JOURNAL OF COMPUTATIONAL BIOLOGY
Volume 16, Number 2, 2009
© Mary Ann Liebert, Inc.
Pp. 181–189
DOI: 10.1089/cmb.2008.05TT

A Network-Based Method for Predicting Disease-Causing Genes

SHAUL KARNI,[1] HERMONA SOREQ,[2] and RODED SHARAN[1]

ABSTRACT

A fundamental problem in human health is the inference of disease-causing genes, with important applications to diagnosis and treatment. Previous work in this direction relied on knowledge of multiple loci associated with the disease, or causal genes for similar diseases, which limited its applicability. Here we present a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under a condition of interest. The latter are used to derive a set of disease-related genes which is assumed to be in close proximity in the network to the causal genes. Our method applies a set-cover-like heuristic to identify a small set of genes that best "cover" the disease-related genes. We perform comprehensive simulations to validate our method and test its robustness to noise. In addition, we validate our method on real gene expression data and on gene specific knockouts. Finally, we apply it to suggest possible genes that are involved in myasthenia gravis.

Key words: gene-disease association, gene expression analysis, myasthenia gravis, protein-protein interaction network.

1. INTRODUCTION

High-throughput technologies such as yeast two-hybrid screens (Fields and Song, 1989) and co-immunoprecipitation (Lee et al., 2002) are routinely used nowadays to map molecular interactions within the cell. Applications of these maps include the prediction of protein function (Sharan et al., 2007) and orthology (Bandyopadhyay et al., 2006), the inference of protein modules (Sharan et al., 2005) and more.

In the last two years, large scale maps of protein-protein interactions (PPIs) have become available for humans (Rual et al., 2005; Stelzl et al., 2005), leading to an array of works aiming at harnessing PPI data to improve the understanding of human disease. In particular, many authors have shown the utility of these networks in inferring *disease-causing genes*. Franke et al. (2006) considered diseases with several associated loci. For such diseases they aimed at identifying a set of genes, spanning the associated loci, whose protein products are connected in a functional network, comprised of PPIs, co-expression relations

[1] Blavatnik School of Computer Science, Tel-Aviv University, Tel-Aviv, Israel.

[2] Interdisciplinary Center for Computational Neuroscience, The Hebrew University of Jerusalem, Jerusalem, Israel.

and gene-ontology similarities. Lage et al. (2007) integrated PPI data with information on the phenotype similarity of different diseases. They developed an algorithm for predicting causal genes that relies on the observation that genes causing similar diseases tend to be connected in a PPI network. Kohler et al. (2007) grouped diseases into families using a random walk method from known genes in its family to prioritize candidate genes. Wu et al. (2008) scores a candidate gene based on the correlation between the vector of similarities to diseases with known causal genes using a propagation method. Mani et al. (2008) used gene expression data in combination with molecular interaction data to identify interactions that exhibit a gain or a loss of expression correlation in a given phenotypic class. They then ranked genes according to the enrichment of their direct neighborhood with such interactions.

Here we present a new algorithm for predicting disease-causing genes. Rather than assuming information on disease loci, or on gene-disease associations, we make use of the abundant information on genes that change their expression levels within the affected tissue under the disease state. We call the latter *disease-related genes*. Our algorithm relies on the assumption that in the disease state, one or more causal genes are disrupted, leading to the expression changes of downstream (disease-related) genes through signaling-regulatory pathways in the network. To uncover the causal genes, we make a parsimonious assumption, seeking the smallest set of genes that could best explain the expression changes of the disease-related genes in terms of probable pathways leading from the causal to the affected genes in a network of physical interactions. Ideally, this network should contain protein-protein and protein-DNA interactions. However, the latter are not available at large scale for humans. Hence, in practice, we use PPI data only.

In simulations, our algorithm attains very high accuracy on a wide range of parameters, including the size of the input affected set, the noise level within the set, the size of the search space, and the number of causal genes simulated. In validation on real expression data from knockout experiments, our algorithm manages to pinpoint the disrupted gene with high accuracy. Further validations on expression data from different types of cancer show high accuracy in pinpointing known oncogenes. Importantly, we show that our method outperforms a naive algorithm that ranks disease-associated genes according to their distances in the network to the directly affected genes. Finally, we apply our method to suggest possible genes that are involved in myasthenia gravis.

