Role of Centrality in Network-Based Prioritization of Disease Genes

Sinan Erten[1] and Mehmet Koyutürk[1,2]

(1) Dept. of Electrical Engineering & Computer Science
(2) Center for Proteomics & Bioinformatics
Case Western Reserve University, Cleveland, OH 44106, USA.

Abstract. High-throughput molecular interaction data have been used effectively to prioritize candidate genes that are linked to a disease, based on the notion that the products of genes associated with similar diseases are likely to interact with each other heavily in a network of protein-protein interactions (PPIs). An important challenge for these applications, however, is the incomplete and noisy nature of PPI data. Random walk and network propagation based methods alleviate these problems to a certain extent, by considering indirect interactions and multiplicity of paths. However, as we demonstrate in this paper, such methods are likely to favor highly connected genes, making prioritization sensitive to the skewed degree distribution of PPI networks, as well as ascertainment bias in available interaction and disease association data. Here, we propose several statistical correction schemes that aim to account for the degree distribution of known disease and candidate genes. We show that, while the proposed schemes are very effective in detecting loosely connected disease genes that are missed by existing approaches, this improvement might come at the price of more false negatives for highly connected genes. Motivated by these results, we develop uniform prioritization methods that effectively integrate existing methods with the proposed statistical correction schemes. Comprehensive experimental results on the Online Mendelian Inheritance in Man (OMIM) database show that the resulting hybrid schemes outperform existing methods in prioritizing candidate disease genes.

1 Introduction

Identification of disease-associated genes is an important step toward enhancing our understanding of the cellular mechanisms that drive human diseases, with profound applications in modeling, diagnosis, prognosis, and therapeutic intervention [1]. Genome-wide linkage and association studies in healthy and affected populations provide chromosomal regions containing up to 300 candidate genes possibly associated with genetic diseases [2]. Investigation of these candidates based on sequencing is an expensive task, thus not always a feasible option. Consequently, computational methods are primarily used to prioritize and identify the most likely disease-associated genes by utilizing a variety of data sources such as gene expression [3], functional annotations [4, 5], and protein-protein interactions (PPIs) [3,6–11]. However, the scope of methods that rely on functional annotations is limited because only a small fraction of genes in the genome are currently annotated.

In recent years, several algorithms are proposed to incorporate topological properties of PPI networks in understanding genetic diseases [3, 6, 10]. These algorithms mostly focus on prioritization of candidate genes and mainly exploit the notion that the products of genes associated with similar diseases are likely to be close to each other and interact heavily in a network of PPIs. However, an important challenge for these applications is the incomplete and noisy nature of the PPI data [12]. Vast amounts of missing interactions and false positives effect the accuracy of methods based on local network information such as direct interactions and shortest distances. Few global methods based on simulation of random walks [6, 10] and network propagation [11] get around this problem to a certain extent by considering multiple alternate paths and whole topology of PPI networks. Nevertheless, as we demonstrate in this paper, these methods favor genes whose products are highly connected in the network and perform poorly in identifying loosely connected disease genes.

Motivated by this observation, we here propose novel statistical correction methods for network-based disease gene prioritization. These methods aim to assess the significance of the connectivity of a candidate gene to known disease genes with respect to a reference model that takes into account the degree distribution of the PPI network. We show that the proposed correction schemes are very effective in detecting loosely connected disease genes which are generally less studied, thus potentially more interesting for many applications in terms of generating novel biological knowledge. However, we observe that these schemes might perform less favorably in identifying highly connected disease genes. Consequently, we develop several uniform prioritization methods that effectively integrate existing algorithms with the proposed statistical adjustment schemes, with a view to delivering high accuracy irrespective of the network centrality of target disease genes. Comprehensive experimental results show that the resulting hybrid prioritization schemes outperform existing approaches in identifying disease-associated genes.

2 Background and Motivation

There exists a wide range of methods based on the analysis of the topological properties of PPI networks. These methods commonly rely on the expectation that the products of genes that are associated with similar diseases interact heavily with each other. It is important to note that the purpose here is to infer functional associations between genes from functional and physical interactions between their products. For this reason, any reference to interactions between genes in this paper refers to the interactions between their products. Existing methods can be classified into two main categories; (i) localized methods, *i.e.*, methods based on direct interactions and shortest paths between known disease genes and candidate genes [3,7,13], (ii) global methods, *i.e.*, methods that model the information flow in the cell to assess the proximity and connectivity between known disease genes and candidate genes. Several studies show that global approaches, such as random walk and network propagation, clearly outperform local approaches [10,11]. For this reason, we focus on global methods in this paper.

