Bioinformatics Advance Access published November 16, 2006

BIOINFORMATICS

Network Neighborhood Analysis With The Multi-Node Topological Overlap Measure

Ai Li [a], Steve Horvath [a,b*]

[a]Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, USA,
[b]Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-7088, USA

Associate Editor: Golan Yona

ABSTRACT

Motivation: The goal of neighborhood analysis is to find a set of genes (the neighborhood) that is similar to an initial 'seed' set of genes. Neighborhood analysis methods for network data are important in systems biology. If individual network connections are susceptible to noise, it can be advantageous to define neighborhoods on the basis of a robust interconnectedness measure, e.g. the topological overlap measure. Since the use of multiple nodes in the seed set may lead to more informative neighborhoods, it can be advantageous to define multi-node similarity measures.

Results: The pairwise topological overlap measure is generalized to multiple network nodes and subsequently used in a recursive neighborhood construction method. A local permutation scheme is used to determine the neighborhood size. Using four network applications and a simulated example, we provide empirical evidence that the resulting neighborhoods are biologically meaningful, e.g. we use neighborhood analysis to identify brain cancer related genes.

Availability: A executable Windows program and tutorial for multi-node topological overlap measure (MTOM) based analysis can be downloaded from the following webpage:
http://www.genetics.ucla.edu/labs/horvath/MTOM/

1 INTRODUCTION

The main focus of this paper is a fundamental screening task: how to define the neighborhood of an initial set of nodes (genes) in a network. Intuitively speaking, a neighborhood is comprised of nodes that are highly connected to a given set of genes. Thus neighborhood analysis facilitates a guilt-by-association screening strategy for finding genes that interact with a given set of biologically interesting genes. To define a neighborhood of an initial gene set, one can make use of a similarity measure. For example, when dealing with gene expression microarray data, it is natural to use the correlation coefficient to measure pairwise gene co-expression similarity (Eisen *et al.*, 1998; Golub *et al.*, 1999).

Here we consider the setting of an undirected network that can be represented by a symmetric adjacency matrix $A = [a_{ij}]$. In an unweighted network, $a_{ij} = 1$ if nodes i and j are connected and 0 otherwise. In a weighted network, $a_{ij} \in [0, 1]$ encodes the pairwise connection strength.

A simple approach for defining a neighborhood of node i is to choose the nodes with highest adjacencies a_{ij}. In an unweighted network, this amounts to choosing the directly connected neighbors of node i.

Erroneous links (adjacencies) can have a strong impact on network topological inference (Lin *et al.*, 2004; Lin and Zhao, 2005). Since spurious or weak connections in the adjacency matrix may lead to 'noisy' neighborhoods, it can be advantageous to use node (dis-)similarity measures that are based on common interacting partners or on topological metrics (Ravasz *et al.*, 2002; Brun *et al.*, 2003; Zhao *et al.*, 2006; Chen *et al.*, 2006; Chua *et al.*, 2006). A comparison of different measures can be found in Chua *et al.* (2006) and Goldberg and Roth (2003).

A limitation of many network similarity measures is that they measure pairwise similarity. While pairwise similarities are useful for clustering procedures and many gene annotation procedures, we will argue that it can be advantageous for neighborhood analysis to introduce multi-node similarity measures.

We outline a procedure for generalizing pairwise network similarity or dissimilarity measures that are based on shared neighbors. The resulting multi-point measures keep track of the numbers of shared neighbors among multiple network nodes. We apply our approach to the topological overlap measure (Ravasz *et al.*, 2002).

1.1 Topological overlap measure

The topological overlap of two nodes reflects their similarity in terms of the commonality of the nodes they connect to. In an unweighted network, the number of shared neighbors of nodes i and j is given by $\sum_{u \neq i,j} a_{iu}a_{ju}$. The topological overlap $T = [t_{ij}]$ is a normalized version of this quantity. Specifically, the following definition of the pairwise topological overlap measure can be found in the supplementary material of Ravasz *et al.* (2002):

$$t_{ij} = \begin{cases} \frac{\sum_{u \neq i,j} a_{iu}a_{ju} + a_{ij}}{\min\{\sum_{u \neq i} a_{iu} - a_{ij}, \sum_{u \neq j} a_{ju} - a_{ij}\} + 1} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (1)$$

The inclusion of the term a_{ij} in the numerator makes t_{ij} explicitly depends on the existence of a direct link between the two nodes in question. An advantage of the quantity 1 in the denominator is that it prevents the denominator from becoming 0 when $\min\{\sum_{u \neq i} a_{iu} - a_{ij}, \sum_{u \neq j} a_{ju} - a_{ij}\} = 0$.

In the following, we use $0 \leq a_{ij} \leq 1$ to prove that $0 \leq t_{ij} \leq 1$. Since $\sum_{u \neq i,j} a_{iu}a_{ju} \leq \sum_{u \neq i} a_{iu} - a_{ij}$ and $\sum_{u \neq i,j} a_{iu}a_{ju} \leq \sum_{u \neq j} a_{ju} - a_{ij}$, which implies $\sum_{u \neq i,j} a_{iu}a_{ju} \leq \min\{\sum_{u \neq i} a_{iu} - a_{ij}, \sum_{u \neq j} a_{ju} - a_{ij}\}$. Along

*to whom correspondence should be addressed (shorvath@mednet.ucla.edu)

with $a_{ij} \leq 1$, we find that the numerator of t_{ij} is smaller than the denominator q.e.d.

2 APPROACH

2.1 Multi-node topological overlap measure

Here we generalize the topological overlap matrix to multiple nodes and show how to use it in neighborhood analysis. Our multi-node TOM is motivated by the observation that formula (1) can be expressed as

$$t_{ij} = \frac{|N(i,j)|+a_{ij}}{\min\{|N(i,-j)|,|N(-i,j)|\}+\binom{2}{2}}, \quad (2)$$

where $N(i,j)$ denotes the set of neighbors shared by i and j, $N(i,-j)$ denotes the set of the neighbors of i excluding j and $|\cdot|$ denotes the number of elements (cardinality) in its argument. Algebraically, we find

$$\begin{aligned} |(N(i,j))| &= \sum_{u \neq i,j} a_{iu}a_{ju} \quad (3) \\ |N(i,-j)| &= \sum_{u \neq i} a_{iu} - a_{ij}. \end{aligned}$$

The binomial coefficient $\binom{2}{2} = 1$ in the denominator of (2) is an upper bound of a_{ij}.

