show that disease genes have discriminatory network topological properties in PPI network that allow their distinction from non-disease genes.

3.2. Performance of different classifier

To highlight the good performance in discrimination of disease and non-disease genes according to network topological properties, we attempted to compare four classification results from KNN, SVM, RF and CF approaches in this work. In the conduction of classifier model, parameter tuning plays a crucial role to build a binary-class model with high prediction accuracy and stability. In each classifier, parameters were optimized according to the prediction performance they eventually achieve. A parameter will be settled when it reaches the best prediction performance.

In KNN model, we search the optimal parameter K value by using a 10-fold cross validation. The prediction errors for a given set of K values are estimated by cross-validation, and then the K value that makes the subsequent KNN classification yield optimal results is selected. Therefore, in this work, 20 K values ($K = 1, 2, \ldots, 20$) were optimized by a 10-fold cross-validation. The KNN classifier achieves an optimal performance when $K = 8$.

In SVM model, we attempt four common kernel functions, including radial basis function (RBF) kernel, linear kernel, polynomial kernel, and sigmoid kernel. Each kernel function has several parameters and proper tuning of these parameters significantly affects the classification performance of SVM model. We use a grid-search technique and 10-fold cross-validation to search the optimal parameter values of SVM with different kernels. The prediction performance results of different kernels in SVM are given in Table 2. For all the four kernels, RBF kernel achieves the best prediction rate. According

Table 2
Kernel optimization in the SVM classifier.

Kernel type	Parameters				Precision (%)	Recall (%)	Accuracy (%)
	C	g	d	r			
Linear	247				70.58	71.77	71.65
RBF	100	0.01			72.48	72.78	73.02
Poly	27	0.1	3		71.83	72.07	72.85
Sigmoid	512	0.01		0.2	72.02	71.86	72.98

Table 3
Parameter optimization details in different classifiers.

Classifier	Parameters	Step size in search	Search range	Optimal value
KNN	K	1	1:20	8
SVM	C	1	1 : 500	100
	$\lg g$	1	$10^{[-6:-1]}$	0.01
Random Forest	*ntree*	10	10:500	130
	mtry	1	1:8	3
CForest	*ntree*	10	10:500	200
	mtry	1	1:8	2

the performance, we chose RBF kernel as the basic kernel function of SVM classifier. There are two important parameters associated with RBF kernels: C and g. The optimal pair of (C, g) is found at $(100, 0.01)$ by a grid-search and 10-fold cross-validation.

In random forest model, two parameters are important: *ntree* and *mtry*. *ntree* is the number of trees used in the forest and *mtry* is the number of variables available for splitting at each tree node. The forest error rate depends on two things: (i) The correlation between any two trees in the forest. Increasing the correlation increases the forest error rate. (ii) The strength of each individual tree in the forest. Increasing the strength of the individual trees decreases the forest error rate. Reducing *mtry* reduces both the correlation and the strength. Increasing it increases both. Somewhere in between is an "optimal" range of *mtry*. Larger number of trees produce more stable models and covariate importance estimates, but require more memory and a longer run time. We use a grid-search and 10-fold cross-validation to find the optimal *ntree* and *mtry*. Finally, we found that when $ntree = 130$ and $mtry = 3$, the RF classifier reaches its best performance.

CForest parameter optimization is similar as random forest. A grid-search with cross-validation is adopted to search the best *ntree* and *mtry*. The optimal pair of $(ntree, mtry)$ is found at $(200, 2)$. All the parameter optimization details in KNN, SVM, RF and CF models were shown in Table 3.

With the optimal parameters, we constructed four classifier models to discover disease-associated genes based on KNN, SVM, RF and CF, respectively. We use precision, recall, accuracy and AUC to compare their prediction performance, the results are shown in Fig. 2. Seen from overall classification results in Fig. 2, the non-linear models (SVM, RF and CF) are superior to the linear model (KNN), and the CForest model is the best. These classification results might be explained by some relevant statistical learning theories. Theoretically, non-linear method performs better than linear method in the level of self-learning and self-adjust. Therefore, non-linear model has often a simpler structure and a higher performance of classification. The basic idea of CForest is to ensemble unbiased base classification trees in a conditional inference framework to eliminate errors caused by majority voting rule that used in KNN and RF model. Therefore, CForest embodies the superior in its theory, which may leads to a better prediction performance than other classifier.

3.3. *Feature importance*

CForest model provides an unbiased measure of variable importance, which can be used to evaluate the importance of the features applying in the classification. We obtain the importance scores of every network topological properties calculated by CForest and rank the network topological properties by their importance scores. Fig. 3 shows the eight features ranked according their importance scores generated by CForest classifier. The rank of a feature indicates its importance in discriminating disease genes from non-disease genes.

