1 Network-based algorithms to predict gene-disease associations.

We review the computational methods considered.

**Neighborhood.** A natural approach is to predict for a protein \( p \) the annotations that are associated with more than \( \theta \) percent of \( p \)'s network neighbors. This method has been widely used [17, 12] to predict biological process annotations for proteins with unknown function. It also serves as the basis of many other approaches considered in this study [16, 8, 20, 19]. We refer to this method as ‘Neighborhood.’

**Oti et al.** The method developed by Oti et al. [16] also looks at the annotations of the neighbors of a protein to make gene-disease predictions. For each disease \( d \), let \( G_d \) be the genes known to be associated with \( d \), and let \( I_1, I_2, \ldots, I_n \) be the interaction partners of the products of \( G_d \). Oti et al. [16] predicted disease \( d \) for protein \( I_i \) if \( d \) and \( I_i \) belong in the same linkage interval. We refer to this method as ‘Oti1.’ We also modified the method by adding a constraint that \( I_i \) must have at least two (‘Oti2’) or three (‘Oti3’) neighbors in \( G_d \) for the prediction to be made, which should strengthen the likelihood of association. None of the Oti methods depend on the prediction threshold.

**Random Walks.** Kohler et al. [7] define a random walk starting from genes known to be associated with a query disease \( d \). Let \( p^0 \) be a distribution that gives equal probability of starting from each protein known to be associated with \( q \), and 0 to all other proteins. The probability distribution at time step \( t + 1 \) is recursively computed as \( p^{t+1} = (1 - r)Wp^t + rp^0 \), where \( W \) is the column-normalized adjacency matrix of the network and \( r \) is the restart parameter. We set \( r = 0.75 \) (as done by [7]), meaning that 75% of the time the random walk returns to the starting proteins in \( p^0 \) before continuing. We iteratively calculated \( p^t \) and defined convergence to be when the \( L_2 \) norm between \( p^t \) and \( p^{t-1} \) is \( < 10^{-6} \). After convergence, we predict \( d \) for all proteins that lie in an interval related to \( d \) and that have a visitation probability greater than \( \theta \), which we varied from 0.01 – 9%. Higher thresholds should result in more confident predictions. We refer to this method as ‘RW’.

**Propagation.** Vanunu et al. [19] used a similar flow-based approach that spreads flow starting from proteins known to be associated with a query disease \( d \). The flow at node \( u \) at time step \( t \) is defined as \( F(u)^t = \alpha W'F(u)^{t-1} + (1 - \alpha)Y \). Here, \( W' \) is a weight matrix such that \( W'_{ij} = W_{ij}/\sqrt{D_{ii}D_{jj}} \), where \( W_{ij} \) gives the reliability of the interaction between proteins \( i \) and \( j \). In our
case, each edge is weighted equally, hence $W$ is the adjacency matrix of the network. $D_{ii}$ equals the sum of row $i$ in $W$. $Y$ is a vector of prior-information; its $i$th entry is 1 if the corresponding gene is known to be associated with a disease, and 0 otherwise. Essentially, the algorithm pumps flow starting from proteins known to be associated with $d$ to their neighbors iteratively. The $\alpha$ parameter controls the percentage of new flow added from the starting set of proteins in each iteration. In our implementation of their method, we set $\alpha = 0.6$, because Vanunu et al. [19] found that the algorithm is not sensitive to the choice of $\alpha$ as long as it is set to be greater than 0.5. We stopped iterating when the $L_2$ norm between $F^t$ and $F^{t-1}$ was $< 10^{-6}$ and then applied the propagation step with $\alpha = 1$ to smooth the distribution of $F$, as done by Vanunu et al. [19]. As for RW, we predicted $d$ for all proteins that lie in an interval related to $d$ and that have a visitation probability greater than $\theta$, which we varied from $1 - 90\%$. We refer to this method as ‘Prop’.

