BMC Bioinformatics

Research article

Disease candidate gene identification and prioritization using protein interaction networks

Jing Chen[1,2], Bruce J Aronow[1,2,3] and Anil G Jegga*[1,3]

Address: [1]Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA, [2]Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH, USA and [3]Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA

Email: Jing Chen - Jing.Chen@cchmc.org; Bruce J Aronow - Bruce.Aronow@cchmc.org; Anil G Jegga* - Anil.Jegga@cchmc.org

* Corresponding author

Published: 27 February 2009

BMC Bioinformatics 2009, **10**:73 doi:10.1186/1471-2105-10-73

Received: 2 June 2008
Accepted: 27 February 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/73

Abstract

Background: Although most of the current disease candidate gene identification and prioritization methods depend on functional annotations, the coverage of the gene functional annotations is a limiting factor. In the current study, we describe a candidate gene prioritization method that is entirely based on protein-protein interaction network (PPIN) analyses.

Results: For the first time, extended versions of the PageRank and HITS algorithms, and the K-Step Markov method are applied to prioritize disease candidate genes in a training-test schema. Using a list of known disease-related genes from our earlier study as a training set ("seeds"), and the rest of the known genes as a test list, we perform large-scale cross validation to rank the candidate genes and also evaluate and compare the performance of our approach. Under appropriate settings – for example, a back probability of 0.3 for PageRank with Priors and HITS with Priors, and step size 6 for K-Step Markov method – the three methods achieved a comparable AUC value, suggesting a similar performance.

Conclusion: Even though network-based methods are generally not as effective as integrated functional annotation-based methods for disease candidate gene prioritization, in a one-to-one comparison, PPIN-based candidate gene prioritization performs better than all other gene features or annotations. Additionally, we demonstrate that methods used for studying both social and Web networks can be successfully used for disease candidate gene prioritization.

Background

Most of the current disease candidate gene prioritization methods [1-6] rely on functional annotations. However, the coverage of the gene functional annotations is a limiting factor. Although more than 1,500 human disease genes have been documented, most of them continue to be functionally uncharacterized. Currently, only a fraction of the genome is annotated with pathways and phenotypes [6]. While two thirds of all the genes are annotated by at least one functional annotation, the remaining one third is yet to be annotated.

Analysis of protein-protein interaction networks (PPINs) is important for inferring the function of uncharacterized proteins. Protein-protein interactions refer to the association among the protein molecules and the study of these associations from the perspective of biochemistry, signal transduction and biomolecular networks. Recent biotech-

nological advances like the high-throughput yeast two-hybrid screen facilitated building proteome-wide PPINs or "interactome" maps in humans [7,8]. The shift in focus to systems biology in the post-genomic era has generated further interest in PPINs and biological pathways. Network-based analyses have been developed with a number of goals [9], including protein function prediction [10], identification of functional modules [11], interaction prediction [12,13], identification of disease candidate genes [14,15] and drug targets [16,17], and the study of network structure and evolution [18-22]. While there is a wealth of protein-disease relationships in the published literature and a number of PPIN resources, relatively few studies have actually used PPIN analyses for prioritizing disease genes. Thus, making use of these networks in the context of disease is a relatively new challenge [14]. One of the earliest efforts [23] uses a classifier based on several topological features, including degree (the number of links to the protein), 1N index (proportion of links to disease-related proteins), 2N index (average 1N index in the neighbors), average distance to disease genes, and positive topology coefficient (average neighborhood overlapping with disease genes). Xu et al., built a KNN-based classifier with all disease genes from OMIM and concluded that hereditary disease genes from OMIM in the literature-curated protein-protein interaction network are characterized by a larger degree, a tendency to interact with other disease genes, more common neighbors, and quick communication to each other [23]. A more recent application, Genes2Networks [24], identifies important genes based on a list of "seed" genes. It generates a Z-score for each "intermediate" gene from a binomial proportions test to represent its specificity or significance to the "seed" genes. The former method, independent of known disease-related genes, is used for disease candidate gene identification, especially in cases where there is little or no prior knowledge about the disease. The latter application, on the other hand, uses a "seed" list as training to score the neighboring genes. It avoids bias toward highly connected "hub" genes, but the candidate gene is searched in a local network region, and the user has to provide the size of the neighborhood region in the network.

Recent technological advances in genomic sequencing, gene expression analysis, and other massively parallel techniques, while opening new opportunities, continue to pose a formidable challenge in deriving meaningful information from the large data silos. Typically, such data can be represented as networks in which the nodes (e.g., genes, mRNA, microRNA, proteins or metabolites) are linked by edges (e.g., DNA-protein or protein-protein or miRNA-mRNA interactions or correlations). Structural analysis of these networks can lead to new insights into biological systems and is a helpful method for proposing new and testable hypotheses. Biological networks have in fact been found to be comparable to communication and social networks [25]. For instance, PPINs and communication networks share several common characters such as scale-freeness and small-world properties, suggesting that the algorithms used for social and Web networks are equally applicable to biological networks. Although PPIs have been used widely to identify novel disease candidate genes [14,15,23,26-35], besides Kohler et al. [30] and Wu et al. [34], there have been no reports on using PPIs for disease gene prioritization. Additionally, to the best of our knowledge, this is the first study that uses social- and Web- network analysis-based algorithms to prioritize disease candidate genes.

In the analysis of social networks, Web graphs and telecommunication networks, a common question frequently asked is: Which entities are most important in the network? Although visualization-centered approaches such as graph drawing are useful to gain qualitative intuition about the structure, especially in small graphs, it is not practical to use these approaches for large and more complex networks. As an alternative, a number of other approaches have therefore been developed. For instance, a variety of measures (degree centrality [36], closeness centrality [37] and betweenness centrality [38]) have been proposed by sociologists to determine the "centrality" of a node in a social network. Likewise, in the area of Web graphs, computer scientists have proposed and used several algorithms such as HITS [39] and PageRank [40] for automatically determining the "importance" of Web pages.

In the current study, for the first time, we utilize the above methods to prioritize disease candidate genes by estimating their relative importance in the PPIN to the disease-related genes. Specifically, we determine the optimal parameter values in the methods and record the corresponding performance. The first algorithm that we use is based on White and Smyth's PageRank algorithm. White and Smyth [41] proposed a general framework and a set of algorithms under the framework to measure the relative importance in networks. The first method is an extension of the original PageRank algorithm and is called PageRank with Priors. It mimics the random surfer model wherein a random Internet surfer starts from one of a set of root nodes, R, and follows one of the links randomly in each step. In this process, the surfer jumps back to the root nodes at probability β, thus restarting the whole process. Intuitively, the PageRank with Priors algorithm generates a score that is proportional to the probability of reaching any node in the Web surfing process. This score indicates or measures the relative "closeness" or importance to the root nodes. The second algorithm, named HITS with Priors, is an extension of HITS (Hyperlink-Induced Topic Search), which is a link analysis algorithm developed by

Jon Kleinberg to rate Web pages. It determines two values for a page: its authority, which estimates the value of the content of the page, and its hub value, which estimates the value of its links to other pages [42]. In the Web surfing model, the surfer still starts from one of the root nodes. In the odd steps he/she can either follow a random "out-link" or jump back to a root node, and in the even steps he/she can instead follow an "in-link" or jump back to a root node. Similar to the PageRank with Priors, HITS with Priors also estimates the relative probability of reaching a node in the network. The third algorithm we use is the K-Step Markov method. In a similar Web surfing scenario, this method mimics a surfer who starts with one of the root nodes. The surfer follows a random link in each step, but he/she return to the root node after K steps and restarts surfing.

