What Do We Learn from Network-Based Analysis of Genome-Wide Association Data?

Marzieh Ayati[1], Sinan Erten[1], and Mehmet Koyutürk[1,2]

[1] Department of Electrical Engineering and Computer Science
[2] Center for Proteomics and Bioinformatics

Network based analyses are commonly used as powerful tools to interpret the findings of genome-wide association studies (GWAS) in a functional context. In particular, identification of disease-associated functional modules, i.e., highly connected protein-protein interaction (PPI) subnetworks with high aggregate disease association, are shown to be promising in uncovering the functional relationships among disease-associated genes and proteins. An important issue in this regard is the scoring of subnetworks by integrating two quantities that are not readily compatible: disease association of individual gene products and network connectivity among proteins. Current scoring schemes either disregard the level of connectivity and focus on the aggregate disease association of connected proteins or use a linear combination of these two quantities. However, such scoring schemes may produce arbitrarily large subnetworks which are often not statistically significant, or require tuning of parameters that are used to weigh the contributions of network connectivity and disease association. Here, we propose a parameter-free scoring scheme that aims to score subnetworks by assessing the disease association of pairwise interactions and incorporating the statistical significance of network connectivity and disease association. We test the proposed scoring scheme on a GWAS dataset for type II diabetes (T2D). Our results suggest that subnetworks identified by commonly used methods may fail tests of statistical significance after correction for multiple hypothesis testing. In contrast, the proposed scoring scheme yields highly significant subnetworks, which contain biologically relevant proteins that cannot be identified by analysis of genome-wide association data alone.

1 Introduction

In recent years, there has been an explosion in genome-wide association studies (GWAS) of complex diseases [8, 30]. These studies have successfully revealed many genetic variants conferring susceptibility to disease. However, GWAS have so far explained a small fraction of the heritability of common diseases and provided limited insights into their molecular mechanisms. A commonly cited reason underlying the limitations of GWAS is the complex nature of diseases, i.e., the interplay among multiple genetic variants in driving disease phenotype [23]. Therefore, many computational methods have been developed to integrate the outcome of GWAS and with other biological such as pathways, annotations, and networks to provide a functional context for disease association of multiple genetic variants [4, 15, 28, 33, 37] and the identification of epistatic interactions [29].

Among computational methods that aim to identify multiple genetic variants associated with diseases, identification of disease-associated functional modules has been commonly used as a powerful tool to gain insights into the systems biology of disease mechanisms [15, 27]. These methods aim to identify highly connected subnetworks of the human protein-protein interaction (PPI) network (hence, functional module) that exhibit high aggregate association with the disease as indicated by the GWAS p-values of associated genetic variants (hence, disease-associated). These methods have been shown to be effective in uncovering the functional relationships among disease-associated genetic variants for a number of complex diseases, including multiple sclerosis [4], breast cancer and pancreatic cancer [15], and sleep apnea [18].

In the identification of disease-associated functional modules, an important challenge is to define a scoring function that will accurately assess the "interestingness" of a given subnetwork in terms of functional modularity (network connectivity) and disease association. While scoring subnetworks, many of the existing methods ignore the degree of network connectivity and score connected subnetworks of the human PPI network using an aggregate of the disease association of comprising gene products [5, 14, 15]. Alternately, some methods incorporate network connectivity by using a linear combination of this aggregate score and the density of the induced subnetwork, using a free parameter to adjust the relative contributions of disease association and network connectivity [20, 38]. Subsequently, they identify high-scoring subnetworks using various algorithmic techniques [14, 20] and empirically assess the significance of these subnetworks based on permutation tests [4].

Scoring schemes that are based on an aggregate of individual disease association scores are highly influenced by subnetwork size, i.e., the number of proteins in the subnetwork. More precisely, if the subnetwork score is computed as the average of the individual association scores of constituent proteins, then obviously smaller subnetworks would be favored. To deal with this problem, existing scoring schemes reward larger subnetworks. However, this results in a "large subnetwork effect", i.e., the resulting algorithms may arbitrarily grow subnetworks to maximize the score. Clearly, since the "large subnetwork effect" also applies to randomized data, such arbitrarily large subnetworks are not likely statistically significant. Furthermore, since the objective here is to identify parsimonious sets of interacting proteins that confer susceptibility to disease, such large subnetworks may not be biologically relevant or useful, or they may require further computational analyses for the extraction of their relevant parts [18]. Indeed, Branzini *et al.* [4] systematically show that, if correction for multiple hypothesis testing is handled properly, such scoring schemes do not yield statistically significant subnetworks for many diseases. Scoring schemes that incorporate the degree of network connectivity, on the other hand, require tuning of a free parameter to adjust the relative contributions of disease association and network connectivity, making it difficult to apply these algorithms to cases where no training data is available. Unfortunately, this is the case for many applications since biologically relevant subnetworks for many diseases are largely unknown.