Network-based candidate disease gene prioritization. For a given disease of interest D, the input to the candidate disease gene prioritization problem consists of two sets of genes, seed set $\mathcal{S}$ and candidate set $\mathcal{C}$. The *seed set* $\mathcal{S}$ specifies prior knowledge on the disease, *i.e.*, it is the set of genes known to be associated with D and diseases similar to D. Each gene $v \in \mathcal{S}$ is also associated with a similarity score $\sigma(v, D)$, indicating the known degree of association between v and D. The similarity score for gene v is computed as the maximum similarity between D and any other disease associated with v (a detailed discussion on computation of similarity scores can be found in [14]). The *candidate set* $\mathcal{C}$ specifies the genes, one or more of which are potentially associated with disease D (e.g., these genes might lie within a linkage interval that is identified by association studies). The overall objective of network based disease prioritization is to use a human PPI network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, to compute a score $\alpha(v, D)$ for each gene $v \in \mathcal{C}$ that represents the likelihood of v to be associated with D.

The PPI network $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ consists of a set of gene products $\mathcal{V}$ and a set of undirected interactions $\mathcal{E}$ between these gene products where $uv \in \mathcal{E}$ represents an interaction between $u \in \mathcal{V}$ and $v \in \mathcal{V}$. In this network, the set of interacting partners of a gene product $v \in \mathcal{V}$ is defined as $N(v) = \{u \in \mathcal{V} : uv \in \mathcal{E}\}$. Global prioritization schemes use this network information to compute α by propagating σ over $\mathcal{G}$. Candidate proteins are then ranked according to α and novel genes that are potentially associated with the disease of interest are identified based on this ranking.

Random walk with restarts. This method simulates a random walk on the network to compute the proximity between two nodes by exploiting the global structure of the network [15, 16]. It is used in a wide range of applications, including identification of functional modules [17] and modeling the evolution of social networks [18]. Recently, random walk with restarts has also been applied to candidate disease gene prioritization [6, 10].

In the context of disease gene prioritization, random walk with restarts is applied as follows. A random walk starts at one of the nodes in $\mathcal{S}$. At each step, the random walk either moves to a randomly chosen neighbor $u \in N$ of the current gene v or it restarts at one of the genes in the seed set $\mathcal{S}$. The probability of restarting at a given time step is a fixed parameter denoted by r. For each restart, the probability of restarting at $v \in \mathcal{S}$ is a function of $\sigma(v, D)$, *i.e.*, the degree of association between v and the disease of interest. After a sufficiently long time, the probability of being at node v at a random time step provides a measure of the functional association between v and the genes known to be associated with D [6, 10]. Algorithmically, random-walk based association scores can be computed iteratively as follows:

$$x_{t+1} = (1 - r)P_{\text{RW}} x_t + r\rho. \tag{1}$$

Here, ρ denotes the restart vector with $\rho(u) = \sigma(u, D)/\sum_{v \in \mathcal{S}} \sigma(v, D)$ for $u \in \mathcal{S}$ and 0 otherwise. P_{RW} denotes the stochastic matrix derived from $\mathcal{G}$, *i.e.*, $P_{\text{RW}}(u, v) = 1/|N(v)|$ for $vu \in \mathcal{E}$ and 0 otherwise. For each $v \in \mathcal{V}$, $x_t(v)$ denotes the probability that the random walk will be at v at time t, where $x_0 = \rho$. For each gene v, the resulting random-walk based association score is defined as $\alpha_{\text{RW}}(v, D) = \lim_{t \to \infty} x_t(v)$.

Network propagation. Propagation based models have been previously shown to be effective in network based functional annotation of proteins [19]. In recent work, Vanunu and Sharan [11] propose a network propagation algorithm to compute the association between candidate proteins and known disease genes. They define a prioritization function which models simulation of an information pump that originates at the seed sets. This idea is very similar to that of random walk with restarts, with one key difference. Namely, in network propagation, the flow of information is normalized by not only the total outgoing flow from each node, but also the total incoming flow into each node. In other words, the matrix P_{RW} is replaced my a matrix P_{NP}, in which each entry is normalized with respect to row and column sums. The resulting propagation based model can also be simulated iteratively as follows:

$$y_{t+1} = (1-r)P_{\text{NP}}y_t + r\rho. \tag{2}$$

Here, the propagation matrix P_{NP} is computed as $P_{\text{NP}}(u,v) = 1/\sqrt{|N(u)||N(v)|}$ for $uv \in \mathcal{E}$, 0 otherwise. For each $v \in \mathcal{V}$, $y_t(v)$ denotes the amount of disease association information at node v at step t, where $y_0 = \rho$. For each gene v, the resulting network propagation based association score is defined as $\alpha_{\text{NP}}(v,D) = \lim_{t\to\infty} y_t(v)$. In this model, $0 \leq r \leq 1$ is also a user-defined parameter that is used to adjust the relative importance of prior knowledge and network topology.