Results

Human protein interaction network

The human protein-protein interactions were extracted from the NCBI Entrez Gene FTP site [43] and contained 8340 nodes or vertices (corresponding to 8340 unique genes/proteins) and 27250 edges (corresponding to 27250 unique interactions). This compilation is based on three interaction databases, namely, BIND (2389 genes and 4054 interactions) [44], BioGRID (7683 genes, 23205 interactions) [45], and HPRD (6594 genes and 22802 interactions) [46] (See Additional File 1 for the overlap among these three resources). Although these literature-based or literature-curated interactions are more subjective to research bias, they are less prone to errors. Analysis of this complete human protein interaction network using "NetworkAnalyzer" [47] in Cytoscape [48] revealed 120 connected components. The largest of these has 8075 genes. The remaining 265 genes are separated into 119 smaller components or sub networks of size two to five nodes or genes. Since majority of these smaller sub-networks contain only two genes, we reasoned that it might not be of interest to check the distribution of the disease genes among them.

Evaluation of PPIN for gene prioritization

We used the same training data, from our previous study [6], comprising 19 diseases with 693 associated genes. Of these, 589 genes were used in the cross validation because the remaining 104 genes do not have any known protein-protein interactions (see Additional File 2). The random training dataset, used as a control, was built with 19 random gene lists, with each list comprising 31–38 genes. We used three methods, namely, K-Step Markov (KSMarkov), PageRank with Priors (PRankP), and HITS with Priors (HITSP), to prioritize the disease gene with different parameter values. The random genes were prioritized using PRankP with back probability set to 0.3. The ROC curves of representative cross validation results are shown in the figures 1 and 2.

Based on our results, we observed that in terms of performance, HITSP was similar to PRankP under different back probability values. Therefore, only PRankP was tested for extreme back probability values such as 0.01 and 0.05. The 13 different test conditions (PRankP with back probability 0.01, 0.05, 0.1, 0.3, 0.5, 0.7, 0.9; KSMarkov with k = 1, 2, 4, 6; and HITSP with Priors with back probability 0.3 and 0.5) along with the AUC values from each validation run are listed in Table 1. Each of the methods, with the same parameter settings, was repeated 5 times. The performance values derived from each of the methods with respect to a particular parameter value are summarized in Table 2. The plots of AUC with different parameter values are shown in figure 3. The best performance of each method was selected, namely, PRankP and HITSP with back probability 0.3 and KSMarkov with K = 4, for Analysis of Variance (ANOVA). The p value of 0.5585 suggests that there is no significant difference among the best performance of the three methods.

Cardiac septal defect candidate gene prioritization

Mining the "clinical synopsis" and "allelic variant sections" of NCBI's OMIM (Online Mendelian Inheritance in Man) database [49], we extracted 166 OMIM records that had the terms "atrial septal defect" OR "ASD" OR "ventricular septal defect" OR "VSD" occurring in the text. There were 81 genes mapping to these records (see Additional File 3 for a list of OMIM records and the corresponding genes associated with cardiac septal defects). These 81 genes were used as the training set. Mining the human protein interactome [43] (see Methods) we extracted the 479 immediate interactants (level 1) of these training 81 genes (Additional File 3). We then sought to rank or prioritize these genes using both integrative functional annotation-based methods and PPIN-based methods using our ToppGene server [6]. There was an overlap of 48 genes which were removed leaving 431 genes for ranking. We call this as the test set for cardiac septal defect.

Among the top 20 ranked genes (Table 3), 4 genes (*SRF*, *SMAD1*, *SMAD2*, *SMAD3*) were common to all the methods (Figure 4). Analyzing the results we observed that the performance of both the approaches (functional annotation- versus PPIN- based methods) was comparable. For instance, among the top 20 ranked genes using functional annotations, 15 genes were reported to be previously associated with cardiac development or malformation (indicated with an asterisk in the Table 3). Six (*INSR*, *ERBB2*, *NOTCH2*, *BMPR2*, *TGFBR2* and *SRF)* of these top 20 have been previously reported to be associated with cardiac septal defects. In case of PPIN-based methods, there were 14 genes previously associated with either cardiac development or abnormalities. Of these, 3 genes (*SRF*, *EP300*, *and CREBBP*) have been associated with cardiac septal defect. The genes *EP300* and *CREBBP* have been ranked 11/431 and 15/431 using PPIN-based meth-

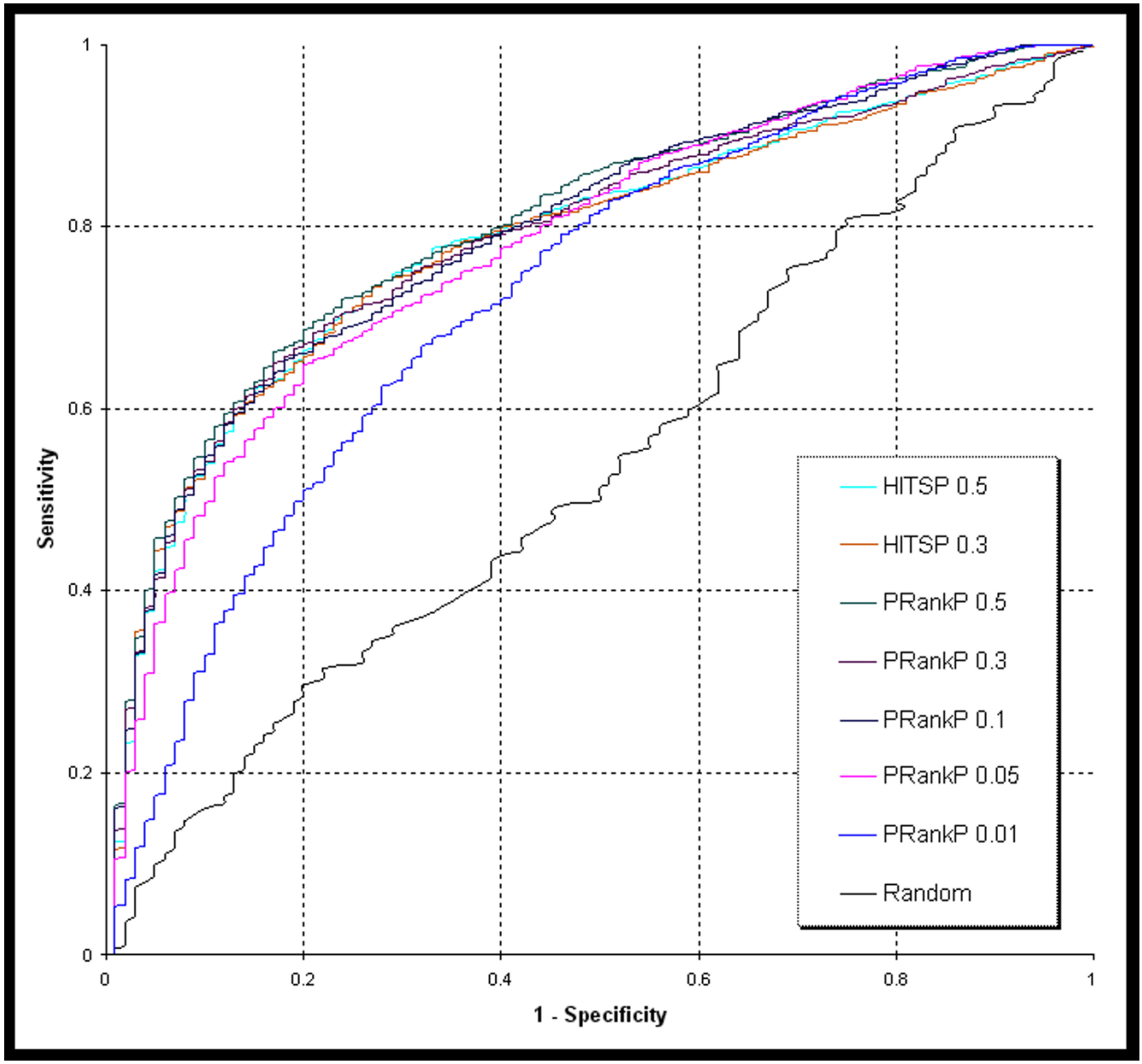

Figure 1
ROC curves from cross validations. This figure shows the representative ROC curves using PageRank with Priors with back probability 0.01, 0.05, 0.1, 0.3 and 0.5, and HITS with Priors with back probability 0.3 and 0.5. The random curve was derived from prioritization of the random training set using the PageRank with Prior method with back probability 0.3.

