Modeling and characterization of disease associated subnetworks in the human interactome using machine learning

Lee T. Sam, MA[1], George Michailidis, PhD[1]
[1]University of Michigan, Ann Arbor, MI

Abstract

The availability of large-scale, genome-wide data about the molecular interactome of entire organisms has made possible new types of integrative studies, making use of rapidly accumulating knowledge of gene-disease associations. Previous studies have established the presence of functional biomodules in the molecular interaction network of living organisms, a number of which have been associated with the pathogenesis and progression of human disease. While a number of studies have examined the networks and biomodules associated with disease, the properties that contribute to the particular susceptibility of these subnetworks to disruptions leading to disease phenotypes have not been extensively studied. We take a machine learning approach to the characterization of these disease subnetworks associated with complex and single-gene diseases, taking into account both the biological roles of their constituent genes and topological properties of the networks they form.

Introduction

Recent advances in gene-disease association and large scale protein interaction have made an unprecedented amount of data available for researchers to study the systems biology of human disease. Particularly, this interest has taken the form of analyses of combined protein interaction data and gene-disease annotations to elucidate the molecular mechanisms underlying human diseases and disorders. These studies suggest the presence of disease-related subnetworks within the larger human protein interaction network. This is consistent with the belief that diseases significantly dysregulate functional biomodules within the interactome. As a result, analysis of these subnetworks may provide insights into the functional modules within the interactome that are responsible for the pathogenesis and progression of human disease.

In this paper, we present a model-driven technique for constructing disease-associated sub-networks based on gene-disease interactions and protein interactions and characterize them using both the topological and biological properties of the constituent genes and the subnetworks they form. Three sets of subnetworks are generated from this process: a group of subnetworks involved in well-defined biological processes, and two groups of subnetworks associated with complex and single gene diseases. We apply unsupervised methods to demonstrate that these three subnetwork sets are poorly separable and train a random forest classifier to delineate between sub-networks specifically associated with disease and those built from *a priori* knowledge from the Gene Ontology in order to better understand the structural and biological characteristics of the biological processes associated with diseases arising from single genes and how they differ from those associated with complex disease through their classification.

Supplementary Methods and Materials for this study are available at http://www.stat.lsa.umich.edu/~gmichail/subnetworks_study/

Background

The advent of high-throughput techniques for determining molecular interactions has opened the door to genome scale evaluation of the molecular interactome of many species due to the quickly growing pool of data. A number of databases have been developed in order to integrate protein interaction data from high throughput experiments such as DIP, BIND, HPRD, and several others. Studies looking at this data across a number of organisms have indicated that these networks are organized into functional biomodules that function at multiple scales (1-3).

Analysis of disease gene knowledge coupled with data from large-scale protein interaction networks to form a phenome-interactome network have revealed that a significant portion of disease-associated genes form small sub-networks. The networks formed by the interactions of known disease genes have been used to relate phenotypically similar inherited diseases together (4). Similarly, subnetworks that represent protein complexes have been used to relate diseases with similar phenotypes and provide novel disease gene candidates when melded to association data (5). The disease-associated genes themselves also seem to possess a number of characteristics within the interactome. Compared to the mean degree values of all proteins, many disease related proteins display relatively elevated degree and tend to interact with other disease-related proteins (6, 7). This property has been used to propose likely candidate genes for disease association (8). Taken together, it suggests that the intermediate nodes in the interactome play a

contributory factor. In addition to the importance of highly interconnected "hub" proteins (9, 10), certain topological features were found to be associated with essentiality/lethality (11). Additional research has suggested that genes expressing proteins of similar importance also share topological characteristics in the interaction network (12). These topological characteristics have been used to explain variable disease outcome (13), making an argument for their role in the progression of disease.

In this study, node count, radius, and diameter are used to measure the size and spread of the networks. In graph-theoretic terms where eccentricity is defined as the greatest distance between a vertex and any other, the diameter and radius are defined as the maximum eccentricity and the minimum eccentricity in a network, respectively. The two degree measurements, clustering coefficient, and observed edge fraction, characterize the density and interconnectivity of the graphs, where degree is defined as the number of connections a vertex has to other vertices. Clustering coefficient analyzes the links in a graph to quantify how close it is to being completely connected with all vertices connected to all other vertices. The observed edge fraction is similar in counting the fraction of edges observed in the subnetwork compared to all possible edges. Cyclicity, defined as the existence of looping paths in the graph, and biconnectivity, defined as the presence of vertices which connect segments of the subgraph, are used to characterize the structure of the graphs.

A number of biological properties characterize the biomodules associated with biological processes and diseases. Genes involved with the same biological process or functional subunit often co-localize on the genome (14) and are often under the control of identical regulatory factors. In consideration of these positional factors, we take into account mean gene start location, mean gene end location, mean length, and mean genomic strand. Mean G-C content fraction is calculated as it affects thermostability of the genetic material and its transcriptional propensity. Similarly, sets of genes with interacting protein products contain motifs for known interacting domains. With this in mind, mean PFAM domain annotation count, mean ProSite annotation count, mean number of signal domains, and mean number of transmembrane domains are considered.

In this case, we applied a random forests ensemble learning method described by Breiman (15). The random forest is composed of a defined set of unpruned decision trees, each trained on a subset of the training data selected with replacement. Each tree chooses a random subset of variables to classify the data at each node, the quantity of which is defined as a parameter. These properties make the classifier extremely robust to overfitting on data.

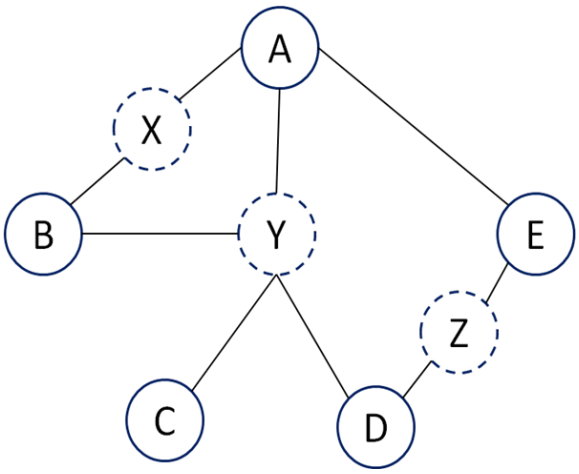

Figure 1. Derivation of an example subnetwork

Methods

Data Extraction. Protein interaction data was retrieved from the Michigan Molecular Interaction Index (MiMi) (16), which integrates interaction and annotation data from BIND , the Gene Ontology, HPRD, DIP, the BioGRID, IntAct, InterPro, IPI, the Max-Delbrueck Center for Molecular Medicine protein interaction database, Pfam, ProtoNet, SwissProt, and RefSeq. This process yielded 12,318 unique protein-protein interactions involving 6199 unique Entrez Gene identifiers. Gene-disease relationships were derived from two sources; the Online Mendelian Inheritance in Man (OMIM) (17) and the PhenoGO database (18). Gene-Disease associations in PhenoGO not using Entrez Gene identifiers were translated using mappings from HUGO (19). Diseases in these two resources were defined in terms of coded Medical Subject Heading (MeSH) (20) and Unified Medical Language System (UMLS) (21) identifiers. The unfiltered, translated data set resulted in 3469 Entrez identifiers associated to 2325 phenotype codes. OMIM mappings found in the mim2gene file supplied by NCBI already employ Entrez Gene identifiers and no translation was necessary for the OMIM data. Entries in the OMIM database were filtered to include only gene-disease references, resulting in 1846 distinct Entrez indentified genes annotated to OMIM-defined diseases. 708 of the identifiers found in the OMIM mappings are also present in the MiMi interaction data set. Gene Ontology (22) data and biological annotation was extracted from BioMart (23) using data from Ensembl version 47 built from the NCBI36 release of the human genome. MeSH and UMLS term descriptors were retrieved directly from the NLM.