In this paper, we propose a scoring scheme that (i) integrates disease association and network connectivity in a parameter-free fashion and (ii) incorporates an approximation of the statistical significance of this integrated score. The key idea of the proposed method is to assess the disease association of each interaction in the network and account for the background disease association as an approximation to statistical significance. In this respect, the proposed approach may be thought of a generalization of Newman's [7] measure of modularity, which was developed for community detection in networks. We test the proposed scoring scheme on a GWAS dataset for type II diabetes (T2D) and compare its performance with two most commonly used scoring methods, namely scoring based on an aggregate of the disease association of individual proteins only and scoring based on a linear combination of network connectedness and aggregate disease association. Our results show that subnetworks that are scored high by the proposed scoring scheme are more likely to be statistically significant as compared to those that are scored high by the other two scoring schemes. We also assess the biological relevance of identified subnetworks in terms of their inclusion of known disease-related proteins that do not exhibit significant disease association based on individual analysis of GWAS data. Our results suggest that the proposed scheme yields parsimonious subnetworks that contain known proteins, as well as those that are not individually significant, but are candidates for further investigation.

In the next section, we describe the proposed scheme for scoring protein subnetworks in terms of their disease association and network connectivity. Subsequently, in Section 3, we present comprehensive experimental results on T2D data obtain from the Wellcome Trust Case-Control Consortium (WTCCC). We conclude our discussion in Section 4.

2 Methods

In this section, we first describe the problem setting for the identification of disease-associated functional modules. Then we describe the three scoring schemes we consider in this study. Subsequently, we describe the algorithms used to identify high-scoring modules according to this scoring scheme. Finally, we discuss our methodology for assessing the statistical significance of identified high-scoring modules.

2.1 Problem Setting

The input to the problem of identifying disease-associated functional modules (DAFM) is a graph $G = (V, E, w)$ that represents the human PPI network. Here, V denotes the set of proteins, E denotes the set of pairwise interactions between these proteins, and $w : E \to \mathbb{R}$ denotes edge weights, where $w(u, v)$ represents the likelihood that proteins $u, v \in V$ interact. The likelihood scores for interactions are usually computed by integrating the outcome of several experimental and computational methods for detecting and predicting protein-protein interactions. In this paper, we use an online tool, MAGNET [17], to score the interactions.

Besides the network, we are given a genome-wide association (GWAS) dataset $D = (C, M, g, f)$, where C denotes the set of genomic loci that are assayed, M denotes the set of samples, $g(c, m)$ denotes the genotype of locus $c \in C$ in sample $m \in M$, and $f(m)$ denotes the phenotype of sample $m \in M$. If the phenotype is dichotomous (i.e., $f : M \to \{0, 1\}$ where 1 denotes case and 0 denotes control), then the disease association of each variant can be computed a standard statistical test, e.g. Cochran-Armitage trend test, Fisher's exact

test, or Cochran-Mantel-Haenszel tests [3]. For quantitative traits (i.e., $f : M \to \mathbb{R}$), association tests such as Breslow-Day or homogeneity of odds ratio can be used [12].

In this paper, our focus is not on assessing the disease association of each variant. We rather assume that the statistical significance of the association of each locus $c \in C$ with the disease is given as a p-value, denoted $p(c)$. From these significance values, we compute the significance of the association of each gene coding for a protein $v \in V$ by taking the most significant association of the variants that lie within the region of interest for that gene. For the experiments reported in this paper, we define the region of interest for a gene, denoted $N(v) \subset C$, as the genomic region within $20kb$ up- and down-stream the coding region for the gene. We further log-transform the significance of disease association for each gene $v \in V$ to obtain disease association score

$$r(v) = \max_{c \in N(v)} \{-\log(p(c))\}. \tag{1}$$

The objective of the disease-associated functional module (DAFM) identification problem is to identify PPI subnetworks such that:

- the subnetwork is enriched in proteins that are associated with the disease,
- the proteins in the subnetwork are functionally associated with each other.

Consideration of these two criteria together enables identification of functionally modular processes that are associated with the disease. An important challenge in this regard is to develop scoring schemes that can achieve a reasonable balance between these two criteria so that the subnetworks that are assigned statistically significant scores are those that are biologically most meaningful and useful.

2.2 Scoring Subnetworks

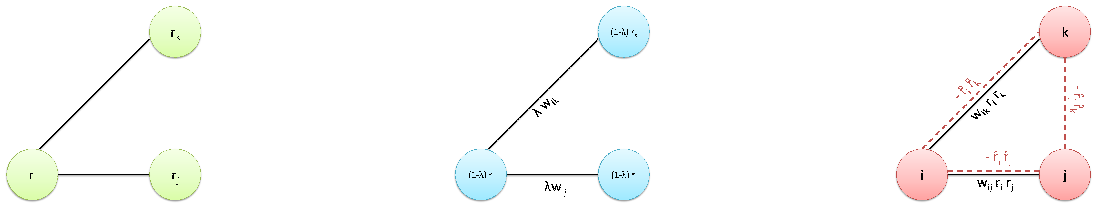


Fig. 1: Illustration of existing and proposed scoring schemes for quantifying the disease association of protein subnetworks: (a) Node-Based scoring, (b) Linear Combination of node scores and edge scores, (c) the proposed Modularity-Based (MoBaS) scoring scheme. For each method, the score of subnetwork is computed as the summation of all quantities in the figure.

In this section, we describe the three scoring schemes that are used in our experimental studies. These scoring schemes are illustrated in Figure 1. Two of these schemes are based on existing methods for the identification of active subnetworks using gene expression data, and these methods are commonly used in integrating GWAS outcome with PPI networks. The third is a novel scoring method that is based on a measure of modularity in networks [24].