ods while the rankings were 41 and 40 respectively using ToppGene. Interestingly, truncated CBP protein (gene *CREBBP*) leads to classical Rubinstein-Taybi syndrome phenotypes in mice characterized by atrial and ventricular septal defects [50]. Likewise, mouse embryos lacking p300 protein (gene *EP300*) show ventricular septal defects [51]. The higher ranking of *EP300* and *CREBBP* in PPIN-based method is because of their direct interactions with training set gene (*CITED2*). Previous studies have reported that the paralogous genes *EP300* and *CREBBP* co-activate *TFAP2A* in the presence of *CITED2* [52]. Similarly, *MYL7* is ranked first in PPIN-based prioritization while it is ranked 122 in functional annotation-based prioritization methods. The higher ranking in the former is because *MYL7* has only one known interactant (*MYH6*), mutation of which is associated with cardiac septal defects [53]. Another noteworthy example is *BMP2*, ranked 6/431 by PPIN-based method while the ToppGene rank was 32/

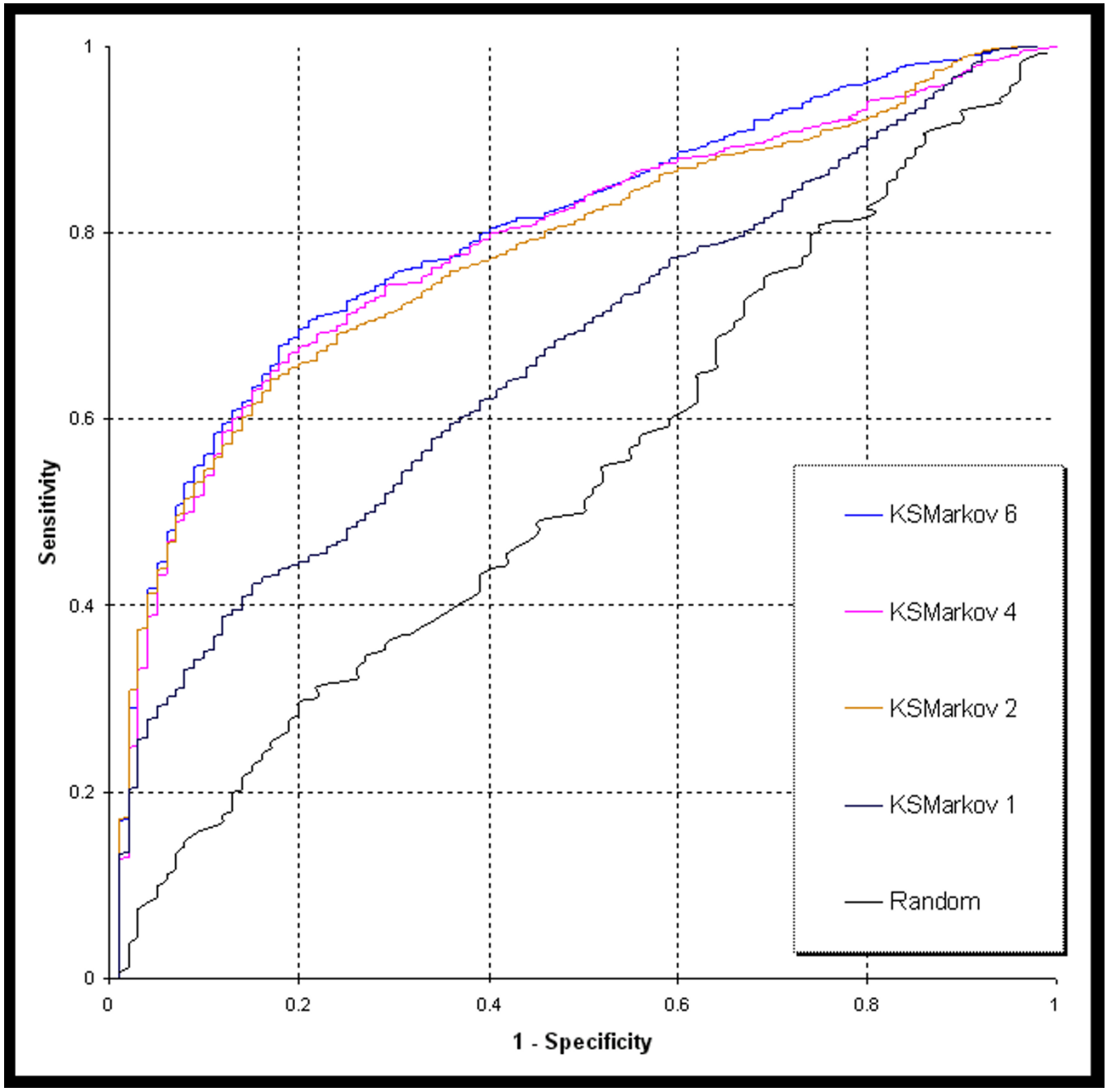

Figure 2
ROC curves from cross validations. This figure shows the representative ROC curves using the K-Step Markov method with K = 1, 2, 4, and 6. The random curve was derived from prioritization of the random training set using the PageRank with Prior method with back probability 0.3.

431. On the other hand there were examples of potential candidate genes which the PPIN-based prioritization methods ranked low while ToppGene ranked them higher. For instance, *ERBB2* was ranked 112/431 by PPIN-based method while functional annotation-based gene prioritization (ToppGene) ranked it as eight. Mice with a ventricular-restricted deletion of *Erbb2* show ventricular septal defect (VSD) [54,55] suggesting that the human ortholog *ERBB2* could be a potential candidate gene for VSD. Thus, while integrative functional annotation-based methods are superior in prioritizing disease candidate genes, PPIN-based methods certainly have their own

Table 1: AUC values from each cross validation run.