Subnetwork Generation. The subnetworks associated to human diseases and biological processes were built by the determination of all shortest pairs paths between all distinct associated genes found in the protein interaction network. Shortest paths in the interaction subnetwork are determined using Dijkstra's shortest paths algorithm (24). For example, **Figure 1** illustrates a hypothetical disease of interest associated to UMLS concept 'UMLS:000000', associated with genes A, B, C, D, and E. The shortest path between pairs {A,B}, {A,C}, {A,D}, {A,E}, {B,C}, {B,D}, {B,E}, {C, D}, {C, E}, and {D, E} would be analyzed, noting the identities of the original nodes, the original node also found in the protein interaction network (as many nodes are not represented within the network), the intermediate connecting nodes, and the respective counts of each class. This process discovers intermediate nodes X, Y, and Z in the process of deriving the subnetwork and associates these nodes.

The generated results were split into three distinct classes. A "background" set was generated from *a priori* knowledge from the Gene Ontology, consisting of the subnetworks formed by the classes represented in the "Biological Process" and "Molecular Function" trees of the Gene Ontology. This process resulted in the generation of 6,606 GO-associated subnetworks. A "single gene disease" (**SGD**) subnetwork set was generated from the contents of OMIM, producing 2,079 subnetworks. A "complex disease" (**CD**) set was built from the PhenoGO annotations, composed of 2,317 subnetworks in total.

Data Characterization and Filtering. Resulting subnetworks in each of the three data sets was topologically characterized using a set of Perl scripts employing the Boost Graph Library interface. Subnetworks are topologically characterized based on node count, clustering coefficient, observed edge fraction, average degree, maximum degree, radius, diameter, cyclicity, and biconnectivity. Biological characteristics noted for each subgraph include mean gene start location, mean gene end location, mean length, strand, mean PFAM domain annotation count, mean ProSite annotation count, mean number of signal domains, mean number of transmembrane domains, and mean G-C content fraction. The networks are filtered for size, imposing a minimum of three nodes found in the interaction network. 79 and 278 subnetworks passed this filter from the SGD and CD sets, respectively. 2590 of the subnetworks generated from the Gene Ontology passed this filter. This final filtered set was used to train and test the classifier.

Machine Learning and Classification. The Waikato Environment for Knowledge Analysis (Weka), version 3.4.12 (25) was used to train and test a random forest classifier with a stratified 10-fold cross validation methodology. In this case, the cross-validation approach was chosen due to the relative paucity of data from the disease subsets. Each random forest was composed of 100 trees, each taking into account four random parameters from the data. In all, a total of nine classifications were done in an attempt to discretize the three sets of subnetworks using varying parameter sets and amalgamations of the two disease sets. Because the Weka random forest classifier did not provide variable importance measures, the analysis was repeated using the randomForest package in R 2.7.1, which provided nearly identical results. Principal components analysis of the data was done using PAST (26).

		Assigned to Cluster		
		GO	**SGD**	**CD**
Source	**GO**	59	4	16
	SGD	1220	435	932
	CD	158	31	89

Table 1: Unsupervised k-means clustering illustrates the poor separability of the data, with 1631 (55.4%) instances incorrectly clustered

Correctly Classified Instances		2795		94.94 %	
Incorrectly Classified Instances		149		5.06%	
TP Rate	FP Rate	Precision	Recall	f-Measure	class
0.101	0.003	0.5	0.101	0.168	SGD
0.997	0.387	0.949	0.997	0.972	GO
0.752	0.001	0.986	0.752	0.853	CD

Table 2. Classification of CD, SGD, and GO classes using all variables

Results

Subnetwork Charactersitics. As expected, the subnetworks derived from OMIM, the SGD set, demonstrated a smaller range in size in terms of total gene count from 3 to 32 genes with a median of five genes, while the PhenoGO derived complex disease set was composed of networks of size ranging from 3 to 127 genes, with a median of eight genes. The Gene Ontology derived background set had the largest range from 3 to 968 genes. As shown in **Sup. Figures 2a-c**, most subnetworks tended to remain small, generally involving between three and nine genes. The GO background set exhibits a long-tailed distribution with most networks remaining under seventeen genes in size.

Classification Accuracy. Unsupervised Principal Components Analysis and k-means clustering methods were first attempted in order to assess the separability of the three classes of subnetworks. As shown in and

Sup Figures 1a and 1b, clustering mirrored the results of the PCA with high misclassification levels (misclassifying ~55% of the data), further demonstrating the poor separability of the data.

As a result, machine learning techniques must be applied to derive the subtle differences between the CD, SGD, and GO sets. As shown in **Sup. Tables 2a-2i**, the overall misclassification error rate remains relatively low across several subsets of the subnetwork parameter data, never exceeding 5%. Other measures – precision, recall, f-measure- exhibit very satisfactory performance. However, a close inspection of the results for the three class problems (SGD, GO, CD) reveals that the results for the SGD class are not satisfactory. Confusion matrices from these analyses show the classifier tends to assign those subnetworks to the GO class, an issue addressed in the discussion section. Further analysis of the data by breaking down the features into biological and topological characteristics further revealed the similarities between the SGD and GO set, further analyzed in the **Supplementary Methods and Materials**. The separability of the SGD and CD sets as shown in **Sup. Table 2j** demonstrates the differences in subnetwork characteristics between those primary involved with single-gene disorders and those associated with multigenic, complex disorders. A reclassification of all the study data was also done using a GO dataset that included only the "Biological Process" entires, with similar results. The complete results of the classifications as well as additional methods and analyses are available in the **Supplementary Methods and Materials**.

The most important variables in the classification of subnetworks to their individual classes is illustrated in **Figure 2** as derived using the reduction in Gini index, a measure of the reduction in misclassification when a particular variable is used.

