A Propagation-based Algorithm for Inferring Gene-Disease Associations

Oron Vanunu* Roded Sharan*

Abstract: A fundamental challenge in human health is the identification of disease-causing genes. Recently, several studies have tackled this challenge via a two-step approach: first, a linkage interval is inferred from population studies; second, a computational approach is used to prioritize genes within this interval. State-of-the-art methods for the latter task are based on the observation that genes causing the same or similar diseases tend to lie close to one another in a network of protein-protein or functional interactions. However, most of these approaches use only local network information in the inference process. Here we provide a global, network-based method for prioritizing disease genes. The method is based on formulating constraints on the prioritization function that relate to its smoothness over the network and usage of prior information. A propagation-based method is used to compute a function satisfying the constraints. We test our method on gene-disease association data in a cross-validation setting, and compare it to extant prioritization approaches. We show that our method provides the best overall performance, ranking the true causal gene first for 29% of the 1,369 diseases with a known gene in the OMIM knowledgebase.

1 Introduction

Associating genes with diseases is a fundamental challenge in human health with applications to understanding disease mechanisms, diagnosis and therapy. Linkage studies are often used to infer genomic intervals that are associated with a disease of interest. Prioritizing genes within these intervals is a formidable challenge and computational approaches are becoming the method of choice for such problems. Prioritization methods are based on comparing a candidate gene to other genes that were implicated with the same or a similar disease. Recently, several methods were suggested that use physical network information for the prioritization task, and these were shown to outperform other approaches to the problem. The basic paradigm underlying these methods is that genes causing the same or a similar disease tend to lie close to one another in a protein-protein interaction (PPI) network.

Previous approaches to prioritizing disease-causing genes can be roughly classified according to whether prior knowledge on some of the genes (or genomic intervals) underlying a disease of interest is assumed or not. Approaches in the first category are based on computing the similarity between a given gene and the known disease genes. Such a similarity can be based on sequence [G^+06], functional annotation [PIBAN07], protein-protein

*School of Computer Science, Tel Aviv University, Tel Aviv 69978. Email: {oronv,roded}@post.tau.ac.il.

interactions [O$^+$06, F$^+$06] and more. The reader is referred to [OB07] for a comprehensive review of these methods.

Approaches in the second category, which is the focus of the current work, are often based on a measure of phenotypic similarity (see, e.g., [vD$^+$06, L$^+$07]) between the disease of interest and other diseases for which causal genes are known. Lage et al. [L$^+$07] scores a candidate protein w.r.t. a disease of interest based on the involvement of its direct network neighbors in a similar disease. Kohler et al. [K$^+$08] group diseases into families. For a given disease, they employ a random walk from known genes in its family to prioritize candidate genes. Finally, Wu et al. [W$^+$08] score a candidate gene g for a certain disease d based on the correlation between the vector of similarities of d to diseases with known causal genes, and the vector of closeness in a protein network of g and those known disease genes.

In this work we suggest a global, network-based approach for predicting disease-causing genes. Our method falls under the second category and is able to exploit information on known genes for the disease of interest or for other similar diseases. The method receives as input a disease-disease similarity measure and a network of protein-protein interactions. It uses a propagation-based algorithm to infer a strength-of-association function that is smooth over the network (i.e., adjacent nodes are assigned similar values) and exploits the prior information (on causal genes for the same disease or similar ones).

Methodologically, we make a three-fold contribution: (i) we suggest a transformation from textual-based disease similarity values to confidence values that are learned automatically from data and better captures the probability that similar diseases share genes that lie close to one another in the network; (ii) we provide a propagation-based method for gene prioritization that takes into account, in a global manner, confidence values for disease similarity and a PPI network in which interactions are weighted by their reliability and the degrees of their end points. (iii) we re-implement three state-of-the-art methods and perform a comprehensive comparison between those methods and ours on the same input data.