Test Type	Test ID	AUC	Test ID	AUC	Test ID	AUC	Test ID	AUC	Test ID	AUC
k1	1	0.66	2	0.66	3	0.87	4	0.66	5	0.66
k2	6	0.78	7	0.78	8	0.78	9	0.78	10	0.78
k4	11	0.8	12	0.8	13	0.8	14	0.8	15	0.8
k6	16	0.8	17	0.8	18	0.8	19	0.8	20	0.8
h3	21	0.8	22	0.8	23	0.8	24	0.8	25	0.8
h5	26	0.8	27	0.8	28	0.8	29	0.8	30	0.8
p01	36	0.73	37	0.73	38	0.73	39	0.73	40	0.73
p05	31	0.78	32	0.78	33	0.78	34	0.78	35	0.78
p1	41	0.79	42	0.79	43	0.79	44	0.79	45	0.79
p3	46	0.8	47	0.8	48	0.8	49	0.8	50	0.8
p5	51	0.8	52	0.8	53	0.8	54	0.8	55	0.8
p7	56	0.8	57	0.8	58	0.8	59	0.8	60	0.8
p9	61	0.79	62	0.8	63	0.79	64	0.8	65	0.8

Column "Test Type" indicates the method and parameter settings of the test. p01 through p9 stand for PageRank with Priors with back probability 0.01 to 0.9, respectively; k1, k2, k4 and k6 represent K-Step Markov with K = 1, 2, 4 and 6, accordingly; h3 and h5 are HITS with Priors with back probability 0.3 and 0.5, respectively. There were 13 test conditions, each repeated 5 times.

advantages. We, therefore, hypothesize that a combined functional annotations- and PPIN- based methods are more effective in identifying and ranking of disease candidate genes. The rankings of all the test (431) genes using different methods (PPIN-based, ToppGene and ENDEAVOUR) are included in the Additional Files 4 and 5. Further, given the continued incomplete annotation coverage of human genes (see Table 4 for a summary of functional annotation coverage of human interactome genes and Additional File 6 for a gene-wise breakdown of all annotations and protein interactions), PPIN-based prioritization is a viable option.

Discussion and Conclusion

Our current study, based on the observation that biological networks share many properties with Web and social networks, is an attempt to extend the successful graph analysis-based algorithms from computer science research to tackle the disease gene prioritization problem. Literature-based and manually curated protein interactions were used to form the base network, and extended versions of the PageRank algorithm and HITS algorithm, as well as the K-Step Markov method, were applied to prioritize disease candidate genes in a training-test schema. For each prioritization, a list of known disease-related genes was used as a training set ("seeds"), and the genes in the test list (candidate genes) were ranked. To evaluate and compare the performance of the methods, a large-scale cross validation was performed. A total of 13 conditions with three algorithms and different parameter settings were tested, each repeated five times. Rank-based ROC curves were plotted, and AUC values were used to quantitatively measure the performance.

Based on our results, we draw the following conclusions: First, under appropriate settings, for example, a back probability of 0.3 for PageRank with Priors and HITS with Priors, and walk length 6 for K-Step Markov method, the three methods achieved the same AUC value and hence similar performance. This suggests that based on the current knowledge of protein-protein interaction networks, even other similar or related methods (e.g., ranking of nodes in an unweighted graph) under the same framework might yield similar results.

Second, the value of back probability in PageRank with Priors and HITS with Priors can be of broad range (e.g., 0.1 to 0.9) and still result in relatively stable performance.

Table 2: Mean and standard deviation (SD) of AUC values with 13 different cross validation conditions.

Method	Parameter	Mean of AUC	SD of AUC
PageRank with Priors	Back probability = 0.01	0.73	0.0008
PageRank with Priors	Back probability = 0.05	0.78	0.0015
PageRank with Priors	Back probability = 0.1	0.8	0.0013
PageRank with Priors	**Back probability = 0.3**	**0.8**	**0.0011**
PageRank with Priors	Back probability = 0.5	0.8	0.0015
PageRank with Priors	Back probability = 0.7	0.8	0.0009
PageRank with Priors	Back probability = 0.9	0.79	0.0007
K-Step Markov	K = 1	0.7	0.096
K-Step Markov	K = 2	0.78	0.0005
K-Step Markov	K = 4	0.8	0.0024
K-Step Markov	**K = 6**	**0.8**	**0.0009**
HITS with Priors	**Back probability = 0.3**	**0.8**	**0.0009**
HITS with Priors	Back probability = 0.5	0.8	0.0004

Highlighted rows correspond to the best parameter value of each method.

However, when the back probability was set to very low (e.g., 0.01), the performance dropped significantly. This is expected because in both the methods (see equations 3 and 4 under Methods), as the back probability reaches 0, the bias toward the "seeds" is eliminated. PageRank/HITS with Priors are same as the original PageRank/HITS algorithm; therefore, the prioritization toward the selected "seeds" fails. The performance of the K-Step Markov method, on the other hand, decreased significantly when the length of random walk K was small (e.g. K = 1). Under this condition, the K-Step Markov method calculates the probability to spend time on each protein from the seeds with a random walk of length 1. The proteins that are not directly interacting with "seeds" will therefore never be reached and scored 0. This suggests that if a true disease candidate gene is not directly interacting with the "seeds", it will be ignored when K is 1. The method converged to the best performance when K was 4. Any further increase in the random walk length did not improve the performance. This can be attributed to the fact that the average shortest path length in the PPIN was only about 4.5.

Third, the overall performance of candidate gene prioritizations based exclusively on protein networks is comparable to functional annotation-based methods [6] since they were all tested using the same cross validation. The AUC value of functional annotation based method, ToppGene [6], was 0.916, and the best AUC value of network-based methods (from the current study) was 0.801. This shows that network-based methods are generally not as effective as the integrated functional annotation-based methods for disease candidate gene prioritization. For a more accurate comparison, we compared PPIN-based methods to the individual functional annotation features used in our previous study [6]. Surprisingly, we found that network-based methods are better than all annotations (see [6] for details). We therefore conclude that PPINs can be a potentially good feature for disease candidate gene prioritization irrespective of whether the genes have other functional annotations or not. Based on our findings that in one-to-one comparison PPIN-based candidate gene prioritization performed better than all other gene features or annotations, we hypothesize that PPINs can be a potentially good feature for disease candidate gene prioritization, especially when the genes lack all other functional annotations or are sparsely annotated

Network-based prioritization methods, however, have certain limitations. Just like functional annotation-based methods, the performance depends on the quality of interaction data. It is an acknowledged fact that the current human protein interactome suffers with incomplete-

