BMC Bioinformatics

BioMed Central

Research article

Gene network interconnectedness and the generalized topological overlap measure

Andy M Yip[1] and Steve Horvath*[2]

Address: [1]Department of Mathematics, National University of Singapore, 2, Science Drive 2, Singapore 117543, Singapore and [2]Department of Human Genetics and Department of Biostatistics, University of California, Los Angeles, CA 90095, USA

Email: Andy M Yip - matymha@nus.edu.sg; Steve Horvath* - shorvath@mednet.ucla.edu

* Corresponding author

Published: 24 January 2007

BMC Bioinformatics 2007, **8**:22 doi:10.1186/1471-2105-8-22

Received: 13 August 2006

Accepted: 24 January 2007

This article is available from: http://www.biomedcentral.com/1471-2105/8/22

Abstract

Background: Network methods are increasingly used to represent the interactions of genes and/or proteins. Genes or proteins that are directly linked may have a similar biological function or may be part of the same biological pathway. Since the information on the connection (adjacency) between 2 nodes may be noisy or incomplete, it can be desirable to consider alternative measures of pairwise interconnectedness. Here we study a class of measures that are proportional to the number of neighbors that a pair of nodes share in common. For example, the topological overlap measure by Ravasz *et al.* [1] can be interpreted as a measure of agreement between the $m = 1$ step neighborhoods of 2 nodes. Several studies have shown that two proteins having a higher topological overlap are more likely to belong to the same functional class than proteins having a lower topological overlap. Here we address the question whether a measure of topological overlap based on higher-order neighborhoods could give rise to a more robust and sensitive measure of interconnectedness.

Results: We generalize the topological overlap measure from $m = 1$ step neighborhoods to $m \geq 2$ step neighborhoods. This allows us to define the m-th order generalized topological overlap measure (GTOM) by (i) counting the number of m-step neighbors that a pair of nodes share and (ii) normalizing it to take a value between 0 and 1. Using theoretical arguments, a yeast co-expression network application, and a fly protein network application, we illustrate the usefulness of the proposed measure for module detection and gene neighborhood analysis.

Conclusion: Topological overlap can serve as an important filter to counter the effects of spurious or missing connections between network nodes. The m-th order topological overlap measure allows one to trade-off sensitivity versus specificity when it comes to defining pairwise interconnectedness and network modules.

Background

We consider undirected, unweighted biological networks that can be represented by a symmetric adjacency matrix $A = [a_{ij}]$. The adjacency a_{ij} between nodes i and j equals 1 if the nodes are connected and 0 otherwise. For notational convenience, we set the diagonal elements to 1. While the adjacency matrix considers each pair of genes in isolation, topological overlap considers each pair of genes in relation to all other genes in the network. More specifically, genes are said to have high topological overlap if they are

connected to roughly the same group of genes in the network (i.e. they share the same neighborhood). To calculate the topological overlap for a pair of genes, their connections with all other genes in the network are compared. If the 2 nodes connect to the same group of other nodes, then they have a high 'topological overlap'. Here we study the properties of the topological overlap measure (TOM) and propose a generalization that enriches TOM's sensitivity to longer ranging connections between nodes.

There is empirical evidence that two substrates having a higher overlap are more likely to belong to the same functional class than substrates having a lower topological overlap [1-5]. Such a finding prompts the question whether a measure of topological overlap based on higher-order neighborhoods would lead to a more sensitive and robust measure of interconnectedness. In this paper, we generalize the topological overlap measure by incorporating information from higher-order neighborhoods and show that it leads to a definition of larger modules. Specifically, the *m*-th order topological overlap measure is constructed by (i) counting the number of *m*-step neighbors that a pair of nodes share and (ii) normalizing it to take a value between 0 and 1. The resulting node similarity measure is a measure of agreement between the *m*-step neighborhoods of 2 input nodes. Such a measure can be applied in a number of ways, for instance, ranking the genes, similarity search, prediction based on *k*-nearest neighbors, multi-dimensional scaling and module identification by clustering.

Results

The algebraic definition of the topological overlap measure can be found in Eq. (7) in the Methods section. Here we provide a more intuitive set theoretic interpretation of the topological overlap measure. Our generalization of the TOM is motivated by the observation that Eq. (7) can be expressed as

$$t_{ij} = \begin{cases} \dfrac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min\{|N_1(i)|, |N_1(j)|\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (1)$$

where $N_1(i)$ denotes the set of direct neighbors of i excluding i itself and $|\cdot|$ denotes the number of elements (cardinality) in its argument. The quantity $|N_1(i) \cap N_1(j)|$ measures the number of common neighbors that nodes i and j share whereas $|N_1(i)|$ gives the number of neighbors of i. The topological overlap t_{ij} assumes a minimal value of 0 if there is no direct linkage between the two nodes and if they share no common direct neighbors. It assumes a maximum value of 1 if there is a direct link between the two nodes and if one set of direct neighbors is a subset of the other. The fact that t_{ij} is bounded between 0 and 1 is used in the definition of the topological overlap based dissimilarity measure (see Eq. 4).