2. RESULTS

We have developed a novel algorithm for identifying disease causing alterations in the pathway of gene expression. Our approach is based on analyzing the network-proximity of candidate proteins within the network to a set of proteins that were implicated in the investigated disease. We used read data to optimize the maximal search depth parameter l used throughout this section, as described below.

Performance on simulated data

To evaluate the performance of the algorithm, we applied it to simulated data. In each simulation, one or more "disease-causing" proteins were taken at random from the network, and artificial loci, consisting of 50–200 genes each, were constructed around the genes they encode. To construct a "disease-related" subset of a certain size (between 30–180), proteins were chosen at random from the set of proteins of distance at most 3 from the "disease-causing" ones. The simulation setting ensures that the random instances follow our assumptions on the disease-causing genes. Thus, the simulations mainly serve to test under what conditions can one recover these genes, overcoming false network signals such as hubs that happen to be close to the disease-related set.

For each locus size and "disease-related subset," 50 random tests were conducted. The results obtained in simulations of a single causal gene are summarized in Figure 1A. Notably, when limiting the search to a certain locus, the algorithm almost always ranks the simulated causal gene first. The accuracy is lower when searching the entire network: the average rank of the causal gene ranges between 3.8 and 5.8, and it is ranked first 22–52% of the time, depending on the locus size.

We compared our performance to that of a naive approach that ranks proteins according to their sum of distances to the input disease-related genes. As can be seen in Figure 1B, our method is considerably more accurate when searching the entire network, achieving performance gains of more than 70%.

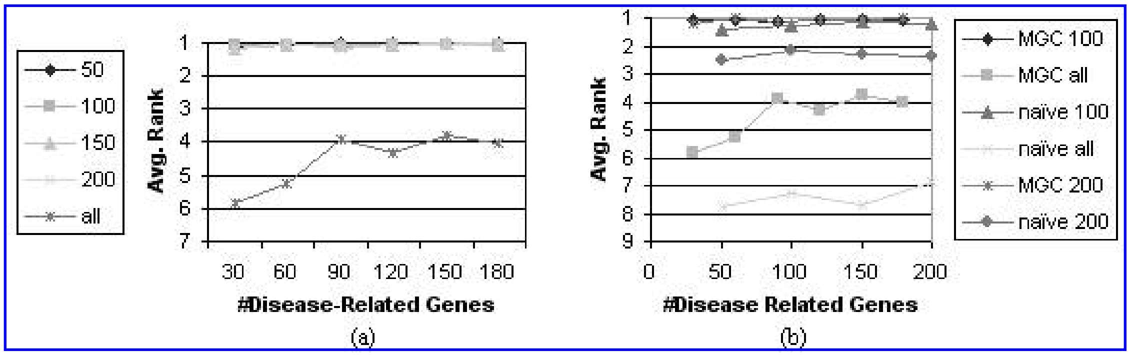

FIG. 1. Performance on simulated data. The average rank of the correct simulated gene as a function of the size of the disease-related set. Each graph corresponds to a different test set. The different plots correspond to loci of different sizes. The "all" plot depicts the results obtained when searching the entire network. **(A)** Success rate tests. **(B)** Comparison to a naive approach.

To test the robustness of the algorithm to noise in the input list of disease-related genes, simulations were carried out in which the size of the disease-related set was fixed at 100, and up to 25% of the proteins in the set were replaced by random proteins. Genes whose "coverage" expectation was equal to that of the simulated causal gene were removed (as they are indistinguishable from it based on our measure). The results are depicted in Figure 2A and show that the algorithm's results are robust even at high noise levels.

Finally, we tested the utility of the algorithm in recovering more than one causal gene. To this end, we conducted experiments in which up to four disease-causing genes were simulated (with a 100-genes sized locus around each one). The results are depicted in Figure 2B. Evidently, the performance is very good for one to two genes, but worsens with three or more genes. For example, with four disease-causing genes, the algorithm detects at least one (as the first ranking) 96% of the time, but identifies all four only 10% of the time.