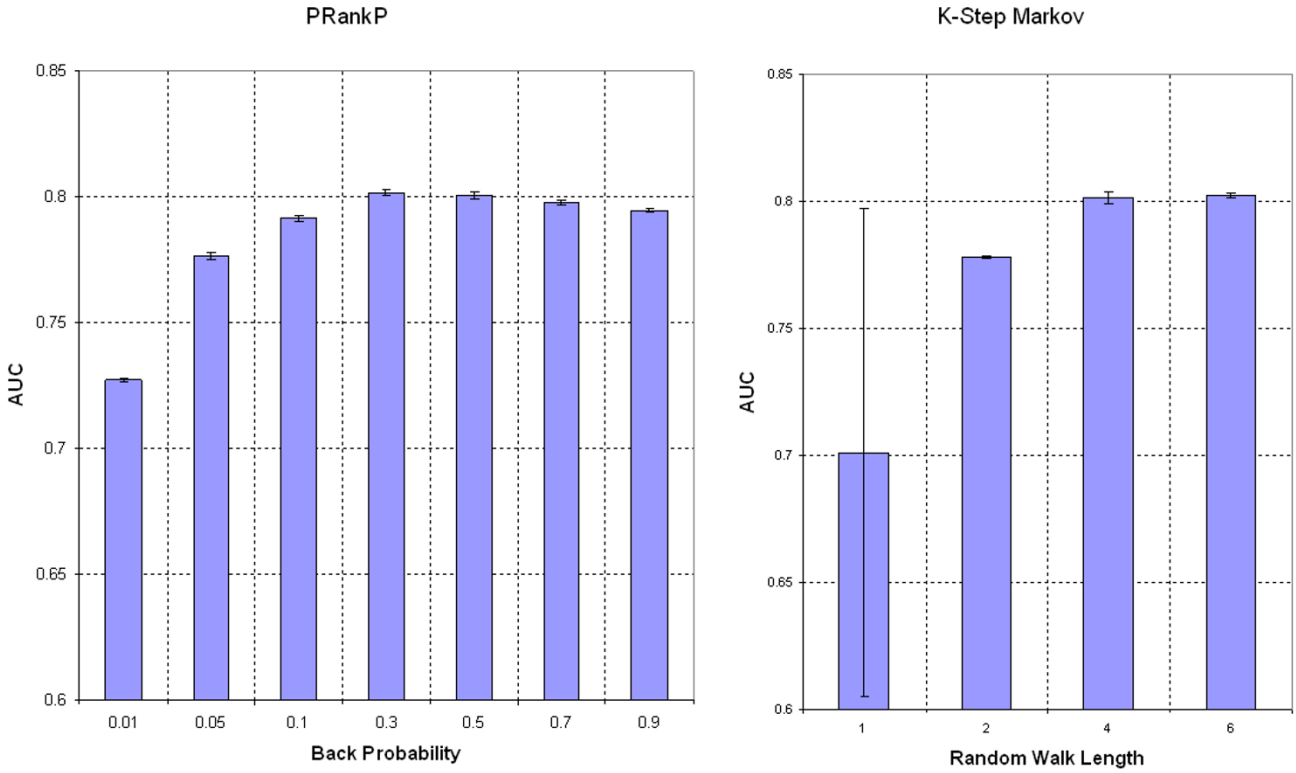

Figure 3
Plots of AUC with different parameter values. The left panel shows the AUC values of PageRank with Priors with back probability varied from 0.01 to 0.9. The right panel shows the AUC values of the K-Step Markov method with random walk length varied from 1 to 6. The vertical bars indicate the standard deviations.

ness and unreliability with missing interactions and false positives. To make reliable candidate gene prioritization – based either on functional annotations or PPINs – we must have reasonably complete datasets that accurately represent the interactions and annotations in the genome and proteome. However, as the quality of these annotations and interactions improves the confidence in candidate gene prioritization approaches based on them will also improve. Certainly, our approach can be improved methodology-wise in the following directions. First, the algorithms used in our current study were originally developed to identify "important" nodes in networks. Although we used extended versions of these algorithms to prioritize nodes to selected "seeds," there still could be a bias toward hubs. Additionally, since these approaches were designed for Web and general networks, there is definitely scope for additional modifications to make them fit better with biological networks (e.g., using weighted nodes (genes or proteins) or edges (interactions)). As future extension, apart from considering weighted nodes and edges, we plan to integrate our method with other methods (e.g., combining results from functional annotation-based methods and expression profiles with network-based approaches). It is expected that using both functional annotations and PPIN-based topological parameters may better facilitate the discovery and prioritization of disease genes.

Methods

Human protein interaction datasets

The human protein interaction dataset (file "interactions.gz"), a compilation of PPIs from BIND [44], BioGRID [45], and HPRD [46], was downloaded from NCBI Entrez Gene FTP site [43]. All of these interactions are derived from large-scale experiments and curated manually. For example, all interactions in BIND are experimentally validated and published in at least one peer-reviewed journal; interactions in BioGRID are entirely derived from manual literature curation, just as in HPRD.

Prioritization methods

In the current study, the protein interaction network is represented as an unweighted, undirected simple graph, G, where proteins (genes) are nodes and interactions are edges. The set of all the proteins in the network is denoted as V and all the interactions as E. The set of known disease

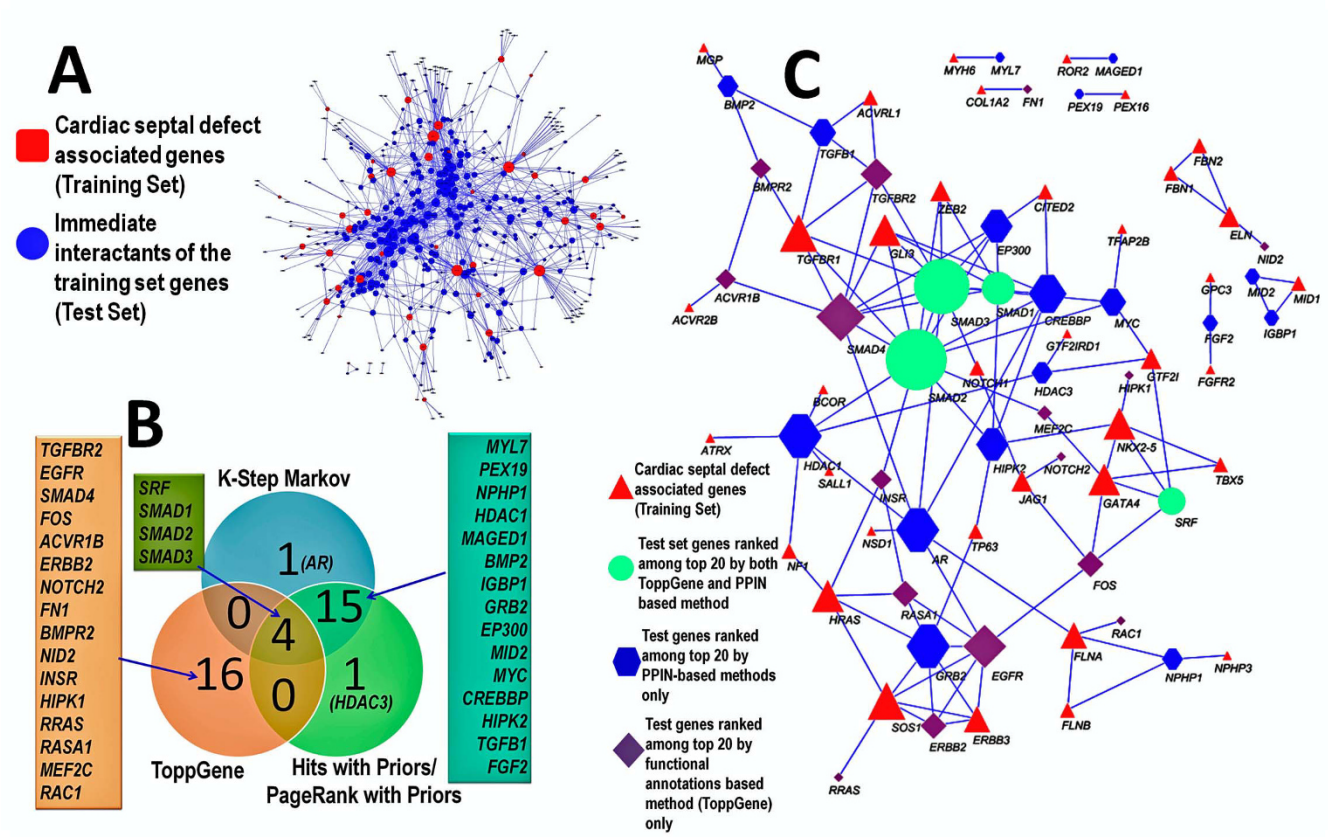

Figure 4

Prioritized candidate genes of cardiac septal defects using both functional annotation- and PPIN- based methods. Panel A shows the sub-network of heart septal defect related genes comprising (i) genes associated with OMIM diseases that have the phenotype of cardiac septal defect (Training set of genes for cardiac septal defect) and their immediate interactants (Test set genes). The size of the nodes is proportional to the degree (number of edges). Panel B shows the intersection among the top 20 ranked cardiac septal defect candidate genes using functional annotation- and PPIN- based methods. Functional annotation-based prioritization was done using ToppGene server. For PPIN-based methods K-Step Markov, Hits with Priors, and PageRank with Priors was used. Panel C shows the top 20 ranked cardiac septal defect genes (generated using PPIN- and functional annotation- based methods) along with their connectivity to training set genes (based on protein-protein interactions).

genes (also called the seeds) is denoted as R. The prioritization approaches are based on the methods of White and Smyth [41], whose general framework, consisting of four successive problem formulations, each building on the next, defines the approach to ranking nodes in an unweighted digraph G(V, E):

1. Relative importance of a node t with respect to a root node r: Given G and r and t, where r and t are both nodes in G and r is the root, compute the "importance" of t with respect to r. This importance is denoted as I(t|r), a non-negative quantity.