Graph partitioning is another promising technique for predicting disease-gene associations because it has been shown to uncover modules that can effectively predict biological processes, and phenotypically similar diseases are often caused by proteins that have similar biological processes [5, 20]. Hence, functionally enriched modules obtained through graph partitioning should also correspond well to disease subunits. We tested this module-based approach using three graph partitioning algorithms, graph summarization [13], MCL [18], and VI-Cut, which were recently shown [14, 2, 15] to find the most biologically relevant modules in PPI networks.

**Graph summarization.** Graph summarization [13, 14] is a MDL-based graph compression technique. It tries to minimize the size or cost of representing a graph by forming a summary that highlights the graph’s dominant patterns, and a list of corrections to the summary that handles noisy edges and allows for lossless reconstruction of the original graph. One advantage of graph summarization over many other graph partitioning algorithms is that the compressed summary structure is an aggregate graph created by merging together nodes with similar neighbors into supernodes connected by superedges. This aggregate graph can also be compressed in the same manner, effectively creating a hierarchical compression. We considered three levels in this hierarchy (the original summary of the input graph, the summary of the summary, and the summary of the summary of the summary), and in each case we took the supernodes outputted as our biological clusters. We considered the clustering outputted by each compression individually (referred to by ‘GS1’, ‘GS2’, and ‘GS3’, respectively), and also present results of merging the predictions made by each compression together (referred to by ‘GS-All’). In each subsequent compression supernodes from the previous step are merged together, and therefore, the number of supernodes decreases. Combining supernodes across each of these methods is therefore an indirect way of forming overlapping clusters, which are useful to make predictions for genes associated with more than one genetic disease. As an implementation step, we added self-loops to each protein in the network to reduce the cost of merging proteins that interact.

**MCL.** The Markov Clustering (MCL [18]) algorithm is a popular graph partitioning technique based on random walks on a graph. It attempts to find regions in the graph with high flow concentration, separated by bridge-edges. MCL requires setting an parameter (inflation) by the user. We performed a parameter sweep that tried to optimize the F1-measure of the predictions made by the resulting clusters using a prediction threshold of 5%. The value maximizing the F1-measure was 2.0, which is also the default inflation value.
VI-Cut. The previous two clustering methods described only use the known annotations to make predictions; the known annotations are not used as part of the graph partitioning process itself. VI-Cut [15] is a semi-supervised clustering approach that takes as input a hierarchical tree decomposition (derived by some bottom-up hierarchical clustering process) and a set of known annotations for some of the leaves in the tree. It outputs a clustering induced by a node-cut in the tree that optimally matches the known annotations, as measured by the variation of information metric [11]. VI-Cut has a binary parameter that controls how to break ties in the case when different clusters equally match the known annotations. We can either choose smaller clusters or larger clusters, which we refer to by ‘VI-CutS’ and ‘VI-CutL’, respectively. The hierarchical tree decomposition we used comes from the greedy graph summarization merging process, which was modified to continue merging nodes even if the cost of representing the graph increases after doing so.

2 Experiments with the OPHID human protein interaction network

To ensure that our results held for more than just the HPRD network, we created an additional human protein interaction network from the Online Predicted Human Interaction Database (OPHID [3]). OPHID incorporates data from BIND, HPRD, MINT, and predictions based on interologs from other model species. After converting to the relevant namespace and removing all but the main component the resulting network contained 9,842 proteins and 73,130 interactions, more than twice as many interactions as HPRD. Figure 1 shows the precision and recall of each method using leave-one-out cross validation. Each method’s performance is very similar to its performance on the HPRD network. This suggests that the quality of predictions was not simply an artifact of the HPRD network, but rather consistent across multiple human networks.

