Network-Based Disease Candidate Gene Prioritization: Towards Global Diffusion in Heterogeneous Association Networks

Joana P. Gonçalves[1,2,3,*], Sara C. Madeira[1,2], and Yves Moreau[3]

[1] Knowledge Discovery and Bioinformatics group (KDBIO), INESC-ID, Rua Alves Redol 9, 1000-029 Lisboa, Portugal
[2] IST, Technical University of Lisbon, Av. Rovisco Pais 1, 1049-001 Lisboa, Portugal
[3] BIOI, ESAT-SCD, Department of Electrical Engineering, KULeuven, Kasteelpark Arenberg 10, 3001 Leuven-Heverlee, Belgium
{jpg,smadeira}@kdbio.inesc-id.pt, yves.moreau@esat.kuleuven.be,

Abstract. Disease candidate gene prioritization addresses the association of novel genes with disease susceptibility or progression. Network-based approaches explore the connectivity properties of biological networks to compute an association score between candidate and disease-related genes. Although several methods have been proposed to date, a number of concerns arise: (i) most networks used rely exclusively on curated physical interactions, resulting in poor coverage of the Human genome and leading to sparsity issues; (ii) most methods fail to incorporate interaction confidence weights; (iii) in some cases, relevance scores are computed as local measures based on the direct interactions with the disease-related genes, ignoring potentially relevant indirect interactions. In this study, we seek a robust network-based strategy by evaluating the performance of selected prioritization strategies using genes known to be involved in 29 different diseases.

Keywords: protein-protein interaction, network, random walk, disease candidate genes, prioritization

1 Introduction

Biomarkers play a crucial role in modern medical practice as a means of improving accuracy in diagnosis, prognosis and treatment. In particular, research has been actively devising associations of novel genes with disease susceptibility or progression, relying on high-throughput technologies and the proliferation of accessible resources of biological data to enable large-scale genome-wide studies.

Most computational methods proposed for disease gene prioritization aim to identify putative candidates based on their similarity with genes known to be involved in the occurrence of a particular phenotype, according to: intrinsic properties, functional annotations, coherent transcriptional responses via expression data analysis, orthologous relations with genes from model organisms or even

co-occurrence in the literature [22]. Alternative strategies adopt a systemic approach and explore the topology of biological networks, including protein-protein interactions, regulatory data or metabolic pathways. These approaches rely on the assumption that genes co-occurring in a particular network substructure or interacting tend to participate together in related biological processes to identify novel genes based on their linkage with the known disease genes [22].

Integrative network-based analysis has been addressed [8,11,15,16,20,23,26], combining knowledge from distinct resources in association networks to unravel novel disease genes. However, most of these approaches rely solely on physical interactions [8,23], potentially inferred via orthologous relations with model organisms [26], often resulting in insufficient coverage of the Human genome. Others include additional interactions predicted from coexpression, pathway, functional or literature data, but still devise sparse networks [11,15]. Although the risk for false positive interactions may rise, the integration of knowledge from heterogeneous sources generates denser networks which tend to be less biased toward a particular evidence, more robust to noise and thus able to perform better in the prioritization task [16].

Network-based prioritization methods further differ in how they define the ranking of the candidates from the known disease-related genes. Local measures are usually computed based on the direct links or shortest paths between the candidates and the disease-related genes [15, 16], while global strategies diffuse or smooth a disease-related signal through the network. In this work, we evaluate whether the latter should be preferred over the former, as the inclusion of indirect associations is able to compensate for missing linkage, ultimately mitigating sparsity and "small world" effect issues [20], and global similarities have recently been shown to outperform local measures [15].

Random walks or diffusion kernels arise as natural candidates for the diffusion approach and their application to prioritization has been proven effective [6, 8, 15, 23]. Not only they compute fast using iterative methods, even for large networks [6], they are also able to straightforwardly establish a ranking of the candidates based on the global connectivity of the network. Nevertheless, some of the proposed methods [8, 15] ignore or fail to incorporate weights expressing the confidence on the evidence of every particular association [16]. Furthermore, their scores are based on the steady-state probability obtained after a large number of iterations or upon convergence. In this study, we assess the claim that limited diffusion is usually sufficient for ranking purposes [9, 10] and on our intuition which leads us to expect the prior knowledge to be somehow lost or of very little importance to the ranking after diffusing to a large extent.