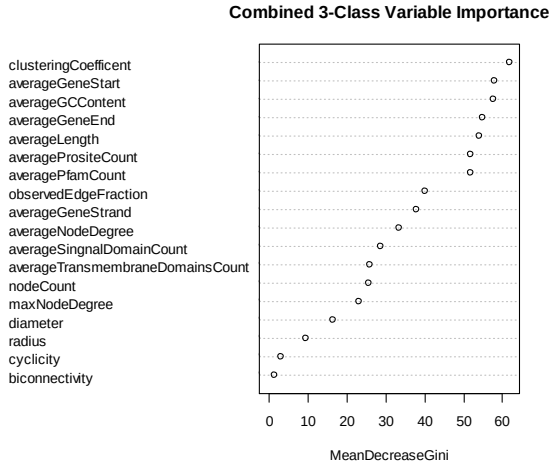

Figure 2: Variables ranked by importance in classification

Discussion and Conclusion

The relative paucity of data describing disease-associated subnetworks continues to present a serious challenge in the analysis of the functional biomodules underlying human disease. While the classification of complex disease-associated subnetworks appears to achieve reasonable results, the underlying heterogeneity of human disease, as evidenced by the SGD set, will always present a problem in classification.

It is notable that the variables with the highest influence are a mix of both topological and biological factors, confirming previous findings that characteristics from both categories play an important role in the susceptibility to biological disruption and resulting disease. The relative importance of clustering coefficients confirms recent results examining the differences between disease-associated genes and essential genes (27). The inclusion of mean gene start locus and GC content confirm the relative importance of genomic localization and transcriptional propensity (28). While the examination of individual factors increases confidence in the findings through recapitualation of established study results, the random forest is able to capture the interaction between these variables. These inter-variable interactions are a prime target for continued study.

It is not completely surprising that the SGD subnetworks appear to bear a strong resemblance to the GO background considering the pathogenesis of diseases that arise from anomalies in a single gene. In many cases, the GO-derived subnetworks can be considered functional biomodules of the interactome. The disruption of certain genes in these functional biomodules is likely to manifest in the form of disease phenotypes if they are not serious enough to result in lethality. This can result in failures of protein complex assembly and complementation such as in Xeroderma Pigmentosum, a single gene disease that can arise from any one of the seven known genes in the XPA-XPG complementation group associated with nucleotide excision repair. As such, these two classes are relatively poorly separable even in a supervised machine learning context.

As we expected, the differences between the networks formed by sets of genes associated with biological processes and those associated with human disease are subtle and not easily derived as they are, by definition, intimately linked. The similarity between the single gene disease-associated subnetworks and those derived

from the Gene Ontology demonstrates the multiscale behavior of a single disruption in a functional biomodule, and its ability to cause debilitating effects. The need for additional data and high specificity data is made abundantly clear in this study, as demonstrated by the propensity for misclassification of complex disease-associated subnetworks as well as the limited number of subnetworks derived from the data due to lack of representation in the interaction network. The limited availability of interaction propensity or data quality measures associated with individual interactions in the particular version of the interaction database we employed led us to treat all interactions as equally probable and equally correct. This may be a source of error in the process that may be ameliorated in the future with additional data and quantitative measures associated with the interactions. As more gene-disease association data becomes available, the effectiveness of this method should be re-evaluated.

References

1. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, et al. Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. Science. 2001 May 4;292(5518):929-34.
2. Rives AW, Galitski T. Modular organization of cellular networks. Proc Natl Acad Sci U S A. 2003 Feb 4;100(3):1128-33.
3. Petti AA, Church GM. A network of transcriptionally coordinated functional modules in Saccharomyces cerevisiae. Genome Res. 2005 Sep;15(9):1298-306.
4. Sam L, Liu Y, Li J, Friedman C, Lussier YA. Discovery of protein interaction networks shared by diseases. Pac Symp Biocomput. 2007:76-87.
5. Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, Rigina O, et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotechnol. 2007 Mar;25(3):309-16.
6. Xu J, Li Y. Discovering disease-genes by topological features in human protein-protein interaction network. Bioinformatics. 2006 Nov 15;22(22):2800-5.
7. Goh KI, Cusick ME, Valle D, Childs B, Vidal M, Barabasi AL. The human disease network. Proc Natl Acad Sci U S A. 2007 May 22;104(21):8685-90.
8. Krauthammer M, Kaufmann CA, Gilliam TC, Rzhetsky A. Molecular triangulation: bridging linkage and molecular-network information for identifying candidate genes in Alzheimer's disease. Proc Natl Acad Sci U S A. 2004 Oct 19;101(42):15148-53.
9. He X, Zhang J. Why do hubs tend to be essential in protein networks? PLoS Genet. 2006 Jun 2;2(6):e88.
10. Jeong H, Mason SP, Barabasi AL, Oltvai ZN. Lethality and centrality in protein networks. Nature. 2001 May 3;411(6833):41-2.
11. Yu H, Kim PM, Sprecher E, Trifonov V, Gerstein M. The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. PLoS Comput Biol. 2007 Apr 20;3(4):e59.
12. Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. Global network analysis of phenotypic effects: protein networks and toxicity modulation in Saccharomyces cerevisiae. Proc Natl Acad Sci U S A. 2004 Dec 28;101(52):18006-11.
13. Tuck DP, Kluger HM, Kluger Y. Characterizing disease states from topological properties of transcriptional regulatory networks. BMC Bioinformatics. 2006;7:236.
14. Makino T, McLysaght A. Interacting gene clusters and the evolution of the vertebrate immune system. Mol Biol Evol. 2008 Sep;25(9):1855-62.
15. Breiman L. Random forests. Machine Learning. 2001 Oct;45(1):5-32.
16. Jayapandian M, Chapman A, Tarcea VG, Yu C, Elkiss A, Ianni A, et al. Michigan Molecular Interactions (MiMI): putting the jigsaw puzzle together. Nucleic Acids Res. 2007 Jan;35(Database issue):D566-71.
17. Hamosh A, Scott AF, Amberger J, Bocchini C, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. 2002 Jan 1;30(1):52-5.
18. Lussier Y, Borlawsky T, Rappaport D, Liu Y, Friedman C. PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. Pac Symp Biocomput. 2006:64-75.
19. Eyre TA, Ducluzeau F, Sneddon TP, Povey S, Bruford EA, Lush MJ. The HUGO Gene Nomenclature Database, 2006 updates. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D319-21.
20. National Library of Medicine. Medical Subject Headings (MeSH®) Fact Sheet. 1999 [updated 1999 27 May 2005; cited]; Available from: http://www.nlm.nih.gov/pubs/factsheets/mesh.html.
21. National Library of Medicine. Unified Medical Language System® Fact Sheet. 2006 [updated 2006 23 March 2006; cited]; Available from: http://www.nlm.nih.gov/pubs/factsheets/umls.html.
22. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.
23. Kasprzyk A, Keefe D, Smedley D, London D, Spooner W, Melsopp C, et al. EnsMart: a generic system for fast and flexible access to biological data. Genome Res. 2004 Jan;14(1):160-9.
24. Dijkstra EW. A note on two problems in connection with graphs. Numerische Mathematik. 1959(1):83–9.
25. Witten IH, Frank E. Data Mining: Practical machine learning tools and techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
26. Hammer Ø, Harper DAT, Ryan PD. PAST: Paleontological Statistics Software Package for Education and Data Analysis. Palaeontologia Electronica. 2001;4(1):9.
27. Feldman I, Rzhetsky A, Vitkup D. Network properties of genes harboring inherited disease mutations. Proc Natl Acad Sci U S A. 2008 Mar 18;105(11):4323-8.
28. Semon M, Mouchiroud D, Duret L. Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance. Hum Mol Genet. 2005 Feb 1;14(3):421-7.