In light of formula (2), it is natural to define the multi-node topological overlap measure (MTOM) for three different nodes i,j,k as follows

$$t_{ijk} = \frac{|(N(i,j,k)|+a_{ij}+a_{ik}+a_{jk}}{\min\{|N(i,j,-k)|,|N(i,-j,k)|,|N(-i,j,k)|\}+\binom{3}{2}} \quad (4)$$

where

$$\begin{aligned} |(N(i,j,k))| &= \sum_{u \neq i,j,k} a_{iu}a_{ju}a_{ku} \quad (5) \\ |N(i,j,-k)| &= \sum_{u \neq i,j} a_{iu}a_{ju} - a_{ik}a_{jk}. \end{aligned}$$

Here $N(i,j,-k)$ can be regarded as the set of the neighbors shared by i and j excluding k. The binomial coefficient $\binom{3}{2} = 3$ in the denominator of (4) is an upper bound of $a_{ij} + a_{ik} + a_{jk}$ and equals the number of connections that can be formed between i, j, and k. Analogous to the proof for 2 nodes, one can prove that $0 \leq t_{ijk} \leq 1$. It is straightforward to extend the definition of the topological overlap measure to four or more nodes.

Generalizing MTOM to weighted networks: The algebraic formulas for MTOM do not require that the adjacencies a_{ij} take on binary values, i.e. that they encode an unweighted network. Even for a weighted network with $0 \leq a_{ij} \leq 1$, MTOM takes on values in the unit interval. Therefore, we use the algebraic formulation of the topological overlap matrix to define MTOM for weighted networks. Two simple examples illustrating the MTOM computation for four nodes are presented in Figure 1.

2.2 MTOM-based neighborhoods

We consider two basic approaches for defining a neighborhood based on the concept of multi-node topological overlap. The default approach is to build the neighborhood recursively. The non-recursive alternative is computationally faster but produces less interconnected neighborhoods.

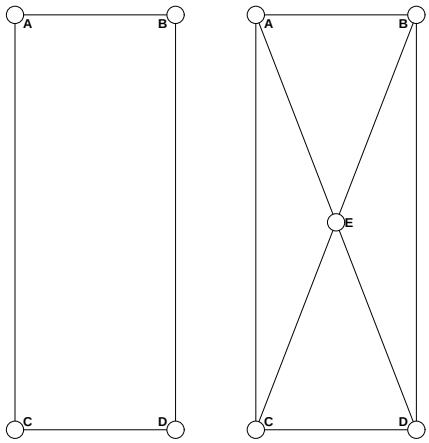


Fig. 1. Computing the four node topological overlap measures for nodes A,B,C,D in two simple networks *1)* $t_{A,B,C,D} = \frac{0+4}{0+6} = 0.667$ and *2)* $t_{A,B,C,D} = \frac{1+4}{1+6} = 0.714$.

The MTOM-based neighborhood analysis requires as input an initial seed neighborhood comprised of S_0 node(s) ($S_0 \geq 1$) and the requested final size of the neighborhood $S_t = S_0 + S \geq S_0$, where the S is the total number of nodes that will add to the initial neighborhood.

1. **Recursive approach**
   a. For each node outside of the current neighborhood, compute the MTOM value of the combined set of this node and the node(s) in the current neighborhood.
   b. Add the node associated with the highest MTOM value to the current neighborhood to reach the updated neighborhood.
   c. Repeat steps a) and b) S times until the final neighborhood size S_t is reached.

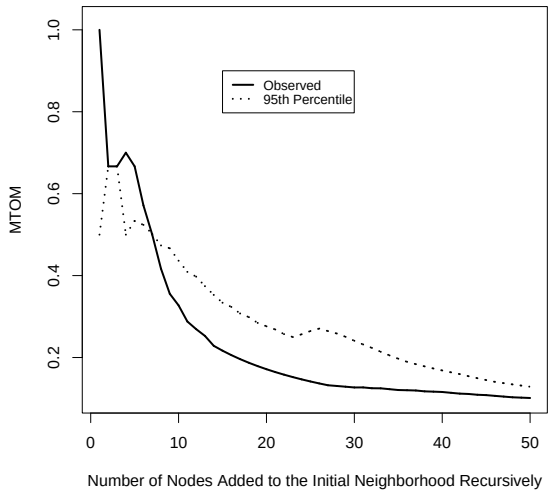
2. **Non-recursive approach**
   a. For each node outside of the initial neighborhood, compute the MTOM value of the combined set of this node and the node(s) in the initial neighborhood.
   b. Choose the S nodes associated with the highest MTOM values and combine with the initial neighborhood as the final neighborhood.

Since the recursive approach leads to neighborhoods with higher MTOM values, it is preferable over the computationally faster, non-recursive approach.

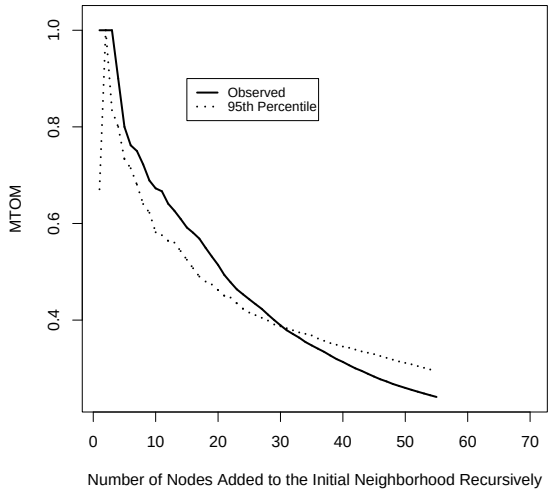
2.3 Local permutations for choosing the neighborhood size S

An obvious challenge is to choose the number $S = S_t - S_0$ of nodes to be added to the initial neighborhood. While prior knowledge of the pathway size may guide this choice, this information is not always available. We propose a permutation test based guideline to assist with the choice of S. The permutation test compares