Throughout this paper, we address the aforementioned topics by analyzing the performance of different prioritization strategies in three case studies: (i) Integrative heterogeneous protein association network *vs* integrative protein-protein physical interaction network (PPPIN); (ii) Global ranking measure *vs* local ranking measure; (iii) Confidence weights, degree of diffusion and parameter variation.

2 Methods

A protein-protein association network can be described as a weighted undirected graph, a special case of a weighted directed graph, defined as $G = (V, E)$, where V is the set of vertices and E is the set of edges. Each vertex in V and edge in E correspond to a gene and an association between two genes, respectively. Let A and D denote the adjacency and diagonal matrices of G, respectively. A_{uv} is the weight $w(u, v)$ of the edge (u, v) between source u and target v. Also, $D_{uu} = \sum_{(u,v)\in E} A_{uv}, \forall u \in V$, that is, the sum of the weights of the edges for which u is the source. Prioritizing disease candidates thus formulates as obtaining a ranking on V given a set $S \in V$ of seed genes. For the local scoring scheme Endeavour's measure was used [2]. As global network-based strategies, the PageRank with priors and Heat Diffusion random walks were applied: an initial signal expressing the relevance of the genes in the context of the disease in the form of a preference vector, $p^{(0)}$, is diffused over the network by performing a limited number of iterations, N.

2.1 Endeavour's Measure: Intersection of Interactors

Endeavour computes a local network-based measure, whereby the score of each gene is computed as the overlap between the sets of genes interacting with the seed genes and those interacting with the candidate gene itself [2]:

$$S_v = \sum_{(u,v)\in E} intSeeds(u) \qquad intSeeds(u) = \begin{cases} 1 & \text{, if } \exists z \in S : (u, z) \in E \\ 0 & \text{, otherwise} \end{cases}$$

2.2 Heat Diffusion and PageRank with Priors

Heat Diffusion is a discrete approximation of the heat kernel [28] first introduced in [9], in which the rate of diffusion is controlled by a non-negative parameter, the heat diffusion coefficient t. The iterative equation is given by

$$p_v^{(i+1)} = \left(1 - \frac{t}{N}\right) \cdot p_v^{(i)} + \frac{t}{N} \sum_{(u,v)\in E} p_u^{(i)} \cdot \frac{A_{uv}}{D_{uu}}.$$

PageRank with priors is an extension of the original PageRank algorithm to consider the original probability distribution of the scores [25]. A parameter β, called "back probability" expresses the probability of jumping to the initial node at each iteration. The iterative equation is

$$p_v^{(i+1)} = \beta \cdot p_v^{(0)} + (1 - \beta) \cdot \sum_{(u,v)\in E} p_u^{(i)} \frac{A_{uv}}{D_{uu}}.$$

3 Results

Evaluation sudies were performed using Human data from the STRING database [12] and a PPPIN from Entrez Gene [1] as representatives of protein-protein heterogeneous association and physical interaction networks, respectively. 620 genes known to be related with 29 diseases were used as prior knowledge to prioritize candidates in a leave-one-out cross-validation scheme.

3.1 Data and Preprocessing

Networks The STRING database [12, 18] integrates physical interactions and predicted associations based on knowledge obtained from heterogeneous sources of transcriptional, functional, metabolic, literature and orthology data. For a fair comparison with Endeavour, we downloaded and parsed version 7.1 of STRING [18], including evidences from MINT [7], HPRD [19], BIND [4], DIP [27], BioGRID [5], KEGG [13] and Reactome [24] databases. Associations from STRING v8.2 [12] were also retrieved to assess to which extent the additional knowledge integrated from IntAct [14], PID [21] and GO [3] protein complexes would improve the prioritization performance relative to the previous release. A PPPIN was downloaded from the NCBI Entrez Gene FTP repository [1]. 130797 Human interactions were selected from 448534 entries, for which both interactant genes were tagged with tax ID 9606. From these, 4611, 51275 and 74911 were originally from BIND [4], BioGRID [5] and HPRD [19], respectively. Genes' identifiers followed Entrez Gene nomenclature. Preprocessing of these networks involved filtering redundant edges and devising an explicit representation of a directed graph. In the case of the STRING releases, original weights were used to express the confidence of every association, while in the PPPIN all edges were attributed weight 1. STRING v7.1 contained 16050 genes and 698534 unique associations. STRING v8.2 covered 17448 Human genes with 1256016 non-redundant associations. Finally, the PPPIN had 47873 physical interactions between 10175 genes.