On the practical side, we apply our method to analyze disease-gene association data from the Online Mendelian Inheritance in Man (OMIM) [H$^+$02] knowledgebase. We test, in a cross-validation setting, two possible applications of our method: (i) prioritizing genes for diseases with at least two known genes; (ii) prioritizing genes for all diseases (with at least one known gene). We compare the performance of our method to two state-of-the-art, recently published methods [K$^+$08, W$^+$08], as well as to a simple shortest-path based prioritization method. In all our tests the propagation-based method outperforms the other methods by a significant margin.

2 Our Algorithmic Approach

Preliminaries. The input to a gene prioritization problem consists of a set A of gene-disease associations; a query disease q; and a protein-protein interaction network $G = (V, E, w)$, where V is the set of proteins, E is the set of interactions and w is a weight function denoting the reliability of each interaction. The goal is to prioritize all the proteins in V (that are not known to be associated with q) w.r.t. q.

For a node $v \in V$, denote its direct neighborhood in G by $N(v)$. Let $F : V \rightarrow \Re$ represent a prioritization function, i.e., $F(v)$ reflects the relevance of v to q. Let $Y : V \rightarrow [0,1]$ represent a prior knowledge function, which assigns positive values to proteins that are known to be related to q, and zero otherwise.

Intuitively, we wish to compute a function F that is both smooth over the network, i.e., adjacent nodes are assigned with similar values, and also respects the prior knowledge, i.e., nodes for which prior information exists should have similar values of F and Y. These requirements often conflict with each other, e.g., when two adjacent nodes have very different Y values. Formally, we express the requirements on F as a combination of these two conditions:

$$F(v) = \alpha[\sum_{u \in N(v)} F(u)w'(v,u)] + (1-\alpha)Y(v) \qquad (1)$$

where w' is a normalized form of w, such that $\sum_{u \in N(v)} w'(v,u) \leq 1$ for every node $v \in V$. Here, the first term expresses the smoothness condition, while the second term expresses the prior information constraint. The parameter $\alpha \in (0,1)$ weighs the relative importance of these constraints w.r.t. one another.

Computing the prioritization function. The requirements on F can be expressed in linear form as follows:

$$F = \alpha W'F + (1-\alpha)Y \Leftrightarrow F = (I - \alpha W')^{-1}(1-\alpha)Y \qquad (2)$$

where W' is a $|V| \times |V|$ matrix whose values are given by w', and F and Y are viewed here as vectors of size $|V|$. Since W' is normalized, its eigenvalues are in $[0,1]$. Since $\alpha \in (0,1)$, the eigenvalues of $(I - \alpha W')$ are in $(0,1]$; in particular, all its eigenvalues are positive and, hence, $(I - \alpha W')^{-1}$ exists.

While the above linear system can be solved exactly, for large networks an iterative propagation-based algorithm works faster and is guaranteed to converge to the system's solution. Specifically, we use the algorithm of Zhou et al. [Z$^+$03] which at iteration t computes

$$F^t := \alpha W'F^{t-1} + (1-\alpha)Y$$

where $F^0 := 0$. This iterative algorithm can be best understood as simulating a process where nodes for which prior information exists pump information to their neighbors. In addition, every node propagates the information received in the previous iteration to its neighbors. In practice, as a final iteration we apply the propagation step with $\alpha = 1$ to smooth the obtained prioritization function F.

We chose to normalize the weight of an edge by the degrees of its end-points, since the latter relate to the probability of observing an edge between the same end-points in a random network with the same node degrees. Formally, define a diagonal matrix D such that $D(i,i)$ is the sum of row i of W. We set $W' = D^{-1/2}WD^{-1/2}$ which yields a symmetric matrix with row sums ≤ 1, where $W'_{ij} = W_{ij}/\sqrt{D(i,i)D(j,j)}$.

Incorporating disease similarity information. As observed by several authors [L^{+}07, OB07], similar diseases are often caused by proteins in the same complex or signalling pathway; therefore, such proteins tend to lie close to one another in the network. This empirical observation motivated us to use disease similarity information to determine the prior information vector Y.