3.4. *Comparing weighted and unweighted networks*

To demonstrate the superiority of weighted network, we compare the precision, recall, accuracy and AUC of weighted and unweighted networks by using CForest classifier. Table 4 shows the results of this comparison, which indicates that weighted network-based method is significantly better than unweighted network.

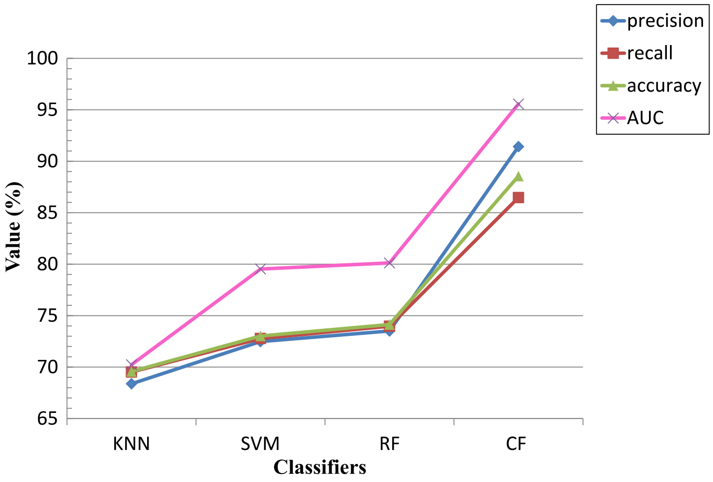

Fig. 2. Prediction performance comparison of different classifiers.

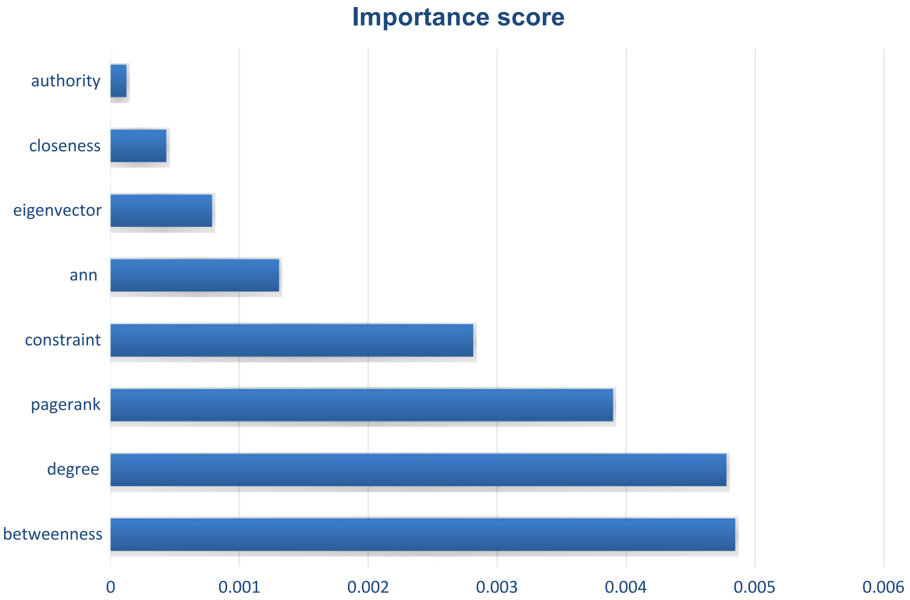

Fig. 3. Importance rank of topological properties.

Table 4
Performance comparison of weighted and unweighted networks.

Network	Precession (%)	Recall (%)	acc(%)	AUC
weighted	91.42	86.47	88.56	95.56
unweighted	85.29	82.63	83.57	91.88

4. Conclusion

This paper was motivated by the need for discovering disease-associated genes based on the topological properties of PPI network with the weight of interactions taken into consideration. We downloaded protein–protein interactions with confidence scores from HIPPE and gene–disease relations from OMIM to constructed a weighted PPI network for the discovery of disease-associated genes. Topological properties including degree, average nearest neighbor degree, authority centrality, betweenness centrality, closeness centrality, eigenvector centrality, Burt's constraint and PageRank were statistically analyzed to evaluate the topological importance of a gene in the PPI network from various perspectives. We then comparatively analyzed these topological properties by different groups and found that there is a significant discrimination between disease-related genes and non-disease genes. In contrast to non-disease genes, disease genes tend to get higher scores score in degree, average nearest neighbor degree, betweenness centrality, and PageRank, but lower in authority centrality, eigenvector centrality, closeness centrality and constraint.

