BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 8 2010, pages 1057–1063
doi:10.1093/bioinformatics/btq076

Systems biology

Advance Access publication February 24, 2010

The power of protein interaction networks for associating genes with diseases

Saket Navlakha and Carl Kingsford*

Center for Bioinformatics and Computational Biology, Institute for Advanced Computer Studies and Department of Computer Science, University of Maryland College Park, College Park, MD 20742, USA

Associate Editor: Thomas Lengauer

ABSTRACT

Motivation: Understanding the association between genetic diseases and their causal genes is an important problem concerning human health. With the recent influx of high-throughput data describing interactions between gene products, scientists have been provided a new avenue through which these associations can be inferred. Despite the recent interest in this problem, however, there is little understanding of the relative benefits and drawbacks underlying the proposed techniques.

Results: We assessed the utility of physical protein interactions for determining gene–disease associations by examining the performance of seven recently developed computational methods (plus several of their variants). We found that random-walk approaches individually outperform clustering and neighborhood approaches, although most methods make predictions not made by any other method. We show how combining these methods into a consensus method yields Pareto optimal performance. We also quantified how a diffuse topological distribution of disease-related proteins negatively affects prediction quality and are thus able to identify diseases especially amenable to network-based predictions and others for which additional information sources are absolutely required.

Availability: The predictions made by each algorithm considered are available online at http://www.cbcb.umd.edu/DiseaseNet

Contact: carlk@cs.umd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 25, 2009; revised on February 16, 2010; accepted on February 22, 2010

1 INTRODUCTION

To understand the molecular basis of genetic diseases, it is important to discover their causal genes. Typically, a disease is associated with a linkage interval on the chromosome if single nucleotide polymorphism (SNPs) in the interval are correlated with an increased susceptibility to the disease (Birnbaum *et al.*, 2009; Kathiresan *et al.*, 2009). These linkage intervals define a set of candidate disease-causing genes. Genes related to the same disease are also known to have protein products that physically interact (Goh *et al.*, 2007; Ideker and Sharan, 2008; Kann, 2007; Oti and Brunner, 2007). A class of computational approaches have recently been proposed that exploit these two sources of information—physical interaction networks and linkage intervals—to predict associations between genes and diseases (Chen *et al.*, 2006, 2009; Kohler *et al.*, 2008; Lage *et al.*, 2007; Oti *et al.*, 2006; Vanunu and Sharan, 2008; Wu *et al.*, 2008, 2009). Previous studies (Kohler *et al.*, 2008; Lage *et al.*, 2007; Oti *et al.*, 2006; Wu *et al.*, 2009) typically begin with an artificial disease subinterval and test how well they can identify a known causal gene from among a fixed number of nearby genes in the query subinterval. In this article, instead of ranking only genes in the subinterval, we rank all genes in all intervals related to a query disease. This more stringent approach is advantageous because it allows us to find disease-causing genes that lie in existing disease intervals but that were previously not associated with the disease. Consequently, we can gauge a gene's relatedness to any query disease.

Several techniques for uncovering gene–disease associations take an integrative approach, leveraging Gene Ontology annotations (Aerts *et al.*, 2006; Franke *et al.*, 2006; Gaulton *et al.*, 2007; Perez-Iratxeta *et al.*, 2007; Radivojac *et al.*, 2008; Sam *et al.*, 2007), gene expression (Aerts *et al.*, 2006; Franke *et al.*, 2006; Gaulton *et al.*, 2007; Karni *et al.*, 2009; Ma *et al.*, 2007), protein sequence (George *et al.*, 2006; Perez-Iratxeta *et al.*, 2007; Radivojac *et al.*, 2008), biological pathways (Aerts *et al.*, 2006; Franke *et al.*, 2006; George *et al.*, 2006), text mining (Ozgur *et al.*, 2008; van Driel *et al.*, 2006), transcription factor binding sites (Aerts *et al.*, 2006) and various phenotypic traits of diseases (Freudenberg and Propping, 2002). Recent studies (Lage *et al.*, 2007; Wu *et al.*, 2008) have suggested that network-based predictions can be of comparable quality with current integrative approaches. We focus here on isolating protein–protein interaction (PPI) networks and linkage intervals to determine how much information is readily extractable from them for predicting gene–disease associations. Any improved network-based analysis can subsequently be incorporated into a more comprehensive, integrative system (Linghu *et al.*, 2009).