We used the similarity metric computed by van Driel et al. [vD^{+}06], which spans $5,080$ diseases in the OMIM [H^{+}02] knowledgebase. Each disease entry in OMIM was scanned for terms taken from the anatomy (A) and the disease (C) sections of the medical subject headings vocabulary (MeSH). A disease was then represented by a binary vector specifying the terms associated with it. Similarity between diseases was computed by taking the cosine of the angle between the corresponding vectors.

van Driel et al. also tested the predictive power of different ranges of similarity values by calculating the correlation between the similarity of two diseases and the functional relatedness of their causative genes. According to their analysis, similarity values in the range $[0, 0.3]$ are not informative, while for similarities in the range $[0.6, 1]$ the associated genes show significant functional similarity. These empirical findings motivated us to represent our confidence that two diseases are related using a logistic function $L(x) = \frac{1}{1+e^{(cx+d)}}$. We constrained $L(0)$ to be close to zero (0.0001) which determines d (as $\log(9999)$), and tuned the parameter c using cross validation (see Parameter Tuning Section below). We used L to compute the prior knowledge Y in the following way: for a query disease q and a protein v associated with a disease d, we set $Y(v) := L(s)$, where s is the similarity between q and d. If v is associated with more than one disease, we set s to be the maximal similarity between q and any of those diseases.

3 Experimental Setup

We extracted $1,600$ known disease-protein associations from GeneCards[R^{+}97] spanning $1,369$ diseases and $1,043$ proteins. We considered only disease-protein relations that included proteins from the network and such that the relations are known to be causative to avoid associations made by circumstantial evidence.

We constructed a human PPI network with $9,998$ proteins and $41,072$ interactions that were assembled from three large scale experiments [R^{+}05, S^{+}05b, E^{+}07] and the Human Protein Reference Database (HPRD) [P^{+}04]. The interactions were assigned confidence scores based on the experimental evidence available for each interaction using a logistic regression model adapted from [S^{+}05a]. We used the obtained scores to construct the adjacency matrix W.

To simulate the case of prioritizing proteins encoded by genes inside a linkage interval, we followed [K^{+}08] and artificially constructed for each protein associated with a disease an interval of size 100 around it. We used the F values obtained from the output of the algorithm to prioritize proteins residing in that interval.

Comparison to other methods. In order to perform a comprehensive comparison of our approach to extant ones on the same input data, we re-implemented two state-of-the-art approaches for gene prioritization: the random-walk based method of [K$^+$08] and the CIPHER [W$^+$08] algorithm. In addition we implemented a simple shortest-path based approach for the problem. We describe the implementation details below. We note that we could not compare our method to that of Lage et al. [L$^+$07], as code or input data for the latter method were not readily available.

The random-walk based approach requires disease grouping information. To allow it to run on the more comprehensive disease similarity data we had, we generalized the approach to use these similarities (transformed by the logistic function L) as initial probabilities for the random walk. The parameter r of the method, which controls the probability for a restart, as well as our transformation parameter c, were optimized using cross-validation (as in the Parameter Tuning Section below). Note that Kohler et al. suggested a second, diffusion-kernel based approach, which was shown to be inferior to the random walk one, hence we did not include it in our comparison. Also note that our propagation-based method reduces to a random walk under appropriate transformations of the edge weights and prior information.

The CIPHER method [W$^+$08] is based on computing protein closeness in a PPI network. Two variants of the algorithm were suggested: CIPHER-DN, which considers only direct neighbors in the closeness computation, and CIPHER-SP, which is based on a shortest path computation. The former was shown to outperform the latter, and hence we implemented this variant (CIPHER-DN) only.

In addition, we implemented a simple shortest-path (SP) based approach, in which a candidate protein is scored according to the most probable path to a disease-related protein. Formally, define the probability of a path connecting a candidate protein to a causal protein v, as the product of the normalized weights w' of the edges along the path and $Y(v)$. The score of a candidate protein is then the score of its best path.