We combined these topological properties in tandem to serve as an input of classifier. We attempted four different classifiers including KNN, SVM, Random Forest and CForest, to complement the discrimination task. We used grid-search and 10-fold cross validation to find the optimal parameters for every classification model, trying to make each of them achieve its best performance. Finally, the improved random forest classifier model called CForest outperforms the other three models and reaches the best prediction accuracy. With the unbiased variable importance measure supplied in CForest model, we ranked the importance of network topological properties, which is supposed to offer us a deeper understanding on the sophisticated genotype–phenotype associations. In addition, we presented a comparison of the prediction accuracy between weighted and unweighted networks to validate the superior of weighted network adoption in discovering disease-associated genes. It is obvious that our work contributes an improved method for disease gene discovery and a deeper understanding of how network topological properties affects gene–disease relations.

Acknowledgments

This work was funded by grants from Natural Science Foundation of Shaanxi Province of China (No. 2016JQ6072), National Natural Science Foundation of China (No. 71501153), Xi'an Science and Technology Program (No. XDKD001), and China Scholarship Council (201506965039, 201606965057).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.physa.2017.12.080.

References

[1] R. Albert, A.L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (1) (2002) 47.
[2] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.
[3] B. Kang, K. Goh, D. Lee, D. Kim, Complex networks: Structure and dynamics, Sae Mulli 48 (2) (2004) 115–141.
[4] A.L. Barabási, N. Gulbahce, J. Loscalzo, Network medicine: A network-based approach to human disease, Nature Rev. Genet. 12 (1) (2011) 56–68.
[5] M. Oti, H.G. Brunner, The modular nature of genetic diseases, Clin. Genet. 71 (1) (2007) 1–11.
[6] K.I. Goh, M.E. Cusick, D. Valle, B. Childs, M. Vidal, A.L. Barabási, The human disease network, Proc. Natl. Acad. Sci. 104 (21) (2007) 8685–8690.
[7] I. Feldman, A. Rzhetsky, D. Vitkup, Network properties of genes harboring inherited disease mutations, Proc. Natl. Acad. Sci. 105 (11) (2008) 4323–4328.
[8] J. Xu, Y. Li, Discovering disease-genes by topological features in human protein–protein interaction network, Bioinformatics 22 (22) (2006) 2800–2805.
[9] J. Chen, B.J. Aronow, A.G. Jegga, Disease candidate gene identification and prioritization using protein interaction networks, BMC Bioinform. 10 (1) (2009) 73.
[10] S.Y. Wu, F.J. Shao, R.C. Sun, Y. Sui, Y. Wang, J.L. Wang, Analysis of human genes with protein–protein interaction network for detecting disease genes, Physica A 398 (2014) 217–228.
[11] Y. Li, J.C. Patra, Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network, Bioinformatics 26 (9) (2010) 1219–1224.
[12] Y. Chen, T. Jiang, R. Jiang, Uncover disease genes by maximizing information flow in the phenome–interactome network, Bioinformatics 27 (13) (2011) i167–i176.
[13] O. Vanunu, O. Magger, E. Ruppin, T. Shlomi, R. Sharan, Associating genes and protein complexes with disease via network propagation, PLoS Comput. Biol. 6 (1) (2010) e1000641.
[14] S. Zickenrott, V. Angarica, B. Upadhyaya, A. Del Sol, Prediction of disease–gene–drug relationships following a differential network analysis, Cell death dis. 7 (1) (2016) e2040.
[15] J.C. Chen, M.J. Alvarez, F. Talos, H. Dhruv, G.E. Rieckhof, A. Iyer, K.L. Diefes, K. Aldape, M. Berens, M.M. Shen, et al., Identification of causal genetic drivers of human disease through systems-level analysis of regulatory networks, Cell 159 (2) (2014) 402–414.
[16] E. Khurana, Y. Fu, J. Chen, M. Gerstein, Interpretation of genomic variants using a unified biological network approach, PLoS Comput. Biol. 9 (3) (2013) e1002886.
[17] S. van Dam, U. Võsa, A. van der Graaf, L. Franke, J.P. de Magalhães, Gene co-expression analysis for functional classification and gene–disease predictions, Brief. Bioinform. (2017) bbw139.
[18] M. Ernst, Y. Du, G. Warsow, M. Hamed, N. Endlich, K. Endlich, H.M. Escobar, L.M. Sklarz, S. Sender, C. Junghanß, et al., Focusheuristics–expression-data-driven network optimization and disease gene prediction, Sci. Rep. 7 (2017) 42638.
[19] P. Maji, E. Shah, S. Paul, Relsim: An integrated method to identify disease genes using gene expression profiles and ppin based similarity measure, Inform. Sci. 384 (2017) 110–125.
[20] C.J. Mungall, N.L. Washington, J. Nguyen-Xuan, C. Condit, D. Smedley, S. Köhler, T. Groza, K. Shefchek, H. Hochheiser, P.N. Robinson, et al., Use of model organism and disease databases to support matchmaking for human disease gene discovery, Hum. Mutat. 36 (10) (2015) 979–984.
[21] J. Li, X. Lin, Y. Teng, S. Qi, D. Xiao, J. Zhang, Y. Kang, A comprehensive evaluation of disease phenotype networks for gene prioritization, PLoS One 11 (7) (2016) e0159457.
[22] U.M. Singh-Blom, N. Natarajan, A. Tewari, J.O. Woods, I.S. Dhillon, E.M. Marcotte, Prediction and validation of gene-disease associations using methods inspired by social network analyses, PLoS One 8 (5) (2013) e58977.
[23] S. Navlakha, C. Kingsford, The power of protein interaction networks for associating genes with diseases, Bioinformatics 26 (8) (2010) 1057–1063.
[24] H.G. Brunner, M.A. Van Driel, From syndrome families to functional genomics, Nature Rev. Genet. 5 (7) (2004) 545–551.
[25] G. Alanis-Lobato, M.A. Andrade-Navarro, M.H. Schaefer, Hippie v2.0: Enhancing meaningfulness and reliability of protein–protein interaction networks, Nucleic Acids Res. 45 (D1) (2017) D408–D414.
[26] A. Chatr-Aryamontri, B.J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'donnell, et al., The biogrid interaction database: 2015 update, Nucleic Acids Res. 43 (D1) (2014) D470–D478.