We compare approaches based on direct network neighbors (Oti *et al.*, 2006, and neighborhood), unsupervised graph partitioning [graph summarization (GS; Navlakha *et al.*, 2008) and Markov clustering (MCL; Van Dongen, 2008)], semi-supervised graph partitioning (VI-Cut; Navlakha *et al.*, 2009a), random walks (Kohler *et al.*, 2008) and network flow (Vanunu and Sharan, 2008), plus several of their variants (see Table 1 and Section 2). Trends in the precision and recall achieved by these computational methods yield several insights about the utility of PPI networks for uncovering gene–disease associations. We find that random walk

*To whom correspondence should be addressed.

Table 1. The primary methods compared in this study

Method	Type of analysis
Neighborhood	Network neighbors
Oti1	Network neighbors
GS	Unsupervised clustering
MCL	Unsupervised clustering
VI-Cut	Semi-supervised clustering
Random walks (RW)	Random walks with restarts
Flow propagation (Prop)	Network flow with priors

Several variations are also considered, including requiring more corroborating interactions than Oti1 (Oti2, Oti3), hierarchically summarizing the graph (GS2, GS3, GS-All) and choosing larger (VI-CutL) and smaller (VI-CutS) clusters.

approaches outperform all other tested classes of methodologies, with performance ranging from high precision and low recall (92% and 1%, respectively) to low precision and mediocre recall (17% and 38%, respectively) based on the parameters used. The graph clustering methods, which have not previously been tested in this domain, mostly perform better than the neighborhood approaches. When only using linkage intervals (without the network), we find substantially lower performance, as is the case when using only the network (without linkage intervals). However, in this latter scenario, graph clustering methods can be more precise than the other methodologies. This suggests that the proper choice of method and parameters depends on the setting.

We also quantify the relationship between the quality of predictions for a disease and the topological distribution of its related proteins in the network. As one would expect, we obtain better results for diseases whose proteins are situated near one another in the network. The measured relationship between closeness (homophily) and performance can be used to estimate precision and recall per disease a priori. The lower precision observed on diseases whose genes are spread apart in the network also suggests that making high-quality predictions for these diseases warrants the integration of more information sources and is where future computational efforts should be directed. We compare the actual predictions made by each method and find that most methods make some correct predictions not made by any other method, and that there are very few incorrect predictions made by multiple methods. Consequently, we show that combining these methods using a consensus Random Forest classifier results in Pareto optimal performance. Given the wide range of approaches considered, the consensus method may be considered the current performance of the network itself for determining gene–disease associations.

2 METHODS

2.1 Protein interaction network and gene–disease annotations

We constructed a PPI network from the Human Protein Reference Database (HPRD Release 7; Keshava Prasad *et al.*, 2008). The entire network contained 9182 proteins and 36 169 interactions. We considered only its main component, which consisted of 8776 proteins and 35 820 interactions. A second network was constructed from the Online Predicted Human Interaction Database (OPHID; Brown and Jurisica, 2005). This larger network contained 9842 proteins and 73 130 interactions. Neither of these databases provided weights associated with their interactions, hence we considered them unweighted.

Diseases were associated with genes and linkage intervals using annotations from the Online Mendelian Inheritance in Man (OMIM; McKusick, 2007) morbid-map file. Diseases that roughly shared the same first name were grouped into disease families as previously done (Kohler *et al.*, 2008; Oti *et al.*, 2006). In the remainder of this text, we refer to a 'disease family' simply by 'disease'. Diseases currently associated with only one gene were discarded in order to facilitate cross-validation testing. Loci for 8470 of the 8776 genes were obtained from UniProt (The UniProt Consortium, 2008). In the HPRD network, 1415 genes were associated with at least 1 of the 450 diseases. There were 189 genes associated with diseases according to OMIM, but which did not lie in any of the disease's recorded linkage intervals according to UniProt. We resolved these incompatibilities by assigning those genes to some linkage interval associated with the disease. Of the annotated genes, an average of 4.60 genes were associated with each disease, and on average 1.46 diseases were associated with each annotated gene. Each disease defined a set of intervals which covered an average of 397 genes.

2.2 Network-based algorithms to predict gene–disease associations

A widely used (Nabieva *et al.*, 2005; Schwikowski *et al.*, 2000) network-based approach ('Neighborhood') predicts for a protein p the annotations that are associated with more than θ percent of p's network neighbors. The method of Oti *et al.* (2006) associates a gene with a disease if it lies within a linkage interval associated with the disease and interacts with ≥ 1 gene annotated with the disease. Our variants ('Oti2' and 'Oti3') require ≥ 2 and ≥ 3 such genes, respectively.