Generalizing TOM to m*-step neighborhoods*

By denoting $N_m(i)$ (with $m > 0$) the set of nodes (excluding i itself) that are reachable from i within a path of length m, i.e.,

$$N_m(i) := \{j \neq i | \text{dist}(i, j) \leq m\} \quad (2)$$

where dist(i, j) is the geodesic distance (shortest path distance) between i and j, we obtain a very natural generalization of the TOM, which reads as follows

$$t_{ij}^{[m]} = \begin{cases} \dfrac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min\{|N_m(i)|, |N_m(j)|\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (3)$$

We define the matrix $T^{[m]} = [\, t_{ij}^{[m]} \,]$ as the *m-th order generalized topological overlap measure* (GTOM*m*). Thus, GTOM*m* measures the agreement of the *m*-step neighborhoods between 2 nodes. When $m = 1$, this definition reduces to the original TOM in Eq. (7).

We find it convenient to define the zeroth order GTOM0 as the adjacency matrix, i.e. $T^{[0]} \equiv A$. Since $T^{[m]}$ is symmetric and non-negative, $T^{[m]}$ can be considered as a *similarity* measure [6]. To turn $T^{[m]}$ into a dissimilarity measure for use in clustering, we make use of the fact that $t_{ij}^{[m]}$ is bounded by 1. The generalized topological overlap-based dissimilarity measure is defined by

$$d_{ij}^{T,[m]} = 1 - t_{ij}^{[m]}. \quad (4)$$

Predicting essential proteins in a Drosophila protein network

Knock-out experiments in lower organisms (e.g. yeast, fly, worm) have shown that essential proteins tend to be more highly connected than non-essential proteins [7-9]. To illustrate the biological usefulness of the proposed interconnectedness measures, we set out to test the ability of GTOM to predict essential proteins using a Drosophila (fly) protein-protein interaction network from the BioGrid Database [10]. In our version of the database, the largest connected component was comprised of 2294 proteins with 21217 pairwise interactions. Knock-out experiments had implicated 282 essential proteins.

To illustrate the use of GTOM*m*, we will show that proteins that are highly interconnected with an essential protein have an increased chance of being essential as well.

Toward this end, we make use of the following terminology. A GTOMm neighborhood of size S around node i is defined as the set $GTOMhood_S^{[m]}(i)$ of S genes with highest GTOM value with i. For example, node j is in the size S = 1 GTOMm neighborhood around node i if it has the highest topological overlap $t_{ij}^{[m]}$ across all nodes. For simplicity, we ignore ties and define $GTOMhood_S^{[0]}(i)$ as the set of genes that directly link to node i (irrespective of the neighborhood size S).

Since essential proteins may participate in the same pathway, it is biologically plausible that the GTOMm neighborhoods of an initial essential protein contain essential proteins as well. Below, we show that for a fixed neighborhood size S and a fixed essential protein, the GTOM2 neighborhood contains a higher proportion of essential proteins than the corresponding GTOM0 or GTOM1 neighborhoods. This provides indirect empirical evidence that the proposed higher order GTOM measure (m = 2) outperforms the standard GTOM measure (m = 1) in this application.

Specifically, we considered the neighborhoods around each of the 30 essential proteins with highest nodal degrees (referred to as essential hub proteins). Since on average the GTOM0 neighborhoods of these essential hub proteins contain 40 proteins, we considered the following neighborhood sizes S = 1,..., 40. For each of the essential hub proteins, we determined the GTOMm neighborhood for m = 0,1,2,3. For a given neighborhood of size S, we determined the proportion of essential proteins. Next we averaged these proportions across the 30 essential hub proteins. Figure 1 reports the average proportion of essential proteins (y-axis) versus different neighborhood sizes (x-axis).