Seed sets 620 disease genes were selected from the OMIM [17] database spanning 29 disease-specific sets, with an average of 21 genes per set. As genes were identified according to Ensembl nomenclature, the seeds could be directly used with STRING. For the PPPIN, however, we performed a conversion between Ensembl and Entrez Gene identifiers. A mapping was parsed from a file downloaded from the NCBI Entrez Gene FTP repository [1] and used to generate the corresponding seed sets using Entrez Gene names. Additionally, we filtered the genes absent from at least one of the networks or for which the conversion between Ensembl and Entrez did not succeed. In total, 94 seed genes were lost (14 with no conversion, 80 absent from the PPPIN). A single occurrence of a gene with several Entrez aliases happened. In this case, only the alias present in the PPPIN was kept. For validation purposes, seed sets containing randomly selected genes were generated. The number of seeds in each set was randomly chosen in the range $[5, 100]$ and the genes were randomly selected from the Human STRING v8.2 network. 546 genes were retrieved.

3.2 Evaluation Measures and Experimental Setting

Evaluation measures Ideally, in a leave-one-out cross-validation scheme, we would expect the prioritization strategy to rank the left-out gene known to be related with the disease at the top. Under this assumption, we assess the performance of the scoring methods overall and per disease based on four evaluation measures: the number of left-out genes ranked in the top 10 and 20 positions, the Area under the ROC curve (AUC) score, and the mean average precision.

For a given combination of diffusion parameter α and number of iterations N, n rankings are generated (one per left-out gene). The AUC score is given by

$$S_{AUC_{\alpha,N}} = \frac{n - \sum_{k=1}^{n} \frac{r_k^{(N)}}{m_k^{(N)}}}{n},$$

where $r_k^{(N)}$ is the ranking position of the k^{th} left-out gene in the k^{th} ranked list and $m_k^{(N)}$ is the number of ranked genes in the k^{th} list.

Mean average precision (MAP) is an evaluation measure that combines precision and recall. Essentially, MAP averages the precisions computed by truncating the list after each of the relevant entities is found. Only one relevant entity must be found, the left-out gene. Thus, precision at rank r is either 0, before it has been found, or $\frac{1}{r}$. Moreover, in our setting the ranked lists contain equal number of genes, allowing us to simplify our MAP score for n lists with the same size to:

$$S_{MAP_{\alpha,N}} = \frac{\sum_{k=1}^{n} \frac{1}{r_k^{(N)}}}{n}$$

Experimental setting In each validation run, one different gene was deleted from the set of seed genes and added to 99 randomly selected candidate genes. A ranking method was then applied to compute a score for every gene in the network. Finally, the ranking of the 100 candidate genes was defined according to the retrieved scores. In the case of the Heat Diffusion, the scores of the seed genes were initialized to 1. For PageRank, an initial seed score of $1/|S|$ was used. Performance was assessed by computing AUC and MAP scores, and counting the number of left-out genes ranked in the top 10 and top 20 positions, both overall and per disease. We sought the best performance of each method using several combinations of parameters. Heat diffusion coefficients t and back probabilities β of 0.1, 0.3, 0.5, 0.7 and 0.9 with 2, 5, 10, 15 and 20 iterations using STRING and 2, 5, 10, 20, 100 iterations using the PPPIN were tried. In the case studies, results are shown only for the parameter settings which achieved the best performance in each case. We further ranked the randomly generated seed sets using the leave-one-out cross-validation in STRING v8.2 to assess whether the Heat Diffusion method was able to take advantage of the information contained in the seed sets to improve the identification of the left-out seeds. Overall, AUC and MAP scores of 0.501 and 0.05 were achieved and only 57 and 92 genes were ranked in the top 10 and 20 positions. Similar results were obtained per seed set (data not shown), in accordance with what would have been expected for random seed sets.