2. Rank of importance of a set of nodes T with respect to a root node r: Given G and a root node r in G, rank all vertices in T, a subset of vertices in G. For each node t in T, the value of I(t|r) can be computed. Then the nodes can be ranked so that the largest values correspond to the highest importance.

3. Rank of importance of a set of nodes T with respect to a set of root nodes R: Given G and a set of root node R in G, rank all vertices in T, a subset of vertices in G. The importance of node t to R is defined as the average sum of importance of t to each node in R:

$$I(t|R) = (1/|R|)(\text{sum}(I(t|r)). \quad (1)$$

4. Given G, rank all nodes: This is a special case where R = T = V.

Based on White and Smyth's framework, the solution to problem 3 is what is needed in this study. To recap it in the context of disease gene prioritization, the problem is

Table 3: Cardiac septal defect candidate gene prioritization.

Rank	Integrative functional annotation based ranking (using ToppGene)	PPIN-based ranking (K-Step Markov)	PPIN-based ranking (Hits with Priors)	PPIN-based ranking (PageRank with Priors)
1	*TGFBR2**#	*MYL7**	*MYL7**	*MYL7**
2	*EGFR**	*PEX19*	*PEX19*	*PEX19*
3	*SMAD4**	*NPHP1*	*NPHP1*	*NPHP1*
4	***SRF****#	*HDAC1**	*MAGED1*	*MAGED1*
5	*FOS*	*MAGED1*	*BMP2**	*BMP2**
6	*ACVR1B*	*BMP2**	*HDAC1**	*HDAC1**
7	***SMAD2****	***SRF****#	*IGBP1*	*IGBP1*
8	*ERBB2**#	*IGBP1*	*MID2*	*MID2*
9	*NOTCH2**#	*GRB2**	***SMAD3****	***SMAD3****
10	*FN1**	***SMAD3****	***SRF****#	***SRF****#
11	*BMPR2**#	*EP300**#	*GRB2**	*GRB2**
12	*NID2*	*MID2*	*MYC**	*MYC**
13	***SMAD1****	*MYC**	*HIPK2*	*HIPK2*
14	***SMAD3****	***SMAD2****	*FGF2**	*FGF2**
15	*INSR**#	*CREBBP**#	***SMAD2****	***SMAD2****
16	*HIPK1*	***SMAD1****	***SMAD1****	***SMAD1****
17	*RRAS*	*HIPK2*	*HDAC3*	*HDAC3*
18	*RASA1**	*TGFB1**	*EP300**#	*EP300**#
19	*MEF2C**	*FGF2**	*CREBBP**#	*CREBBP**#
20	*RAC1**	*AR**	*TGFB1**	*TGFB1**

The cardiac septal defect sub-network was created using known cardiac septal defect genes (from OMIM) and their immediate interactants, and was prioritized using functional annotation and PPIN based methods. Functional annotation based prioritization was done using ToppGene server. The PPIN based rankings were obtained using 3 methods: K step Markov, Hits with Priors, and PageRank with Priors. The highlighted genes are those occurring in all of the prioritized top 20 genes generated using different methodologies. Note that the Hits with Priors and PageRank with Priors gave identical results (see Additional Files 3, 4 and 5 for the list of genes and prioritization results). The genes marked with * are associated with abnormal heart morphology (ToppGene: 15/20; K-Step Markov: 14/20; and Hits with Priors and PageRank with Priors: 13/20) while those marked with # have been reported to be associated with cardiac septal defects (6/20 and 3/20 in ToppGene and PPIN prioritized top 20 candidate genes for cardiac septal defects).

to prioritize a set of genes in the network based on their importance to a set of root genes (e.g., genes known to be associated with a disease). The importance of a gene to the set of root genes is just the average sum of the importance of it to each individual root gene. Although this framework was proposed for directed networks, it can also be applied to the undirected networks because the latter is just a special case of the former. In this study, the undirected protein interaction network was converted to an equivalent directed network, when necessary.

With the problem formulation defined, the key of the solution is to find I(t|r), the importance of node t with respect to a root node r. For this, we use three algorithms

Table 4: Summary of functional annotation coverage of human interactome genes.

Interactome genes with 3 annotations (2440)	
GO + MP + Pathways	2440
Interactome genes with any 2 annotations (2866)	
GO + Pathways	1630
GO+MP	1232
MP + Pathways	4
Interactome genes with only 1 annotation (2505)	
GO only	2448
MP only	10
Pathways only	47
Interactome genes with no annotations (223)	223
Total Interactome genes	**8034**

About 1/3rd (2505/8034) genes of the interactome are sparsely annotated (GO – Gene Ontology; MP – Mammalian Phenotype; and Pathway annotations).

from White and Smyth [41]: a) PageRank with Priors, b) HITS with Priors, and c) K-step Markov.

The PageRank with Priors method is an extension of the original PageRank algorithm. The iterative stationary probability equation is:

$$\pi(v)^{(i+1)} = (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} p(v \mid u)\pi^{(i)}(u)\right) + \beta p_v \quad (2)$$

In this equation, p_v represents the "prior bias". $p_v = 1/|R|$ for v in R, the root node set; $p_v = 0$ otherwise. β, empirically defined on [0, 1], represents a "back probability." $d_{in}(v)$ is the in-degree of v. p(v|u) is the probability of arriving vertex v from u. With the surfing model described earlier taken in consideration, "Prior bias" represents the probability to start with a particular node. In this case, all root nodes are considered equally important; therefore prior bias is 1/|R| for all root nodes. The prior bias in case of non-root nodes is set to 0 to eliminate the probability of starting with a non-root node. The "back probability" represents the probability to jump back to the root node in each step.

The HITS with Priors is an extension of the original HITS algorithm. The iterative equations are defined as:

$$\begin{aligned} a^{(i+1)}(v) &= (1-\beta)\left(\sum_{u=1}^{d_{in}(v)} \frac{h^{(t)}(u)}{H^{(i)}}\right) + \beta p_v \\ h^{(i+1)}(v) &= (1-\beta)\left(\sum_{u=1}^{d_{out}(v)} \frac{a^{(t)}(u)}{A^{(i)}}\right) + \beta p_v \end{aligned} \quad (3)$$

where $d_{in}(v)$ and $d_{out}(v)$ are the in-degree and out-degree of v, respectively, and $H^{(i)}$ and $A^{(i)}$ are defined as:

$$\begin{aligned} H^{(i)} &= \sum_{v=1}^{|V|} \sum_{u=1}^{d_{in}(v)} h^{(i)}(u) \\ A^{(i)} &= \sum_{v=1}^{|V|} \sum_{u=1}^{d_{out}(v)} a^{(i)}(u) \end{aligned} \quad (4)$$

For definitions of prior bias p_v and "back probability" β refer to the earlier sections under PageRank with Priors. The authority score is set as the importance of the node.

The K-Step Markov approach computes the relative probability that the system will spend time at any particular node given that it starts in a set of roots R and ends after K steps. According to White and Smyth [41], the value of K controls the relative trade-off between a distribution "biased" toward R, and when K gets larger the steady-state distribution will converge to the PageRank result. The equation to compute the K-Step Markov importance is:

$$I(t|R) = [Ap_R + A^2p_R \ldots A^Kp_R]_t \quad (5)$$

where A is the transition probability matrix of size n × n, p_R is an n × 1 vector of initial probabilities for the root set R, and I(t|R) is the t-th entry in this sum vector.