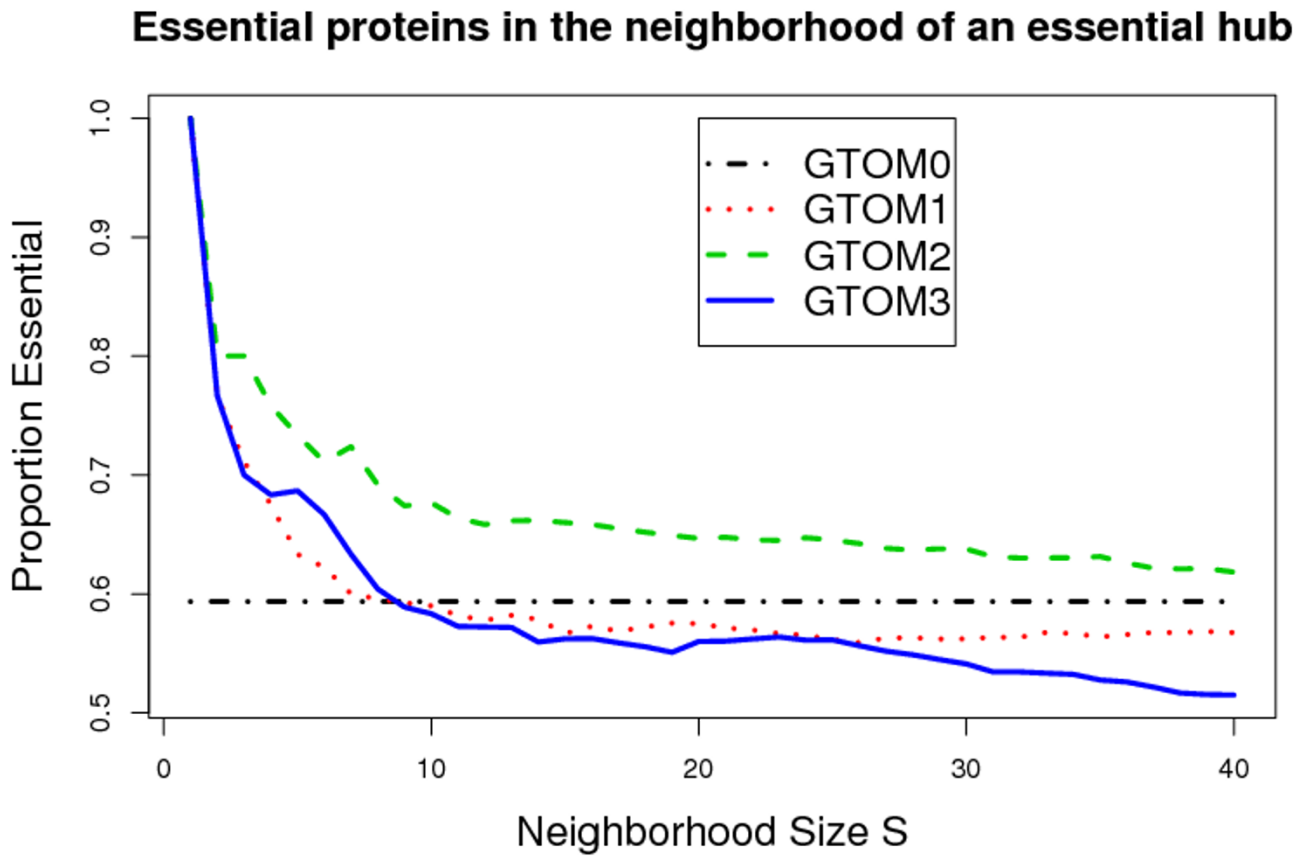

Figure 1
Proportions of essential proteins among the S proteins that are most highly interconnected with a given essential hub protein in the Drosophila protein-protein interaction network. For a given neighborhood size S (x-axis) we averaged the results over 30 essential hub proteins (y-axis). The black horizontal line (GTOM0) represents the average proportion of essential proteins among the directly linked neighbors (adjacency = 1) of an essential hub protein.

As can be seen from Figure 1, GTOM2 performs better than the other measures in this application involving a relatively sparse network. For example, when considering neighborhoods comprised of 10 proteins (rank 10) based on GTOM0, GTOM1, GTOM2, GTOM3 the proportion of essential proteins is given by 0.59, 0.59, 0.68, and 0.58, respectively. We find that neighborhood analysis with GTOM2 leads to significantly better results than GTOM0 (Wilcoxon p-value = 0.034), GTOM1 (p-value = 0.015) and GTOM3 p-value = 0.02).

Module detection in a yeast co-expression network

There is evidence that genes and their protein products carry out cellular processes in the context of functional modules [11]. Thus, an important task in biological network analysis is to identify groups, or 'modules', of densely interconnected genes. Here we focus on module identification methods that are based on using a node dissimilarity measure in conjunction with a clustering method. Further, we assume that the nodes in a network module have high topological overlap with their neighbors. A review of alternative module detection methods is beyond the scope of this article, see e.g. [12-21].

To demonstrate the usefulness of the GTOM dissimilarity measures for module detection, we apply the proposed measures to gene co-expression networks constructed based on a microarray dataset recording gene expression levels during different stages of cell cycle in yeast [22]. Because the transcriptional response of cells to changing conditions involves the coordinated co-expression of genes encoding interacting proteins, studying co-expression patterns can provide insights into the underlying cellular processes [23,24]. As detailed in the Methods section, the co-expression network was constructed by thresholding the absolute pair-wise (Pearson) correlation coefficient between the expression profiles. We cluster the genes into modules using the average linkage hierarchical clustering with different choices of dissimilarity measures. Modules correspond to branches of the resulting clustering tree. While there is evidence that this clustering procedure leads to biologically meaningful modules in several applications [1-5,25], we do not claim that this clustering method is optimal. Since our interest lies in the performance of topological overlap based dissimilarity measures but not the clustering procedure, comparing different clustering procedures is beyond the scope of this article. In our applications, modules correspond to branches of a hierarchical clustering dendrogram [6]. Figure 2 shows the modules (as branches of the dendrogram) detected by applying the average linkage hierarchical clustering with 3 different similarity measures: The adjacency matrix (GTOM0), Ravasz *et al.*'s TOM (GTOM1) and a generalized TOM (GTOM2) presented in the Results section. Genes that belong to the functional class 'protein biosynthesis' are grouped together when the GTOM2 measure is used. However, they are separated into two distinct subgroups if GTOM0 or GTOM1 are used. This suggests that GTOM2 is a more sensitive measure for detecting the higher order connections between the nodes in this large module. Thus membership in the protein biosynthesis module is more robust when neighborhoods of step size 2 is used for measuring topological overlap.