Validation on real data

To further evaluate the algorithm's performance, we applied it also to real data sets in which a causal gene is known. To this end, we used gene expression data for diseases where the genetic origin is known, or knockout data sets where a gene was knocked out and as a result other genes changed their (wild-type)

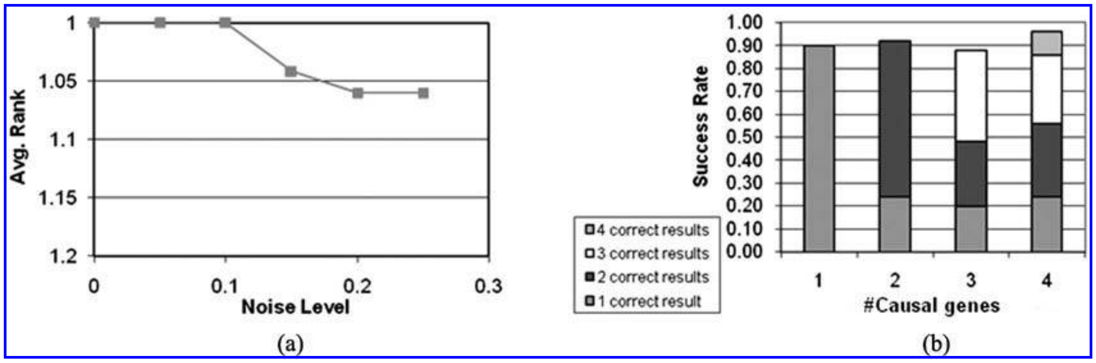

FIG. 2. **(A)** Performance in noisy simulations. **(B)** Performance on simulated data with multiple disease-causing genes. The success rate measures the percentage of runs where the simulated causal genes were ranked first, for one to four target genes. In all cases, the size of the disease-related genes was fixed to 100.

expression levels. To simulate partial knowledge on the location of the causal gene, we used information on the chromosomal segment in which the gene is located from the OMIM database (Hamosh et al., 2002).

In order to choose an optimal maximal depth with which to run our algorithm, we applied it with maximal depth of 1–5 to three of the data sets (ATM knockout, NFκB knockout, and MLL gene expression data), which are described in detail below. For depths 1 and 2, correct results were found only when considerably narrowing the search segments (to a few dozens of genes, data not shown). The results for depths 3–5 are shown in Figure 3A. As evident from the figure, the best results (except for the ATM knockout data) were attained at maximal depth 3, which was subsequently used in all our runs.

The first set of experiments was performed on knockout data from Elkon et al. (2005), where the knockout effect of several genes was investigated under DNA damage conditions. In response to knocking out the transcription factor NFκB, 48 genes changed their expression levels. This gene is located in chromosomal segment $4q24$, which contains 31 genes, 15 of which appear in the PPI network. Reassuringly, NFκB was ranked first in all our tests. When knocking out the signaling protein ATM, 47 genes changed their expression levels. ATM lies within segment $11q22$, which contains 75 genes, 31 of which appear in our network. Overall, ATM ranked third, with an average rank of 3.12.

Next, we used data on acute lymphoblastic leukemia (ALL) (Armstrong et al., 2001), consisting of expression profiles for a subset of acute leukemias involving chromosomal translocation of the mixed leukemia gene (MLL). Overall, 67 genes were found to be differentially expressed and appeared in our network. The MLL gene is located at segment $11q23$, which contains 168 genes, 61 of which are in the network. When applying the algorithm to this data, MLL scored best with an average rank of 1.5. The second highest ranking gene, with an average rank of 2.86, was matrix metallopeptidase 7 (MMP-7), a member of the matrix metalloproteinase family. This gene has been linked before to leukemia (Lynch and McDonnell, 2000), and many other forms of cancer (Liu et al., 2007; Rome et al., 2007). Three additional proteins that ranked among the top 10 are involved in phosphorylation signaling cascades known to be involved in the leukemic processes (Perry et al., 2002).

The results on the different validation sets are summarized in Figure 3B, which plots average ranks under different loci sizes for each of the diseases. As evident from the figure, our method significantly outperforms the naive one. Notably, in these three data sets, the known causal gene was not differentially expressed, hence the network information was essential for its discovery.