For additional details of the methods, the readers are referred to the original paper by White and Smyth [41].

PPIN analysis and derivation of topological parameters

The basic network statistics and topological parameters were derived using NetworkAnalyzer [47]. NetworkAnalyzer is a *JAVA* plug-in for Cytoscape [48], a software platform for the analysis and visualization of molecular interaction networks. The version of Cytoscape was 2.5.2 and NetworkAnalyzer was 2.5.1.

The implementation of the prioritization methods, PageRank with Priors, HITS with Priors, and the K-Step Markov approach are all available in the JUNG (*JAVA* Universal Network/Graph) [56] framework. It is a *JAVA* package that provides a common and extendible language for the modeling, analysis, and visualization of data that can be represented as a graph or network. Version 2.0 was used and integrated with other in-house programs through APIs to perform all the required functions. Fur-

ther details of JUNG can be obtained from the web site [56].

Evaluation of prioritization methods

Cross-validations to test the performance of the prioritization methods were done as described earlier [6]. Briefly, 19 diseases from OMIM [57] and GAD [58] were used as training sets. For each disease, the associated genes (with the one under test removed) were used as "seeds" and leave-one-out random cross-validation was performed. Random sets of genes were used as the control training sets. The rank-based sensitivity and specificity followed the previous definitions. ROC curves were plotted to visualize the performance with AUC values as quantitative measures. For further details refer to our previous publication [6].

All of the three node ranking methods require pre-determined parameters. For PageRank with Priors and HITS with Priors, the "back probability" is needed. It represents the bias toward the seeds, and the recommended value, according to White and Smyth [41], is 0.3. For the K-Step Markov approach, the only parameter is the length of the random walk, which controls the relative trade-off between a distribution "biased" toward the "seeds" and the steady-state distribution, which is independent of the "seeds." As K gets bigger, the final state is moving toward the steady state. The recommended K value was 6. In order to evaluate the effect of different values of the parameters on the performance, different values of parameters were used in the cross-validations and a test of each parameter setting was repeated five times to estimate the mean and standard deviation. Comparison of the performance of each of the three methods was done through analysis of variance.

Cardiac septal defect gene network

To obtain a list of all diseases that have a phenotype cardiac septal defect, we queried the "Clinical Synopsis" and "Allelic Variants" sections of OMIM database with the terms "atrial septal defect" or "ventricular septal defect" or "ASD" or "VSD". We then downloaded the associated genes (Training set) and their immediate interactants (Test set) based on PPIN. The test genes were then ranked using (a) functional annotation-based prioritization (ToppGene server [6]); and (b) PPIN-based ranking (as described earlier). The network view of the top 20 ranked genes along with their interactions with the training set genes was generated using Cytoscape (version 2.6.1) and the plug-in "NetworkAnalyzer" [47,48].

Authors' contributions

JC, BA, and AJ conceived the study design, which was coordinated by AJ. JC designed and implemented the gene prioritization algorithms and along with AJ participated in the analysis and interpretation of results. JC and AJ drafted the manuscript. All the authors have read and approved the final manuscript.

Additional material

Additional file 1

Venn diagrams of unique genes and interactions from the three (BIND, BioGRID, and HPRD) PPIN data source.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S1.pdf]

Additional file 2

Training set data used for evaluation of PPIN in disease candidate gene prioritization, comprising 19 diseases with 693 associated genes. *Of these, 589 genes were used in the cross validation because the rest (104 genes) had no reported interactions.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S2.pdf]

Additional file 3

Cardiac septal defect associated OMIM records, genes and their immediate interactants (based on protein-protein interactions).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S3.pdf]

Additional file 4

Prioritized candidate genes for cardiac septal defects using PPIN- and functional annotation- based methods.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S4.xls]

Additional file 5

Prioritized candidate genes for cardiac septal defects using three PPIN-based methods and two functional annotation- based methods (ToppGene and ENDEAVOUR).
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S5.xls]

Additional file 6

Annotation and PPIN coverage of the human genes.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-73-S6.xls]

Acknowledgements

This research was supported in part by the State of Ohio Computational Medicine Center (ODD TECH 04-042) and Digestive Health Center (DHC), Cincinnati (PHS Grant P30 DK078392). This study is a partial fulfillment of Jing Chen's requirements toward his Ph.D. thesis at the University of Cincinnati, Cincinnati, USA. We thank Eric Bardes and Vivek Kaimal, Division of Biomedical Informatics, CCHMC, Ohio, U.S.A., for their help in generating the annotation coverage map of human gene annotations. We also acknowledge the help of Ron Bryson, Technical Writer, Division of Biomedical Informatics, CCHMC, Ohio, U.S.A., in editing the manuscript.