For the sake of brevity, we only present our analysis of the protein biosynthesis module in this methodological paper. Since the protein biosynthesis pathway is relatively large, it makes sense to use a relatively sensitive dissimilarity measure (GTOM2) since it favors the discovery of large modules. However, when considering a functional category that involves few genes, it would be better to use a dissimilarity measure with higher specificity (GTOM0 or GTOM1) since it favors the discovery of smaller modules. A more detailed biological analysis of a related yeast co-expression network can be found in [3].

Comparing GTOM to the correlation coefficient

Since the absolute value of the Pearson correlation coefficient is widely used for clustering gene expression profiles, we compare it here to the GTOM measures. Specifically, we consider the following class of correlation based dissimilarities:

$$d_{ij}^{C,[p]} = 1 - |\rho_{ij}|^p . \qquad (5)$$

Here, ρ_{ij} is the (Pearson) correlation coefficient between the expression profiles of i and j. Setting $p = 1$ yields the absolute correlation coefficient which is widely used for clustering genes. Setting $p > 1$ has the effect of emphasizing larger values of $|\rho_{ij}|$ while deemphasizing the smaller ones. We consider $p = 6$ since we find that the resulting distance is highly related to GTOM1 in the yeast dataset. Such a setting has also been used in [26] for functional annotation.

Figure 3 shows the relationship between six dissimilarity measures, GTOM-based $d^{T,[m]}$ for $m = 0,1,2,3$ and correlation-based $d^{C,[p]}$ for $p = 1,6$. For the yeast co-expression network, we arrive at the following results. First, $d^{C,[6]}$ is highly correlated (> 0.8) with the lower-order GTOM dissimilarities, $d^{T,[0]}$ and $d^{T,[1]}$. The correlation-based measure $d^{C,[1]}$ is highly correlated (0.79) with $d^{T,[2]}$. Second, the higher-order GTOM dissimilarities $d^{T,[2]}$ and $d^{T,[3]}$ show a high correlation of 0.78. Third, two GTOM-based dissimilarities are moderately correlated (< 0.4) if their orders differ by 2 or more. Finally, the frequency distribution of $d^{T,[3]}$ is concentrated around 0 while that of the others are concentrated around 1. This illustrates that increasing m leads to increased sensitivity but decreased specificity when defining interconnectedness.

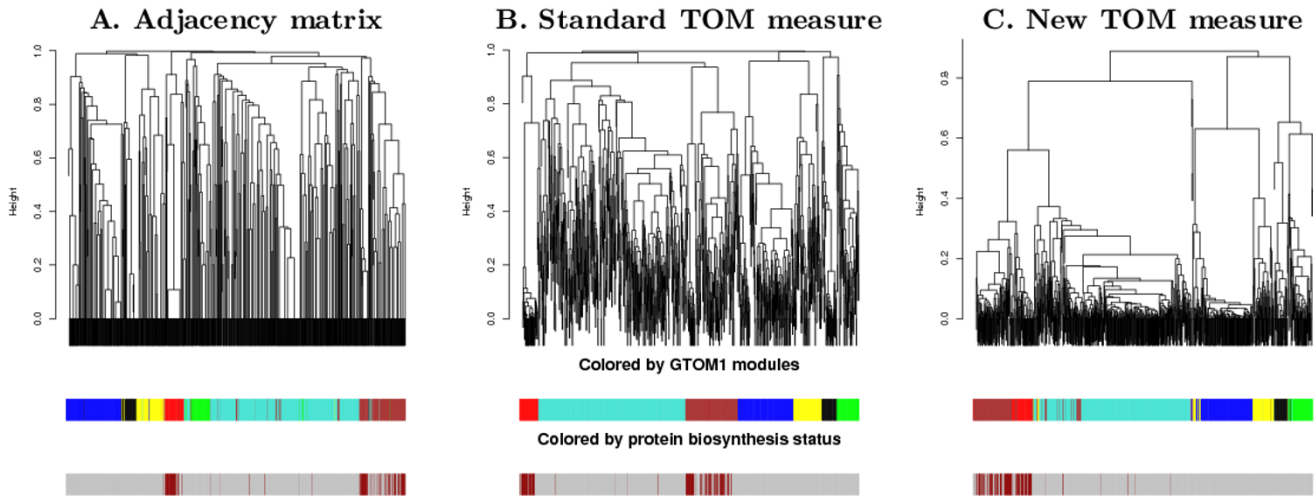

Figure 2
Yeast network modules and protein biosynthesis genes for different GTOM*m*. **A**. The adjacency matrix (GTOM0). **B**. Standard Ravasz *et al*.'s TOM (GTOM1). **C**. Our new generalized TOM (GTOM2). In each column, the top row shows the dendrogram obtained by applying the average linkage hierarchical clustering to the corresponding GTOM dissimilarity, the middle row shows the color bar ordered by the corresponding dendrogram but colored by the module assignment with respect to the TOM measure in **B**, the bottom shows the color bar ordered by the corresponding dendrogram but colored in dark red if the gene belongs to the class 'protein biosynthesis'. The modules defined by the TOM are the branches of the dendrogram in **B** at the cutoff 0.95. Almost all protein biosynthesis genes are grouped together by the proposed new TOM measure whereas the other two measures tend to distribute the class over two modules. The modules defined by GTOM2 are more pronounced in the sense that they are separated by larger distances.