3.3 Case Studies

Heat Diffusion and PageRank with priors achieved similar results in both networks (Table 1). For this reason, we abstain ourselves of comparing the results of both random walks, considering the results equivalent when applied to the same network. Throughout this section, we will always refer to one of them as a representative of a global measure. A brief description of the prioritization performances obtained for each case study follows.

Method	Network	Parameters	AUC	MAP	TOP 10	TOP 20	#BRM	#BRN
HeatDiffusion	STRING8	$t = 0.3, N = 10$	0.962	0.711	484	502	26%	68%
PageRank	STRING8	$\beta = 0.7, N = 2$	0.961	0.693	485	502	20%	69%
HeatDiffusion	PPPIN	$t = 0.5, N = 2$	0.862	0.352	301	373	40%	11%
PageRank	PPPIN	$\beta = 0.5, N = 2$	0.861	0.349	304	384	38%	10%

Table 1. Results of Heat Diffusion and PageRank using both STRING v8.2 and the PPPIN. '#BRM' (better ranked by method in each network) shows the percentage of genes with a higher rank in a one-to-one comparison of the ranks per gene for both methods in each network. '#BRN' (better ranked by network for each method) shows the percentage of genes with a higher rank in a one-to-one comparison of the ranks per gene for both networks using the same method. Total number of genes: 526.

Global measure *vs* Local measure A network-based global ranking was obtained using the Heat Diffusion method with $t = 0.3, N = 10$, while Endeavour [2] was used to score the genes using its local measure. Both rankings were based on STRING v7.1, the version included in Endeavour. Overall, the random walk global measure outperformed the local interaction overlap in all evaluation measures (see Table 2), that is, the higher number of left-out genes was ranked on the top positions, also achieving better ranks in general, using the latter.

Method	Network	AUC	MAP	TOP 10	TOP 20
HeatDiffusion ($t = 0.3, N = 10$)	STRING v7.1	0.942	0.643	536	569
Endeavour	STRING v7.1	0.806	0.326	393	464

Table 2. Overall results of Heat Diffusion and Endeavour using STRING v7.1. Total number of genes: 620.

Regarding the AUC scores per disease (see Table 3), the Heat Diffusion method outperformed Endeavour in all diseases except Ehlers-Danlos syndrome (0.944 opposed to 0.948, respectively). This was also the only disease for which

the number of genes ranked in the top 20 positions was higher using the local measure (Endeavour was able to rank one more gene in the top 20). However, the MAP score was better for the Heat Diffusion method and, in fact, 9 of the 10 seed genes ranked in the top 10 positions by both methods scored higher using the global measure.

For the remaining diseases, Heat Diffusion was always able to rank the same or a higher number of genes in both the top 10 and the top 20 positions. Regarding the MAP scores, Heat Diffusion outperformed Endeavour in every disease and was able to rank all genes of both amyotrophic lateral sclerosis and Usher syndrome in the first position.