MTOM values based on the original adjacency matrix with their corresponding values in permuted versions of the adjacency matrix. We find that global (whole network) permutations often lead to a network without any module structure and to unrealistically large estimates of the neighborhood size (thousands of nodes). Therefore, we propose to permute only those rows of the adjacency matrix that correspond to nodes in the initial seed neighborhood. Next the corresponding columns are permuted so that the resulting permuted adjacency matrix remains symmetric. After performing multiple permutations, one can estimate the 95th percentile of the permuted MTOM values. Figure 2 shows 1) the original MTOM value as a function of S and 2) the 95th percentile of the MTOM values calculated on the basis of locally permuted versions of the adjacency matrix. In our applications, we find that there is a value S_c such that if more than S_c nodes are added to the initial neighborhood recursively MTOM value curve dips below the 95th percentile of the permuted MTOM value curves. Since for neighborhood sizes smaller than S_{t_0}, where $S_{t_0} = S_0 + S_c$, the neighborhood is more interconnected than 95 percent of the locally permuted neighborhoods, we chose a neighborhood size close to S_{t_0} in our applications. The proposed local permutation test for choosing a neighborhood size is meant as a heuristic. In practice, the user should explore the robustness of the estimate with respect to picking other percentiles, e.g. the 90th percentile. Of course, prior biological knowledge regarding the neighborhood size should take precedent over the rough estimate provided by the local permutation test. As an alternative, we suggest that hierarchical clustering analysis involving the pairwise TOM dissimilarity may also provide some estimate on how large a cluster may surround the initial set. Neighborhood analysis, similar to gene screening strategies, leads to results that require careful validation involving independent data sets and biological validation methods.



(a)



(b)

Fig. 2. Using a local permutation test to choose the neighborhood size, i.e. the number of nodes S to be added to the initial neighborhood. a) yeast cell-cycle example; b) Drosophila protein-protein interaction network. The solid line shows the MTOM values (y-axis) of the observed network as a function of different neighborhood sizes (x-axis). The dashed line shows the 95th percentile of the MTOM values based on locally permuted adjacency matrices. A local permutation only permutates those rows (and columns) of the adjacency matrix that correspond to node of the initial neighborhood. As heuristic, we suggest to choose a value for S close to where the solid line (observed values) crosses the dashed line (95th percentile of permuted values).

3 APPLICATIONS

In the following sections, we apply our methods to gene co-expression networks and simple protein-protein interaction networks.

3.1 Predicting brain cancer genes in a co-expression network

The proposed neighborhood analysis can be used for both unweighted and weighted gene co-expression networks. Here we apply the method to find brain cancer related genes based on different initial seed neighborhoods. The data consisted of 55 brain cancer patients and their survival times. The gene expression profiles of each patient were measured with Affymetrix HG-U133A microarrays as detailed in Horvath *et al.* (2006). The details of the gene co-expression network construction are presented in Zhang and Horvath (2005). Briefly, the network adjacency matrix was defined by raising the Pearson correlation matrix between the gene expression profiles to the 6-th power, i.e. $a_{ij} = |cor(x_i, x_j)|^6$, where x_i and x_j are the expression profiles of gene i and j, respectively. Our findings remain largely unchanged with regard to different choices of the power $\beta = 6$. Further, an unweighted network construction approach leads to similar results (see our online material).

To illustrate the value of taking a multi-node perspective, we applied the MTOM approach to an initial seed neighborhood comprised of five well-known cancer-related genes: TOP2A, Rac1, TPX2, EZH2 and KIF14. Table 1 shows the results from the recursive MTOM analysis. Out of 20 probes in the MTOM neighborhood, we find that 15 are cancer related, which provides empirical evidence that the MTOM approach leads to biologically meaningful results.

Table 1. MTOM neighborhood analysis of an initial neighborhood comprised of five well-known cancer genes: TOP2A, Rac1, TPX2, EZH2 and KIF14. The columns report whether a probe set is known to be neuron related or cancer related according to a Pubmed search

Probe Name	Gene Name	Neuron	Cancer
209642_at	BUB1	unknown	yes
218355_at	KIF4A	unknown	unknown
222077_s_at	RACGAP1	unknown	yes
219918_s_at	ASPM	yes	yes
207828_s_at	CENPF	unknown	yes
202580_x_at	FOXM1	unknown	yes
202870_s_at	CDC20	unknown	yes
202095_s_at	BIRC5	unknown	yes
221591_s_at	FLJ10156	unknown	unknown
218009_s_at	PRC1	unknown	yes
204641_at	NEK2	unknown	yes
209172_s_at	CENPF	unknown	yes
209464_at	AURKB	unknown	yes
212020_s_at	MKI67	unknown	yes
204962_s_at	CENPA	unknown	yes
212023_s_at	MKI67	unknown	yes
204444_at	KIF11	unknown	unknown
212949_at	BRRN1	unknown	unknown
204026_s_at	ZWINT	unknown	unknown
203213_at	CDC2	unknown	yes

Table 2. Neighborhood of survival time based on the recursive MTOM approach. The columns report whether a probe set is known to be neuron related or cancer related according to a Pubmed search. The last column reports the correlation between gene expression profiles and survival times (TTS). Note that MTOM implicates known cancer genes even if their correlation is relatively low

Probe Name	Gene Name	Neuron	Cancer	Correlation
208464_at	GRIA4	yes	unknown	0.62
221623_at	BCAN	yes	unknown	0.406
91920_at	BCAN	yes	unknown	0.226
219107_at	BCAN	yes	unknown	0.212
216476_at	LOC115131	unknown	unknown	0.142
222301_at	CROC4	yes	unknown	0.093
212655_at	BDG29	unknown	unknown	0.297
213768_s_at	ASCL1	yes	yes	0.233
209988_s_at	ASCL1	yes	yes	0.191
209987_s_at	ASCL1	yes	yes	0.129
212265_at	QKI	yes	unknown	0.163
218902_at	NOTCH1	yes	yes	-0.023
202981_x_at	SIAH1	unknown	yes	0.085
221776_s_at	BRD7	unknown	yes	0.068
212615_at	FLJ12178	unknown	unknown	0.123
212616_at	FLJ12178	unknown	unknown	0.126
213891_s_at	TCF4	unknown	yes	0.091
201310_s_at	C5orf13	yes	yes	0.072
214239_x_at	RNF110	unknown	yes	0.076
213551_x_at	RNF110	unknown	yes	0.075

In the following, we provide a limited comparison to the simple (naive) approach of defining a neighborhood of node t based on ranking the remaining network nodes by their adjacencies with node t. For weighted gene co-expression networks constructed with the power function, this naive approach is equivalent to choosing genes based on the absolute values of their correlations with a given gene expression profile x_t.