Finally, we applied our algorithm to input sets from multiple expression studies on breast cancer (Pawitan et al., 2005; Sotiriou et al., 2003; van de Vijver et al., 2002; Wang et al., 2005). We tested the rate at which our algorithm managed to recover BRCA1 (breast cancer 1, early onset) or BRCA2 (breast cancer 2, early onset), two of the major causal genes known. Here our search was conducted on 114 genes of the BRCA-1

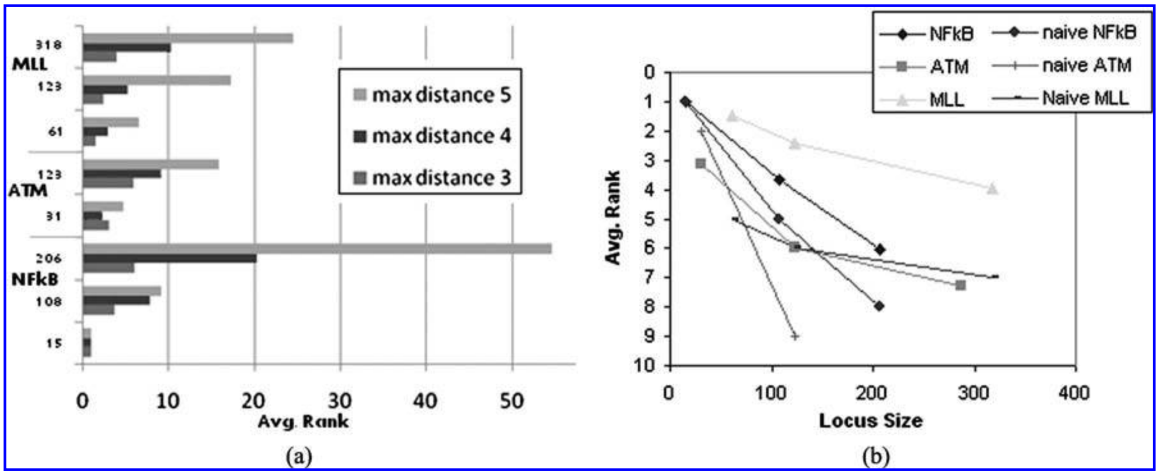

FIG. 3. **(A)** Performance with different maximal search depths (1–5) for three real data sets: ATM and NFκB knockouts and MLL gene expression data. **(B)** Performance comparison to a naive approach on the same data sets as in A using a maximal search depth of 3.

TABLE 1. AVERAGE RANKS OF BRCA1 AND BRCA2 ON BREAST CANCER DATA FROM DIFFERENT STUDIES

Data source	*BRCA1*	*BRCA2*
Pawitan et al. (2005)	5.72	>10
Wang et al. (2005)	2.88	1.76
Sotiriou et al. (2003)	>10	2.5
van de Vijver et al. (2002)	>10	3.06

associated segment ($17q21$) and the 32 genes of the BRCA2 associated segment ($13q12$). The results, summarized in Table 1, show that at least one of BRCA1 or BRCA2 was recovered in each of the data sets. Intriguingly, the BRCA1 set of top 10 genes from both Pawitan et al. (2005) and Wang et al. (2005) were enriched with genes involved in transcriptional regulation, whereas the BRCA2 sets from Sotiriou et al. (2003), van de Vijver et al. (2002), and Wang et al. (2005) were rich in phosphorylation-associated genes, in accordance with the distinct functions of these two genes.

Application to myasthenia gravis

After establishing the utility of our algorithm in predicting disease-causing genes, we sought to apply it to a multi-factorial disease for which the causal genes are not known. As our test case we used myasthenia gravis (MG), a neuromuscular autoimmune disease. We used data from Gilboa-Geffen et al. (2007), which contains a list of genes that are significantly expressed in the thymus of patients with mild and severe cases of the disease. First, we applied the algorithm to each severity class separately, using 391 genes for mild MG and 354 genes for severe MG. Then, we composed a list of 63 genes which appear in the severe cases but not in the mild ones. In all these applications the search for causal genes was conducted on the entire set of proteins in the network. The results are summarized in Table 2.