Node-Based Scoring A popular method for scoring subnetworks is implemented in JactiveModules [14], a Cytoscape plug-in for the identification of "active subnetworks". This method was originally designed to integrate gene expression data with PPI networks to identify PPI subnetworks that are differentially expressed (or "active") under a certain condition or phenotype. However, it has also found common application in integrating GWAS data with PPI networks to identify disease-associated subnetworks. It takes as input the individual disease association scores for all proteins and aims to identify connected subgraphs of the

PPI network with high aggregate association score. More precisely, under this scoring scheme, a subnetwork $Q \subset V$ that induces a connected subgraph in the PPI network is scored as follows:

$$\sigma_N(Q) = \frac{1}{\sqrt{|Q|}} \sum_{u \in Q} r(u). \tag{2}$$

The NODE-BASED scoring scheme is illustrated in Figure 1(a). Since this scoring scheme is based on the individual disease association of the proteins composing the subnetwork, we refer to it as NODE-BASED scoring. Under this scheme, the connectivity of the subnetwork is imposed as a qualitative constraint to ensure that the proteins in the subnetwork are functionally related. However, the degree of connectivity, hence the degree of functional association among the proteins, is not quantified.

Many studies in the context of a related problem, candidate disease gene prioritization, have shown that the degree of network connectivity provides valuable information on the functional relationships between individual proteins in terms of their association with disease [11, 36]. To this end, taking into account the degree of network connectivity may lead to the identification of more relevant networks since it may better account for the noise in the network, as well as the modularity of the processes in which the proteins are involved together.

Linear Combination of Node and Edge Scores Disease association and the degree of connectivity in the network are two criteria that are not readily comparable. Therefore, incorporation of these two criteria into a single scoring scheme is rather challenging. To this end, scoring subnetworks based on a linear combination of the two criteria is reasonable, in that it provides a framework in which the relative contributions of the two criteria can be adjusted using a single parameter. Indeed, Ma et al. [20] propose the following LINEAR COMBINATION based scoring scheme for the identification of disease associated subnetworks:

$$\sigma_L(Q) = \lambda \frac{\sum_{u,v \in Q} w(u,v)}{\sqrt{\binom{|Q|}{2}}} + (1-\lambda) \frac{\sum_{u \in Q} r(u)}{\sqrt{|Q|}}. \tag{3}$$

This approach has been shown to be more effective than NODE-BASED scoring in the context of identifying "active subnetworks" [20]. However, to the best of our knowledge, it has not found application in the identification of disease-associated subnetworks based on GWAS outcome. An important drawback of this approach is its dependence on a tunable parameter, since the objective of DAFM is to find subnetworks that exhibit statistically significant association with the disease in an unsupervised manner, and training data (i.e., "known" disease-associated subnetworks) are rarely available.

Modularity Based Scoring (MOBAS) The objective in any pattern discovery problem for biological applications is to discover patterns that are *statistically significant*. To this end, it is important to note that "high scoring" does not necessarily mean "statistically significant" and a scoring scheme should not be overly conservative or overly relaxed, since a conservative scoring scheme may not produce any non-trivial high-scoring patterns and a relaxed scoring scheme may produce high scoring patterns that are not significant. Here, we argue (and show in Section 3) that both NODE-BASED and LINEAR-COMBINATION based scoring schemes are overly relaxed in that they can lead to the identification of very large subnetworks that will achieve high scores just because of their size, since these scoring schemes do not explicitly penalize for the inclusion of more proteins in the subnetwork.

In this paper, we propose a novel scoring scheme that integrates degree of network connectivity with disease association in a parameter-free manner by assessing the disease association of each pair of proteins (a potential interaction) in the network. Further, building on Newman's [24] measure of modularity for community detection in networks, the proposed scoring scheme incorporates (an approximation of) statistical significance into the scoring of subnetworks by taking into account the background disease association scores.

Namely, we define the disease association of a pair of proteins $u, v \in V$ as follows:

Recall that r_u indicates the likelihood that protein u is associated with the disease of interest. Therefore, s_{uv} provides a measure of the disease association of the interaction between u and v with the disease; i.e., the higher s_{uv}, the more likely it is that u and v are two functionally related proteins that are both associated with the disease. Note that, if proteins u and v do not interact, we have $s_{uv} = 0$; i.e., the disease association of a non-extant interaction is zero.

We then define the disease association score of a given subnetwork $Q \subseteq vertexset$ as follows:

$$\sigma_M(Q) = \sum_{u,v \in Q} s_{uv} - \hat{r}_u \hat{r}_v, \tag{4}$$

where $\hat{r}_u$ and $\hat{r}_v$ respectively denote the "background" disease association scores for proteins u and v. We compute these background scores empirically for each protein. For this purpose, we randomize the original GWAS data by permuting the labels of the samples to break the relationship between the genotype and phenotype, while preserving the distribution of genotypes for each locus. We repeat the permutation multiple (N) times such that the number of samples derived from the distribution is sufficiently large and the computation is feasible (we use $N = 100$ in our experiments). For each randomized instance $1 \leq i \leq N$, we compute the disease association of gene u on that instance i as $r_u^{(i)}$ using Equation 1. Subsequently, we compute the background disease association of each individual gene u as

$$\hat{r}_u = \sum_{i=1}^{N} r_u^{(i)} / N. \tag{5}$$

In other words, the disease association of subnetwork $Q \subseteq V$ is defined as the linear combination of the differences between the observed and background disease association scores of all potential pairwise interactions in the subnetwork. Note that, it is assumed that an interaction exists between every pair of proteins in the background, therefore any pair of proteins in the subnetwork that do not interact with each other are penalized by a factor of the multiplication of their background association scores. For this reason, groups of proteins that induce a heavily connected (hence, functionally modular [34]) subgraph in the PPI network are favored by this scoring scheme.