3 Quantifying homophily

We defined two measures to quantify the homophily of a disease $d$. The neighborhood homophily of $d$ is the average percentage of network neighbors of a disease-$d$ gene that are also known to be associated with $d$. The average pairwise distance of $d$ is the average number of interactions separating two genes associated with $d$. For each measure, we plotted how the precision and recall of predictions for a disease vary as a function of the disease’s homophily. Figure 2 shows how precision and recall vary as a function of neighborhood homophily (top) and average pairwise distance (bottom) for Neighborhood, VI-CutL, GS-All, Prop, and RW. These plots serve as a way to gauge the quality of predictions for a disease given only the disease’s homophily.

4 Diseases most amenable to network-based prediction

The following 19 diseases were predicted for by at least half (7) of the 13 methods, and with minimum precision $\geq 90\%$:

- Albinism
- Aldosteronism
- Bradyopsia
- C1r/C1s deficiency
- Chronic granulomatous disease
- Cold-induced sweating syndrome
- Dysfibrinogenemia
- Exostoses
- Gaucher disease
- GM2-gangliosidosis
- Griscelli syndrome
- Hemochromatosis
- Liddle’s syndrome
- Meckel syndrome
- Nephronophthisis
- Omenn syndrome
- Persistent Mullerian duct syndrome
- Thyrotropin-releasing hormone deficiency/resistance
- Trichothiodystrophy
References

[1] S. Birnbaum, K. U. Ludwig, H. Reutter, S. Herms, M. Steffens, M. Rubini, C. Baluardo, M. Ferriani, N. Almeida de Assis, M. A. Alblas, S. Barth, J. Freudenberg, C. Lauster, G. Schmidt, M. Scheer, B. Braumann, S. J. Berge, R. H. Reich, F. Schießke, A. Hemprich, S. Potzsch, R. P. Steegers-Theunissen, B. Potzsch, S. Moebus, B. Horsthemke, F. J. Kramer, T. F. Wienker, P. A. Mossey, P. Propping, S. Cichon, P. Hoffmann, M. Knapp, M. M. Nothen, and E. Mangold. Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. Nat. Genet., 41(4):473–477, 2009.

[2] S. Brohee and J. van Helden. Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics, 7:488+, 2006.

[3] K. R. Brown and I. Jurisica. Online predicted human interaction database. Bioinformatics, 21(9):2076–2082, May 2005.

[4] E. F. Firoz, M. Warycha, J. Zakrzewski, D. Pollens, G. Wang, R. Shapiro, R. Berman, A. Pavlick, P. Manga, H. Ostrer, J. T. Celebi, H. Kamino, F. Darvishian, L. Rohnitzky, J. D. Goldberg, I. Osman, and D. Polsky. Association of mdm2 snp309, age of onset, and gender in cutaneous melanoma. Clin. Cancer Res., 15(7):2573–2580, 2009.

[5] H. B. Fraser and J. B. Plotkin. Using protein complexes to predict phenotypic effects of gene mutation. Genome Biol., 8:R252+, November 2007.

[6] S. Kathiresan, B. F. Voight, S. Purcell, K. Musunuru, D. Ardissino, et al. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet., 41(3):334–341, 2009.

[7] S. Kohler, S. Bauer, D. Horn, and P. N. Robinson. Walking the interactome for prioritization of candidate disease genes. Am. J. Human Genet., 82(4):949–958, 2008.

[8] K. Lage, O. E. Karlberg, Z. M. Storling, Páll, A. G. Pedersen, O. Rigina, A. M. Hinsby, Z. Tümer, F. Pociot, N. Tommerup, Y. Moreau, and S. Brunak. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol., 25(3):309–316, March 2007.

[9] D. Li, H. Suzuki, B. Liu, J. Morris, J. Liu, T. Okazaki, Y. Li, P. Chang, and J. L. Abbruzzese. Dna repair gene polymorphisms and risk of pancreatic cancer. Clin. Cancer Res., 15(2):740–746, 2009.

[10] V. McKusick. Mendelian inheritance in man and its online version, OMIM. Am. J. Hum. Genet., 80(4):588–604, 2007.

[11] M. Meila. Comparing clusterings—an information based distance. J. Multivariate Anal., 98(5):873–895, 2007.