Supplementary Tables and Figures

Supplementary Figure 1. Principal Components Analysis demonstrates the poor separability of the data

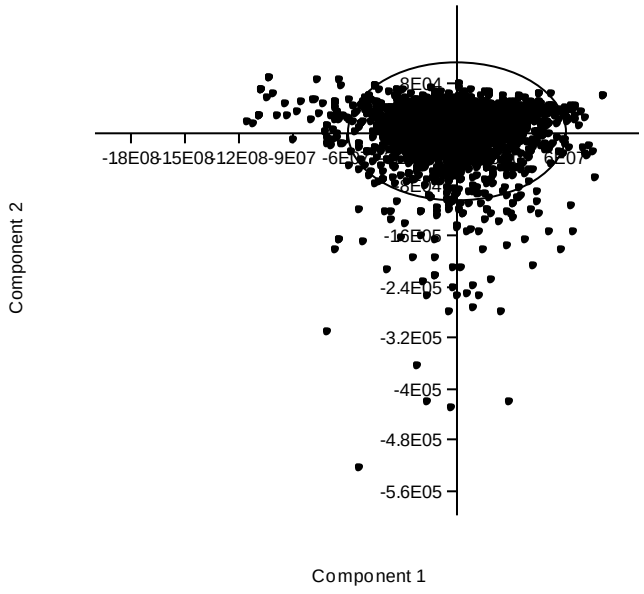

Figure 1a. Principal components analysis of all sets using all parameters. 95% of data points fall within the ellipse.

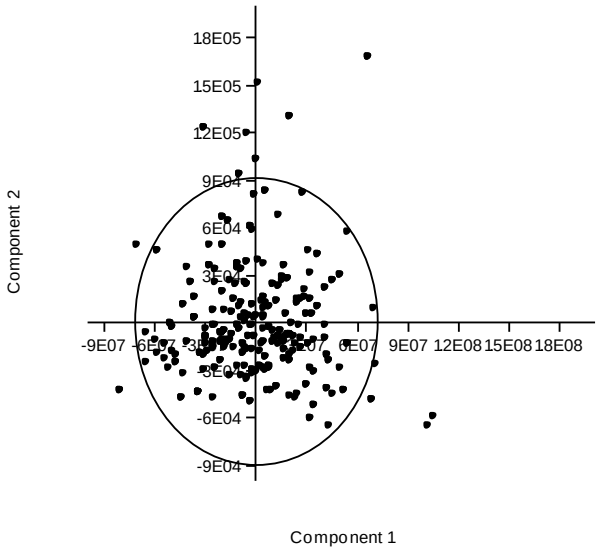

Figure 1b. Principal components analysis of SGD and CD sets using all parameters. 95% of data points fall within the ellipse.

A principal components analysis of the combined sets using all the parameters, suggests that the difference between disease-related subnetworks and the GO baseline subnetworks are subtle and not easily derived. When the PCA is done over just the CD and SGD sets, we see a similar pattern where there is no clear separation. However the non-continuous nature of the features may be a confounding factor when applying the PCA approach. With that in mind, a simple k-means clustering approach was taken where k = 3 to represent the three source types.

Supplementary Table 1. Complete results of unsupervised k-means clustering of the data