2.3 Searching for High Scoring Subnetworks

Subnetwork search queries with combinatorial objective functions often lead to NP-hard problems. For this reason, existing methods for identifying disease-associated functional modules use approximation algorithms or heuristics, such as greedy algorithms, simulated annealing [14], genetic algorithms [20], or linear programming based on a continuous approximation [38]. Since our focus here is on the development of a sound scoring function, the algorithm we use to search for high scoring subnetworks should be compatible with those implemented by existing methods, so that the scoring functions can be compared without any algorithmic bias. Here, for simplicity, we implement a greedy algorithm as well. Namely, to find all high-scoring subnetworks, we search the PPI network by starting from the protein with most significant disease association, repeatedly examining the proteins in the neighborhood of the proteins so far in the subnetwork, and adding to the subnetwork the protein that provides the best improvement of the subnetwork score. We repeat these steps until we cannot find any neighboring protein that improves the subnetwork score. We further reduce the computational complexity of the search algorithm by constraining the search space to a locality in the network (i.e. within two jumps of the first protein added to the subnetwork). Once a subnetwork with maximal score is found, we save it as a high-scoring subnetwork and remove its constituent proteins from the network. We then repeat the procedure to find other high scoring modules, until the entire network is exhausted. Finally, we sort all identified modules according to their score and assess the statistical significance of their scores.

2.4 Assessment of Statistical Significance

The proposed scoring scheme approximates the statistical significance of subnetworks by accounting for the background distribution of disease association. However, the distributions used in this approximation do not take into account multiple hypothesis testing, since each subnetwork is scored independently. Furthermore, only sample means are incorporated in the scoring function, which may not account for the variability in the distributions of network connectivity and disease association. Consequently, high-scoring modules identified using the proposed scoring scheme are not necessarily significant. For this reason, for all the three scoring schemes that are considered, we assess the statistical significance of all identified subnetworks using empirical distributions generated by running the algorithm on multiple randomized datasets.

We generate the randomized datasets using two different approaches:

1. Random permutation of the phenotypes of samples, with a view to testing the hypothesis that the the high score of each identified subnetwork arises from the correlations between genotype and phenotype.
2. Random permutation of the PPI network, with a view to testing the hypothesis that each high-scoring subnetwork are composed of functionally associated proteins. To randomize the PPI network while preserving the degree of each protein, we use a standard algorithm that repeatedly swaps the two ends of randomly selected pairs of interactions [22].

Observe that, since the number of hypotheses being tested is equal to the number of potential connected subnetworks of the PPI network, multiple hypothesis testing poses an important challenge in evaluating the significance of identified subnetworks. We tackle this challenge by using the ranking of subnetworks identified on random datasets to generate a null distribution for each subnetwork based on its rank on the original dataset. Namely, for the subnetwork that has the ith highest score on the original dataset, we test the hypothesis that the algorithm could discover at least i subnetworks with higher or equal score even if the phenotypes and the interactions in the network were assigned at random.

To be more precise, we generate a sufficiently large number (M) of randomized datasets for each type of permutation (i.e., randomized genotype and randomized PPI). Then we identify and rank all high-scoring subnetworks on each dataset. We then assess the statistical significance of each subnetwork identified on the original data by comparing its score against the scores of the subnetworks that are ranked at least as high as itself on the randomized datasets. Namely, for subnetwork Q_i that is ranked ith in the original dataset, we take the highest scoring i subnetworks from each of the M datasets and compute the fraction of subnetworks among these Mi modules whose scores are at least as high as that of Q_i. We call this fraction the q-value of the module, since it implicitly accounts for multiple hypothesis testing. We call a subnetwork is significant only if its q-value for both types of permutation is below a preset q-value threshold. We use $M = 100$ (as a trade off between feasibility and statistical power) and a $q - value$ threshold of 0.05 in our experiments.

3 Results

In this section, we first describe the datasets used in our experiments. Subsequently, we investigate the statistical significance of the subnetworks identified by the proposed scoring scheme, as well as those identified by aggregation of node scores (Node-Based) and linear combination of node and edge scores (Linear Combination). We assess the biological relevance of the identified subnetworks using a literature-driven list of genes and processes that have been reported to be associated with T2D. We also perform pathway enrichment analysis to identify the biological processes and pathways potentially associated with T2D. Finally, we investigate the biological relevance of the "novel genes" identified by the scoring gene, namely those that are not known to be associated by the disease, do not show significant disease association according to GWAS data, but are recruited in the significant subnetworks identified by the proposed scoring scheme.

3.1 Datasets and Preprocessing

GWAS dataset: To evaluate the performance of the proposed method, we use a Type 2 Diabetes (T2D) case-control dataset, obtained from Wellcome Trust Case-Control Consortium (WTCCC) [8]. The T2D data contains SNP microarray data for 500000 SNPs on 1999 case and 1504 control samples (1958 British Birth Cohort). For this dataset, we use the genotype calls provided by WTCCC, which were obtained by using CHIAMO.SNPs with $> 10\%$ missing genotypes are excluded from the analyses.