It is worth emphasizing that when dealing with microarray data, one can also determine the neighborhood of a quantitative microarray sample trait, e.g. cancer survival time. To accomplish this mathematically, one considers the sample trait as an additional, idealized gene expression profile when constructing the co-expression network. For our weighted network example, the adjacency between the sample trait T and the i-th gene expression profile is given by $a_{Ti} = |cor(T, x_i)|^\beta$. In Table 2, we report the results of using MTOM to find a neighborhood with 20 gene neighbors around the sample trait 'brain cancer survival time'. We find that the MTOM-based neighborhoods are enriched with cancer and neuron related genes. Out of the 20 probe sets, 11 are related to neuron cells and 10 are related to cancer. Since the brain cancer microarray data were based on neuronal tissue samples, finding neuron or cancer related genes provides indirect (but only tentative) evidence that the resulting neighborhoods are biologically meaningful. Note that several of the probe sets in Table 2 correspond to the same genes, but the correlation between gene expression profiles and survival time varies greatly across the different probe sets of a gene.

In contrast, a standard, naive approach, which simply selects a neighborhood on the basis of the absolute values of the correlations between gene expression profile and survival time, leads to a neighborhood with fewer cancer and neuron related genes. Out of the 20 most highly correlated probe sets in Table 3, only 4 are related to neuron cells and only 6 are related to cancer. Comparing Tables 2 and 3 provides indirect empirical evidence that the MTOM neighborhood analysis leads to biologically more meaningful results than the standard approach in this application.

3.2 Neighborhood analysis for predicting cell cycle proteins in yeast

Here we use a MTOM neighborhood analysis to predict cell cycle related proteins. Numerous protein annotation methods have been presented in the literature, e.g. recent papers include Deng *et al.* (2006) and Carroll and Pavlovic (2006). Our limited analysis is meant to illustrate the value of taking a multi-node perspective. A comprehensive comparison to other methods is beyond the scope of this article. The protein identifiers of the open reading frames (ORF) were obtained from the Saccharomyces Genome Database (SGD) and the yeast protein-protein interactions (PPI) were retrieved from the Munich Information Center for Protein Sequences (MIPS) (Guldener *et al.*, 2006). We restricted the analysis to the largest connected component comprised of 3858 proteins with 7196 pairwise interactions. To compare different neighborhood analysis approaches, we studied the neighborhoods of subsets of 101 cell cycle related proteins found in the Kyoto Encyclopedia of Genes and Genomes (KEGG). A local permutation test suggested $S = 10$. Within each neighborhood, we determined the number C of cell cycle related proteins. We found that C is significantly correlated

Table 3. Correlation based neighborhood of the survival time (TTS). The columns report whether a probe set is known to be neuron related or cancer related according to a Pubmed search. The last column lists the correlation between the genes and TTS.

Probe Name	Gene Name	Neuron	Cancer	Correlation
208464_at	GRIA4	yes	unknown	0.62
204529_s_at	TOX	unknown	unknown	0.601
206170_at	ADRB2	unknown	unknown	0.539
216247_at	RPS20	unknown	unknown	0.537
214028_x_at	TDRD3	unknown	unknown	0.526
213447_at	IPW	unknown	unknown	0.522
207113_s_at	TNF	yes	yes	0.514
218036_x_at	CGI-07	unknown	unknown	0.504
209160_at	AKR1C3	unknown	yes	0.5
206107_at	RGS11	unknown	unknown	0.496
209782_s_at	DBP	unknown	unknown	0.495
211653_x_at	AKR1C1	unknown	yes	0.494
213778_x_at	ZFP276	unknown	unknown	0.49
215119_at	MYR8	yes	unknown	0.486
202753_at	p44S10	unknown	yes	-0.481
209292_at	ID4	unknown	yes	0.481
204530_s_at	TOX	unknown	unknown	0.481
221974_at	SNRPN	unknown	unknown	0.481
219188_s_at	LRP16	unknown	yes	0.48
205630_at	CRH	yes	unknown	0.478

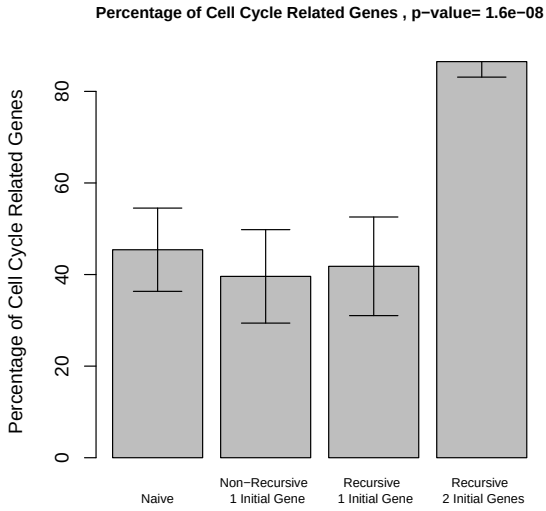


Fig. 3. Comparing the percentage of cell cycle proteins R (y-axis) in neighborhoods constructed in different ways for the Yeast Protein-Protein Physical Interaction Network (MIPS Data). The recursive approach involving an initial neighborhood of two cell cycle related 'hub' proteins performs better than approaches based on an initial set comprised of a single protein. In this application, the recursive and the non-recursive MTOM neighborhood analysis involving a *single* initial protein do not lead to better results than the naive approach of building a neighborhood on the basis of direct connections (adjacency=1) with the initial protein. We also report the p-values of the Kruskal-Wallis rank sum test, which is a non-parametric multi-group comparison test.

with the network connectivity k of the initial protein (Spearman correlation $r = 0.36$, p-value ≤ 0.001) across the 101 cell cycle genes. We focused the neighborhood analysis on subsets of the 50 most highly connected 'hub' cell cycle related proteins. These proteins had a connectivity greater than or equal to 4, i.e. each initial protein had at least 4 known interactions. Our results are largely unchanged with regards to using more highly connected genes in the initial neighborhood set. However, using less connected proteins ($k < 3$) leads to neighborhoods that contain very few cell cycle related proteins.