[27] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, et al., The intact molecular interaction database in 2012, Nucleic Acids Res. 40 (D1) (2011) D841–D846.

[28] L. Licata, L. Briganti, D. Peluso, L. Perfetto, M. Iannuccelli, E. Galeota, F. Sacco, A. Palma, A.P. Nardozza, E. Santonico, et al., Mint, the molecular interaction database: 2012 update, Nucleic Acids Res. 40 (D1) (2011) D857–D861.

[29] T. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, et al., Human protein reference database–2009 update, Nucleic Acids Res. 37 (suppl_1) (2008) D767–D772.

[30] L. Salwinski, C.S. Miller, A.J. Smith, F.K. Pettit, J.U. Bowie, D. Eisenberg, The database of interacting proteins: 2004 update, Nucleic Acids Res. 32 (suppl_1) (2004) D449–D451.

[31] R. Isserlin, R.A. El-Badrawi, G.D. Bader, The biomolecular interaction network database in psi-mi 2.5, Database 2011.

[32] P. Pagel, S. Kovac, M. Oesterheld, B. Brauner, I. Dunger-Kaltenbach, G. Frishman, C. Montrone, P. Mark, V. Stümpflen, H.W. Mewes, et al., The mips mammalian protein–protein interaction database, Bioinformatics 21 (6) (2004) 832–834.

[33] V.A. McKusick, Mendelian inheritance in man and its online version, omim, Am. J. Hum. Genet. 80 (4) (2007) 588–604.

[34] X. Wang, N. Gulbahce, H. Yu, Network-based methods for human disease gene prediction, Brief. Funct. Genom. 10 (5) (2011) 280–293.

[35] R. Diestel, Graph Theory, Springer, 2000.

[36] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, The architecture of complex weighted networks, Proc. Natl. Acad. Sci. USA 101 (11) (2004) 3747–3752.

[37] J.M. Kleinberg, Authoritative sources in a hyperlinked environment, J. ACM 46 (5) (1999) 604–632.

[38] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2) (2001) 163–177.

[39] L.C. Freeman, S.P. Borgatti, D.R. White, Centrality in valued graphs: A measure of betweenness based on network flow, Social Networks 13 (2) (1991) 141–154.

[40] P. Bonacich, Power and centrality: A family of measures, Am. J. Sociol. 92 (5) (1987) 1170–1182.

[41] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1) (1998) 107–117.

[42] R.S. Burt, Structural holes and good ideas, Am. J. Sociol. 110 (2) (2004) 349–399.

[43] L.E. Peterson, K-nearest neighbor, Scholarpedia 4 (2) (2009) 1883.

[44] V. Vapnik, The Nature of Statistical Learning Theory, Springer science & business media, 2013.

[45] Y. Cui, J. Han, D. Zhong, R. Liu, A novel computational method for the identification of plant alternative splice sites, Biochem. Biophys. Res. Commun. 431 (2) (2013) 221–224.

[46] L. Breiman, Random forests, Mach. Learn. 45 (1) (2001) 5–32.

[47] Y. Cui, C. Zhang, M. Cai, Prediction and feature analysis of intron retention events in plant genome, Comput. Biol. Chem. 68 (2017) 219–223.

[48] T. Hothorn, K. Hornik, A. Zeileis, Unbiased recursive partitioning: A conditional inference framework, J. Comput. Graph. stat. 15 (3) (2006) 651–674.

[49] C. Strobl, A.L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis, Conditional variable importance for random forests, BMC Bioinform. 9 (1) (2008) 307.