Association analysis for individual SNPs: We compute the statistical significance of the association of each SNP with T2D using PLINK[26], a well-established toolkit for whole-genome association analysis. We assess the disease-association of SNPs based on minor allele frequency, obtaining a p-value for the association of each SNP with the disease. We use raw p-values instead of correcting for multiple hypothesis testing, since the p-values are used for scoring the nodes of the PPI network and not for assessing the statistical significance of individual SNPs.

SNP-gene mapping and association analysis for individual genes: To compute the disease-association for individual genes, we map SNPs to genes by defining the region of interest (ROI) for a gene as the genomic region that extends from 20kb upstream to 20kb downstream of the coding region for that gene. We compute the disease association of each gene as the minimum of the p-values of the SNPs in the region of interest for that gene, that is the p-value of the most significant SNP associated with the gene. We log-transform these values to obtain a disease association score for each gene, as described by Equation 1.

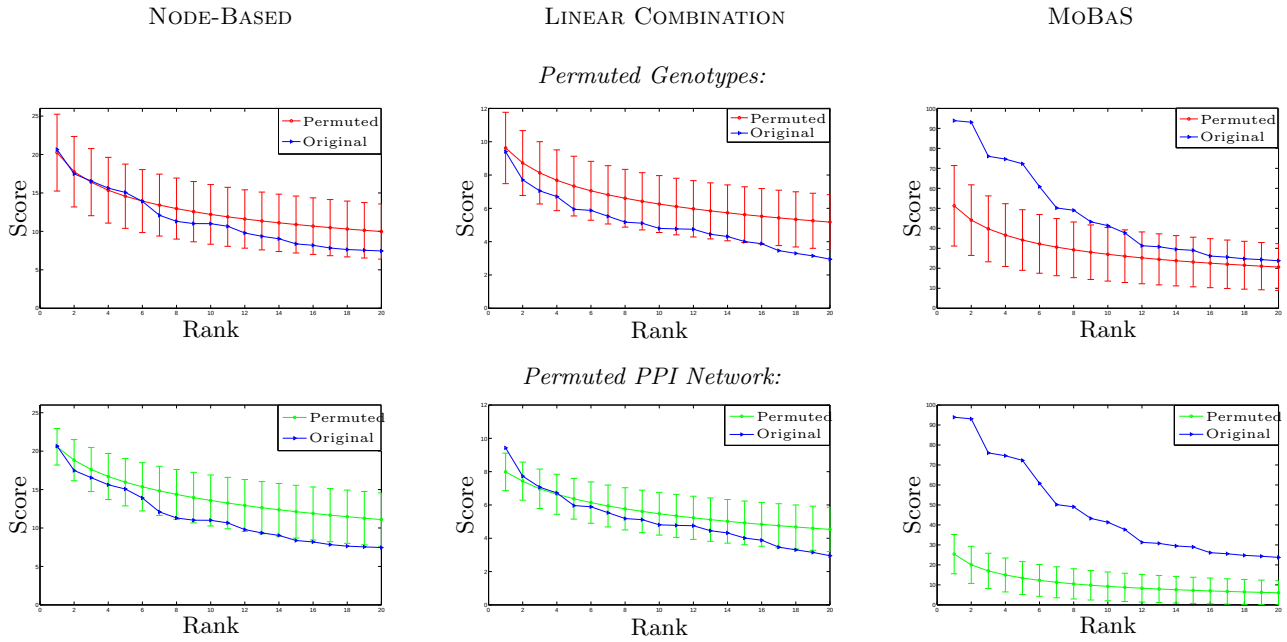


Fig. 2: Statistical significance of high-scoring subnetworks identified using Node-Based scoring (first column), Linear Combination of node scores and edge scores (second column), and Modularity-Based (MoBaS) scoring (third column). The highest scoring 20 subnetworks identified using each scoring scheme are shown. The x-axis shows the rank of each subnetwork according to their score, the y-axis shows its score. The blue curve shows the scores of the subnetworks identified on the WTCCC-T2D dataset. For each i on the x-axis, the red curve and error bar in the first row (the green curve and error bars in the second row) show the distribution of the scores of i highest scoring subnetworks in 100 datasets obtained by permuting the genotypes of the samples (permuting the interactions in the PPI networks while preserving node degrees).

Protein-protein interaction (PPI) dataset: We use a comprehensive human PPI network downloaded from NCBI Entrez Gene Database [21]. This database integrates interaction data from several PPI databases, including HPRD, BioGrid, and BIND. The PPI network contains 56110 interactions among 7692 proteins. We assess the reliability of each interaction in this dataset using MAGNET [17], a web service that uses logistic regression to assign reliability scores to PPIs. For this purpose, MAGNET utilizes the type of experimental platform, mRNA-level co-expression, and co-localization, as well as gold standard interactions from MIPS.

Genes reported to be associated with T2D: In order to assess the biological relevance of identified subnetworks, we use a manually curated database of genes that are reported to be associated with T2D in the literature [16]. This list contains 286 genes. We also use a second database that is generated by using seven independent computational disease gene prioritization methods [31], namely GeneSeeker [10], POCUS [35], G2D [25], PROSPECTR [1], eVOC annotation [32], DGP [19] and SUSPECTS [2].

Pathway enrichment analysis: We also evaluate the subnetworks that are found to be significantly associated with T2D using pathway enrichment analysis. For this purpose, we use Ingenuity Pathway Analysis (IPA), a commercial software that uses a manually curated and highly reliable database of pathway associations to perform pathway enrichment analysis.