Role of network centrality. In order to motivate our approach, we evaluate here the performance of random walk with restarts and network propagation with respect to the network degree (number of known interactions) of candidate genes. As shown in Figure 1(a), these methods are clearly biased toward scoring highly connected proteins higher. In this figure, the performance measure is the average rank of the true candidate protein among other 99 proteins in the same linkage interval. As evident in the figure, existing global methods work very well in predicting highly connected proteins, whereas they perform quite poorly for loosely connected proteins, especially for those with degree less than 6. Furthermore, as seen in Figure 1(b), the degree distribution of known disease genes is slightly biased toward highly connected genes, however there exist many disease genes that are loosely connected as well. For this reason, it is at least as important to correctly identify loosely connected disease genes as to identify those that are highly connected, in order to remove the effect of ascertainment bias in PPI data and known disease associations.

The dependency of performance on network degree can be understood by carefully inspecting the formulation of random walk and network propagation models. Random walk with restarts is actually a generalization of Google's well-known page-rank algorithm [20], such that for $r = 0$, α is solely a measure of network centrality. Therefore, for any $r > 0$ (in our experiments, we observe that $r = 0.3$ is optimal for the performance of both algorithms after running the algorithms with small increments of r values; this is also the setting used in Figure 1), $\alpha(v,D)$ contains a component that represents the network centrality of v, in addition to its association with D. Network propagation alleviates this problem by normalizing the incoming flow into a gene, therefore provides a slightly more balanced performance compared to random walk with restarts. However, as evident in the figure, its performance is still influenced heavily by node degrees. Motivated by these insights, we argue that the association scores computed by these algorithms have to be statistically adjusted with respect to reference models that take into account the degree distribution of the network.

3 Methods

In this section, we propose several reference models for assessing the significance of network-based disease association scores. Subsequently, we discuss how these models can be used in conjunction with existing methods to obtain uniform prioritization schemes that can deliver high accuracy regardless of centrality of candidate genes.

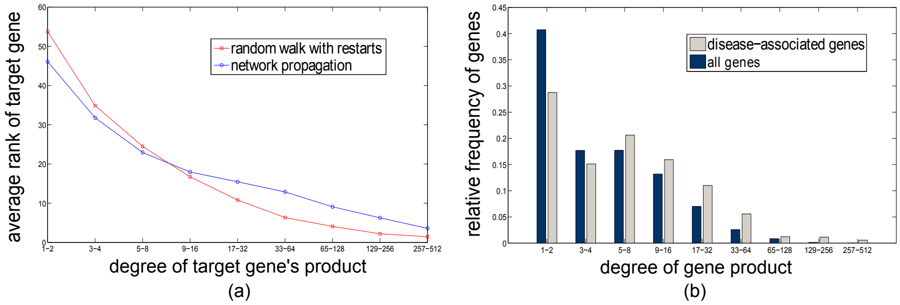


Fig. 1. (a) The effect of degree to the performance of existing global approaches. x-axis is the degree range while y-axis represents the average rank for the true disease genes. (b) Histogram of the degrees of disease genes and all genes in the network.

3.1 Reference Models for Statistical Adjustment

Here, we consider three alternate reference models for assessing the significance of disease association scores obtained by random walk with restarts or network propagation: (i) a model that generates a separate background population for each candidate gene based on the degree distribution of the seed set, (ii) a model that generates a background population for each group of candidates with similar degree for a fixed seed set, (iii) a model that assesses the log-likelihood of the association of a gene with the seed set with respect to its network centrality. Here, for the sake of clarity, we describe each model assuming that random walk based restarts is used to compute raw association scores (we also drop the subscript RW from our notation for simplicity).

Reference model based on seed degrees. The objective here is to generate a reference model that captures the degree distribution of seed proteins accurately. To this end, we compare the association score $\alpha(v, D)$ for each protein with scores computed using random seed sets (by preserving the degree distribution of the seed genes). The expectation here is that false positives that correspond to centralized and highly connected proteins will have high association scores even with respect to these randomly generated seed sets. Furthermore, this model aims to balance the effect of highly connected known disease genes with that of loosely connected ones.

Given a disease D, seed set $\mathcal{S}$, and candidate set C, this reference model is implemented as follows:

- We first compute network-based association scores $\alpha(v, D)$ for the original seed set $\mathcal{S}$, using the procedure described in Equation 1.
- Then, based on the original seed set $\mathcal{S}$, we generate a random instance $\mathcal{S}^{(i)}$ that represents $\mathcal{S}$ in terms of degree distribution. $\mathcal{S}^{(i)}$ is generated as follows:
  - First, a bucket $\mathcal{B}(u)$ is created for each protein $u \in \mathcal{S}$.
  - Then, each protein $v \in \mathcal{V}$ is assigned to bucket $\mathcal{B}(u)$ if $|N(v) - N(u)| < |N(v) - N(u')|$ for all $u' \in \mathcal{S}$, where ties are broken randomly.
  - Subsequently, $\mathcal{S}^{(i)}$ is generated by choosing a protein from each bucket uniformly at random. It can be observed that each protein in $\mathcal{S}$ is represented by exactly one protein in $\mathcal{S}^{(i)}$, thus the total degree of proteins in $\mathcal{S}^{(i)}$ is expected to be very close to that of $\mathcal{S}$.