Random walks have been used to transfer annotations within networks (Chen *et al.*, 2009; Kohler *et al.*, 2008). Kohler *et al.* (2008) define a random walk ('RW') starting from genes known to be associated with a query disease d. At each time step, the walk has a probability r of returning to the initial nodes. We set $r = 0.75$, as was done by Kohler *et al.* (2008). Once the process converged (L_2-distance between probability vectors in consecutive time steps $< 10^{-6}$), a prediction was made for all genes in relevant intervals with visitation probability greater than θ. A similar flow propagation algorithm was given by Vanunu and Sharan (2008), which we refer to as 'Prop'.

Graph partitioning is a promising technique for predicting gene–disease associations because it can uncover functional modules in PPI networks, and phenotypically similar diseases are often caused by proteins that have similar biological processes (Fraser and Plotkin, 2007; Wu *et al.*, 2008). We tested three graph partitioning algorithms that were recently shown (Brohee and van Helden, 2006; Navlakha *et al.*, 2009a, b) to find the most biologically relevant modules: GS (Navlakha *et al.*, 2008), MCL (Van Dongen, 2008) and VI-Cut (Navlakha *et al.*, 2009a). GS1 losslessly compresses the input network, producing a smaller summary network and a list of corrections to over-generalizations in the summary. The nodes in this summary correspond to modules in the input network. The summary graph can be further compressed by discarding the list of corrections and applying GS again, resulting in larger modules ('GS2'). This process can be repeated i times, yielding a 'GSi' method. The 'GS-All' method makes the union of the predictions made by GS1, GS2 and GS3. VI-Cut is a semi-supervised clustering method that uses annotations in the training set when creating modules. We test two variants of VI-Cut, dubbed 'VI-CutS' and 'VI-CutL', which break ties by favoring smaller and larger modules, respectively. A complete description of the algorithms is provided in Supplementary Section 1.

Since code is not available for the machine learning methods of Wu *et al.* (2008) and Lage *et al.* (2007), we were unable to test their algorithms on our framework. Both methods predict human genetic diseases drawn from the OMIM database, but differ slightly in the exact diseases, interactions and validation methodology used. Each of them also define a similarity measure

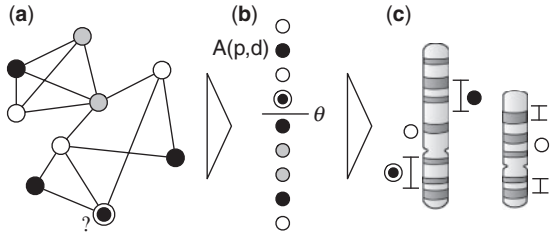

Fig. 1. (**a**) The disease annotations (if any) are discarded from one protein p (double-circled node), and an attempt is made to predict these annotations as follows. (**b**) For each disease d, an algorithm A is used to give a score $A(p,d)$ measuring how much p appears to be associated with disease d. If $A(p,d) \geq \theta$, the p-d association is considered as a candidate. (**c**) Finally, candidates are filtered based on genetic intervals known to be associated with disease. A p-d association is predicted if $A(p,d) \geq \theta$ and p lies in a chromosomic interval known to be associated with disease d.

between diseases, which allows them to include diseases in the test set for which only one causative gene is known.

Finally, we consider a consensus method that incorporates all 13 tested methods into a Random Forest classifier (Breiman, 2001; Witten and Frank, 1999). For each tested gene–disease pair, a 13D vector was created with entries containing each method's score for the pair. A vector was classed as *yes* if its gene was known to be associated with its disease, otherwise it was classed as *no*. To predict a gene–disease association, we required a minimum *yes* probability of θ, which we varied from 0.5 (default) to 0.9.

2.3 Testing framework

To test each potential protein–disease association p-d, we used leave-one-out cross-validation (Fig. 1). The algorithms described above were used to compute a score $A(p,d)$ for each possible disease d that is associated with an interval containing p. When scoring p-d, all disease associations known for p are discarded. The score $A(p,d)$ was then compared with a specified threshold θ, with higher thresholds yielding more conservative predictions.

True positives (TP) are those p-d associations with $A(p,d) \geq \theta$, where protein p is contained within an interval known to be associated with disease d and for which p is known to be associated with d. False positives (FP) are those p-d associations for which $A(p,d) \geq \theta$, with p contained in an appropriate interval, but for which p is not currently known to be associated with d. We conservatively considered predictions made for any of the 7361 unannotated genes in the network as incorrect, even though some of these predictions might in fact be novel associations. False negatives (FN) are p-d associations for which p is known to be associated with d but $A(p,d) < \theta$. Precision is TP/(TP+FP), the number of correct predictions made divided by the total number of predictions. Recall is TP/(TP+FN), the number of correct predictions divided by the total number of possible correct protein–disease associations.