Table 1: Statistical significance(q-value) of top two subnetworks identified using each scoring scheme according to the permuted genotype and PPI for WTCCC-T2D

Scoring Method	Size	q-value in Permuted Genotype	q-value in Permuted PPI networks
Node-Based	187	0.37	0.45
	190	0.70	0.92
Linear Combination	41	0.46	0.09
	17	0.79	0.52
MoBaS	14	0.04	< 0.01
	14	0.05	< 0.01

3.2 Significance of Identified Subnetworks

In this section, we investigate the statistical significance of the subnetworks identified by each scoring scheme. For this purpose, we compare the scores of highest-scoring subnetworks identified on the WTCCC dataset with that of the highest-scoring subnetworks identified on 100 randomized datasets in which (i) the sample phenotypes are permuted, (ii) PPIs are randomly permuted while preserving the number of interactions for each protein. The results of this analysis are shown in Figure 2.

In the figure, the identified subnetworks are sorted according to their score and the scores of the top 20 subnetworks identified by each method are shown with blue curves. In each plot, the red and green curves with error bars show the distribution of the scores of the modules that have at least the respective rank in the 100 randomized datasets. Randomized permutation of samples captures the significance of the score of the subnetworks in terms of the the component of the score that represents disease association. On the other hand, randomized permutation of PPIs assesses the significance of the score in terms of the network connectivity component. Comparison to the highest ranked subnetworks identified on permuted data by the same method corrects for multiple hypothesis testing, since the null distribution represents the scores of highest scoring subnetworks that can be found by the method, as opposed to the score of the subnetwork being tested on the randomized datasets.

The null distribution displayed in Figure 2 is precisely the distribution used to compute the q-values of each identified subnetworks, as described in Section 2.4.

As seen in top row of Figure 2, the nine highest scoring subnetworks identified using MoBaS have scores at least one standard deviation above the mean of the top subnetworks identified on randomized datasets. At a q-value threshold of 0.05, two of these subnetworks are detected to be statistically significant. In contrast, all subnetworks identified by Linear Combination and Node-Based scoring are within one standard deviation of the average score of the top subnetworks identified on randomized datasets. In other words, when the existing genotype-phenotype relationship in the dataset is broken via randomization of samples, Linear Combination and Node-Based can still detect subnetworks that score high. The respective q-values are shown in Table 1.

We observe a similar pattern when we compare the subnetworks identified on the original data to those identified on randomly permuted PPI networks. The bottom row of Figure 2 shows that the top-scoring 20 subnetworks identified by MoBaS on the original dataset have higher scores compared to those identified on the permuted dataset. The respective q-values are shown in table 1. Baranzini *et al.* [4] also investigate this issue systematically on a number of complex diseases and show that, while the subnetworks identified by jActiveModules (Node-Based scoring) on some diseases (including multiple sclerosis and rheumatoid arthritis) are significant, many subnetworks that are identified for other diseases are not, including those for T2D. Our results stand as a reproduction of these results and suggest that the proposed modularity-based scoring scheme does not suffer from this problem.

To choose significant subnetworks for further investigation, we require statistical significance in terms of both disease association and network connectivity. For this purpose, we compute the q-value of each subnetwork as the maximum of its q-values with respect to permuted genotype and permuted PPI. Consequently, only the two subnetworks identified by the proposed method are deemed statistically significant at a false discovery rate of $q < 0.05$.

3.3 Biological Relevance

In this section, we investigate the biological relevance of the two statistically significant subnetworks ($q < 0.05$) identified by the proposed method. These two subnetworks are shown in Figure 3. According to Ingenuity Pathway Analysis (IPA) software, the top subnetwork (Figure 3(a)) is significantly enriched in Estrogen Receptor Signaling ($p < 3.42E-12$) and Glucocorticoid Receptor Signaling ($p < 1.19E-3$). The second subnetwork (Figure 3(b)) is significantly enriched in Wnt/β-catenin Signaling ($p < 0.01$) and Cell Cycle Regulation by BTG Family Proteins ($p < 2.2E-4$).

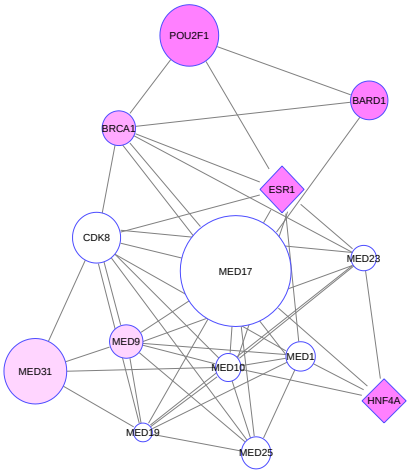
The association between a region of the estrogen receptor-α (ESR1) gene and T2D is reported in the literature [13]. Although the p-value of its association with T2D according to GWAS data before correction for multiple hypotheses is moderate ($p < 0.003$), this gene appears in the most significant subnetwork identified by the proposed algorithm. This subnetwork is significantly enriched in Estrogen receptor signaling pathway, which is known to play a crucial role on insulin resistance syndrome [9]. Glucocorticoid excess in vivo has been shown to cause decreased insulin sensitivity and insulin receptor binding in target tissues [6]. The first subnetwork is also enriched in Glucocorticoid Receptor Signaling. As shown in 3(a), this subnetwork contains

nine subunit of mediator complex which has an important role in regulating lipid metabolism linked to major human diseases including type 2 diabetes [39].