As can be seen from Figure 3, the neighborhoods of cell cycle genes tend to be enriched with other cell cycle genes as well. A major advantage of the MTOM screening approach is the ability to input multiple initial nodes as seed set. Figure 3 shows that an initial seed neighborhood comprised of two cell cycle related hub proteins leads to far better results than using a single protein as input. But as Figure 4 indicates, this is only true for protein pairs that have high topological overlap. Note that pairs of proteins resulting in neighborhoods with high percentages of cell cycle related proteins are comprised of proteins with high topological overlap measure.

3.3 Neighborhood analysis for predicting essential yeast proteins

Networks are a natural framework for understanding protein-protein interactions, see e.g. Jeong *et al.* (2001), Yook *et al.* (2004) and Deng *et al.* (2006). Knock-out experiments in lower organisms (e.g. yeast, fly, worm) have shown that essential proteins tend to be highly connected 'hub' proteins in protein-protein interaction networks (Jeong *et al.*, 2001, 2003; Hahn and Kern, 2005). Here we use MTOM-based neighborhood analysis to predict essential proteins in a yeast protein-protein interaction network (BioGrid data) (Breitkreutz *et al.*, 2003). The largest connected component contained 3332 proteins that include 877 essential proteins. We find that proteins that are in the neighborhood of essential, highly connected hub proteins have an increased chance of being essential as well. Specifically, we picked essential seed genes from among the 200 most highly connected essential proteins. Based on our local permutation test, we chose $S = 30$ for MTOM analysis. The percentage of essential proteins in the neighborhoods constructed by different methods are reported in Figure 5. Apart from seed sets comprised of a single gene, we also considered seeds involving two and three essential hub proteins with high topological overlap. Note that as the initial neighborhood size increases, so does the biological signal in the resulting neighborhoods. In this application, neighborhoods built on the basis of multiple interconnected initial proteins lead to better results than standard methods that can only input a single protein.

3.4 Neighborhood analysis for predicting essential proteins in Drosophila

Here we apply MTOM based neighborhood analysis to predict essential proteins in a Drosophila (fly) protein-protein interaction network (BioGrid Data) (Breitkreutz *et al.*, 2003). The largest connected component contained 2294 proteins that include 282 known essential proteins. Since essential genes tend to be highly connected, we chose subsets of the 100 most highly connected

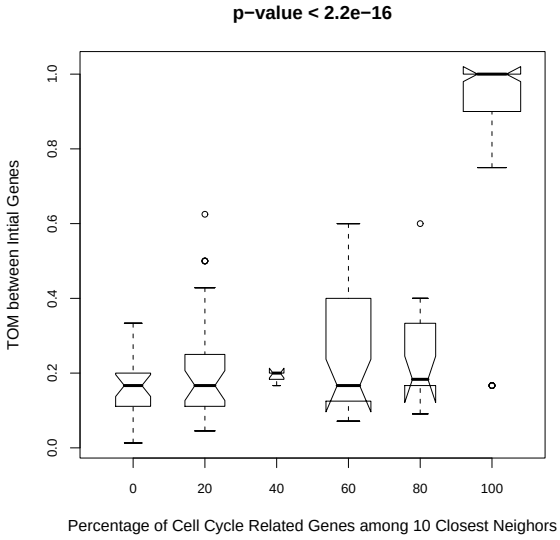


Fig. 4. Boxplots for visualizing the distribution of the topological overlap (y-axis) of initial protein pairs that lead to neighborhoods with a high percentage of cell cycle genes (x-axis). A boxplot consists of the most extreme values (the whiskers) in the data set (maximum and minimum values), the lower and upper quartiles (lower and upper boundary of the box), and the median value (horizontal line inside the notch). A notch is drawn in each side of the box. If the notches of two plots do not overlap, the two medians differ significantly.

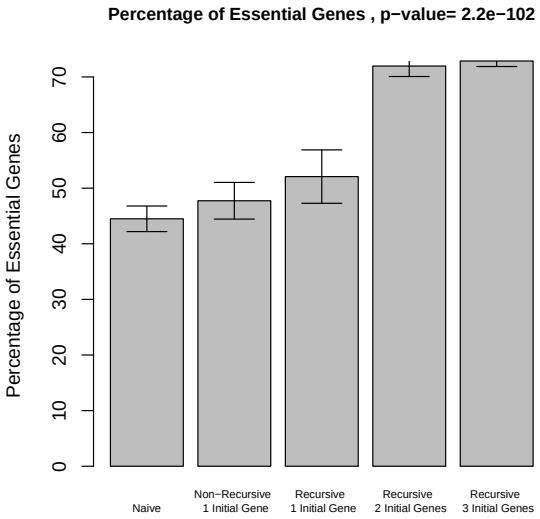


Fig. 5. Comparing the percentage of essential proteins R (y-axis) in neighborhoods constructed in different ways for the yeast protein-protein interaction network (BioGrid Data). The neighborhoods initialized by sets of two or 3 hub proteins contain more essential proteins that those constructed from a single protein. We also report the p-values of the Kruskal-Wallis rank sum test, which is a standard non-parametric multi-group comparison test.

essential proteins as initial seeds. Our local permutation test suggested $S = 30$. Figure 6 reports the percentages of essential proteins in neighborhoods constructed using alternative methods. In this application, the recursive MTOM neighborhood analysis involving a single initial seed protein leads to a better result than both the naive and the non-recursive MTOM approaches. Further, Figure 6 demonstrates the value of choosing multiple nodes as seeds for neighborhood analysis.

4 SIMULATION

To evaluate out method, we simulated a network model motivated by our yeast and cancer co-expression network applications. While this simple model was motivated by our unpublished research on the structure of co-expression networks, it is beyond the scope of this article to discuss the relationship of this simple model to actual weighted gene co-expression networks. Here we use the model to argue that MTOM leads to more meaningful results than standard neighborhood analysis methods.