For neighborhood and clustering algorithms, $A(p,d)$ was the percentage of p's neighbors or co-clustered proteins that were associated with disease d, with the threshold θ varying between 5% and 90%. For random walk methods (RW and Prop), $A(p,d)$ was the visitation probability of p in the random walk started from seed genes associated with d. For RW, we varied θ between 0.01% and 9% (Supplementary Material).

2.4 Quantifying homophily

We quantified the relationship between predictive performance and the topological distribution of the disease proteins in the network using two measures. These measures are designed to assess whether a set of proteins (that are associated with a given disease) is located in dense pockets in the network or is more uniformly distributed. The first, average pairwise distance, is the average number of interactions separating two proteins associated with a disease. A similar idea was recently used by Radivojac *et al.* (2008) as one of many integrative features in an support vector machine (SVM) to predict disease annotations, and by Lavallee-Adam *et al.* (2009) to quantify the distribution of Gene Ontology (Ashburner *et al.*, 2000) annotations in a PPI network. This measure is reasonable when all proteins are in one dense region, but is incorrectly large in instances where the nodes are located in several dense but well-separated regions in the network. A second measure, neighborhood homophily, does not suffer from this problem. The neighborhood homophily of disease d is the average percentage of network neighbors of a disease d protein also known to be associated with d.

3 DISCUSSION

3.1 The quality of network-based predictions

Predictions were made by each computational method (Table 1) as described in Section 2. The precision and recall for each method on the HPRD (Keshava Prasad *et al.*, 2008) network is shown in Figure 2a. Lines connect the performance for the same method using different prediction thresholds. Performance points where recall dropped <1% were removed. There was a wide range of performance among all the methods tested, with precision ranging between 17.0% and 92.3% and recall between 1.2% and 37.6% (Fig. 2a and b).

The random walk methods [RW (Kohler *et al.*, 2008) and Prop (Vanunu and Sharan, 2008)] show a clear dominance over the clustering and neighborhood methods. The similar performance of RW and Prop is not surprising because the prior-evidence vector of Vanunu and Sharan (2008) is similar in principle to the restart probability in the random walk of Kohler *et al.* (2008). Thus, although couched in different terms, RW and Prop are closely related. The slight advantage to RW might be attributed to the fact that Prop's prior-evidence vector pumps one unit of flow along each edge, instead of normalizing by a node's degree. Hence, there may be a bias toward annotations from high-degree nodes. As the threshold increases both methods gain in precision, with RW plateauing at 92.3% precision, the highest of any single method. The generally superior performance of the random walk methods suggests that the clustering and neighborhood methods are too restrictive when defining their locality.

The clustering methods [MCL (Van Dongen, 2008), VI-Cut (Navlakha *et al.*, 2009a), and GS1 (Navlakha *et al.*, 2008) and its variants GS2, GS3 and GS-All], which have not previously been appraised for the task of predicting gene–disease associations, performed slightly worse than the random walk methods, but better than the neighborhood approaches. They achieve between 18.4% and 68.6% precision and 1.1% and 17.9% recall. The performance of GS1 and GS2 was similar, though GS2 covered a wider range of precision and recall. GS3 created too few modules and performed relatively poorly by itself. Taking the union of GS1, GS2 and GS3 (GS-All) improved over GS1 and GS2 by yielding a higher recall, and improved over GS3 in both precision and recall. This suggests that iteratively compressing the PPI network yields informative modules. MCL extended the range of precision and recall further than GS2, but still fell within a tight linear band along which most clustering methods lie. VI-Cut incorporates known annotations when finding clusters, unlike GS and MCL, which are unsupervised approaches. VI-CutS breaks ties by choosing smaller, more homogeneous clusters, and, as a result, yielded a high precision