In mild MG, the highest ranking proteins contribute to general housekeeping functions: cell growth and cell-cell interactions, transcriptional activity and peroxisome properties. In severe MG, we also observed impairments in hematopoietic differentiation compatible with the lymphocytic hyperproliferation which is characteristic of MG thymuses (Gilboa-Geffen et al., 2007).

When looking at the set of genes that were differentially expressed in the severe cases, but not in the mild ones, the highest ranking protein was major histocompatibility complex, class I, B (HLA-B). This protein is part of the HLA class I heavy chain paralogs, which play a central role in the immune system and are expressed in nearly all cells. The linkage between HLA and MG is supported by previous studies (Donmez et al., 2004; Huang et al., 1999; Vandiedonck et al., 2005). The second highest ranking protein was cAMP-dependent protein kinase catalytic subunit alpha isoform 1 (PRKACA). This protein is known to phosphorylate and inhibit acetylcholine receptor functioning and affect the disease (Li et al., 1996; Plested et al., 2002).

TABLE 2. A SUMMARY OF THE RESULTS OF RUNNING THE MGC ALGORITHM ON MYASTHENIA GRAVIS DATA

Mild	*Severe*	*Severe but not mild*
LGALS3BP	INSR	HLA-B
PEX6	ZMYM2	PRKACA
INSR	GJA1	PPP1R2
CD46	GUCY2C	HLA-A
POU4F2	CD46	CALCOCO1
ITGAL	EBF1	GRLF1
ZDHHC4	GATA3	GYS1

3. CONCLUSION

We presented a new approach to causal gene prediction that is based on integrating protein-protein interaction network data with gene expression data under disease conditions. The latter are used to highlight a set of disease-related genes that are assumed to be in close proximity to the causal genes in the PPI network. Based on this assumption, we apply a greedy heuristic that recovers putative causal genes as those admitting pathways to a maximal number (in expectation) of disease-related genes. We comprehensively validated the accuracy of our algorithm in pinpointing causal genes, both in simulations and on real network data. By applying our algorithm to data on myasthenia gravis, we were able to suggest candidate causal genes and gain insights about their roles in the progression of the disease.

While our results are encouraging, several enhancements could be introduced to our framework. First, it would be revealing to integrate protein-DNA interactions into the network and study the impact of such interactions on the identified genes and pathways. To date, no experimental large-scale transcriptional network is available for human, although recent computational efforts have aimed at inferring it (Adler et al., 2007). As our results indicate, the algorithm exhibits high retrieval rate when using PPI data only. This may be explained by the known correlation between PPI and gene expression data (Deng et al., 2003). Second, it could be beneficial to analyze different stages of a certain disease to obtain clues on its progression. As suggested by the MG example, and as further indicated by our initial results on other diseases, advanced stages of a disease tend to imply larger sets of causal genes that are more widespread in the network.

4. METHODS

Problem definition. We study the problem of predicting one or more disease-causing genes given a set of genes that are implicated in a disease. We propose a network-based framework for it, which relies on the assumption that the protein product of a disease-causing gene should be highly connected in a network of physical interactions to the protein products of genes affected by it. Formally, the basic problem we consider is defined as follows:

Definition 1 (Gene Cover (GC)). *Given a graph $G = (V, E)$, a subset $U \subseteq V$ and a distance threshold l, find a subset of vertices D of minimum size such that for each $u \in U$ there exists a vertex in D of distance at most l from u.*

As we show below, GC is NP-complete as it is polynomially equivalent to Set Cover.

Theorem 1. *The decision versions of GC and Set Cover are polynomially equivalent.*

Proof. Let (S, C, k) be an instance of Set Cover where S is the set of elements, C is a collection of subsets of S and k is a parameter. W.l.o.g., we assume that C covers S. We can easily transform this instance into an instance $(G, S, 1, k)$ of (the decision version of) GC as follows: we construct a bipartite graph $G = (S, C, E)$ with vertices on one side representing elements and vertices on the other side representing subsets. For every $T \in C$ and $s \in T$ we add an edge (s, T) to E. It is trivial to observe that the Set Cover instance admits a solution iff the GC instance admits a solution (the only problematic case is when an GC solution contains an element from S, but such an element can always be substituted by a subset containing it).