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 3 -S 10
Relation:     combined_data
Instances:    2944
Attributes:   20
              average gene start
              average gene end
              average length
              average gene strand
              average pfam count
              average prosite count
```

average # of singnal domains
average # transmembrane domains
average GC content
observed edges/total possible edges
average node degree
max node degree
radius
diameter
node count
cyclicity
biconnectivity
clustering coefficent

Ignored:

source
phenotype code

Test mode: Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

======

Number of iterations: 6

Within cluster sum of squared errors: 1660.859140812153

Cluster centroids:

Variables	Cluster 0		Cluster 1		Cluster 2	
Variable	**Mean/Mode**	**Std Devs**	**Mean/Mode**	**Std Devs**	**Mean/Mode**	**Std_Devs**
average gene start	**70562607**	21895436	**72069986**	8353007	**71199760**	12743762
average gene end	**70623972**	21898365	**72141696**	8355198	**71264921**	12743801
average length	**61364.07**	39100.1	**71710.6**	34510.64	**65160.52**	32673.26
average gene strand	**0.2259**	0.4311	**0.0898**	0.1649	**0.1195**	0.2538
average pfam count	**26.3999**	48.5588	**26.908**	16.715	**25.0051**	20.9655
average prosite count	**26.3999**	48.5588	**26.908**	16.715	**25.0051**	20.9655
average # of singnal domains	**0.1312**	0.1977	**0.156**	0.1162	**0.1235**	0.1314
average # transmembrane domains	**0.1335**	0.2008	**0.1715**	0.1121	**0.1415**	0.1412
average GC content	**43.1182**	3.0456	**41.7223**	1.2326	**42.2927**	2.0194
observed edges/total possible edges	**0.318**	0.1051	**0.0559**	0.0477	**0.1336**	0.0608
average node degree	**2.173**	0.6153	**4.417**	1.2629	**3.3774**	0.9891
max node degree	**4.2338**	1.9778	**60.4362**	70.3251	**12.8042**	13.2817
radius	**2**	N/A	**4**	N/A	**3**	N/A
diameter	**3.199**	0.8274	**7.0021**	1.1488	**5.0434**	0.724
node count	**5.6166**	3.0982	**151.7489**	185.4197	**26.2334**	29.3995
cyclicity	**0.7564**	0.4294	**0.9936**	0.0797	**0.9711**	0.1677
biconnectivity	**0.0237**	0.152	**0.0064**	0.0797	**0.0222**	0.1473
clustering coefficent	**0.0207**	0.0386	**0.0204**	0.0391	**0.0202**	0.0387
Clustered Instances	**1437 (49%)**		**470 (16%)**		**1037 (35%)**	

Class attribute: source

		Assigned to Cluster		
		Cluster 0 <-- GO	**Cluster 1 <-- OMIM**	**Cluster 2 <-- PhenoGO**
Source	**SGD/OMIM**	59	4	16
	GO	1220	435	932
	CD/PhenoGO	158	21	89

Incorrectly clustered instances : 1631.0 55.4008 %

Supplementary Figure 2. Size distribution of subnetworks in each category

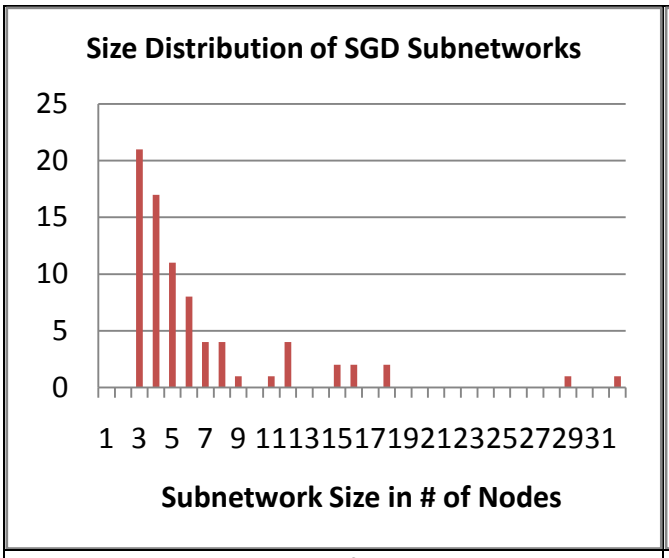

Figure 2a: Size Distribution of SGD Subnetworks

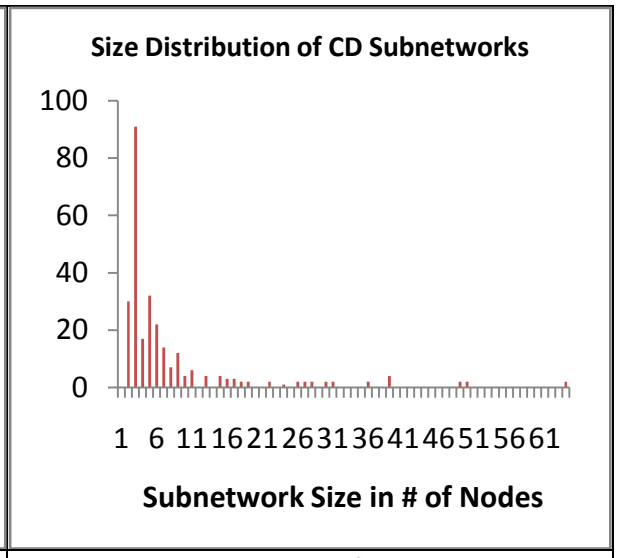

Figure 2b: Size Distribution of CD Subnetworks

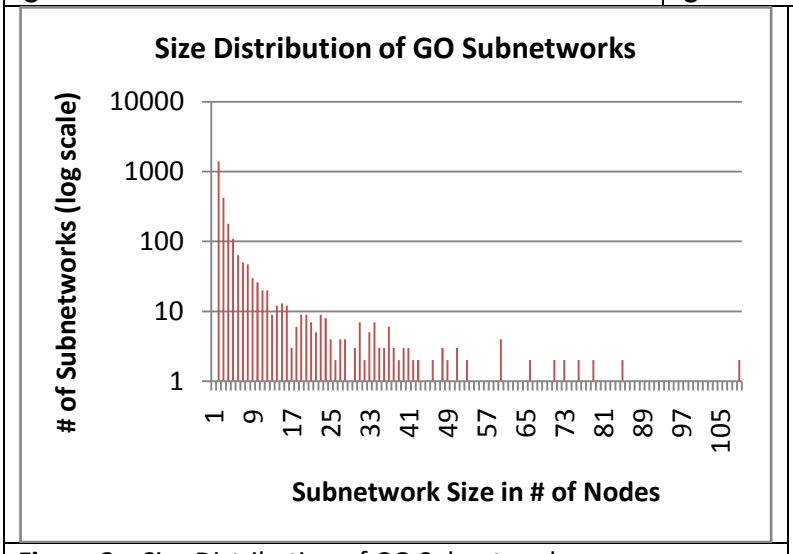

Figure 2c: Size Distribution of GO Subnetworks

Supplementary Table 2: Classification results from each of nine classification attempts using complete GO set

Biological Parameters Only

Table 2a. Biological parameters only: dataset split into "disease" and "normal" classes

Out of bag error: 0.0309

Correctly Classified Instances		2836		96.2988 %	
Incorrectly Classified Instances		109		3.7012 %	
Kappa statistic		0.8064			
Mean absolute error		0.1287			
Root mean squared error		0.216			
Relative absolute error		60.339 %			
Root relative squared error		66.1667 %			
Total Number of Instances		2945			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.995	0.272	0.964	0.995	0.979	GO
0.728	0.005	0.956	0.728	0.827	Disease

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2576	12	**a = GO/Normal**
97	260	**b = Disease**

Table 2b. Biological parameters only: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0309

Correctly Classified Instances		2832		96.163 %	
Incorrectly Classified Instances		113		3.837 %	
Kappa statistic		0.8008			
Mean absolute error		0.0893			
Root mean squared error		0.1801			
Relative absolute error		61.2569 %			
Root relative squared error		66.7931 %			
Total Number of Instances		2945			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.165	0.003	0.565	0.165	0.255	SGD
0.867	0.001	0.992	0.867	0.925	CD
0.996	0.283	0.962	0.996	0.979	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
2578	1	9	**a = GO**
36	241	1	**b = CD**
65	1	13	**c = SGD**

Table 2c. Biological parameters only: SGD and GO classes

Out of bag error: 0.0274

Correctly Classified Instances		2590		97.1129 %	
Incorrectly Classified Instances		77		2.8871 %	
Kappa statistic		0.1974			
Mean absolute error		0.0527			
Root mean squared error		0.1661			
Relative absolute error		91.1176 %			
Root relative squared error		97.9961 %			
Total Number of Instances		2667			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.127	0.003	0.556	0.127	0.206	SGD
0.997	0.873	0.974	0.997	0.985	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2580	8	**a = GO**
69	10	**b = SGD**

Topological Parameters Only

Table 2d. Topological Parameters Only: dataset split into "disease" and "normal" classes

Out of bag error: 0.0853

Correctly Classified Instances		2675		90.8628 %	
Incorrectly Classified Instances		269		9.1372 %	
Kappa statistic		0.4646			
Mean absolute error		0.1475			
Root mean squared error		0.2732			
Relative absolute error		69.1481 %			
Root relative squared error		83.7012 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.392	0.02	0.729	0.392	0.51	Disease
0.98	0.608	0.921	0.98	0.95	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2535	52	**a = GO/Normal**
217	140	**b = Disease**

Table 2e. Topological Parameters Only: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0832

Correctly Classified Instances		2688		91.30%	
Incorrectly Classified Instances		256		8.70%	
Kappa statistic		0.4863			
Mean absolute error		0.1016			
Root mean squared error		0.2241			
Relative absolute error		69.7015 %			
Root relative squared error		83.1102 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.038	0.004	0.214	0.038	0.065	SGD
0.493	0.011	0.83	0.493	0.619	CD
0.985	0.608	0.922	0.985	0.952	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
2548	28	11	**a = GO**
141	137	0	**b = CD**
76	0	3	**c = SGD**

Table 2f. Topological Parameters Only: SGD and GO classes

Out of bag error: 0.0315

Correctly Classified Instances		2581		96.81%	
Incorrectly Classified Instances		85		3.19%	
Kappa statistic		0.0586			
Mean absolute error		0.0543			
Root mean squared error		0.1716			
Relative absolute error		93.8315 %			
Root relative squared error		101.201 %			
Total Number of Instances		2666			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.038	0.003	0.25	0.038	0.066	SGD
0.997	0.962	0.971	0.997	0.984	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2578	9	**a = GO**
76	3	**b = SGD**

Combined Parameterization

Table 2g. All parameters: dataset split into "disease" and "normal" classes

Out of bag error: 0.0452

Correctly Classified Instances		2791		94.803 %	
Incorrectly Classified Instances		153		5.197 %	
Kappa statistic		0.7128			
Mean absolute error		0.1269			
Root mean squared error		0.2191			
Relative absolute error		59.5021 %			
Root relative squared error		67.1287 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.611	0.005	0.94	0.611	0.74	Disease
0.995	0.389	0.949	0.995	0.971	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
218	139	**a = Disease**
14	2573	**b = GO/Normal**

Table 2h. All parameters: dataset split into CD, SGD, and GO classes

Out of bag error: 0.0438

Correctly Classified Instances		2795		94.9389 %	
Incorrectly Classified Instances		149		5.0611 %	
Kappa statistic		0.7225			