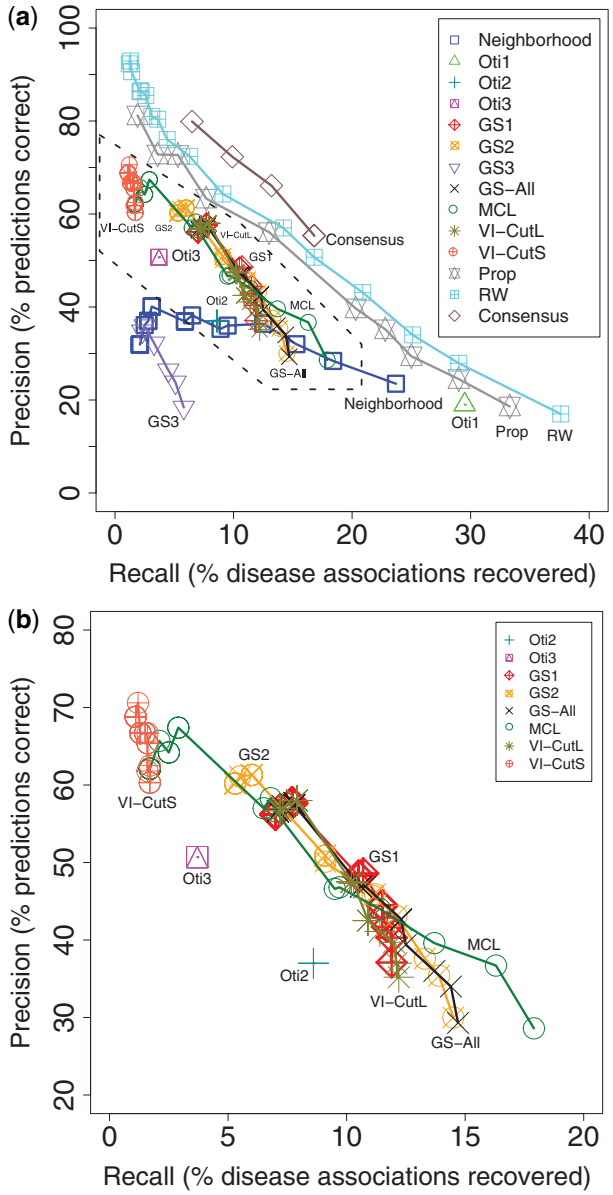

Fig. 2. Performance of the methods. (**a**) Precision and recall for each method using leave-one-out cross-validation on the HPRD network. The random walk methods individually perform the best, followed by the clustering and neighborhood approaches. The consensus method, which combines predictions made by all methods using a Random Forest classifier, outperforms all other methods. (**b**) A magnification of the dashed region corresponding to the clustering methods.

(average of 66.0%), albeit a very low recall (1.4% on average). VI-CutL breaks ties by choosing larger clusters and therefore yielded a lower precision but a higher recall. Across all clustering methods, smaller clusters produced more precise predictions. The similar performance of the many clustering algorithms tested suggests that their utility for predicting gene–disease associations lies within a well-defined range.

The Neighborhood and Oti methods each make predictions by only considering the annotations of the neighbors of a protein. Predictions made by Neighborhood ranged in precision from 23.5% to 40.1% and 2.1% to 23.7% recall, depending on the prediction threshold θ used. The Oti methods do not vary with respect to θ and are therefore shown as single points in Figure 2. Oti1 yielded a recall of 29.5% with still a relatively high precision (19.0%). Oti2 and Oti3 both drive up the predictive confidence by requiring more seed proteins to interact with the candidate protein. Both not only have successively higher precision than Oti1, but also have successively lower recall. Oti4 showed no improvement over Oti3. For the clustering and neighborhood approaches, precision improved as θ increased from 0% to 50%, but remained relatively stable for $\theta \geq 50\%$, indicating that a 50% prediction threshold is appropriate and that there are few competing majority annotations among the cluster or network neighbors of a protein.

The same experiments on the OPHID (Brown and Jurisica, 2005) network yielded similar performance for all methods (Supplementary Fig. S1).

3.2 Interplay between linkage intervals and interaction information

A disease is typically associated with a linkage interval if SNPs in that interval result in an increased susceptibility to the disease. The actual causal genes for the disease could lie anywhere in the interval. To understand how much added benefit the network provides in identifying the target genes, we considered two baseline genomic methods that only used linkage intervals, ignoring the network entirely. The first method predicted, for each disease d, x random genes within linkage intervals known to be associated with d, where x is the known number of d-causing genes. This resulted in 1.6% precision and 1.6% recall on average. The second method predicted a disease for all genes contained within the disease's intervals [i.e. $A(p,d)=\infty$ for all p,d in related intervals]. This resulted in 100.0% recall but only a 1.2% precision. We also tested the quality of the predictions made by each network-based method assuming linkage interval information is not available. Again, we found a large drop in performance compared to using linkage intervals and PPI networks in conjunction. The random walk methods achieved precision ranging from 1.3% to 37.1%, and recall ranging from 1.2% to 29.0%. Some clustering methods were more precise in this scenario [precision between 4.4% (MCL) and 48.0% (VI-CutS) and recall between 1.1% (VI-CutS) and 17.9% (MCL)], which suggests that some clusters found represent true disease modules, and that the random walk methods benefit more from the filter that linkage intervals provide. Undoubtedly, linkage intervals or networks by themselves are not sufficient to make high-quality predictions; however, such predictions can be anecdotally useful. Recent literature (Birnbaum *et al.*, 2009; Firoz *et al.*, 2009; Kathiresan *et al.*, 2009) reports several gene-disease associations that are not currently in OMIM but which were uncovered by one of the tested methods when run without using linkage intervals. A table summarizing these predictions is in Supplementary Table T1.