[12] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics, 21:302–310, June 2005.
[13] S. Navlakha, R. Rastogi, and N. Shrivastava. Graph summarization with bounded error. In Proceedings of the 2008 ACM SIGMOD Conference, pages 419–432, 2008.

[14] S. Navlakha, M. C. Schatz, and C. Kingsford. Revealing biological modules via graph summarization. J. Comp. Biol., 16(2):253–264, 2009.

[15] S. Navlakha, J. White, N. Nagarajan, M. Pop, and C. Kingsford. Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In Proceedings of RECOMB 2009, LNCS, volume 5541, pages 400–417, 2009.

[16] M. Oti, B. Snel, M. A. Huynen, and H. G. Brunner. Predicting disease genes using protein-protein interactions. J. Med. Genet., 43(8):691–698, August 2006.

[17] B. Schwikowski, P. Uetz, and S. Fields. A network of protein-protein interactions in yeast. Nat. Biotechnol., 18(12):1257–1261, December 2000.

[18] S. van Dongen. A cluster algorithm for graphs. Technical Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam, 2000.

[19] O. Vanunu and R. Sharan. A propagation-based algorithm for inferring gene-disease associations. In Proceedings of the German Conference on Bioinformatics, pages 54–63, 2008.

[20] X. Wu, R. Jiang, M. Q. Zhang, and S. Li. Network-based global inference of human disease genes. Mol. Syst. Biol., 4:189, May 2008.
| New association                        | Reference | Gene          | Computational method                                                                 |
|---------------------------------------|-----------|---------------|--------------------------------------------------------------------------------------|
| Myocardial Infarction to 21q22        | [6]       | PRMT2 (21q22.3) | Neighborhood (0.1), Oti1 (1), GS3 (0.04), MCL (0.03), Prop (0.12), RW (9.8e^{-5})   |
| Myocardial Infarction to 1p13         | [6]       | PSMA5 (1p13)  | GS1 (0.5), GS2 (0.5), MCL (0.8), VI-CutL (0.1), Prop (0.18), RW (2.3e^{-4})           |
| Myocardial Infarction to 1p13         | [6]       | BCAS2 (1p13.2)| Neighborhood (0.25), Oti1 (1), GS3 (0.04), MCL (0.03), Prop (0.17), RW (9.3e^{-5})   |
| Myocardial Infarction to 10q11        | [6]       | ALOX5 (10q11.2)| Neighborhood (0.2), Oti1 (1), GS1 (1.0), GS2 (1.0), MCL (0.5), VI-CutL (1.0), Prop (0.18), RW (0.02) |
| Cleft lip to 8q24                     | [1]       | MYC (8q24.12-q24.13)| Neighborhood (0.02), Oti1 (1), Prop (0.009), RW (6.1e^{-4})                           |
| Melanoma to MDM2                      | [4]       |               | Neighborhood (0.02), Oti1 (1), Prop (0.016), RW (8.2e^{-4})                           |
| Pancreatic cancer to ATM              | [9]       |               | Neighborhood (0.08), Oti1 (1), Oti2 (2), Oti3 (3), Prop (0.044), RW (2.7e^{-3})       |

Table 1: New associations from the literature which concur with computational network-based predictions made by methods considered in this study, and which are not currently in OMIM[10]. The second column refers to the study which linked a disease to either an interval or specific gene. The third column shows the gene prediction made by the computational method(s), listed in the fourth column along with its score in parenthesis.
Figure 1: Precision and recall on the OPHID network.
Figure 2: Performance changes with respect to different measures of homophily. The top two plots correspond to neighborhood homophily. Bars indicate the average precision (left) and recall (right) across the five methods for all diseases within the homophilic range. For example, the average precision of predictions for diseases with neighborhood homophily values between $60 - 69\%$ is about 95%. Least-squares lines fit the performance points of each method, with regression values shown in the legend next to the method’s name. Vertical bars indicate variance.