The second subnetwork is enriched in Wnt/β-catenin Signalling, which is a well-known pathway related to T2D. STRN, STRN4 and PPP2CA are previously reported to be associated with T2D, but do not have significant p-value according to the association analysis for individual variants (respectively 0.16, 0.19 and 0.12 before correction for multiple hypothesis testing). The subnetwork discovered using the proposed scoring scheme reveals the involvement of these genes in T2D-related processes, demonstrating that network analysis can provide information beyond what can be detected by GWAS data alone.

(a)

Gene	p-value
MED17	7.88E-08
MED31	0.000167
POU2F1	0.000292
CDK8	0.002154
ESR1	0.003988
HNF4A	0.004325
BARD1	0.0141
BRCA1	0.02167
MED9	0.03583
MED25	0.04457
MED1	0.05521
MED10	0.08962
MED23	0.09706
MED19	0.3629



(b)

Gene	p-value
MCC	0.001308
STRN3	0.001537
TRAF3IP3	0.001746
FAM40A	0.002053
CTTNBP2NL	0.004266
STK25	0.01041
FGFR1OP2	0.01574
CTTNBP2	0.023
STK24	0.03524
PPP2R1A	0.06095
PDCD10	0.06578
PPP2CA	0.1232
STRN	0.1642
STRN4	0.1934

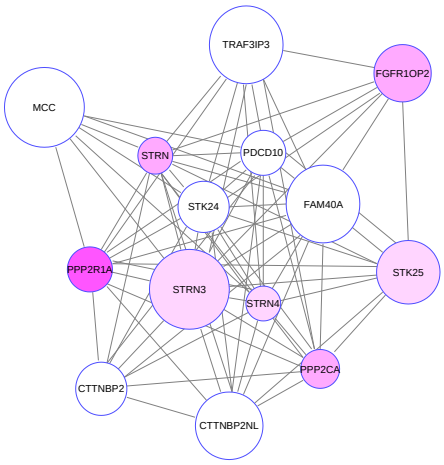


Fig. 3: Two subnetworks that are found to be significantly associated with T2D. The size of each node indicates the significance of the association of the corresponding protein with T2D (r_v, as defined by Equation 1). The diamond nodes are those that are previously reported to be associated with T2D in the literature [16]. The intensity of purple coloring in the nodes indicates the number of computational disease gene prioritization methods [31] that identified the respective gene to be associated with T2D. The individual p-values of each gene in the subnetwork are shown in the table left of the subnetwork. The genes with insignificant p-value ($p > 0.05$) that are known to be related to T2D are highlighted in yellow. The genes with insignificant p-value and are not reported to be related to T2D are highlighted in blue. These genes can be candidates for further investigation.

4 Conclusion

In this paper, with a view to facilitating the identification of disease-associated functional modules, we propose a novel methodology for scoring PPI subnetworks in terms of their association with a complex disease of

interest and their network connectivity. Our experimental studies show that objective criteria for scoring subnetworks have to be selected carefully to ensure that the algorithms can detect parsimonious subnetworks that are statistically significant. In particular, we show that, with a carefully designed scoring scheme, network analysis can extract knowledge from GWAS data beyond the scope of the data itself. Namely, the subnetworks identified by the proposed method contain genes that do not exhibit significant association with the disease based on analysis of GWAS data, but are known to have mechanistic role in the disease. Furthermore, the subnetworks identified by the proposed method include genes that are not yet reported to have a role in the disease, are not detected to be significant by GWAS, but have molecular functions that indicate potential involvement in the disease.

The method presented in this paper focuses on a single network pattern: dense subgraphs of the PPI network. However, investigation of different network patterns may provide additional insights on the relationships between different disease-associated genes and molecular mechanisms of these associations. The results reported here are limited to a single disease (T2D) based on a single large scale GWAS. In future work, application of the proposed method to various diseases and reproducibility analyses based on data from multiple cohorts will be crucial in establishing the generalizability of these promising results.

Acknowledgments

We would like to thank Thomas LaFramboise, Yu Liu, Pamela Clark, and Mark Chance for useful discussions. This work was supported in part by US National Science Foundation (NSF) award CCF-0953195 and US National Institutes of Health (NIH) award R01-LM011247. This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from `www.wtccc.org.uk`. Funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

Supplementary Material

The source code of MoBaS is freely available at `http://compbio.case.edu/mobas/`.