3.3 Prediction quality per disease

Performance varied widely when assessed on a per-disease basis. For each disease, we computed the maximum precision and maximum recall separately across the 13 methods. The number of diseases for which performance is within each precision–recall range is given in Figure 3. Row and column sums are shown at the margins. Very

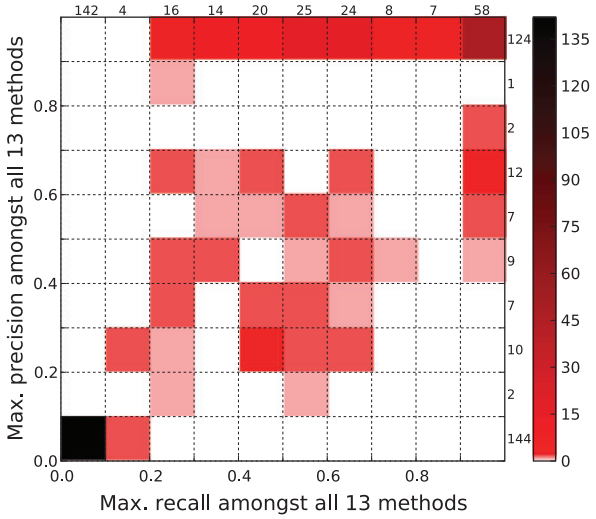

Fig. 3. Upper bound on achievable performance. Each (x, y) square is colored by the number of diseases that had maximum recall x and maximum precision y across all 13 methods using the prediction threshold for each method corresponding to roughly 10% recall.

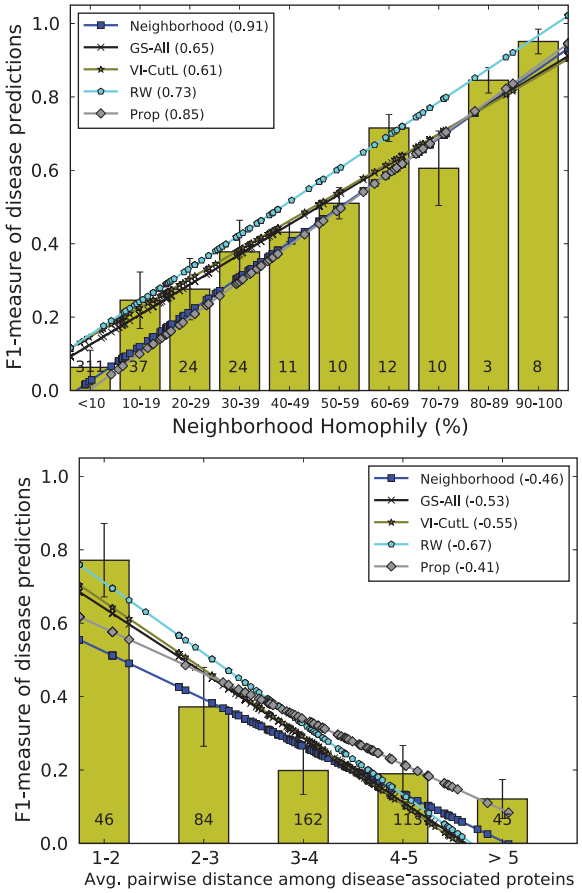

Fig. 4. Disease homophily versus prediction quality. The effect of disease homphily on the quality of the predictions made for that disease. The x-axes correspond to homophily, measured via (**a**) neighborhood homophily, and (**b**) the average pairwise distance of a disease. The y-axes are the F_1-measure (harmonic mean of precision and recall) of the predictions for the disease. Least squares fit lines are shown for each method, with regression values in the legend. Vertical bars indicate variance. The trends uniformly indicate that the lower the average pairwise distance and higher the percentage of similarly annotated neighbors, the better the predictions. Numbers in bars give the count of diseases with the given level of homophily.

good performance can be achieved for many diseases using some method. In particular, there are 124 diseases for which the maximum precision is >90%. There were 19 diseases for which at least half of the 13 methods achieved precision >90% (Supplementary Section 4). Assuming the optimal method is chosen per disease, this is the upper bound on the best performance possible.