Performance evaluation. To evaluate the performance of the different methods we tested, we used a leave-one-out cross validation procedure. In each cross-validation trial, we removed a single disease-protein association $< d, p >$ from the data, and in addition all other associations involving protein p. An algorithm was evaluated by its success in reconstructing the hidden association, i.e. by the rank it assigned to protein p when querying disease d. The reason we hid all associations of p was to avoid "easy" cases in which p is also associated with other diseases that are very similar to d.

We evaluated the performance of an algorithm in terms of overall precision vs. recall when varying the rank threshold $1 \leq k \leq 100$. *Precision* is the fraction of gene-disease associations that ranked within the top $k\%$ at some trials and are true associations. In other words, it is the number of trials in which a hidden association was recovered as one of the top $k\%$ scoring ones, over the total number of trials times $k\%$ of the interval size. *Recall* is the fraction of trials in which the hidden association was recovered as one of the top $k\%$ scoring ones.

In addition, we used two other measures for quality evaluation. The first, is the *enrichment* measure [$\mathrm{L}^{+}07$] which is defined as follows: If the correct gene is ranked in the top $m\%$ in $n\%$ of the trials then there is a n/m-fold enrichment. For example, if the algorithm ranks the correct gene in the top 10% in 50% of the cases, a 5-fold enrichment is achieved, while random prioritization of the genes is expected to rank the correct gene in the top 10% only in 10% of the cases, yielding a 1-fold enrichment. The second, is the *average rank* of the correct gene throughout the cross-validation trials. Note that when $m = 1$, recall, precision and enrichment measures are all equal.

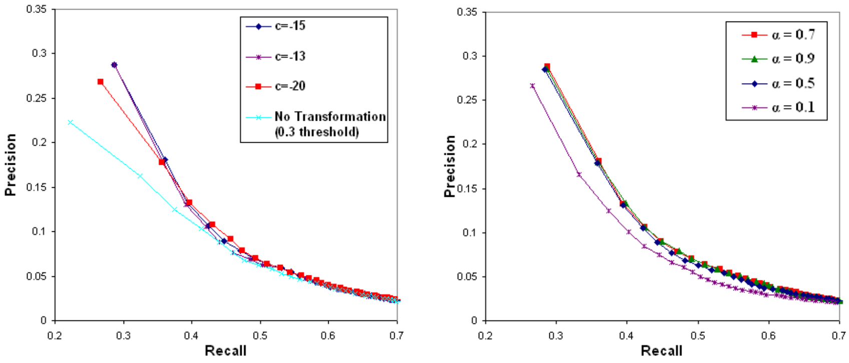
4 Results

We implemented our propagation algorithm and tested its performance in recovering known disease-gene association both on 150 diseases for which more than one causal gene is known, and the entire collection of 1,369 diseases. We report these results and compare our algorithm to previous state-of-the-art algorithms for the prioritization problem.

Parameter tuning. Our algorithm has three parameters that should be tuned: (i) c – the parameter controlling the logistic regression transformation; (ii) α – controlling the relative importance of prior information in the association assignment; and (iii) the number of propagation iterations employed. We used the cross validation framework to test the effect of these parameters on the performance of the algorithm. The precision-recall plots for the general disease case are depicted in Figure 1. By Figure 1(a) the optimal regression coefficient is $c = -15$, implying that similarity values below 0.3 are assigned with very low probability(< 0.002), in accordance with the analysis of [$\mathrm{vD}^{+}06$]. The algorithm is not sensitive to the actual choice of α as long as it is above 0.5 (Figure 1(b)). Finally, the algorithm shows fast convergence, achieving optimal results after only ten iterations (data not shown). Similar results were obtained in the tuning process for diseases with more than one known gene.