- For $1 \leq i \leq n$, the association scores $\alpha^{(i)}$ for seed set $\mathcal{S}^{(i)}$ are computed using Equation 1. Here, n is a sufficiently large number that is used to obtain a representative sampling $\{\alpha^{(1)}, \alpha^{(2)}, \alpha^{(3)}, ..., \alpha^{(n)}\}$ of the population of association scores for seed sets that match the size and degree distribution of $\mathcal{S}$ (we use $n = 1000$ in our experiments).
- We then estimate the mean of this distribution as $\mu_{\mathcal{S}} = \sum_{1 \leq i \leq n} \alpha^{(i)}/n$ and the standard deviation as $\sigma_{\mathcal{S}}^2 = \sum_{1 \leq i \leq n}((\alpha^{(i)} - \mu_{\mathcal{S}})(\alpha^{(i)} - \mu_{\mathcal{S}})^T)/(n-1)$.
- Finally, we compute the seed degree adjusted association score for each gene v as $\alpha_{\text{SD}}(v, D) = (\alpha(v, D) - \mu_{\mathcal{S}})/\sigma_{\mathcal{S}}$.

Note that, since the multiple hypotheses being tested here are compared and ranked against each other (as opposed to accepting/rejecting individual hypotheses), it is not necessary to perform correction for multiple hypothesis testing.

Reference model based on candidate degree. This reference model aims to assess the statistical significance of the association score $\alpha(v, D)$ of a gene $v \in \mathcal{V}$ with respect to seed set $\mathcal{S}$ based on a population of association scores that belong to genes with degree similar to that of v. This reference model is generated as follows:

- First, we compute the network-based association vector α with respect to the given seed set $\mathcal{S}$, again using Equation 1.
- Then, for each candidate gene $v \in \mathcal{C}$, we select the n genes in the network with smallest $|N(v) - N(u)|$ to create a representative set $\mathcal{M}(v)$ that contains the n genes most similar to v in terms of their degree ($n = 1000$ in our experiments).
- Subsequently, for each gene $v \in \mathcal{C}$, we estimate the mean association score of its representative population as $\mu(v) = \sum_{u \in \mathcal{M}(v)} \alpha(u)/|\mathcal{M}(v)|$ and the standard deviation of association scores as $\sigma^2(v) = \sum_{u \in \mathcal{M}(v)} (\alpha_{\mathcal{S}}(u) - \mu(v))/(|\mathcal{M}(v)| - 1)$.
- Finally, we compute the candidate degree adjusted association score of each candidate gene v as $\alpha_{\text{CD}}(v, D) = (\alpha_{\mathcal{S}}(v, D) - \mu(v))/\sigma(v)$.

Likelihood-ratio test using eigenvector centrality. Here, we assess the association of a gene with the seed set using a likelihood-ratio test. More precisely, considering $\alpha_{\text{RW}}(v, D)$ as the likelihood of v being associated with the seed set $\mathcal{S}$ for disease D, we compare this likelihood with the likelihood of v being associated with any other gene product in the network. To compute the likelihood of v's association with any other gene in the network, we use eigenvector centrality [20], which is precisely equal to the random walk based association score of v for zero restart probability ($r = 0$). Indeed, setting $r = 0$ corresponds to the case where the seed set is empty, thereby making the resulting association score a function of the gene's network centrality. For each $v \in \mathcal{C}$, the eigenvector centrality based log-likelihood score is computed as:

$$\alpha_{\text{EC}}(v, D) = \log \frac{\alpha^{(r>0)}(v, D)}{\alpha^{(r=0)}(v, D)}. \tag{3}$$

3.2 Uniform Prioritization

As we demonstrate in the next section, the adjustment strategies presented improve the performance of global prioritization algorithms in identifying loosely connected disease genes. However, this comes at the price of increased number of false negatives for highly connected disease genes. Motivated by this observation, we propose several hybrid scoring schemes that aim to take advantage of both raw and statistically adjusted association scores. The idea here is to derive a uniform prioritization method that uses the adjusted scores for loosely connected candidate genes, while using the raw scores for highly connected candidate genes.

For this purpose, we first sort the raw crosstalk scores (α_{RW} or α_{NP}) of candidate genes in descending order. Let $R_{\text{RAW}}(v)$ denote the rank of gene $v \in \mathcal{C}$ in this ordering. Clearly, for $u, v \in \mathcal{C}$, $R_{\text{RAW}}(v) < R_{\text{RAW}}(u)$ indicates v is more likely to be associated with the disease than u is. Similarly, we sort the statistically adjusted association scores (α_{SD}, α_{CD}, or α_{EC}) in descending order, to obtain a rank $R_{\text{ADJ}}(v)$ for each gene $v \in \mathcal{C}$. We propose three alternate strategies for merging these two rankings to obtain a uniform ranking R_{UNI}, where the objective is to have $R_{\text{UNI}}(v) < R_{\text{UNI}}(u)$ if gene v is associated with the disease, while gene u is not. Once $R_{\text{UNI}}(v)$ is obtained using one of the following methods, we map it into the interval $[1, |\mathcal{C}|]$ in the obvious way.