Fourteen diseases have at least three associated proteins and achieved maximum precision >90% and maximum recall >90% for some combination of methods (Bare lymphocyte syndrome, Bernard–Soulier syndrome, Dysfibrinogenemia, Elliptocytosis, Epidermolysis, Griscelli syndrome, Heinz body anemia, Hemochromatosis, Mismatch repair cancer syndrome, MODY diabetes, Nephronophthisis, Ovarioleukodystrophy, Thalassemia and Trichothiodystrophy). Fanconi anemia, which has been experimentally shown (Macé *et al.*, 2005; Pan, 2008) to have protein products that interact, had a maximum precision and recall of 100% and 69.2%, respectively. There were also 144 diseases that had a maximum precision <10%. These are the diseases for which the network seems to provide little information and for which new computational methods or additional data are absolutely required.

The existence of proteins implicated in the same disease that do not interact has been shown to adversely affect predictive performance for a disease (Lage *et al.*, 2007). We can quantify the degree to which proteins associated with the same disease tend to be located near each other in the network using measures of homophily, such as neighborhood homophily and average pairwise distance (Section 2). Predictions made for more homophilic diseases were typically of higher quality than those made for diseases that do not exhibit strong homophily. Figure 4a shows how predictive performance varies as a function of neighborhood homophily for five representative methods (Neighborhood, GS-All, VI-CutL, Prop and RW) using the prediction threshold for each method that corresponds to roughly 10% recall. The bars in the plot indicate the average F_1-measure (harmonic mean of precision and recall) of predictions made by all five methods for diseases with neighborhood homophily in the given range. Error bars show the variance across the five methods. Five least squares lines fit the performance points of each method, with regression values shown in the legend. (Separate precision and recall plots for each homophily measure are in Supplementary Fig. S2). Even the methods that do not directly use network neighbors (i.e. GS-All, VI-CutL and RW) showed a significant correlation with neighborhood homophily.

A similar dependence was seen for average pairwise distance (Fig. 4b). On average, as the distance between disease-related proteins grew, performance worsened. Thus, homophily can be used to provide an a priori estimate of the quality of network-based predictions for a given disease.

3.4 Consensus classifier improves predictions

The methods considered here use a variety of techniques to extract predictions from the PPI network, and consequently each might be expected to make successful predictions for genes not correctly

handled by other methods. To quantify this, we define the uniqueness of method M to be the percentage of correct predictions made by M that were not made by any other method. When more methods are included in such an analysis, the uniqueness for each method will generally decrease.

We considered the five representative methods (Neighborhood, VI-CutL, GS-All, Prop and RW), using the prediction threshold for each method that corresponds to roughly 10% recall. All five methods made predictions that were not made by the other four. In particular, 18.9%, 7.6%, 5.0%, 3.1% and 30.7% of the correct predictions made by these five methods, respectively, were unique. The incorrect predictions were also not shared across the methods. Among the five methods, 976 total predictions were made, yet only 19 (1.9%) were both wrong and made by all five methods.

This implies that, although random walks individually perform the best, an aggregate method that combines several of the network analysis strategies will be useful. In particular, in Figure 2a, we show the performance of a consensus method using an ensemble of decision trees (Section 2.2). All five points are Pareto optimal over all other methods (meaning no other method has higher precision and recall). The superior performance of the consensus method indicates that many of the individual methods capture different kinds of structure in the network and that these individual abilities can be used in tandem to make higher quality predictions.

4 CONCLUSION

The classes of network-based methods considered here each approached the task of predicting gene–disease associations using very different philosophies. Although random walk approaches are superior to clustering and neighborhood approaches, we showed that all methods make unique predictions and can be used together to increase performance. We also quantified the relationship between disease homophily and prediction quality, and found certain diseases for which high-throughput PPI networks were an especially useful source from which to make high-quality predictions. Diseases that have little correlation with the interaction network call for higher quality networks or an integrative approach that considers sequence, functional annotations, expression data or other additional information.

Funding: National Science Foundation (0812111 and 0849899 to C.K.).

Conflict of Interest: none declared.

REFERENCES

Aerts,S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nat. Biotechnol.*, **24**, 537–544.

Ashburner,M. *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Birnbaum,S. *et al.* (2009) Key susceptibility locus for nonsyndromic cleft lip with or without cleft palate on chromosome 8q24. *Nat. Genet.*, **41**, 473–477.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Brohee,S. and van Helden,J. (2006) Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics*, **7**, 488–507.

Brown,K.R. and Jurisica,I. (2005) Online predicted human interaction database. *Bioinformatics*, **21**, 2076–2082.

Chen,J.Y. *et al.* (2006) Mining Alzheimer disease relevant proteins from integrated protein interactome data. *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 367–378.

Chen,J. *et al.* (2009) Disease candidate gene identification and prioritization using protein interaction networks. *BMC Bioinformatics*, **10**, 73–87.

Firoz,E.F. *et al.* (2009) Association of mdm2 snp309, age of onset, and gender in cutaneous melanoma. *Clin. Cancer Res.*, **15**, 2573–2580.