References

1. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **Speeding disease gene discovery by sequence based candidate prioritization.** *BMC Bioinformatics* 2005, **6:**55.
2. Adie EA, Adams RR, Evans KL, Porteous DJ, Pickard BS: **SUSPECTS: enabling fast and effective prioritization of positional candidates.** *Bioinformatics* 2006, **22(6):**773-774.
3. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, De Smet F, Tranchevent LC, De Moor B, Marynen P, Hassan B, *et al.*: **Gene prioritization through genomic data fusion.** *Nat Biotechnol* 2006, **24(5):**537-544.
4. Tiffin N, Adie E, Turner F, Brunner HG, van Driel MA, Oti M, Lopez-Bigas N, Ouzounis C, Perez-Iratxeta C, Andrade-Navarro MA, *et al.*: **Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes.** *Nucleic Acids Res* 2006, **34(10):**3067-3081.
5. Turner FS, Clutterbuck DR, Semple CA: **POCUS: mining genomic sequence annotation to predict disease genes.** *Genome Biol* 2003, **4(11):**R75.
6. Chen J, Xu H, Aronow BJ, Jegga AG: **Improved human disease candidate gene prioritization using mouse phenotype.** *BMC Bioinformatics* 2007, **8:**392.
7. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N, *et al.*: **Towards a proteome-scale map of the human protein-protein interaction network.** *Nature* 2005, **437(7062):**1173-1178.
8. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, *et al.*: **A human protein-protein interaction network: a resource for annotating the proteome.** *Cell* 2005, **122(6):**957-968.
9. Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison.** *Nat Biotechnol* 2006, **24(4):**427-433.
10. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M: **Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps.** *Bioinformatics* 2005, **21(Suppl 1):**i302-310.
11. Lubovac Z, Gamalielsson J, Olsson B: **Combining functional and topological properties to identify core modules in protein interaction networks.** *Proteins* 2006, **64(4):**948-959.
12. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, Emili A, Snyder M, Greenblatt JF, Gerstein M: **A Bayesian networks approach for predicting protein-protein interactions from genomic data.** *Science* 2003, **302(5644):**449-453.
13. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, *et al.*: **Combining biological networks to predict genetic interactions.** *Proc Natl Acad Sci USA* 2004, **101(44):**15682-15687.
14. Sam L, Liu Y, Li J, Friedman C, Lussier YA: **Discovery of protein interaction networks shared by diseases.** *Pac Symp Biocomput* 2007:76-87.
15. Goehler H, Lalowski M, Stelzl U, Waelter S, Stroedicke M, Worm U, Droege A, Lindenberg KS, Knoblich M, Haenig C, *et al.*: **A protein interaction network links GIT1, an enhancer of huntingtin aggregation, to Huntington's disease.** *Mol Cell* 2004, **15(6):**853-865.
16. Ruffner H, Bauer A, Bouwmeester T: **Human protein-protein interaction networks and the value for drug discovery.** *Drug Discov Today* 2007, **12(17–18):**709-716.
17. Neduva V, Linding R, Su-Angrand I, Stark A, de Masi F, Gibson TJ, Lewis J, Serrano L, Russell RB: **Systematic discovery of new recognition peptides mediating protein interaction networks.** *PLoS Biol* 2005, **3(12):**e405.
18. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286(5439):**509-512.
19. Berg J, Lassig M, Wagner A: **Structure and evolution of protein interaction networks: a statistical model for link dynamics and gene duplications.** *BMC Evol Biol* 2004, **4(1):**51.
20. Eisenberg E, Levanon EY: **Preferential attachment in the protein network evolution.** *Phys Rev Lett* 2003, **91(13):**138701.
21. Rzhetsky A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome.** *Bioinformatics* 2001, **17(10):**988-996.
22. Wagner A, Fell DA: **The small world inside large metabolic networks.** *Proc Biol Sci* 2001, **268(1478):**1803-1810.
23. Xu J, Li Y: **Discovering disease-genes by topological features in human protein-protein interaction network.** *Bioinformatics* 2006, **22(22):**2800-2805.
24. Berger SI, Posner JM, Ma'ayan A: **Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases.** *BMC Bioinformatics* 2007, **8:**372.
25. Junker BH, Koschutzki D, Schreiber F: **Exploration of biological network centralities with CentiBiN.** *BMC Bioinformatics* 2006, **7:**219.
26. Bortoluzzi S, Romualdi C, Bisognin A, Danieli GA: **Disease genes and intracellular protein networks.** *Physiol Genomics* 2003, **15(3):**223-227.
27. George RA, Liu JY, Feng LL, Bryson-Richardson RJ, Fatkin D, Wouters MA: **Analysis of protein sequence and interaction data for candidate disease gene prediction.** *Nucleic Acids Res* 2006, **34(19):**e130.
28. Gonzalez G, Uribe JC, Tari L, Brophy C, Baral C: **Mining gene-disease relationships from biomedical literature: weighting protein-protein interactions and connectivity measures.** *Pac Symp Biocomput* 2007:28-39.
29. Kann MG: **Protein interactions and disease: computational approaches to uncover the etiology of diseases.** *Brief Bioinform* 2007, **8(5):**333-346.
30. Kohler S, Bauer S, Horn D, Robinson PN: **Walking the interactome for prioritization of candidate disease genes.** *Am J Hum Genet* 2008, **82(4):**949-958.
31. Limviphuvadh V, Tanaka S, Goto S, Ueda K, Kanehisa M: **The commonality of protein interaction networks determined in neurodegenerative disorders (NDDs).** *Bioinformatics* 2007, **23(16):**2129-2138.
32. Pattin KA, Moore JH: **Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases.** *Hum Genet* 2008, **124(1):**19-29.
33. Wachi S, Yoneda K, Wu R: **Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues.** *Bioinformatics* 2005, **21(23):**4205-4208.
34. Wu X, Jiang R, Zhang MQ, Li S: **Network-based global inference of human disease genes.** *Mol Syst Biol* 2008, **4:**189.
35. Oti M, Snel B, Huynen MA, Brunner HG: **Predicting disease genes using protein-protein interactions.** *J Med Genet* 2006, **43(8):**691-698.
36. Freeman LC: **Centrality in social networks conceptual clarification.** *Social Networks* 1978, **1(3):**215-239.
37. Sabidussi G: **The centrality index of a graph.** *Psychometrika* 1966, **31(4):**581-603.
38. Freeman LC: **A set of measures of centrality based on betweenness.** *Sociometry* 1977, **40(1):**35-41.
39. Jon MK: **Authoritative sources in a hyperlinked environment.** *ACM* 1999, **46:**604-632.
40. Page L, Brin S, Motwani R, Winograd T: **The pagerank citation ranking: Bringing order to the web.** 2001 [http://infolab.stanford.edu/~backrub/pageranksub.ps].
41. White S, Smyth P: **Algorithms for estimating relative importance in networks.** In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* ACM Press; 2003:266-275.
42. Kleinberg J: **Authoritative sources in a hyperlinked environment.** *Journal of the ACM* 1999, **46(5):**604-632.
43. **Entrez Gene** [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene]
44. Bader GD, Betel D, Hogue CW: **BIND: the Biomolecular Interaction Network Database.** *Nucleic Acids Res* 2003, **31(1):**248-250.
45. Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bahler J, Wood V, *et al.*: **The BioGRID Interaction Database: 2008 update.** *Nucleic Acids Res* 2008:D637-640.
46. Peri S, Navarro JD, Kristiansen TZ, Amanchy R, Surendranath V, Muthusamy B, Gandhi TK, Chandrika KN, Deshpande N, Suresh S, *et al.*: **Human protein reference database as a discovery resource for proteomics.** *Nucleic Acids Res* 2004:D497-501.
47. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M: **Computing topological parameters of biological networks.** *Bioinformatics* 2008, **24(2):**282-284.
48. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13(11):**2498-2504.
49. McKusick VA: **Mendelian Inheritance in Man and its online version, OMIM.** *Am J Hum Genet* 2007, **80(4):**588-604.

50. Oike Y, Hata A, Mamiya T, Kaname T, Noda Y, Suzuki M, Yasue H, Nabeshima T, Araki K, Yamamura K: **Truncated CBP protein leads to classical Rubinstein-Taybi syndrome phenotypes in mice: implications for a dominant-negative mechanism.** *Hum Mol Genet* 1999, **8(3):**387-396.
51. Roth JF, Shikama N, Henzen C, Desbaillets I, Lutz W, Marino S, Wittwer J, Schorle H, Gassmann M, Eckner R: **Differential role of p300 and CBP acetyltransferase during myogenesis: p300 acts upstream of MyoD and Myf5.** *Embo J* 2003, **22(19):**5186-5196.
52. Bamforth SD, Braganca J, Eloranta JJ, Murdoch JN, Marques FI, Kranc KR, Farza H, Henderson DJ, Hurst HC, Bhattacharya S: **Cardiac malformations, adrenal agenesis, neural crest defects and exencephaly in mice lacking Cited2, a new Tfap2 co-activator.** *Nat Genet* 2001, **29(4):**469-474.
53. Ching YH, Ghosh TK, Cross SJ, Packham EA, Honeyman L, Loughna S, Robinson TE, Dearlove AM, Ribas G, Bonser AJ, *et al.*: **Mutation in myosin heavy chain 6 causes atrial septal defect.** *Nat Genet* 2005, **37(4):**423-428.
54. Ozcelik C, Erdmann B, Pilz B, Wettschureck N, Britsch S, Hubner N, Chien KR, Birchmeier C, Garratt AN: **Conditional mutation of the ErbB2 (HER2) receptor in cardiomyocytes leads to dilated cardiomyopathy.** *Proc Natl Acad Sci USA* 2002, **99(13):**8880-8885.
55. Crone SA, Zhao YY, Fan L, Gu Y, Minamisawa S, Liu Y, Peterson KL, Chen J, Kahn R, Condorelli G, *et al.*: **ErbB2 is essential in the prevention of dilated cardiomyopathy.** *Nat Med* 2002, **8(5):**459-465.
56. **JUNG** [http://jung.sourceforge.net/]
57. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA: **Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.** *Nucleic Acids Res* 2005:D514-517.
58. Becker KG, Barnes KC, Bright TJ, Wang SA: **The genetic association database.** *Nat Genet* 2004, **36(5):**431-432.