In the other direction, suppose we are given an instance (G, U, l, k) of GC. We transform it into an instance (U, C, k) of Set Cover, where C is defined as follows: for each vertex v in G we create a subset $T \subseteq U$ composed of all vertices that are of distance at most l from v (including v if it is part of U). If $T \neq \emptyset$ we add it to C. Again there is a solution to the GC instance iff there is a solution to the Set Cover instance. ■

On the positive side, Set Cover can be efficiently approximated to within a logarithmic factor (Cormen et al., 2001); as the reduction from Gene Cover to Set Cover is approximation preserving, it implies an $O(\log |U|)$ approximation algorithm for GC as well.

A biologically-motivated formulation. The combinatorial formulation presented above treats all edges of the protein network being analyzed in a uniform manner. Since, protein-protein interactions vary greatly in their associated confidence scores, it is desirable to take edge reliabilities into account. A natural extension to the distance-based formulation above is to quantify the relatedness of a protein v to a set U by the *expected* number of proteins in U that can be reached from it by paths of length at most l. Denote this expectation by $E_l(v, U)$ (we defer the details of its computation to the next section), and consider the following formulation of the gene coverage problem:

Definition 2 (Maximum-expectation Gene Cover (MGC)). *Given a graph $G = (V, E)$, a subset $U \subseteq V$, a distance threshold l, and a parameter k, find a subset of vertices D of size k such that $\sum_{v \in D} E_l(v, U)$ is maximal.*

It is possible to approximate MGC to within a factor of $O(\log |U|)$ by adapting the greedy-based approximation algorithm for Weighted Set Cover. Below we provide a practical heuristic to MGC which is based on this approximation strategy.

In many cases, additional information is available that can help us to limit the search space (Wu et al., 2008). Specifically, association studies may provide information on genomic regions which are associated with the investigated disease, reducing the initial search space from thousands of proteins to a few hundred (McCarthy et al., 2008). Similarly, copy number variation data can pinpoint areas of the genome whose copy number is modified in the disease state (McCarroll and Altshuler, 2007); these areas are then good candidates for causal gene searches.

Expectation computation. Let $U = \{u_1, \ldots, u_n\}$. Recall that $E_l(v, U)$ denotes the expected number of vertices in U that are reachable from v by paths of length at most l. From the linearity of expectation,

$$E_l(v, U) = \sum_{i=1}^{n} E_l(v, \{u_i\}) = \sum_{i=1}^{n} P_l(v, u_i) \qquad (1)$$

where $P_l(a, b)$ is the probability of having a path of length at most l between a and b.

For two vertices a and b, let $\Pi_l(a, b) = \{\Pi_1, \ldots, \Pi_m\}$ denote the set of paths of length at most l between a and b. Let π_i be a random variable indicating whether the path Π_i exists. Then $P_l(a, b) = Prob(\cup_{i=1}^{m} \pi_i)$. This probability can be computed using the inclusion-exclusion formula in time that is exponential in m.

To save on running time, one can partition the set of paths into subsets that are edge-disjoint. This is done by constructing a new graph whose vertices represent paths and whose edges connect edge-intersecting paths. The connected components of this graph yield the desired partition. Let $\Delta_1, \ldots, \Delta_t$ denote the resulting subsets of paths, and consider a pair of vertices a, b. Then $P_l(a, b) = 1 - \prod_{i=1}^{t}(1 - Prob(\bigcup_{\pi \in \Delta_i} \pi))$. Each term can be computed by an inclusion-exclusion formula:

$$Prob\left(\bigcup_{\pi \in \Delta_i} \pi\right) = \sum_{k=1}^{|\Delta_i|} (-1)^{k-1} \sum_{\substack{\Delta \subseteq \Delta_i \\ |\Delta| = k}} Prob\left(\bigcap_{\pi \in \Delta} \pi\right) \qquad (2)$$

where the probability of an intersection of paths is simply the product of the probabilities of the edges in the intersection.