Diseases with more than one known gene. Our first set of tests focused on 150 diseases for which more than one causal gene is known. For such diseases we first checked whether our algorithm gains in performance when incorporating information on similar diseases, compared to when using information on the disease of interest d alone. For the latter case we set $Y(v) := 1$ if protein v is associated with d and $Y(v) := 0$ otherwise. As evident from Figure 2 the disease similarity information improves the quality of predictions.

Next, we compared the performance of our algorithm to those of the random-walk and CIPHER methods, as well as to our SP variant. The results are depicted in Figure 3 and summarized in Table 1. Our algorithm achieved the best performance, ranking the correct gene as the top-scoring one in 50.9% of the cases. Interestingly, SP was the second-best performer with 43.7% correct top-1 predictions, while the method of [$\mathrm{K}^{+}08$] and CIPHER attained lower success rates of 40.9% and 37.5%, respectively.



(a) (b)

Figure 1: Effect of parameters on our algorithm's performance, as measured in cross-validation on the set of 1,369 diseases with a known gene. (a) Precision vs. recall plots for different c values, as well as for a simple identity transformation in which values below 0.3 are ignored. (b) Performance comparison for different α values.

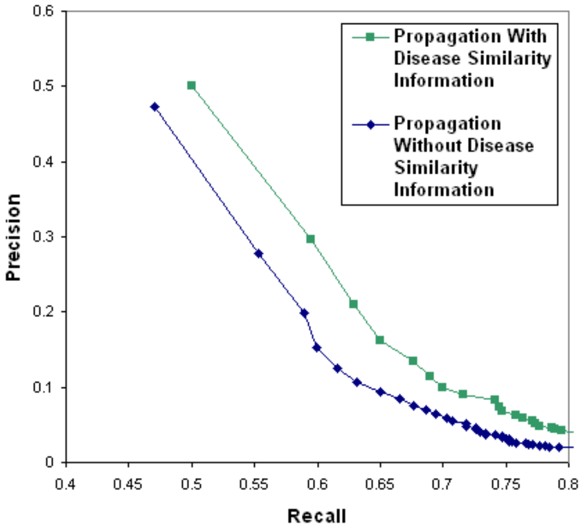


Figure 2: The effect of incorporating disease similarity information on prioritizing genes for 150 diseases with more than one known gene.

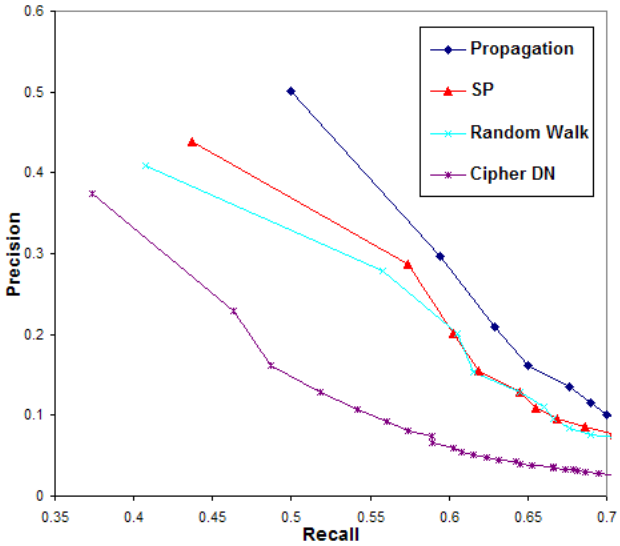


Figure 3: Performance comparison on 150 diseases with more than one known gene.

General diseases. Our second set of tests concerned all 1,369 diseases with a known gene in the OMIM database. The results of applying the different methods are depicted in Figure 4 and summarized in Table 1. Again, our algorithm achieved the best performance, ranking the correct gene as the top-scoring one in 28.7% of the cases. SP, CIPHER and random-walk methods all achieved inferior results with 26.8%, 22.6% and 21.7% success rates, respectively.