Uniform prioritization based on the degree of candidate gene. This uniform prioritization scheme chooses the ranking of each candidate gene based on its own degree. Namely, for a given user-defined threshold λ, we define $R_{\text{UNI}}^{(\text{C})}$ as:

$$R_{\text{UNI}}^{(\text{C})}(v) = \begin{cases} R_{\text{RAW}}(v) & \text{if } |N(v)| > \lambda \\ R_{\text{ADJ}}(v) & \text{otherwise} \end{cases} \tag{4}$$

for each $v \in \mathcal{C}$. Thus the ranking of a highly-connected gene is based on its raw association score, while that of a loosely-connected gene is based on the statistical significance of its association score. Note that, with respect to this definition, the ranking of two genes can be identical (but there cannot be more than two genes with identical ranking). In this case, the tie is broken based on the unused ranking of each gene.

Optimistic uniform prioritization. This approach uses the best available ranking for each candidate gene, based on the expectation that a true disease gene is more likely to show itself in at least one of the rankings as compared to a candidate gene that is not associated with the disease. Namely, we define $R_{\text{UNI}}^{(\text{O})}$ as:

$$R_{\text{UNI}}^{(\text{O})}(v) = \begin{cases} R_{\text{RAW}}(v) & \text{if } R_{\text{RAW}}(v) < R_{\text{ADJ}}(v) \\ R_{\text{ADJ}}(v) & \text{otherwise} \end{cases} \tag{5}$$

for each $v \in \mathcal{C}$. Again, ties are broken based on the unused rankings.

Uniform prioritization based on degree of known disease genes. Based on the notion that some diseases are studied more in detail compared to other diseases, we expect the degrees of genes associated with similar diseases to be somewhat close to each other. Statistical tests on disease associations currently available in the OMIM (Online Mendelian Inheritance in Man) database confirms this expectation (data not shown). We take advantage of this observation to approximate the network degree of the unknown disease gene in terms of the degrees of the known disease genes. This enables having a global criterion for choosing the preferred ranking for all genes, as opposed to the gene-specific (or "local") criteria described above. For a given seed set $\mathcal{S}$, we first compute $\overline{d}(\mathcal{S}) = (\sum_{u \in \mathcal{S}} |N(u)|)/|\mathcal{S}|$. Subsequently, if $\overline{d}(\mathcal{S}) > \lambda$ (where λ is defined as above), we set $R_{\text{UNI}}^{(\text{S})}(v) = R_{\text{RAW}}(v)$ for all $v \in \mathcal{C}$, otherwise, we set $R_{\text{UNI}}^{(\text{S})}(v) = R_{\text{ADJ}}(v)$.

4 Results

In this section, we comprehensively evaluate the performance of the methods presented in the previous section.

4.1 Datasets

In our experiments, we use the human PPI data obtained from NCBI Entrez Gene Database [21]. This database integrates interaction data from several other databases available, such as HPRD, BioGrid and BIND. After the removal of nodes with no interactions, the final PPI network contains 8959 proteins and 33528 distinct interactions among these proteins.

We obtain disease information from Online Mendelian Inheritance in Man (OMIM) database. OMIM provides a publicly accessible and comprehensive database of genotype-phenotype relationship in humans. We map genes associated with diseases to our PPI network and remove those diseases for which we are unable to map more than two associated genes. After this step, we have a total of 206 diseases with at least 3 associated genes. Number of genes associated with these diseases ranges from 3 to 36, with the average number of associations for each disease being approximately 6.

4.2 Experimental Setting

In order to evaluate the performance of different methods in terms of accurately prioritizing disease-associated genes, we apply leave-one-out cross-validation. For each gene that is associated with a disease, we conduct the following experiment:

- We remove a gene from the set of genes associated with a particular disease.
- We generate an artificial linkage interval, containing this removed gene with other 99 genes located nearest in terms of the genomic distance. Note that, according to our experiments, the size of candidate set does not have a significant effect on the performance gap between different methods as long as it is greater than 20 (data not shown).
- Using each of the methods described in the previous section, we obtain a ranking of candidate genes and use this ranking to predict disease genes. Note that, due to space considerations, we only use random walk with restarts in conjunction with the proposed statistical correction and uniform prioritization methods, however, these methods can also be applied to network propagation straightforwardly.

In order to systematically compare the performance of different methods, we use the following evaluation criteria:

Average rank. Average rank of the correct disease gene among all candidate genes, computed across all disease. Clearly, a lower average rank indicates better performance.