The MGC algorithm. We focus on the biologically motivated MGC. Our algorithmic approach is motivated by the greedy approximation algorithm to Weighted Set Cover. Given a protein network G and a subset of disease-related proteins U, we apply an iterative algorithm to infer the disease-causing genes.

Intuitively, at each iteration the protein, whose "coverage" expectation with respect to the current subset U is maximal, is chosen and the diseased proteins that it "covers" are removed from U. However, the expectation computation gives an advantage to high degree proteins. To circumvent this problem, we compare the original expectation to that obtained w.r.t. 100 random disease-related subsets of the same size as U. The results of the random runs are used to derive a z-score for each vertex, and the highest-scoring vertex is chosen at each iteration. The algorithm terminates when the highest score attained is below a predefined threshold (1.65, corresponding to a p-value of 0.05), or when all the disease-related genes have been "covered." Due to the randomized nature of the algorithm (in computing the z-score), the results may change slightly between runs. Hence, each experiment is repeated 50 times, and the genes are ranked based on their average ranks in these 50 runs.

Network construction. PPI data were collected from the HPRD database (Peri et al., 2003; Mishra et al., 2006) and two large scale yeast-two-hybrid experiments (Rual et al., 2005; Stelzl et al., 2005). The constructed network consists of 28,972 interactions among 7,915 proteins. The interactions were assigned confidence scores based on the experimental evidence available for each interaction using a logistic regression model adapted from Sharan et al. (2005). From this network we then removed the most substantial hubs, which had over 150 network connections. Thus, the network used in this paper consists of 27,707 interactions among 7,870 proteins.

ACKNOWLEDGMENTS

We thank Nir Yosef for providing us with the PPI network data and Igor Ulitsky for providing us with the breast cancer data. S.K. is a fellow of the Edmond J. Safra Bioinformatics Program at Tel-Aviv University. This research was supported by a grant from the Israel Science Foundation (grant 385/06 to R.S.) and a grant from the Ministry of Science and Technology, Israel (to R.S. and H.S.).

DISCLOSURE STATEMENT

No competing financial interests exist.

REFERENCES

Adler, A.S., Sinha, S., Kawahara, T.L., et al. 2007. Motif module map reveals enforcement of aging by continual NFkB activity. *Genes Dev.* 21, 3244–3257.

Armstrong, S.A., Staunton, J.E., Silverman, L.B., et al. 2001. MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nat. Genet.* 30, 41–47.

Bandyopadhyay, S., Sharan, R., and Ideker, T. 2006. Systematic identification of functional orthologs based on protein network comparison. *Genome Res.* 16, 428–435.

Cormen, T.H., Leiserson, C.E., Rivest, R.L., et al. 2001. *Introduction to Algorithms*, 2nd ed. The MIT Press, Cambridge, MA.

Deng, M., Sun, F., and Chen, T. 2003. Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac. Symp. Biocomput.* 140–151.

Donmez, B., Ozakbas, S., Oktem, M.A., et al. 2004. HLA genotypes in turkish patients with myasthenia gravis: comparison with multiple sclerosis patients on the basis of clinical subtypes and demographic features. *Hum. Immunol.* 65, 752–757.

Elkon, R., Rashi-Elkeles, S., Lerenthal, Y., et al. 2005. Dissection of a DNA-damage-induced transcriptional network using a combination of microarrays, RNA interference and computational promoter analysis. *Genome Biol.* 6, R43.

Fields, S., and Song, O. 1989. A novel genetic system to detect protein-protein interactions. *Nature* 340, 245–246.

Franke, L., Bakel, H., Fokkens, L., et al. 2006. Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* 78, 1011–1025.

Gilboa-Geffen, A., Lacoste, P.P., Soreq, L., et al. 2007. The thymic theme of acetylcholinesterase splice variants in myasthenia gravis. *Blood* 109, 4383–4391.

Hamosh, A., Scott, A.F., Amberger, J., et al. 2002. Online Mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* 30, 52–55.

Huang, D.R., Pirskanen, R., Matell, G., et al. 1999. Tumour necrosis factor-alpha polymorphism and secretion in myasthenia gravis. *J. Neuroimmunol.* 94, 165–171.