Specifically, we simulated a gene expression data set $\{x_{ij}\}$ comprised of 2000 genes ($1 \leq i \leq 2000$) and 60 microarray samples ($1 \leq j \leq 60$). The network was simulated to be comprised of 6 modules with sizes $n_1 = 400$, $n_2 = 400$, $n_3 = 400$, $n_4 = 300$, $n_5 = 300$ and $n_6 = 200$. Within the k-th module, the i-th gene had the following expression value in the j-th microarray sample.

$$x_{ij}^{(k)} = m_j^{(k)} \times (\frac{i}{n_k})^{1/4} + \epsilon_{ij}^{(k)}$$

where the stochastic noise $\epsilon_{ij}^{(k)}$ was simulated to follow a normal distribution with mean 0 and variance 6. The vector $m_j^{(k)}$ was given below and turned out to be highly correlated ($r > 0.95$) with the first principal component of the corresponding module expression matrix (also known as module eigengene or metagene).

$$\begin{aligned}
m_j^{(1)} &= 1.5 \times I_{30<j\leq 45} + 1 \times I_{45<j\leq 60}, \\
m_j^{(2)} &= 1 \times I_{0<j\leq 15} + 1 \times I_{15<j\leq 30}, \\
m_j^{(3)} &= 1 \times I_{0<j\leq 15} + 1 \times I_{15<j\leq 30} + 1 \times I_{30<j\leq 45}, \\
m_j^{(4)} &= 1 \times I_{0<j\leq 15} + 1 \times I_{30<j\leq 45}, \\
m_j^{(5)} &= 1.5 \times I_{0<j\leq 15} + 1.5 \times I_{30<j\leq 45} + 0.5 \times I_{45<j\leq 60}, \\
m_j^{(6)} &= 0,
\end{aligned}$$

where the indicator function $I_{30<j\leq 45}$ equals 1 if the condition is satisfied and 0 otherwise. To quantify co-expression, we correlated the simulated gene expression with each other, which resulted in a 2000×2000 dimensional correlation matrix. To arrive at a simulated weighted gene co-expression network (adjacency matrix), we raised the entries of the correlation matrix to the power of $\beta = 6$, i.e. $a_{ij} = |cor(x_i, x_j)|^6$,where x_i and x_j are the expression profiles of gene i and j, respectively.

The goal of our neighborhood analysis was to determine membership in the first module that contained $n_1 = 400$ genes. We considered initial neighborhoods comprised of 1 or 2 genes out of the 50 most highly connected module genes. We considered $S = 30$. For each neighborhood, the percentage of module 1 genes represents the simulated biological signal. Figure 7 shows the results from averaging the signal over 50 MTOM analyses corresponding

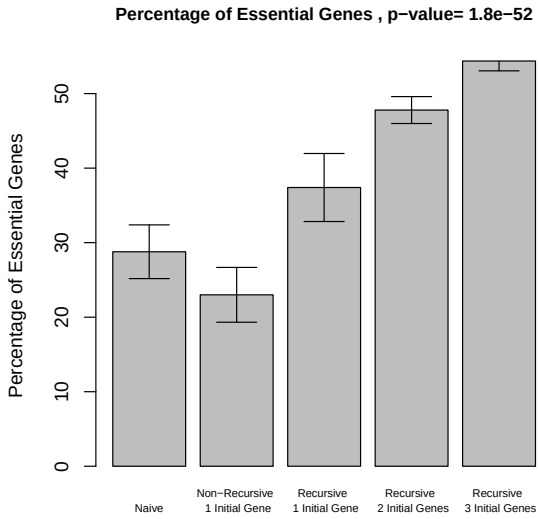


Fig. 6. Comparing the percentages of essential proteins R (y-axis) in neighborhoods constructed in the Drosophila protein-protein interaction network. The recursive approach involving an initial neighborhood of a single essential protein performs better than the non-recursive and naive approaches. As the initial neighborhood size increases, so does the biological signal in the resulting neighborhoods.

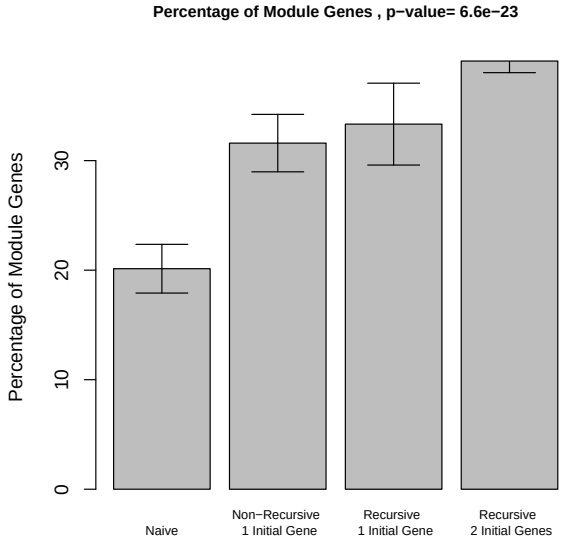


Fig. 7. Comparing the percentage of module 1 genes (y-axis) that are retrieved by different neighborhood construction methods for the simulated network. The recursive approach involving an initial neighborhood of two 'hub' genes in the first module leads to the best neighborhoods. In this application, the recursive and the non-recursive MTOM neighborhood analysis involving a *single* initial gene outperform the naive approach of simply using the 30 genes with highest adjacency with the initial gene. Further, an initial neighborhood comprised of 2 genes (with high topological overlap) leads to better results than initial neighborhoods comprised of a single gene.

to a single initial hub gene and 500 MTOM analyses corresponding to pairs of genes with high topological overlap.

5 COMPARING MTOM TO THE AVERAGE PAIRWISE TOM

One can easily define a multi-node similarity measure by the average of the pairwise similarities between the nodes. Since the average pairwise similarity measure is computationally much simpler, it is important to argue that a multi-node TOM measure performs better than the average pairwise topological overlap measures. To facilitate such a comparison, we study here the performance of the **averaged TOM neighborhood** construction method which recursively adds nodes based on average pairwise topological overlap measure, i.e. at each step, it adds the node with the maximum average pairwise similarity to the current neighborhood.

To compare our proposed recursive MTOM method and with the averaged TOM neighborhood construction method, we carried out 3 comparisons.

The first comparison involves comparing the simulated or biological signal in the resulting neighborhoods for the different applications. Using simulated and biological applications, we find that the MTOM method outperforms the averaged TOM method. In Figure 8, we report three representative comparisons.