ROC curves. We also plot ROC curves, *i.e.*, *sensitivity vs.* 1-*specificity*, by thresholding the rank to be considered a “predicted disease gene” from 1 to 100. *Sensitivity* (recall) is defined as the percentage of true disease genes that are ranked above the particular threshold, whereas *specificity* is defined as the percentage of all genes that are ranked below the threshold. The area under ROC curve (AUC) is used as another measure to assess the performance of different methods.

Percentage of the disease genes ranked in top 1% and 5%. Percentages of true disease genes that are ranked as one of the genes in the top 1% (practically, the top gene) and also in the top 5% among all candidates are listed separately.

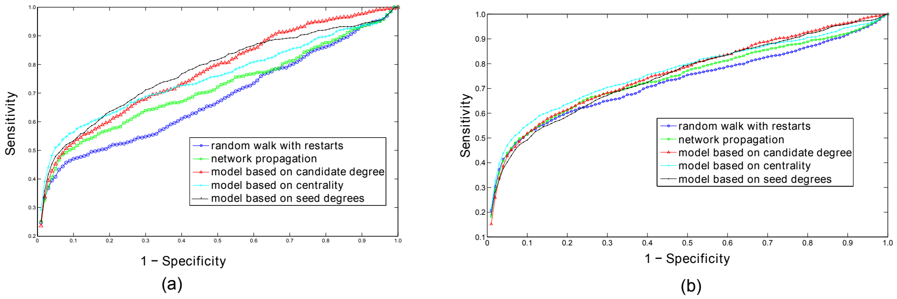


Fig. 2. ROC curves of proposed statistical adjustment schemes and existing methods for (a) cases in which true disease gene has degree at most five, (b) all disease genes. All of the proposed adjustment schemes outperform existing methods.

4.3 Performance of Statistical Adjustment Schemes

As mentioned before, the performance of the global methods is highly biased with the degree of the true candidate protein. The effect of the degree of true disease gene on the performance of global methods is demonstrated in Figure 1. To investigate the effect of the proposed statistical correction schemes on accurate ranking of low-degree proteins, we first compare the ROC curves achieved by different methods by considering the true disease genes with degree $\leq\ 5$. These results are shown in Figure 2(a) and Table 1. As seen in the figure, all of the three statistical adjustment schemes outperform existing methods for these genes. Furthermore, as evident in Figure 2(b), when all genes are considered, the statistical adjustment schemes still perform better than existing methods. However, as seen in the figure, the performance difference is minor because of the relatively degraded performance of statistical adjustment schemes for highly connected genes. Next, we investigate how the proposed uniform prioritization methods improve the performance of these statistical adjustment schemes.

4.4 Performance of Uniform Prioritization

Here, we systematically investigate the performance of the proposed uniform prioritization methods, by considering the combination of each of these methods with each of the three statistical adjustment methods (a total of nine combinations). In these experiments, the degree threshold λ is set to 5. For convenience, we refer to each uniform prioritization method using the corresponding ranking symbol introduced in the previous section ($R^{(\mathrm{C})}_{\mathrm{UNI}}$, $R^{(\mathrm{O})}_{\mathrm{UNI}}$, or $R^{(\mathrm{S})}_{\mathrm{UNI}}$).

Table 1. The effect of statistical adjustment on performance. Average Rank of the true disease genes and AUC values are listed. To demonstrate the effect of connectivity, we also provide separate results for the cases in which the degree of true disease gene is $\leq\ 5$ and > 5.

	All Genes		Degrees$\leq\ 5$		Degrees$>\ 5$	
Method	Avg. Rank	AUROC	Avg. Rank	AUROC	Avg. Rank	AUROC
Network Propagation	26.32	0.74	33.12	0.61	18.29	0.83
Random walk w/ restarts	28.02	0.73	37.73	0.62	17.56	0.84
Based on seed degree	25.55	0.75	26.10	0.73	24.43	0.78
Based on candidate degree	24.62	0.76	26.46	0.72	23.66	0.79
Based on centrality	24.55	0.76	26.27	0.73	23.16	0.79

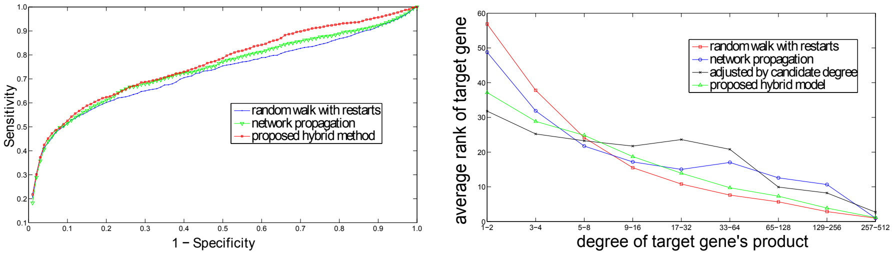
Table 2. Performance of all combinations of uniform prioritization methods.