Hierarchical clustering and GTOM plots

In networks involving few nodes, modules can easily be identified by inspecting the network but for large networks involving hundreds of nodes, it is useful to provide a 'reduced' view of the network. For example, one can visualize the topological overlap dissimilarity using classical multi-dimensional scaling plots [27], see the Multi-dimensional Scaling Plots section. Alternatively, it can be useful to visualize the topological overlap dissimilarity matrix $[\, d_{ij}^{T,[m]} \,]$ directly using a TOM plot. As an example, consider the four GTOM plots corresponding to the zeroth- to third-order GTOM in Figure 4. The dataset used here is the same as the one in Figure 2. Red/yellow indicate low/high values of $d_{ij}^{T,[m]}$. Both rows and columns of $d_{ij}^{T,[m]}$ have been sorted using the hierarchical clustering tree. Since $d_{ij}^{T,[m]}$ is symmetric, the GTOM plot is also symmetric around the diagonal. Since modules are sets of nodes with high (generalized) topological overlap, modules correspond to red squares along the diagonal.

Figure 4 shows that modules are more pronounced and larger with increasing values of m. This illustrates that higher values of m increase the sensitivity of measuring interconnectedness at the expense of specificity. This is further discussed in the section on the asymptotic behavior of GTOM below. For comparison purposes, a color bar is shown on the top on each GTOM plot. The color bar is ordered by the respective dendrogram and colored by the GTOM1 module assignment (c.f. Figure 2B.) The fact that the module colors stay together for different choices of m provides evidence that the module assignment is fairly robust with respect to the dissimilarity measure. One advantage of our proposed general class of dissimilarity measures is that they allow one to verify that module assignment is robust with respect to different network dissimilarities. If there is a strong biological signal, one would hope that the results are robust with respect to different choices of statistical methods.

But a more subtle analysis provides indirect empirical evidence of the usefulness of GTOM2 for module definition. Note that a second color bar is included on the left of the heatmap. Here, dark red indicates the membership to the class 'protein biosynthesis'. Genes that belong to other classes (or are unknown) are depicted by a gray color in the bar. We observe that protein biosynthesis genes are grouped together in the GTOM2 and GTOM3 plots.