Kohler, S., Bauer, D., snd Horn, S., et al. 2007. Walking the interactome for prioritization of candidate disease genes. *Am. J. Human Genet.* 82, 949–958.

Lage, K., Karlberg, O.E., Størling, Z.M., et al. 2007. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.* 25, 309–316.

Lee, T.I., Rinaldi, N.J., Robert, F., et al. 2002. Transcriptional regulatory networks in saccharomyces cerevisiae. *Science* 298, 799–804.

Li, Z., Forester, N., and Vincent, A. 1996. Modulation of acetylcholine receptor function in TE671 (rhabdomyosarcoma) cells by non-AChR ligands: possible relevance to seronegative myasthenia gravis. *J. Neuroimmunol.* 64, 179–183.

Liu, D., Nakano, J., Ishikawa, S., et al. 2007. Overexpression of matrix metalloproteinase-7 (MMP-7) correlates with tumor proliferation, and a poor prognosis in non-small cell lung cancer. *Lung Cancer* 58, 384–391.

Lynch, C.C., and McDonnell, S. 2000. The role of matrilysin (MMP-7) in leukaemia cell invasion. *Clin. Exp. Metastasis* 18, 401–406.

Mani, K.M., Lefebvre, C., Wang, K., et al. 2008. A systems biology approach to prediction of oncogenes and molecular perturbation targets in b-cell lymphomas. *Mol. Syst. Biol.*, 4, 169.

McCarroll, S.A., and Altshuler, D.M. 2007. Copy-number variation and association studies of human disease. *Nat. Genet.* 39(7 Suppl), S37–S42.

McCarthy, M.I., Abecasis, G.R., Cardon, L.R., et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* 9, 356–369.

Mishra, G., Suresh, M., Kumaran, K., et al. 2006. Human protein reference database–2006 update. *Nucleic Acids Res.* 34, D411–D414.

Pawitan, Y., Bjöhle, J., Amler, L., et al. 2005. Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Res.* 7, 953–964.

Peri, S., Navarro, J.D., Amanchy, R., et al. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Res.* 13, 2363–2371.

Perry, C., Eldor, A., and Soreq, H. 2002. Runx1/AML1 in leukemia: disrupted association with diverse protein partners. *Leukemia Res.* 26, 221–228.

Plested, C.P., Tang, T., Spreadbury, I., et al. 2002. AChR phosphorylation and indirect inhibition of AChR function in seronegative MG. *Neurology* 59, 1682–1688.

Rome, C., Arsaut, J., Taris, C., et al. 2007. MMP-7 (matrilysin) expression in human brain tumors. *Mol. Carcinogen.* 46, 446–452.

Rual, J.F., Venkatesan, K., Hao, T., et al. 2005. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 437, 1173–1178.

Sharan, R., Suthram, S., Kelley, R.M., et al. 2005. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci. USA* 102, 1974–1979.

Sharan, R., Ulitsky, I., and Shamir, R. 2007. Network-based prediction of protein function. *Mol. Syst. Biol.* 3, 88.

Sotiriou, C., Neo, S.Y., McShane, L.M., et al. 2003. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proc. Natl. Acad. Sci. USA* 100, 10393–10398.

Stelzl, U., Worm, U., Lalowski, M., et al. 2005. A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122, 957–968.

van de Vijver, M.J., He, Y.D., van't Veer, L.J., et al. 2002. A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.* 347, 1999–2009.

Vandiedonck, C., Giraud, M., and Garchon, H.J. 2005. Genetics of autoimmune myasthenia gravis: the multifaceted contribution of the HLA complex. *J. Autoimmun.* 25, Suppl 1, 6–11.

Wang, Y., Klijn, J.G., Zhang, Y., et al. 2005. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* 365, 671–679.

Wu, X., Jiang, R., Zhang, M.Q., et al. 2008. Network-based global inference of human disease genes. *Mol. Syst. Biol.*, 4, 189.

Address reprint requests to:
Dr. Roded Sharan
Blavatnik School of Computer Science
Tel-Aviv University
Tel-Aviv 69978, Israel

E-mail: roded@tau.ac.il