	Candidate deg.			Seed deg.			Centrality		
	$R^{(\mathrm{C})}_{\mathrm{UNI}}$	$R^{(\mathrm{O})}_{\mathrm{UNI}}$	$R^{(\mathrm{S})}_{\mathrm{UNI}}$	$R^{(\mathrm{C})}_{\mathrm{UNI}}$	$R^{(\mathrm{O})}_{\mathrm{UNI}}$	$R^{(\mathrm{S})}_{\mathrm{UNI}}$	$R^{(\mathrm{C})}_{\mathrm{UNI}}$	$R^{(\mathrm{O})}_{\mathrm{UNI}}$	$R^{(\mathrm{S})}_{\mathrm{UNI}}$
Avg. Rank	23.22	24.33	23.30	25.01	25.29	25.42	24.95	24.92	24.02
AUROC	0.76	0.76	0.77	0.75	0.75	0.76	0.75	0.75	0.76
Perc. ranked in top 1%	21.7	19.4	14.7	18.4	18.5	19.3	20.0	20.5	21.3
Perc. ranked in top 5%	45.1	44.4	42.1	45.5	44.1	41.2	46.3	45.7	47.0

The average rank and AUC for the performance of the nine combinations of proposed methods are listed in Table 2. As seen in the table, while all methods improve upon the performance of raw statistical adjustment schemes, it is difficult to choose between the proposed methods. We suggest that the hybrid method based on candidate degree ($R^{(\mathrm{C})}_{\mathrm{UNI}}$), combined with statistical adjustment based on candidate degree, can be considered the "winner", since this approach provides the best accuracy in correctly predicting the disease protein as the top candidate (21.7%) and it provides the lowest average rank of the true candidate gene (23.22). We compare this combination of proposed algorithms with existing global methods in Table 3 and Figure 3(a). These results clearly show that our final uniform prioritization scheme outperforms existing methods with respect to all performance criteria. Furthermore, careful inspection of average rank with respect to the degree of true disease gene in Figure 3(b) shows that, this method almost matches the performance of the best performing algorithm for each degree regime. Namely, if the target gene has low degree, our uniform prioritization method performs close to statistical adjusted random walk, while it performs close to raw random walk for high-degree target genes.

4.5 Case Example

Here, we provide a real example to demonstrate the power of the proposed method in identifying loosely connected disease genes. We focus on *Microphthalmia* which is a disease that has 3 genes directly associated with it in our PPI network, namely *SIX6*, *CHX10* and *BCOR*. In our experiments, we remove *SIX6* and try to predict this gene using the other two genes, as well genes associated with diseases similar to Microphthalmia. This experiment is illustrated in Figure 4. The figure shows the 2-neighborhood of proteins *SIX6*, *CHX10* and *BCOR*. As seen in the figure, the global methods fail be-



(a) (b)

Fig. 3. (a) ROC curves to compare the proposed method with existing global approaches. (b)The effect of the degree of target gene on the performance of existing global approaches, adjusted method based on candidate degrees as well as the our final uniform prioritization method.

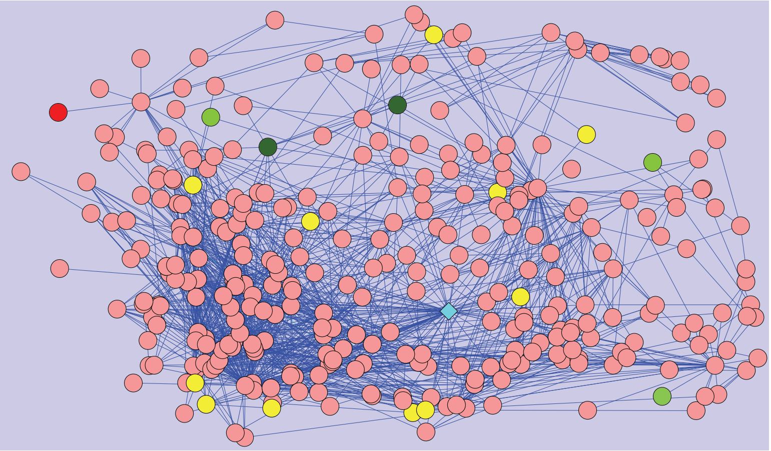


Fig. 4. Case example for the Microphthalmia disease. Products of genes associated with Microphthalmia or a similar disease are shown by green circles, where the intensity of green is proportional to the degree of similarity. The target disease gene that is left out in the experiment and correctly ranked first by our algorithm is represented by a red circle. The gene that is incorrectly ranked first for both of the existing global approaches is shown by a diamond. Other candidate genes that are prioritized are shown by yellow circles.

cause the product of *SIX6* is not a centralized protein with a degree of only 1. Thus, random walk with restarts model ranks this true gene as 26^{th} and network propagation ranks it 16^{th} among 100 candidates. On the other hand, our method is able to correctly rank this gene as the 1^{st} candidate. Both random walk and network propagation rank the gene *AKT1* top among all candidates, which, not surprisingly, is a high degree node (78), also connected to other hub gene products.