Mean absolute error		0.0886			
Root mean squared error		0.1815			
Relative absolute error		60.7398 %			
Root relative squared error		67.2984 %			
Total Number of Instances		2944			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.101	0.003	0.5	0.101	0.168	SGD
0.997	0.387	0.949	0.997	0.972	GO
0.752	0.001	0.986	0.752	0.853	CD

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
8	70	1	**a = SGD**
7	2578	2	**b = GO**
1	68	209	**c = CD**

Table 2i. All parameters: SGD and GO classes

Out of bag error: 0.0281

Correctly Classified Instances		2591		97.1868 %	
Incorrectly Classified Instances		75		2.8132 %	
Kappa statistic		0.2332			
Mean absolute error		0.0498			
Root mean squared error		0.1594			
Relative absolute error		86.0831 %			
Root relative squared error		93.9883 %			
Total Number of Instances		2666			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.152	0.003	0.6	0.152	0.242	SGD
0.997	0.848	0.975	0.997	0.986	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
12	67	**a = SGD**
8	2579	**b = GO**

Table 2j. All parameters: SGD and CD classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: OMIM-PhenoGO-weka.filters.unsupervised.attribute.Remove-R2
Instances: 357
Attributes: 19

- source
- average gene start
- average gene end
- average length
- average gene strand
- average pfam count
- average prosite count
- average # of signal domains
- average # transmembrane domains
- average GC content
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1232

Correctly Classified Instances		315		88.2353 %	
Incorrectly Classified Instances		42		11.7647 %	
Kappa statistic		0.5965			
Mean absolute error		0.1785			
Root mean squared error		0.2972			
Relative absolute error		51.6603 %			
Root relative squared error		71.5991 %			
Total Number of Instances		357			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.519	0.014	0.911	0.519	0.661	SGD
0.986	0.481	0.878	0.986	0.929	CD

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
274	4	**a = CD**
38	41	**b = SGD**

The first classification was done on a set combining all SGD and CD subnetworks into a single larger disease class in comparison to the GO-derived background set. The second classification used only the SGD subset of the data in comparison to the GO data. The third classification used each subset of data in its own discrete class. These subsets were further separated into three groups depending on the underlying parameters available to the classifier. These groups used parameters exclusively from the topological and biological parameter sets, as well as the combined parameterization.

It can be seen that overall the biological characteristics prove more informative than the topological ones and achieve a lower misclassification error rate, ranging between 2.89 and 3.70%. On the other hand, for the topological characteristics the misclassification error rate was around 10% for the three class problem. However, when the CD class was excluded, the topological characteristics matched the performance of the biological ones. Further, an inspection of **Sup. Tables 2e and 2f** suggests that the presence of the SGD class is the source of the significantly higher misclassification error rate with respect to the topological features. In most cases, the presence of the large number of representative GO subnetworks leads to a high classification accuracy. However, it is useful to examine the true positive (**TP**) rate of classification between the combined "disease" set, a combination of the SGD and CD sets, and the GO background. In the combined parameterization and biological parameter only cases, the TP rate of this combined set is relatively good, at 61% and 72%, respectively. Examination of the TP rates for classifying into the three distinct classes revels that the subnetworks in the SGD set appear to be poorly distinguishable from the background GO set. However, the CD set appears to have predictive power setting it apart from the GO background. This similarity between the GO and SGD sets likely leads to the poor classification accuracy seen between the two sets as reflected in the poor TP values for the SGD set in **Sup. Tables 2e, 2f, 2h, and 2i**.

Supplementary Table 3: Ranked features by parameter type

Table 3a: Biological Parameters Only

	GO	SGD/OMIM	CD/PhenoGO	MeanDecreaseAccuracy	MeanDecreaseGini
averageGeneStart	0.2783482	1.0280783	0.9059960	0.2757494	84.56684
averageGeneEnd	0.2768157	0.9394527	0.8925733	0.2747467	82.32455
averageLength	0.2644807	1.2301754	0.9510359	0.2876197	89.97404
averageGeneStrand	0.1758904	0.1357294	0.9539724	0.2776031	63.51283
averagePfamCount	0.2730130	0.5254745	0.8856997	0.2717815	68.71366
averagePrositeCount	0.2732054	0.7780531	0.8667791	0.2729219	71.44485
averageSingnalDomainCount	0.2126032	1.1321489	0.9215645	0.2744301	46.04487
averageTransmembraneDomainsCount	0.2369126	0.7511460	0.9107138	0.2746473	41.26618
averageGCContent	0.2527932	1.1863229	0.9633071	0.2872784	90.52120

Table 3b: Topological Parameters Only

	GO	SGD/OMIM	CD/PhenoGO	MeanDecreaseAccuracy	MeanDecreaseGini
observedEdgeFraction	0.23001163	0.5940764	0.90312482	0.24675347	93.847995
averageNodeDegree	0.18907358	-0.1896722	0.92494854	0.25118579	73.325193
maxNodeDegree	0.23248537	-0.0195584	0.75146118	0.23964507	45.595834
radius	0.14363009	0.3341730	0.73797260	0.17620126	10.558500
diameter	0.16504637	0.3258433	0.89950612	0.21990106	24.283709
nodeCount	0.24716779	0.1174077	0.62814917	0.24756213	47.349672
cyclicity	0.07668406	0.1599157	0.05666838	0.08233893	2.229017
biconnectivity	0.05281318	0.2182699	0.47637630	0.10961336	3.538654
clusteringCoefficent	0.28966769	0.9925351	0.96101890	0.28810431	97.553541

Table 3c: Combined Parameterization

	GO	SGD/OMIM	CD/PhenoGO	MeanDecreaseAccuracy	MeanDecreaseGini
averageGeneStart	0.25577147	0.6187922	0.8782965	0.2631096	58.025555
averageGeneEnd	0.24189366	0.8649050	0.8823725	0.2517155	54.866536
averageLength	0.21860181	1.0476172	0.9157395	0.2702029	53.928221
averageGeneStrand	0.21222727	0.4779712	0.8899448	0.2613027	37.971447
averagePfamCount	0.24589871	0.7138733	0.8139401	0.2557329	51.837923
averagePrositeCount	0.24653767	0.8026352	0.8288924	0.2553449	51.873560
averageSingnalDomainCount	0.17608440	0.8725259	0.8494207	0.2504462	28.695867
averageTransmembraneDomainsCount	0.17643006	0.8016404	0.8587388	0.2398903	25.758543
averageGCContent	0.20630777	1.0249891	0.9042456	0.2621500	57.568889
observedEdgeFraction	0.22721854	0.9567640	0.8553682	0.2424491	39.992423
averageNodeDegree	0.24357245	0.6044357	0.8350696	0.2586311	33.451690
maxNodeDegree	0.23311884	0.5687222	0.7704013	0.2418791	23.089282
radius	0.19372018	0.5507024	0.5725879	0.1942285	9.303571
diameter	0.22263432	0.7683270	0.7232573	0.2295851	16.473967
nodeCount	0.23954530	0.7925986	0.8041791	0.2430081	25.501844
cyclicity	0.11050759	0.2201386	0.5157050	0.1559355	3.013125
biconnectivity	0.07642597	0.1890160	0.2993280	0.1074229	1.420956
clusteringCoefficent	0.26042896	1.4008805	0.8991804	0.2705517	61.914586

Supplementary Table 4. Classification results from each of nine classification attempts using "Biological Process" only GO set

Biological Parameters Only

Table 4a. Biological parameters only: dataset split into "disease" and "normal" classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Instances: 1706
Attributes: 10
source
average gene start
average gene end
average length
average gene strand
average pfam count
average prosite count
average # of signal domains
average # transmembrane domains
average GC content
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0569