References

1. E.A. Adie, R.R. Adams, and et al. Speeding disease gene discovery by sequence based candidate prioritization. *BMC Bioinformatics*, 6, 2005.
2. E.A. Adie, R.R. Adams, and et al. SUSPECTS: enabling fast and effective prioritization of positional candidates. *Bioinformatics*, 22, 2006.
3. A. Agresti. *Categorical Data Analysis*. New York: John Wiley and Sons, 1990.
4. S. E. Baranzini, N. W. Galwey, J. Wang, P. Khankhanian, and et al. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum. Mol. Genet.*, 18:2078–2090, 2009.
5. C. Brorsson, N. T. Hansen, K. Lage, R. Bergholdt, S. Brunak, and F. Pociot. Identification of t1d susceptibility genes within the mhc region by combining protein interaction networks and snp genotyping data. *Diabetes Obes Metab*, pages 60–66, 2009.
6. E. V. Obberghen C. Grunfeld, K. Baird and C. R. Kahn. Glucocorticoid-induced insulin resistance in vitro: Evidence for both receptor and postreceptor defects. *Endocrinology*, 109:1723–1730, 1981.
7. Aaron Clauset, M. E. J. Newman, and Cristopher Moore. Finding community structure in very large networks. *Phys. Rev*, E 70, 2004.
8. W. T. C. C. Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.
9. J.Y. Deng, P.Sh. Hsieh, J.P. Huang, and et al. Activation of estrogen receptor is crucial for resveratrol-stimulating muscular glucose uptake via both insulin-dependent and -independent pathways. *Diabetes*, 57:1814–1823, 2008.
10. M.A. Driel, K. Cuelenaere, P.P. Kemmeren, and et al. GeneSeeker: extraction and integration of human disease-related information from web-based genetic databases. *Nucleic Acids Res.*, 33, 2005.
11. S. Erten, G. Bebek, and M. Koyutürk. Vavien: an algorithm for prioritizing candidate disease genes based on topological similarity of proteins in interaction networks. *Journal of Computational Biology*, 18, 2011.
12. J. L. Fleiss. *Statistical Methods for Rates and Proportions*. New York: Wiley., 1981.
13. C. J. Gallagher, C. D. Langerfeld, C. J. Gordon, and et al. Association of the estrogen receptor- gene with the metabolic syndrome and its component traits in african-american families. *Diabetes*, 56:2135–2141, 2007.

14. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18:S233–240, 2002.
15. P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao. dmGWAS: dense module searching for genome-wide association studies in protein-protein interaction networks. *Bioinformatics*, 27:95–102, 2011.
16. J. Lim, K. Hong, H. Jin, Y. Kim, H. Park, and B. Oh. Type 2 diabetes genetic association database manually curated for the study design and odds ratio. *BMC Medical Informatics and Decision Making*, 2010.
17. George C. Linderman, Mark R. Chance, and Gurkan Bebek. MAGNET: MicroArray Gene expression and Network Evaluation Toolkit. *Nucl. Acids Res.*, 2012.
18. Y. Liu, S. Patel, R. Nibbe, S. Maxwell, S. A. Chowdhury, M. Koyutürk, and et al. Systems biology analyses of gene expression and genome wide association study data in obstructive sleep apnea. *Pacific Symposium on Biocomputing*, pages 14–25, 2011.
19. N. Lopez-Bigas and C.A. Ouzounis. Genome-wide identification of genes likely to be involved in human genetic disease. *Nucleic Acids Res.*, 32, 2004.
20. H. Ma, E. Schadt, L. M. Kaplan, and H. Zhao. COSINE: COndition-SpecIfic sub-NEtwork identification using a global optimization method. *Bioinformatics*, 2011.
21. D. Maglott, J. Ostell, K. D. Pruitt, , and T. Tatusova. Entrez gene: gene-centered information at NCBI. *Nucl. Acids Res.*, 35, 2007.
22. S. Maslov and K. Sneppen. Specificity and stability in topology of protein networks. *Science*, 296, 2002.
23. Jason H. Moore, Folkert W. Asselbergs, and Scott M. Williams. Bioinformatics challenges for genome-wide association studies. *Bioinformatics*, 26(4):445–455, February 2010.
24. M. E. J. Newman. Fast algorithm for detecting community structure in networks. *Phys. Rev*, E 69(066133), 2004.
25. C. Perez-Iratxeta, M. Wjst, P. Bork, and M.A. Andrade. G2D: a tool for mining genes associated with disease. *BMC Genet.*, 6, 2005.
26. S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, and et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81:559–575, 2007.
27. Vijay K. Ramanan, Li Shen, Jason H. Moore, and Andrew J. Saykin. Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genetics*, 28:323–332, 2012.
28. M. D. Ritchie. Using prior knowledge and genome-wide association to identify pathways involved in multiple sclerosis. *Genome Medicine*, 1:65, 2009.
29. M. D. Ritchie. Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies. *Annals of Human Genetics*, 75(1):172–182, 2011.
30. L.J. Scott. A Genome-Wide Association Study of Type 2 Diabetes in Finns Detects Multiple Susceptibility Variants. *Science*, 316(5829):1341–1345, June 2007.
31. N. Tiffin, E. Adie, F. Turner, and et al. Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.*, 2006.
32. N. Tiffin, J.F. Kelso, and et al. Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Res.*, 33, 2005.
33. A. Torkamani, E. J. Topol, and N. J. Schork. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics*, 92:265–272, 2008.
34. S. Tornow and HW. Mewes. Functional modules by relating protein interaction networks and gene expression. *Nucleic Acids Res.*, 31, 2003.
35. F.S. Turner, D.R. Clutterbuck, and C.A. Semple. POCUS: mining genomic sequence annotation to predict disease genes. *Genome Biol.*, 4, 2003.
36. O. Vanunu, O. Magger, E. Ruppin, and et al. Associating genes and protein complexes with disease via network propagation. *PLoS Comput Biol.*, 2010.
37. K. Wang, M. Li, and M. Bucan. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet*, 81:1278–1283, 2007.
38. Y. Xia Y. Wang. Condition specific subnetwork identification using an optimization model. In *Proceedings of The Second International Symposium on Optimization and Systems Biology*, pages 333–340, 2008.
39. Y. Zhang, X. Zhao, and F. Yang. The mediator complex and lipid metabolism. *Journal of Biochemical and Pharmacological Research*, 1:51–55, 2013.