		Heat Diffusion STRING v7.1				Endeavour STRING v7.1			
Disease	#Genes	AUC	MAP	Top10	Top20	AUC	MAP	Top10	Top20
Alzheimer's disease	8	0.934	0.586	7	7	0.930	0.376	6	7
amyotrophic lateral sclerosis	4	0.990	1.000	4	4	0.975	0.550	4	4
anemia	44	0.928	0.499	36	40	0.718	0.187	21	30
breast cancer	24	0.930	0.608	21	22	0.782	0.214	13	19
cardiomyopathy	22	0.973	0.812	21	21	0.862	0.579	18	18
cataract	20	0.890	0.693	15	16	0.883	0.363	13	17
Charcot-Marie-Tooth disease	14	0.889	0.752	12	12	0.738	0.361	8	8
colorectal cancer	21	0.961	0.697	19	20	0.918	0.389	17	20
deafness	42	0.941	0.642	37	40	0.732	0.186	17	25
diabetes	26	0.967	0.731	22	26	0.820	0.232	17	21
dystonia	5	0.986	0.867	5	5	0.938	0.381	4	5
Ehlers-Danlos syndrome	10	0.944	0.650	9	9	0.948	0.296	9	10
emolytic anemia	13	0.965	0.683	12	12	0.737	0.269	8	8
epilepsy	15	0.989	0.933	15	15	0.749	0.612	10	10
ichthyosis	9	0.881	0.598	8	8	0.778	0.226	6	6
leukemia	112	0.922	0.428	88	100	0.807	0.203	68	86
lymphoma	31	0.920	0.420	24	25	0.796	0.275	19	22
mental retardation	24	0.918	0.629	21	21	0.624	0.110	7	11
muscular dystrophy	24	0.981	0.780	24	24	0.869	0.390	19	21
myopathy	41	0.961	0.594	37	39	0.885	0.535	34	34
neuropathy	18	0.965	0.671	14	17	0.648	0.205	8	9
obesity	13	0.931	0.796	12	12	0.918	0.559	12	12
Parkinson's disease	9	0.903	0.728	7	7	0.661	0.158	4	4
retinitis pigmentosa	30	0.957	0.882	27	28	0.845	0.470	22	23
spastic paraplegia	7	0.930	0.860	6	6	0.927	0.586	5	6
spinocerebellar ataxia	7	0.959	0.863	6	6	0.816	0.250	3	6
Usher syndrome	8	0.990	1.000	8	8	0.988	0.917	8	8
xeroderma pigmentosum	10	0.987	0.850	10	10	0.785	0.704	7	7
Zellweger syndrome	9	0.989	0.944	9	9	0.823	0.513	6	7

Table 3. Results of the Heat Diffusion ($t = 0.3$, 10 iterations) and Endeavour methods using STRING v7.1, per disease. Total number of genes: 620.

Protein-Protein Associations *vs* Protein-Protein Physical Interactions Heat Diffusion achieved better performance using STRING v8.2, with AUC score 0.962, opposed to 0.862 using the PPPIN (see Table 1). Furthermore, STRING enabled to rank more than 90% of the genes in the top 10 positions, while using the PPPIN less than 60% were in top 10. In a one-to-one comparison, Heat Diffusion ranked 68% of the genes better using STRING, while only 11% of the ranks were better using the PPPIN. Table 4 compares the results obtained for the Heat Diffusion method using STRING v8.2 with PageRank with priors in a PPPIN, one of the best performing strategies in [8], per disease.

		Heat Diffusion STRING v8.2				PageRank PPPIN			
Disease	#Genes	AUC	MAP	Top10	Top20	AUC	MAP	Top10	Top20
Alzheimer's disease	8	0.929	0.877	7	7	0.668	0.456	5	5
amyotrophic lateral sclerosis	4	0.990	1.000	4	4	0.530	0.028	0	1
anemia	37	0.967	0.599	35	36	0.679	0.268	15	19
breast cancer	22	0.952	0.618	20	20	0.877	0.427	17	18
cardiomyopathy	19	0.986	0.904	19	19	0.789	0.383	12	13
cataract	16	0.980	0.781	16	16	0.751	0.485	10	11
Charcot-Marie-Tooth disease	10	0.934	0.735	9	9	0.665	0.251	3	3
colorectal cancer	29	0.969	0.785	19	19	0.912	0.382	15	19
deafness	28	0.950	0.623	23	27	0.547	0.210	7	8
diabetes	25	0.966	0.743	23	24	0.838	0.422	17	20
dystonia	5	0.986	0.800	5	5	0.700	0.316	2	2
Ehlers-Danlos syndrome	8	0.990	1.000	8	8	0.850	0.613	6	7
emolytic anemia	12	0.978	0.772	12	12	0.793	0.149	4	6
epilepsy	13	0.989	0.962	13	13	0.803	0.454	8	8
ichthyosis	7	0.954	0.768	6	6	0.651	0.367	3	3
leukemia	98	0.948	0.520	86	93	0.811	0.209	50	67
lymphoma	26	0.930	0.476	21	22	0.850	0.270	15	18
mental retardation	19	0.926	0.727	16	17	0.739	0.303	8	12
muscular dystrophy	20	0.983	0.790	20	20	0.893	0.524	15	15
myopathy	35	0.969	0.702	33	35	0.731	0.272	20	24
neuropathy	17	0.951	0.699	15	15	0.636	0.201	5	8
obesity	12	0.988	0.917	12	12	0.892	0.621	10	10
Parkinson's disease	8	0.935	0.878	7	7	0.754	0.465	5	5
retinitis pigmentosa	23	0.981	0.883	22	23	0.736	0.310	11	12
spastic paraplegia	5	0.990	1.000	5	5	0.490	0.083	1	1
spinocerebellar ataxia	7	0.957	0.768	6	6	0.726	0.095	3	4
Usher syndrome	4	0.990	1.000	4	4	0.880	0.631	3	3
xeroderma pigmentosum	10	0.988	0.900	10	10	0.980	0.811	10	10
Zellweger syndrome	8	0.990	1.000	8	8	0.871	0.814	7	7