Correctly Classified Instances		1595		93.4936 %	
Incorrectly Classified Instances		111		6.5064%	
Kappa statistic		0.7938			
Mean absolute error		0.195			
Root mean squared error		0.2682			
Relative absolute error		58.8853 %			
Root relative squared error		65.9355 %			
Total Number of Instances		1706			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.976	0.218	0.944	0.976	0.96	GO
0.782	0.024	0.894	0.782	0.834	Disease

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
1316	33	**a = GO/Normal**
78	279	**b = Disease**

Table 4b. Biological parameters only: dataset split into CD, SGD, and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Instances: 1706
Attributes: 10
source
average gene start
average gene end
average length
average gene strand
average pfam count
average prosite count
average # of signal domains
average # transmembrane domains
average GC content
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0557

Correctly Classified Instances		1584		92.8488 %	
Incorrectly Classified Instances		122		7.1512 %	
Kappa statistic		0.7715			
Mean absolute error		0.1409			
Root mean squared error		0.2327			
Relative absolute error		60.966 %			
Root relative squared error		68.5266 %			
Total Number of Instances		1706			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.152	0.006	0.571	0.152	0.24	SGD
0.874	0.01	0.946	0.874	0.908	CD
0.985	0.277	0.931	0.985	0.957	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
1329	13	7	**a = GO**
33	243	2	**b = CD**
66	1	12	**c = SGD**

Table 4c. Biological parameters only: SGD and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 6 -S 1
Relation: filtered_biological_2class_OMIM_omly-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1428
Attributes: 10
source
average gene start
average gene end
average length
average gene strand
average pfam count
average prosite count
average # of signal domains
average # transmembrane domains
average GC content
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 6 random features.
Out of bag error: 0.0539

Correctly Classified Instances		1353		94.7479 %	
Incorrectly Classified Instances		75		5.2521 %	
Kappa statistic		0.2391			
Mean absolute error		0.089			
Root mean squared error		0.2166			
Relative absolute error		84.6926 %			
Root relative squared error		94.745 %			
Total Number of Instances		1428			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.165	0.007	0.591	0.165	0.257	SGD
0.993	0.835	0.953	0.993	0.973	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
1340	9	**a = GO**
66	13	**b = SGD**

Topological Parameters Only

Table 2d. Topological Parameters Only: dataset split into "disease" and "normal" classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_2class_topological_data-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1705
Attributes: 10
- state
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1466

Correctly Classified Instances		1433		84.0469 %	
Incorrectly Classified Instances		272		15.9531 %	
Kappa statistic		0.4677			
Mean absolute error		0.2218			
Root mean squared error		0.3381			
Relative absolute error		66.9557 %			
Root relative squared error		83.1077 %			
Total Number of Instances		1705			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.49	0.067	0.66	0.49	0.563	Disease
0.933	0.51	0.874	0.933	0.902	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
1258	90	**a = GO/Normal**
182	175	**b = Disease**

Table 2e. Topological Parameters Only: dataset split into CD, SGD, and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_3class_topological_data-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1705
Attributes: 10
source
observed edges/total possible edges
average node degree
max node degree
radius
diameter
node count
cyclicity
biconnectivity
clustering coefficient
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1455

Correctly Classified Instances		1436		84.2229 %	
Incorrectly Classified Instances		269		15.7771 %	
Kappa statistic		0.4509			
Mean absolute error		0.162			
Root mean squared error		0.289			
Relative absolute error		70.0664 %			
Root relative squared error		85.0661 %			
Total Number of Instances		1705			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.038	0.007	0.2	0.038	0.064	SGD
0.514	0.036	0.737	0.514	0.606	CD
0.957	0.577	0.862	0.957	0.907	GO

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
1290	47	11	**a = GO**
134	143	1	**b = CD**
72	4	3	**c = SGD**

Table 2f. Topological Parameters Only: SGD and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_2class_omimonly_topological_data-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1427
Attributes: 10
source
observed edges/total possible edges
average node degree
max node degree
radius
diameter
node count
cyclicity
biconnectivity
clustering coefficient
Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0589

Correctly Classified Instances		1336		93.623 %	
Incorrectly Classified Instances		91		6.377%	
Kappa statistic		0.0422			
Mean absolute error		0.0938			
Root mean squared error		0.2341			
Relative absolute error		89.1962 %			
Root relative squared error		102.3648 %			
Total Number of Instances		1427			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.038	0.011	0.167	0.038	0.062	SGD
0.989	0.962	0.946	0.989	0.967	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
2578	9	**a = GO**
76	3	**b = SGD**

Combined Parameterization

Table 2g. All parameters: dataset split into "disease" and "normal" classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_combined_data_2class-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1705
Attributes: 19

- source
- average gene start
- average gene end
- average length
- average gene strand
- average pfam count
- average prosite count
- average # of signal domains
- average # transmembrane domains
- average GC content
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0774

Correctly Classified Instances		1555		91.2023 %	
Incorrectly Classified Instances		150		8.7977 %	
Kappa statistic		0.7058			
Mean absolute error		0.1928			
Root mean squared error		0.2775			
Relative absolute error		58.1885 %			
Root relative squared error		68.2136 %			
Total Number of Instances		1705			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.658	0.021	0.894	0.658	0.758	Disease
0.979	0.342	0.915	0.979	0.946	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
235	122	**a = Disease**
28	1320	**b = GO/Normal**

Table 2h. All parameters: dataset split into CD, SGD, and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_combined_data-weka.filters.unsupervised.attribute.Remove-R2
Instances: 1705
Attributes: 19

- source
- average gene start
- average gene end
- average length
- average gene strand
- average pfam count
- average prosite count
- average # of signal domains
- average # transmembrane domains
- average GC content
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation
=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0774

Correctly Classified Instances		1571		92.1408 %	
Incorrectly Classified Instances		134		7.8592 %	
Kappa statistic		0.742			
Mean absolute error		0.1339			
Root mean squared error		0.2244			
Relative absolute error		57.9155 %			
Root relative squared error		66.0713 %			
Total Number of Instances		1705			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.165	0.004	0.65	0.165	0.263	SGD
0.989	0.328	0.919	0.989	0.953	GO
0.809	0.007	0.957	0.809	0.877	CD

Confusion Matrix:

Classified as:			
a	**b**	**c**	**Actual assignment**
13	65	1	**a = SGD**
6	1333	9	**b = GO**
1	52	225	**c = CD**

Table 2i. All parameters: SGD and GO classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_combined_data_2class_omim_only-