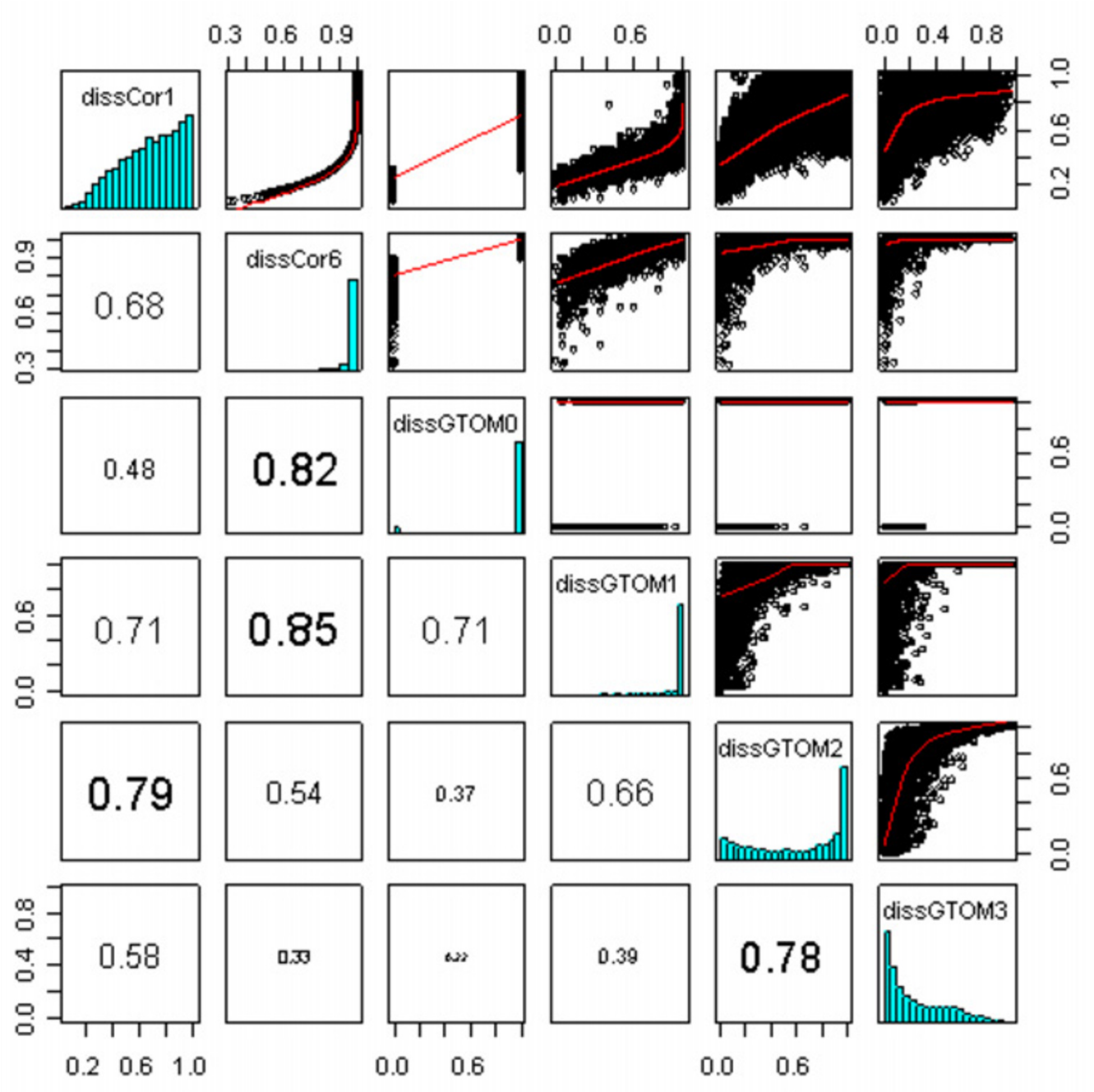

Figure 3
Pair-wise scatter plots between different GTOM*m* dissimilarity measures. The upper triangular panel shows the scatter plots, the lower triangular panel shows the corresponding Pearson correlation coefficients, the diagonal panel shows the frequency distributions of the dissimilarities. Correlation-based dissimilarities $d^{C,[p]}$ are denoted by dissCor*p*. GTOM-based dissimilarities $d^{T,[m]}$ are denoted by dissGTOM*m*. Note that dissGTOM0 (= 1 - ADJ) takes on binary values for the unweighted network.

Multi-dimensional scaling plots

We visualize the dissimilarity measures using classical multi-dimensional scaling (MDS) plots. Classical multidimensional scaling takes as input matrix a dissimilarity matrix (here the GTOM dissimilarity). The result of multi-dimensional scaling are vectors in a low dimensional