5 Conclusion

In this paper, we have shown that approaches based on global network properties in prioritizing disease-associated genes are highly biased by the degree of the candidate gene, thus perform poorly in detecting loosely connected disease genes. We proposed several statistical adjustment strategies that improve the performance, particularly in identifying loosely connected disease genes. We have shown that, when these adjustment schemes are used together with existing global methods, the resulting method outperforms existing approaches significantly. These results clearly demonstrate that, in order to avoid exacerbation of ascertainment bias and propagation of noise, network-

Table 3. Comparison of the proposed method with existing global approaches. The proposed method outperforms others with respect to all performance criteria.

METHOD	Avg. Rank	AUROC	Perc. Ranked in top 1%	Perc. Ranked in top 5%
Proposed Hybrid Method	23.22	0.76	21.7	45.1
Network propagation	26.32	0.74	18.2	43.2
Random walk w/ restarts	28.02	0.73	20.7	43.9

based biological inference methods have to be supported by statistical models that take into account the degree distribution.

6 Acknowledgements

This work is supported in part by NSF CAREER Award CCF-0953195.

References

1. Brunner, H.G., van Driel, M.A.: From syndrome families to functional genomics. Nat Rev Genet **5**(7) (2004) 545–551
2. Glazier, A.M., Nadeau, J.H., Aitman, T.J.: Finding Genes That Underlie Complex Traits. Science **298**(5602) (2002) 2345–2349
3. Lage, K., Karlberg, E., Storling, Z., Olason, P., Pedersen, A., Rigina, O., Hinsby, A., Tumer, Z., Pociot, F., Tommerup, N., Moreau, Y., Brunak, S.: A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Bio. **25**(3) (2007) 309–316
4. Adie, E., Adams, R., Evans, K., Porteous, D., Pickard, B.: SUSPECTS: enabling fast and effective prioritization of positional candidates. Bioinformatics **22**(6) (2006) 773–774
5. Turner, F., Clutterbuck, D., Semple, C.: Pocus: mining genomic sequence annotation to predict disease genes. Genome Biology **4**(11) (2003) R75
6. Chen, J., Aronow, B., Jegga, A.: Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics **10**(1) (2009) 73
7. Oti, M., Snel, B., Huynen, M.A., Brunner, H.G.: Predicting disease genes using protein-protein interactions. J Med Genet (2006) jmg.2006.041376
8. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.L.A.A.: The human disease network. PNAS **104**(21) (2007) 8685–8690
9. Ideker, T., Sharan, R.: Protein networks in disease. Genome research **18**(4) (2008) 644–652
10. Köhler, S., Bauer, S., Horn, D., Robinson, P.N.: Walking the interactome for prioritization of candidate disease genes. Am J Hum Genet **82**(4) (2008) 949–958
11. Vanunu, O., Sharan, R.: A propagation based algorithm for inferring gene-disease associations. Proceedings of German Conference on Bioinformatics (2008)
12. Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., Gerstein, M.: Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends in Genetics **18**(10) (2002) 529–536
13. George, R.A., Liu, J.Y., Feng, L.L., Bryson-Richardson, R.J., Fatkin, D., Wouters, M.A.: Analysis of protein sequence and interaction data for candidate disease gene prediction. Nucl. Acids Res. **34**(19) (2006) e130–
14. van Driel, M.A., Bruggeman, J., Vriend, G., Brunner, H.G., Leunissen, J.A.: A text-mining analysis of the human phenome. EJHG **14**(5) (2006) 535–542
15. Lovász, L.: Random walks on graphs: A survey. Combinatorics, Paul Erdos is Eighty **2** (1996) 353–398
16. Tong, H., Faloutsos, C., Pan, J.Y.: Random walk with restart: fast solutions and applications. Knowledge and Information Systems **14**(3) (2008) 327–346
17. Macropol, K., Can, T., Singh, A.: Rrw: repeated random walks on genome-scale protein networks for local cluster discovery. BMC Bioinformatics **10**(1) (2009) 283
18. Tong, H., Faloutsos, C.: Center-piece subgraphs: problem definition and fast solutions. In: KDD '06: Proceedings of the 12th ACM SIGKDD, NY, USA, ACM (2006) 404–413
19. Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., Singh, M.: Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinf. **21** (2005) i302–310
20. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Computer Networks and ISDN Systems. Volume 30. (1998) 107–117
21. Maglott, D., Ostell, J., Pruitt, K.D., Tatusova, T.: Entrez Gene: gene-centered information at NCBI. Nucl. Acids Res. **35**(suppl-1) (2007) D26–31