5 Acknowledgements

We thank Tomer Shlomi for his help in preprocessing the data and insightful suggestions. We also thank Nir Yosef for providing us with the human PPI network. This research was supported by a German-Israel Foundation grant.

References

[E$^+$07] R. Ewing et al. Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol Syst Biol*, 3, 2007. 10.1038/msb4100134.

[F$^+$06] L. Franke et al. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am J Hum Genet*, 78(6):1011–1025, June 2006.

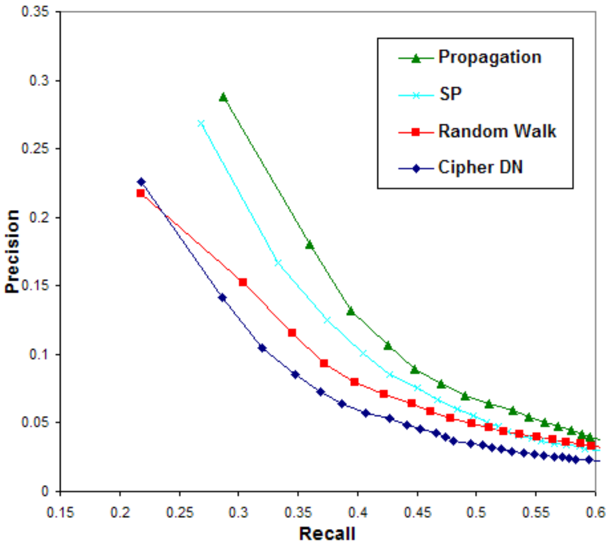


Figure 4: Performance comparison on 1,369 diseases with a known gene.

[G+06] R. A. George et al. Analysis of protein sequence and interaction data for candidate disease gene prediction. *Nucleic Acids Res*, 34(19):e130, 2006.

[H+02] A. Hamosh et al. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.*, 30(1):52–55, January 2002.

[K+08] S. Kohler et al. Walking the Interactome for Prioritization of Candidate Disease Genes. *American journal of human genetics*, 82(4):949–958, 2008.

[L+07] K. Lage et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotech*, 25(3):309–316, 2007.

[O+06] M. Oti et al. Predicting disease genes using protein-protein interactions. *J Med Genet*, 43(8):691–698, August 2006.

[OB07] M. Oti and MG. Brunner. The modular nature of genetic diseases. *Clinical Genetics*, 71(1):1–11, January 2007.

[P+04] S. Peri et al. Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res*, 32(Database issue):D497–501, 2004.

[PIBAN07] C. Perez-Iratxeta, P. Bork, and M. A. Andrade-Navarro. Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res*, 35(Web Server issue):W212–6, 2007.

[R+97] M. Rebhan et al. GeneCards: integrating information about genes, proteins and diseases. *Trends in Genetics*, 13:163, 1997.

[R+05] JF. Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437(7062):1173–1178, 2005.

General diseases		
Method	Enrichment	Avg Rank
Propagation	**28.7**	**24.2**
SP	26.8	25
Kohler et al.	21.7	25.7
Wu et al.	22.6	29.5

Diseases with more than one known gene		
Method	Enrichment	Avg Rank
Propagation	**50.9**	**14.3**
SP	43.7	15
Kohler et al.	40.9	15.4
Wu et al.	37.5	21.2

Table 1: Summary of performance comparison on two collections of diseases.

[S$^+$05a] R. Sharan et al. Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, 102(6):1974–1979, February 2005.

[S$^+$05b] U. Stelzl et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122(6):957–68, 2005.

[vD$^+$06] M. van Driel et al. A text-mining analysis of the human phenome. *Eur J Hum Genet*, 14(5):535–542, 2006.

[W$^+$08] X. Wu et al. Network-based global inference of human disease genes. *Mol Syst Biol*, 4, May 2008.

[Z$^+$03] D. Zhou et al. Learning with local and global consistency, 2003. In 18th Annual Conf. on Neural Information Processing Systems.