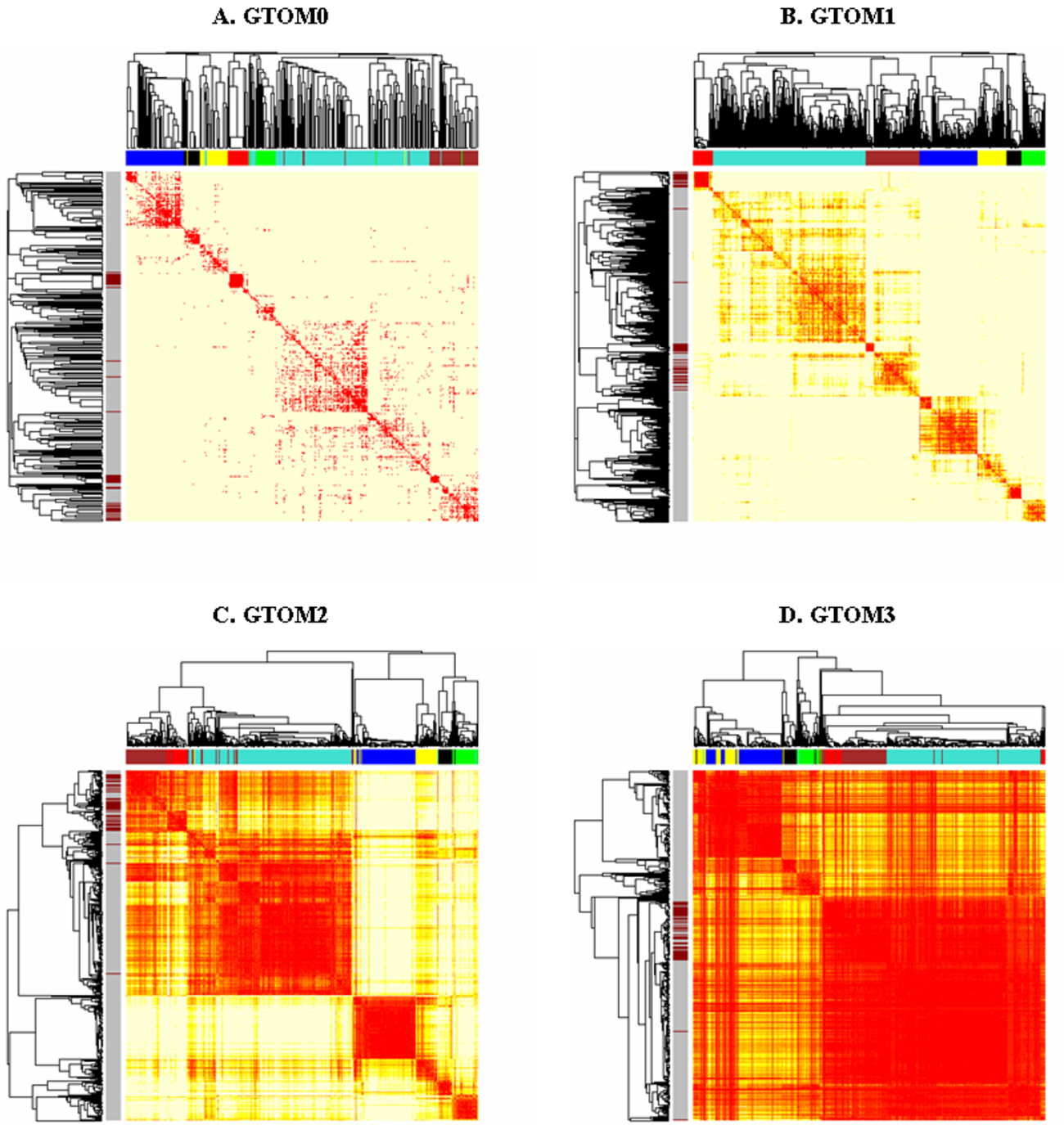

Figure 4
Topological overlap matrix plots for the yeast gene co-expression network. **A**. GTOM0 plot. **B**. GTOM1 plot. **C**. GTOM2 plot. **D**. GTOM3 plot. The color bar on the top of each heatmap shows the module assignment obtained from GTOM1. The color bar on the left of each heatmap shows the functional category of the corresponding genes. Dark red indicates the membership to the class 'protein biosynthesis'. Modules are more pronounced in the GTOM2 and GTOM3 plots (larger contrast between the diagonal blocks and off-diagonal blocks). Smaller modules (as diagonal blocks of red) are more visible in GTOM0 and GTOM1 plots whereas larger modules are more respected in GTOM2 and GTOM3 plots. However, GTOM3 leads to excessively large modules and thus the specificity of the modules is compromised. Protein biosynthesis genes are grouped together in the GTOM2 and GTOM3 plots.

Euclidean space (here the 2 dimensional Euclidean plane) such that the Euclidean distances between the vectors approximate the dissimilarities. To compute these vectors, an eigenvector problem is solved to find the locations that minimize distortions to the dissimilarity matrix [27].

The MDS plots are shown in Figure 5. All the plots are color-coded according to the modules with respect to GTOM1 depicted in Figure 2. The relative position of the points are well-preserved as we can see that points having the same color are almost always clustered together. Genes that belong to the class 'protein biosynthesis' are depicted by the symbol '▲'. Other genes are denoted by a '○'. Interestingly, almost all 'protein biosynthesis' genes are in the vicinity of each other. The plot using the GTOM1 dissimilarity in Figure 5B shows a more clear separation between the red and brown modules.

We observe from the MDS plots that there is a tendency of consolidation as the order m of the GTOM measure increases. This phenomenon can be seen in Figure 5D where GTOM3 is used and a few 'sinks' (points of attraction) have been formed.

Robustness of the results to network perturbations

To demonstrate that GTOM2 may outperform GTOM1 and GTOM0 in the case of noisy network data, we randomly removed connections in the yeast gene co-expression network. Specifically, we randomly set a proportion p = 0, 0.1, 0.25, 0.33, 0.5, 0.67, 0.75, 0.9 of entries in the adjacency matrix to 0. We used the perturbed network to compute the corresponding GTOMm measures (for m = 0,1,2,3). To quantify the ability of GTOMm to separate protein biosynthesis genes ($\mathcal{B}$) from non-protein biosynthesis genes ($\mathcal{NB}$) we defined a measure of separation, which is motivated by the intergroup dissimilarity measures used in average linkage hierarchical clustering. Specifically, we define

$$\text{GTOMdiff}(m) = a_{\mathcal{B}} - a_{\mathcal{NB}} \qquad (6)$$

where

$$a_{\mathcal{B}} = \frac{\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{B}} t_{ij}^{[m]}}{|\mathcal{B}|^2}$$

is the average GTOMm among protein biosynthesis genes and

$$a_{\mathcal{NB}} = \frac{\sum_{i\in\mathcal{B}}\sum_{j\in\mathcal{NB}} t_{ij}^{[m]}}{|\mathcal{B}||\mathcal{NB}|}$$

is the average GTOMm between protein biosynthesis and non-protein biosynthesis genes. Here, $|\mathcal{B}|$ and $|\mathcal{NB}|$ denote the total number of protein biosynthesis genes and non-protein biosynthesis genes respectively. The higher the value of GTOMdiff(m), the better is the performance of GTOMm in this application. For each probability p, we averaged the results across 20 perturbed versions of the network. The results in Figure 6 demonstrate that high values of m counter the effect of misspecified (missing) adjacencies in this application.

A simple example

An example comparing modules detected by the GTOM1 and GTOM2 similarities is given in Figure 7. As a rule of thumb, if many of the nodes in a module are separated by a distance of 1 or 2 from each other, then they form a tight module with respect to the GTOM1 similarity. Likewise, if many of the nodes in a module are separated by a distance of 3 or 4 from each other, then they form a tight module with respect to the GTOM2 similarity.

*The asymptotic behavior of GTOM*m *for large* m

Here we consider the situation when m is larger than or equal to the network diameter, i.e. each pair of nodes can be connected with a path of length $\leq m$. In this case, $|N_m(i)| = n - 1$ and $|N_m(i) \cap N_m(j)| = n - 2$ where n denotes the network size. Then

$$t_{ij}^{[m]} = \frac{n - 2 + a_{ij}}{n - a_{ij}} \approx 1,$$

when n is large. This demonstrates that for sufficiently large values of m all pairs of nodes within a connected network component will be highly interconnected. Thus, large and tight modules result when GTOMm with large m is used as input of a clustering procedure, see Figure 4.

Choosing the order m

How to choose the order m is an important question in many applications. While it seems intuitive that the choice of m has some relationship to the network diameter, it is unclear to us how to use the network diameter to guide the choice of m (other than providing an upper bound).