weka.filters.unsupervised.attribute.Remove-R2
Instances: 1427
Attributes: 19

- source
- average gene start
- average gene end
- average length
- average gene strand
- average pfam count
- average prosite count
- average # of signal domains
- average # transmembrane domains
- average GC content
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.0526

Correctly Classified Instances		1353		94.8143 %	
Incorrectly Classified Instances		74		5.1857 %	
Kappa statistic		0.2424			
Mean absolute error		0.0863			
Root mean squared error		0.2113			
Relative absolute error		82.0074 %			
Root relative squared error		92.3897 %			
Total Number of Instances		1427			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.165	0.006	0.619	0.165	0.26	SGD
0.994	0.835	0.953	0.994	0.973	GO

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
13	66	**a = SGD**
8	1340	**b = GO**

Table 2j. All parameters: SGD and CD classes

=== Run information ===

Scheme: weka.classifiers.trees.RandomForest -I 100 -K 4 -S 1
Relation: filtered_OMIM-PhenoGO-weka.filters.unsupervised.attribute.Remove-R2
Instances: 357
Attributes: 19

- source
- average gene start
- average gene end
- average length
- average gene strand
- average pfam count
- average prosite count
- average # of signal domains
- average # transmembrane domains
- average GC content
- observed edges/total possible edges
- average node degree
- max node degree
- radius
- diameter
- node count
- cyclicity
- biconnectivity
- clustering coefficient

Test mode: 10-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 4 random features.
Out of bag error: 0.1232

Correctly Classified Instances		315		88.2353 %	
Incorrectly Classified Instances		42		11.7647 %	
Kappa statistic		0.5965			
Mean absolute error		0.1785			
Root mean squared error		0.2972			
Relative absolute error		51.6603 %			
Root relative squared error		71.5991 %			
Total Number of Instances		357			
TP Rate	**FP Rate**	**Precision**	**Recall**	**f-Measure**	**class**
0.519	0.014	0.911	0.519	0.661	SGD
0.986	0.481	0.878	0.986	0.929	CD

Confusion Matrix:

Classified as:		
a	**b**	**Actual assignment**
38	41	**a = CD**
274	4	**b = SGD**

Supplementary Methods

Data Extraction

Data was extracted from MiMi using SQL queries for human-specific interactions from the National Center for Integrative Biomedical Informatics SQL server using SQL Server Management Studio Express.

Derivation of disease subnetworks

The disease and biological process associated subnetworks are built from two fundamental components. First, a protein interaction network is used to define the relationships and interactions between the proteins considered in the study.

We separate the OMIM and PhenoGO sets for two reasons. The primary factor for the separation is the drastically different underlying focus of both of these resources, although they do share some commonly annotated diseases. PhenoGO contains data describing both single gene and multi-gene complex disease, whereas OMIM is primary focused on single gene diseases. The secondary factor is curation; the OMIM data is manually curated while PhenoGO is a computationally derived data source.

Derivation of the subnetworks was done using the Boost Library version 1.43.1 (http://www.boost.org/) and version .9 of the Boost Graph Library bindings to Python (http://osl.iu.edu/~dgregor/bgl-python/) using ActiveState ActivePython version 2.4.3 (http://www.activestate.com/).

Subnetworks that resulted in errors in the software were removed from the set, as the memory requirements for processing a number of large, dense networks was beyond the memory capacity of our workstation.

Filtering of Results

Because the data in the PhenoGO resource spans drugs, cell types, and other biological contexts not directly associated with disease, the subnetworks formed by this resource were filtered using the UMLS metathesaurus. Therefore, only genes associated with MeSH and UMLS terms are used to create the subnetworks. To restrict the set, a list of UMLS and MeSH codes was derived using a Perl script containing a total of unique terms. Of the 423,550 terms in the UMLS and MeSH that met these rules, the UMLS composed 419,087 terms and MeSH composed 5,563 terms. This process of restricting the set yielded a dramatic reduction in the number of subnetworks in the disease set.

The data from the biological and topological characterization for each of the classes was then filtered for size using a perl script, constraining the set to networks of size between 3 and 9999 nodes. 79 and 278 subnetworks passed this filter from the OMIM and PhenoGO sets, respectively. 2590 of the subnetworks generated from the Gene Ontology passed this filter.

Parameterization/Characterization of Subnetworks

To characterize subnetworks structurally, we chose a number of well-defined metrics to measure their size, density, and connectivity. Subnetworks are characterized based on node count, clustering coefficient, average degree, maximum degree, radius, diameter, cyclicity, and biconnectivity. Cyclicity and biconnectivity are handled as Boolean variables with values of either 1 (True) or 0 (false). To account for the biological characteristics of the constituent genes of these subnetworks, we use biological characteristics for the constituent genes extracted from BioMart. These factors accounted for positional and orientation effects, biological role of the protein product, and physical stability. Factors include mean gene start location, mean gene end location, mean length, strand, mean PFAM domain annotation count, mean ProSite annotation count, mean number of signal domains, mean number of transmembrane domains, and mean G-C content fraction.

Parameterization of subnetworks was done using a series of Perl scripts using the Perl-Graph library version .84 (http://search.cpan.org/dist/Graph/) as well as the Boost Graph Library Bindings for Perl version 1.4 (http://search.cpan.org/~dburdick/Boost-Graph-1.4/). These libraries were used to determine the topological characteristics of each of the subnetworks. Factors include the average degree, maximum degree, node count, radius, and diameter for each subnetwork. Each subnetwork was also tested for cyclicity and biconnectivity.

During the parameterization process, a number of entries were removed from the set as the subnetworks they formed were not computable within the memory limits of our workstation.

GO:0007218 - neuropeptide signaling pathway
GO:0045893 - positive regulation of transcription, DNA-dependent
GO:0006937 - regulation of muscle contraction

Classification

Classification was done with Weka using the built-in weka.classifiers.trees.RandomForest package . The parameterized data was split into 3 sets for the biological and topological groups. The first set composed of all three data sources comprising three distinct classes. The second set assigned “normal” and “disease” flags to the subnetworks derived from the Gene Ontology, and OMIM and PhenoGO, respectively. The third subset was composed of only disease subnetworks derived from OMIM while maintaining the GO background set.

Feature Analysis

A factor analysis was done using the RandomForest package in R 2.7.1 in each of the biological parameter only, topological parameter only, and combined parameter groups to determine the relative influence of each of the parameters in determining class membership in each of the classification sets. The random forest was set to use 4 variables per tree and 100 total trees for the classification task.