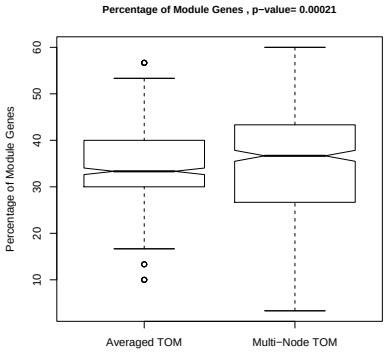
The second comparison involves comparing the MTOM values of the neighborhoods constructed with the different methods. As is to be expected, MTOM based neighborhoods have significantly higher MTOM values than neighborhoods constructed with the averaged TOM method (Figure 9).

The third comparison involves comparing the average pairwise TOM values of the neighborhoods constructed with the different methods. According to this metric, we find that the recursive MTOM method is significantly better than the averaged TOM approach (Figure 10). In summary, we find that the proposed MTOM measure outperforms the average pairwise TOM measure in our applications and simulated example.

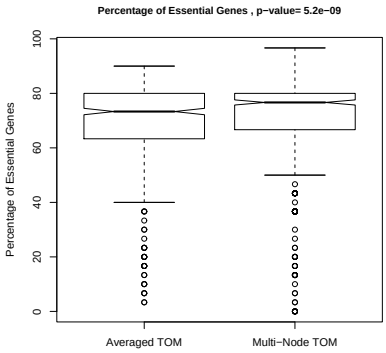
6 CONCLUSION

If individual network connections are susceptible to noise, then it can be advantageous to define neighborhoods on the basis of a more robust measure based on shared neighbors, e.g. the topological overlap measure. To illustrate the value of taking a multi-node perspective when defining neighborhoods, we generalize the standard pairwise topological overlap measure (TOM) to measure the topological overlap of multiple nodes (MTOM). MTOM is a natural extension of the standard pairwise topological overlap measure to multiple nodes. But it should be straightforward to adapt our approach to alternative overlap measures described in Brun *et al.* (2003), Zhao *et al.* (2006), Chen *et al.* (2006) and Chua *et al.* (2006). Since computation time was a concern in our analyses, we presented a recursive and non-recursive approach for constructing neighborhoods. But it may be worth-wile to explore the use of alternative, more time consuming, construction methods. For example, stepwise methods that allow for node deletion at each step may lead to neighborhoods with higher MTOM values.

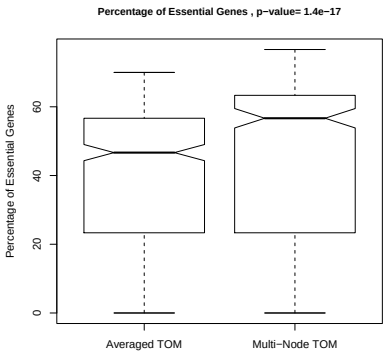
Further, we describe a local permutation scheme for determining the size of a neighborhood.



(a)

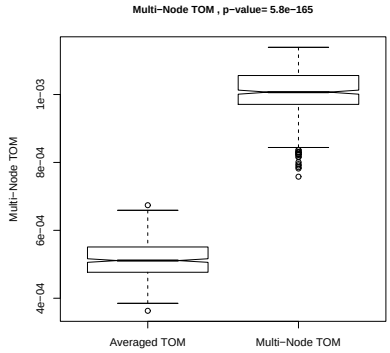


(b)

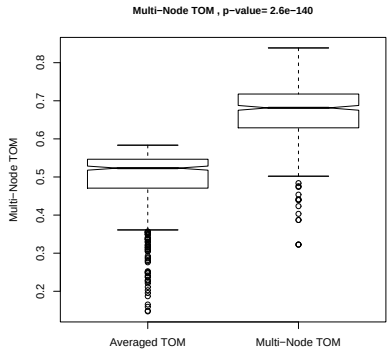


(c)

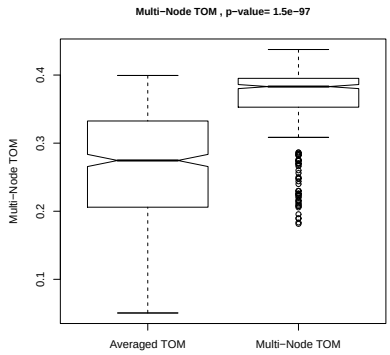
Fig. 8. Recursive MTOM neighborhoods contain a significantly better signal (y-axis) than averaged TOM neighborhoods. Here we report three representative examples: a) the simulated network; b) essential genes in the yeast protein-protein interaction network; c) essential genes in the Drosophila (fly) protein-protein interaction network. We report the Kruskal Wallis p-values for comparing the median values. The median value corresponds to the horizontal line inside the box. The corresponding notch around the median line denotes the 95 percent confidence interval.



(a)



(b)

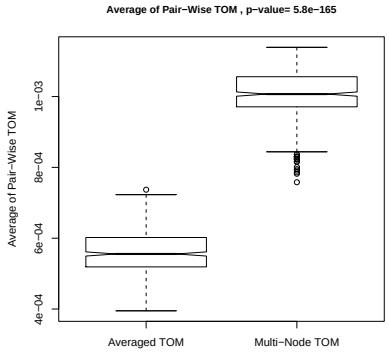


(c)

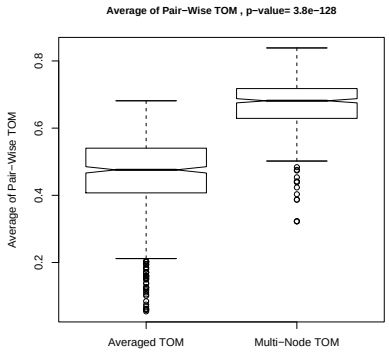
Fig. 9. Recursive MTOM neighborhoods have higher MTOM values (y-axis) than averaged TOM neighborhoods. Here we report three representative examples: a) the simulated network; b) essential genes in the yeast protein-protein interaction network; c) essential genes in the Drosophila (fly) protein-protein interaction network.

Using four network applications and a simulated example, we provide evidence that the MTOM approach yields biologically meaningful results. For example, we use MTOM to identify brain cancer related genes in a co-expression network and to identify essential genes in protein interaction networks. We provide empirical evidence that a neighborhood surrounding an initial set of two or more nodes can be far more informative than the neighborhood of a single node.