Table 4. Heat Diffusion using STRING v8.2 ($t = 0.3$, $N = 10$) vs PageRank with priors using the PPPIN ($\beta = 0.5$, $N = 2$), per disease. Total number of genes: 526.

Regarding the disease-specific scores (see Table 4), the lowest AUC (and MAP) values for the combination Heat Diffusion and STRING v8.2 were of 0.926 (0.727) for mental retardation, and 0.930 (0.476) for lymphoma, which are still good results. For five diseases, namely amyotrophic lateral sclerosis, Ehlers-Danlos syndrome, spastic paraplegia, Usher syndrome and Zellweger syndrome, the heterogeneous association network approach was actually able to rank all the seed genes in the first position of the ranking. On the other hand, the PageRank diffusion in the PPPIN achieved AUC scores above 0.9 only for two diseases: colorectal cancer with 0.912 and xeroderma pigmentosum with 0.98. The lowest AUC and MAP scores were obtained for amyotrophic lateral sclerosis (0.53 and 0.028) and spastic paraplegia (0.49 and 0.083). The PPPIN strategy could not rank any of the seed genes for amyotrophic lateral sclerosis in the top 10 positions and only one was identified in the first 20. Also, only one gene out of the 5 seeds for spastic paraplegia was ranked in the top 10/20. In this case, the performance for both diseases is comparable to the one obtained using the random seed sets (data now shown).

Confidence weights, number of iterations and diffusion rate We assessed the contribution of STRING's weights expressing the degree of confidence in the associations between genes to the performance of the prioritization method by diffusing the initial preference vector using the filtered disease-specific seed sets on the network after setting all associations' weights to 1. Although the resulting AUC and MAP scores (0.957 and 0.662) were not substantially different from the ones obtained using the confidence weights (0.962 ans 0.711), they actually reflected in less 9 genes ranked in the top 10 (data not shown). Overall, the number of genes in the top 20 was the same, with slight variations per disease. From the five diseases achieving maximum performance in the differentially association weighted setting, only for Ehlers-Danlos syndrome, spastic paraplegia and Zellweger syndrome these results could be maintained.

In both random walk approaches, the best results were achieved using a limited number of iterations. STRING v8.2 provided consistent and stable performance when varying the number of diffusion steps. On the PPPIN, the best ranking was always obtained using two iterations. It would then stabilize for larger numbers of steps, although measuring considerably lower in the evaluation, since it was never able to rank more than 289 or 346 genes - out of 526 - in the top 10 and top 20, respectively.

Regarding the parameter controlling the rate of diffusion, the Heat Diffusion method delivered quite similar performance for the set of heat coefficients tried: in STRING v8.2, resulting in AUC scores ranging from 0.960 to 0.962 for each diffusion coefficient, considering equal number of iterations; in the PPPIN, AUC scores ranging between 0.859 and 0.862 with 2 iterations, $N = 2$, and between 0.766 and 0.771 using 5, 10, 20 and 100 iterations.These results indicate its robustness to variations in this parameter. For PageRank with priors, the impact of the back probability value was not neglegible. For the lowest back probabilities (0.01 and 0.05) the scores were unstable leading to considerable performance

variations, even using STRING v8.2. For $\beta = \{0.1, 0.3, 0.5, 0.7, 0.9\}$, the PageRank AUC scores in STRING v8.2 varied between 0.936 and 0.961 considering the results obtained using the same number of iterations. In the PPPIN, PageRank obtained AUC scores between 0.859 and 0.861 using 2 iterations and ranging between 0.758 and 0.775 using 5, 10, 20 and 100 iterations.