Franke,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.

Fraser,H.B. and Plotkin,J.B. (2007) Using protein complexes to predict phenotypic effects of gene mutation. *Genome Biol.*, **8**, R252–R261.

Freudenberg,J. and Propping,P. (2002) A similarity-based method for genome-wide prediction of disease-relevant human genes. *Bioinformatics*, **18** (Suppl. 2), S110–S115.

Gaulton,K.J. *et al.* (2007) A computational system to select candidate genes for complex human traits. *Bioinformatics*, **23**, 1132–1140.

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

Goh,K.-I. *et al.* (2007) The human disease network. *Proc. Natl Acad. Sci. USA*, **104**, 8685–8690.

Ideker,T. and Sharan,R. (2008) Protein networks in disease. *Genome Res.*, **18**, 644–652.

Kann,M.G. (2007) Protein interactions and disease: computational approaches to uncover the etiology of diseases. *Brief. Bioinform.*, **8**, 333–346.

Karni,S. *et al.* (2009) A network-based method for predicting disease-causing genes. *J. Comput. Biol.*, **16**, 181–189.

Kathiresan,S. *et al.* (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat. Genet.*, **41**, 334–341.

Keshava Prasad,T.S. *et al.* (2008) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kohler,S. *et al.* (2008) Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.*, **82**, 949–958.

Lage,K. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.

Lavallee-Adam,M. *et al.* (2009) Detection of locally over-represented GO terms in protein-protein interaction networks. In *Proceedings of RECOMB 2009*, Vol. 5541 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pp. 302–320.

Linghu,B. *et al.* (2009) Genome-wide prioritization of disease genes and identification of disease-disease associations from an integrated human functional linkage network. *Genome Biol.*, **10**, R91.

Macé,G. *et al.* (2005) 3R coordination by Fanconi Anemia proteins. *Biochimie*, **87**, 647–658.

Ma,X. *et al.* (2007) CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, **23**, 215–221.

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

Nabieva,E. *et al.* (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**, 302–310.

Navlakha,S. *et al.* (2008) Graph summarization with bounded error. In *Proceedings of the 2008 ACM SIGMOD Conference*. ACM, NY, pp. 419–432.

Navlakha,S. *et al.* (2009a) Finding biologically accurate clusterings in hierarchical tree decompositions using the variation of information. In *Proceedings of RECOMB 2009*, Vol. 5541 of *Lecture Notes in Computer Science*. Springer, Heidelberg, pp. 400–417.

Navlakha,S. *et al.* (2009b) Revealing biological modules via graph summarization. *J. Comput. Biol.*, **16**, 253–264.

Oti,M. and Brunner,H.G. (2007) The modular nature of genetic diseases. *Clin. Genet.*, **71**, 1–11.

Oti,M. *et al.* (2006) Predicting disease genes using protein-protein interactions. *J. Med. Genet.*, **43**, 691–698.

Ozgur,A. *et al.* (2008) Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, **24**, i277–i285.

Pan,W. (2008) Network-based model weighting to detect multiple loci influencing complex diseases. *Hum. Genet.*, **124**, 225–234.

Perez-Iratxeta,C. *et al.* (2007) Update of the G2D tool for prioritization of gene candidates to inherited diseases. *Nucleic Acids Res.*, **35**, W212–W216.

Radivojac,P. *et al.* (2008) An integrated approach to inferring gene-disease associations in humans. *Proteins*, **72**, 1030–1037.

Sam,L. *et al.* (2007) Discovery of protein interaction networks shared by diseases. In *Pacific Symposium on Biocomputing*. World Scientific, Singapore, pp. 76–87.

Schwikowski,B. *et al.* (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

The UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res.*, **36**, D190–D195.

Vanunu,O. and Sharan,R. (2008) A propagation-based algorithm for inferring gene-disease assocations. In *Proceedings of the German Conference on Bioinformatics*. GI, Germany, pp. 54–63.

Van Dongen,S. (2008) Graph clustering via a discrete uncoupling process. *SIAM J. Matrix Anal. Appl.*, **30**, 121–141.

van Driel,M.A. *et al.* (2006) A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.*, **14**, 535–542.

Witten,I.H. and Frank,E. (1999) *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations (The Morgan Kaufmann Series in Data Management Systems)*, 1st edn. Morgan Kaufmann, San Francisco, CA.

Wu,X. *et al.* (2008)Network-based global inference of human disease genes. *Mol. Syst. Biol.*, **4**, 189.

Wu,X. *et al.* (2009) Align human interactome with phenome to identify causative genes and networks underlying disease families. *Bioinformatics*, **25**, 98–104.