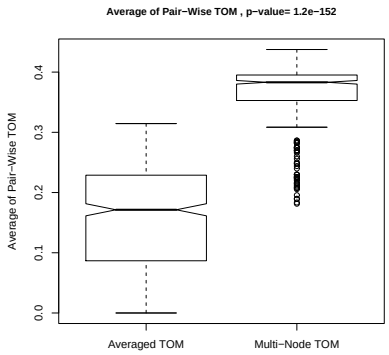
Our approach has several limitations. First and foremost, MTOM-based neighborhood analysis will only be useful in applications that satisfy the following assumption: *The more neighbors are shared*



(a)



(b)



(c)

Fig. 10. Recursive MTOM neighborhoods have higher average pairwise TOM value (y-axis) than averaged TOM neighborhoods. Here we report three representative examples: a) the simulated network; b) essential genes in the yeast protein-protein interaction network; c) essential genes in the Drosophila (fly) protein-protein interaction network.

by multiple nodes, the stronger is the biological relationship among them. The second limitation of our approach is that we assume the setting of an undirected network. While these types of networks are widely used in systems biology, we briefly mention that directed, Bayesian or Boolean network models allow for a probabilistic or even causal analysis of similar data, see e.g. Schaefer and Strimmer (2005) and Carroll and Pavlovic (2006).

The topological overlap measure can serve as a filter that decreases the effect of spurious or weak connections. Our applications and several publications provide empirical evidence that the topological overlap matrix leads to biologically meaningful results (Ravasz *et al.*, 2002; Ye and Godzik, 2004; Carlson *et al.*, 2006; Gargalovic *et al.*, 2006; Ghazalpour *et al.*, 2006; Horvath *et al.*, 2006). But there will undoubtedly be situations when alternative similarity measures are preferable. We expect that the multi-node measures will also be useful for module detection when coupled with a suitable clustering procedure.

ACKNOWLEDGMENT

The authors would like to thank our collaborators Jun Dong, Dan Geschwind, Peter Langfelder, Jake Lusis, Paul Mischel, Stan Nelson, Mike Oldham, Anja Presson, Lin Wang and Wei Zhao.

REFERENCES

Breitkreutz, B.J., Stark, C. and Tyers, M. (2003). The GRID: the general repository for interaction datasets. *Genome Biol*, **4**, R23.

Brun, C., Chevenet, F., D., Martin, Wojcik, J., Guenoche, A. and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a proteinprotein interaction network. *Genome Biol*, **5**(1), R6.

Carlson, M., Zhang, B., Fang, Z., Mischel, P., Horvath, S. and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: Predictions from modular yeast co-expression networks. *BMC Genomics*, **7**(40).

Carroll, S. and Pavlovic, V. (2006). Protein classification using probabilistic chain graphs and the gene ontology structure. *Bioinformatics*, **22**, 1871–78.

Chen, J., Hsu, W., Lee, M.L. and Ng, S. (2006). Increasing confidence of protein interactomes using network topological metrics. *Bioinformatics*, **22**, 1998–2004.

Chua, N. H., Sung, W and Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from proteinprotein interactions. *Bioinformatics*, **22**, 1623–1630.

Deng, M., Tu, Z., Sun, F. and Chen, T. (2006). Mapping gene ontology to proteins based on proteinprotein interaction data. *Bioinformatics*, **20**(6), 895–902.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *PNAS*, **95**(25), 14863–14868.

Gargalovic, P.S., Imura, M., Zhang, B., Gharavi, N.M., Clark, M.J., Pagnon, J., Yang, W.P., He, A., Truong, A., Patel, S., Nelson, S.F., Horvath, S., Berliner, J.A., Kirchgessner, T.G. and Lusis, A.J. (2006). Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *PNAS*, **103**(34), 12741–6.

Ghazalpour, A., Doss, S., Zhang, B., Plaisier, C., Wang, S., Schadt, E.E., Thomas, A., Drake, T.A., Lusis, A.J. and Horvath, S. (2006). Integrating genetics and network analysis to characterize genes related to mouse weight. *PloS Genetics*, **2**(8).

Goldberg, D.S. and Roth, F.P. (2003). Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. USA*, **100**, 4372–4376.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., H. Coller, 1 M. L. Loh, Downing, J. R., Caligiuri, M. A., Bloomfield, C. D. and Lander, E. S. (1999). Molecular

classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guldener, U., Munsterkotter, M., Oesterheld, M., Pagel, P., Ruepp, A., Mewes, HW. and Stumpflen, V. (2006). Mpact: the mips protein interaction resource on yeast. *Nucleic Acids Res.*, **34**(Jan), 436–441.

Hahn, M. W. and Kern, A. D. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Molecular biology and evolution*, **22**(4), 803–806.

Horvath, S., Zhang, B., Carlson, M., Lu, K.V., Zhu, S., Felciano, R.M., Laurance, M.F., Zhao, W., Shu, Q., Lee, Y., Scheck, A.C., Liau, L.M., Wu, H., Geschwind, D.H., Febbo, P.G., Kornblum, H.I., T.F., Cloughesy, Nelson, S.F. and Mischel, P.S. (2006). Analysis of oncogenic signaling networks in glioblastoma identifies aspm as a novel molecular target. *Proc Nat Acad Sci*, **103**(46), 22–29.

Jeong, H., Mason, S. P., Barabasi, A. L. and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, **411**(May), 41.

Jeong, H., Oltvai, Z. and Barabási, A. (2003). Prediction of protein essentiality based on genome data. *ComPlexUs*, **1**, 19–28.

Lin, N. and Zhao, H. (2005). Are scale-free networks robust to measurement errors? *BMC Bioinformatics*, **16**(6), 119.

Lin, N., Wu, B., Jansen, R., Gerstein, M. and Zhao, H. (2004). Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, **4**, 154.

Ravasz, E, Somera, A L, Mongru, D A, Oltvai, Z N and Barabasi, A L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, **297**(5586), 1551–5.

Schaefer, J. and Strimmer, K. (2005). An empirical bayes approach to inferring large-scale gene association networks. *Bioinformatics*, **21**, 754–764.

Ye, Y. and Godzik, A. (2004). Comparative analysis of protein domain organization. *Genome Biology*, **14**(3), 343–353.

Yook, S Y, Oltvai, Z N and Barabasi, A L. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, **4**, 928–942.

Zhang, B. and Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**(1), 17.

Zhao, W., Serpedin, E. and Dougherty, E.R. (2006). Information theoretic method for recovering temporal gene regulations from time series microarray data. *Bioinformatics*, **22**, 2129–2135.