4 Conclusions

Prioritization results confirmed our hypothesis that networks integrating gene associations retrieved or predicted using data from heterogeneous sources should be in general more informative and potentially able to perform better in the identification of genes associated with a given disease when compared to networks containing only physical interactions. Advantages of the former are supported by three key observations: (1) associations derived from the combination of several types of evidence should be more reliable and accurate; (2) heterogeneous data integration enables a better coverage of the genome and larger network density, confering robustness to noise; (3) confidence weights can be devised in order to differentiate associations and mitigate the impact of false positive associations, particularly when based on a limited number of sources.

Nevertheless, our analysis shows that heterogeneous association networks do not present sufficient guarantee for maximum performance by themselves. In fact, the network-based score measuring the degree of relatedness of each candidate gene with a given disease based on a set of known disease-related genes proved to play a major role. Essentially, based on the results we could conclude that in comparison to neighborhood-limited scores a network-based measure able to capture global connectivity properties by considering indirect associations between genes is not only (1) more robust, as it compensates for the sparsity related to direct associations and tackles the "small world" effect issue; but also (2) more informative, deriving a score based on a systemic view of the interactome. This claim has also been previously hinted at in [15, 16].

Propagation schemes tested in the computation of global network-based scores diffused an initial preference vector expressing the distribution of the known disease-related genes through the network using random walks. These methods compute fast using iterative procedures, even for large networks. Furthermore, we could verify that in the context of prioritization in association or physical interaction networks the maximum performance can be achieved using only a limited number of iterations. Heat Diffusion and PageRank with priors delivered high quality results and achieved similar performance under appropriate parameter settings, supporting the claim of equivalence [8, 25] for other approaches of the same kind, namely HITS with priors and K-Step Markov. The importance of confidence weights was inconclusive, as the difference in performance exhibited by our experiments was residual. We believe, however, that appropriate association confidence weights may improve accuracy of network-based prioritization results.

Acknowledgments This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds. JPG is the recipient of a doctoral grant supported by FCT (SFRH/BD/36586/2007).

References

1. NCBI Entrez Gene FTP Repository (Jan 2010), ftp://ftp.ncbi.nih.gov/gene/
2. Aerts, S., Van Loo, P., De Smet, F., Lambrechts, D., Maity, S., Tranchevent, L.C., De Moor, B., Coessens, B., Marynen, P., Hassan, B., Carmeliet, P., Moreau, Y.: Gene prioritization through genomic data fusion. Nature Biotechnology 24(5), 537–44 (2006)
3. Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Others: Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 2529 (2000)
4. Bader, G.D., Betel, D., Hogue, C.W.V.: BIND: the Biomolecular Interaction Network Database. Nucleic Acids Research 31(1), 248–250 (2003)
5. Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V., Dolinski, K., Tyers, M.: The BioGRID Interaction Database: 2008 update. Nucleic Acids Research 36(suppl_1), D637–640 (2008)
6. Can, T., Çamoğlu, O., Singh, A.K.: Analysis of protein-protein interaction networks using random walks. In: Proceedings of the 5th International Workshop on Bioinformatics - BIOKDD '05. p. 61. ACM Press, New York, New York, USA (2005)
7. Chatr-Aryamontri, A., Zanzoni, A., Ceol, A., Cesareni, G.: Searching the protein interaction space through the MINT Database. Methods in Molecular Biology 484, 305–317 (2008)
8. Chen, J., Aronow, B.J., Jegga, A.G.: Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics 10, 73 (2009)
9. Chung, F., Yau, S.: Coverings, heat kernels and spanning trees. Electronic Journal of Combinatorics 6, R12 (1999)
10. Francisco, A.P., Gonçalves, J.P., Madeira, S.C., Oliveira, A.L.: Using personalized ranking to unravel relevant regulations in the Saccharomyces cerevisiae regulatory network. In: Jornadas de Bioinformática 2009. Lisbon, Portugal (2009)
11. Franke, L., Bakel, H., Fokkens, L., Jong, D., E.d, Egmont-petersen, M., Wijmenga, C.: Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. The American Journal of Human Genetics 78, 1011–1025 (2006)
12. Jensen, L.J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., Bork, P., von Mering, C.: STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research 37(Database issue), D412–6 (2009)
13. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., Hattori, M.: The KEGG resource for deciphering the genome. Nucleic Acids Research 32(suppl_1), D277–280 (2004)
14. Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., Kohler, C., Khadake, J., Leroy, C., Liban, A., Lieftink, C., Montecchi-Palazzi, L., Orchard, S., Risse, J., Robbe, K., Roechert, B., Thorneycroft, D., Zhang, Y., Apweiler, R., Hermjakob, H.:

IntAct–open source resource for molecular interaction data. Nucleic Acids Research 35(suppl_1), D561–565 (2007)
15. Köhler, S., Bauer, S., Horn, D., Robinson, P.: Walking the interactome for prioritization of candidate disease genes. The American Journal of Human Genetics 82(4), 949958 (2008)
16. Linghu, B., Snitkin, E.S., Hu, Z., Xia, Y., Delisi, C.: Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. Genome biology 10(9), R91 (2009)
17. McKusick, V.A.: Mendelian Inheritance in Man and Its Online Version, OMIM. The American Journal of Human Genetics 80(4), 588–604 (2007)
18. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Krüger, B., Snel, B., Bork, P.: STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Research 35(Database issue), D358–62 (2007)
19. Mishra, G.R., Suresh, M., Kumaran, K., Kannabiran, N., Suresh, S., Bala, P., Shivakumar, K., Anuradha, N., Reddy, R., Raghavan, T.M., Menon, S., Hanumanthu, G., Gupta, M., Upendran, S., Gupta, S., Mahesh, M., Jacob, B., Mathew, P., Chatterjee, P., Arun, K.S., Sharma, S., Chandrika, K.N., Deshpande, N., Palvankar, K., Raghavnath, R., Krishnakanth, R., Karathia, H., Rekha, B., Nayak, R., Vishnupriya, G., Kumar, H.G.M., Nagini, M., Kumar, G.S.S., Jose, R., Deepthi, P., Mohan, S.S., Gandhi, T.K.B., Harsha, H.C., Deshpande, K.S., Sarker, M., Prasad, T.S.K., Pandey, A.: Human protein reference database–2006 update. Nucleic Acids Research 34(suppl_1), D411–414 (2006)
20. Nitsch, D., Tranchevent, L.C., Thienpont, B., Thorrez, L., Van Esch, H., Devriendt, K., Moreau, Y.: Network analysis of differential expression for the identification of disease-causing genes. PloS ONE 4(5), e5526 (2009)
21. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: PID: the Pathway Interaction Database. Nucleic Acids Research 37(suppl_1), D674–679 (2009)
22. Tiffin, N., Andrade-Navarro, M.A., Perez-Iratxeta, C.: Linking genes to diseases: it's all in the data. Genome Medicine 1(8), 77 (Jan 2009)
23. Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., Sharan, R.: Associating Genes and Protein Complexes with Disease via Network Propagation. PLoS Computational Biology 6(1) (2010)
24. Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E., Stein, L.: Reactome: a knowledge base of biologic pathways and processes. Genome Biology 8(3), R39 (2007)
25. 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. pp. 266–275. ACM, New York, NY, USA (2003)
26. Wu, X., Jiang, R., Zhang, M.Q., Li, S.: Network-based global inference of human disease genes. Molecular Systems Biology 4(189), 189 (2008)
27. Xenarios, I., Rice, D.W., Salwinski, L., Baron, M.K., Marcotte, E.M., Eisenberg, D.: DIP: the Database of Interacting Proteins. Nucleic Acids Research 28(1), 289–291 (2000)
28. Yang, H., King, I., Lyu, M.: Diffusionrank: a possible penicillin for web spamming. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 438. ACM (2007)