In general, the optimal choice of m will depend on the data quality and the goals of the analysis. Roughly speaking, if the adjacency matrix contains very few errors and if the goal is to determine which nodes are linked to a given node then m = 0 is the obvious choice. But if many adjacencies have falsely been set to 0 (since the corresponding connections are unknown) and/or if the goal is to detect possibly longer ranging interactions then relatively large

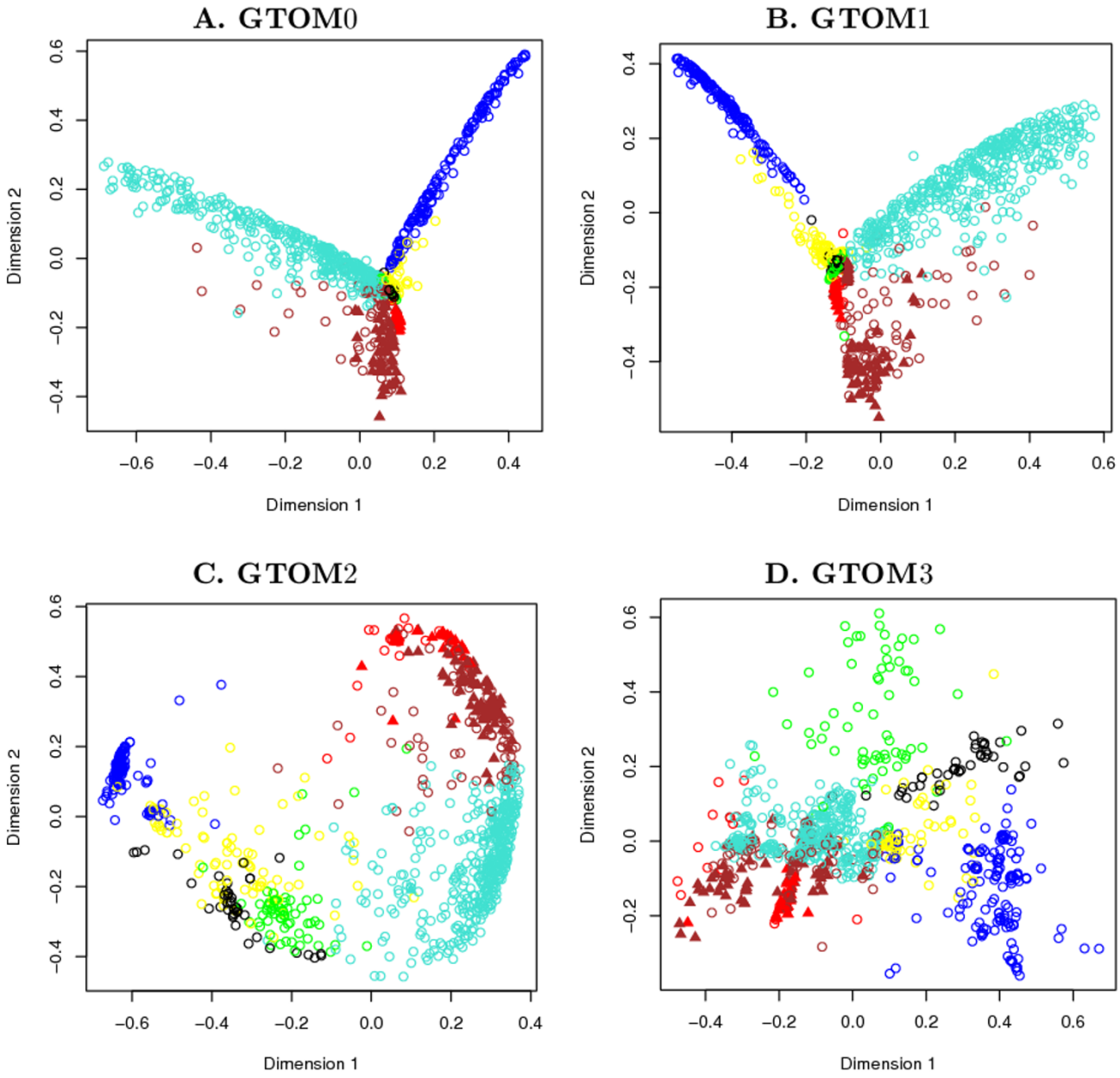

Figure 5
Multi-dimensional scaling plots of the yeast gene co-expression network. MDS plots using **A**. GTOM0, **B**. GTOM1, **C**. GTOM2, and **D**. GTOM3. The coloring scheme is used to reflect the 7 modules shown in Figure 2**B** detected by using hierarchical clustering with the GTOM1-based dissimilarity. The symbol '▲' denotes genes that belong to the functional category 'protein biosynthesis'. Genes that belong to other classes are denoted by a '○'. In general, the module assignment is preserved across the different GTOM measures. But the spatial distributions of the points vary to a large extent. Genes in the 'protein biosynthesis' class appear to be closer together.

values of m may uncover that 2 nodes are interconnected even if the corresponding adjacency is 0.

When an external label is available for at least some of the nodes then one can use it to inform the choice of m. For example, when the external node label y encodes group membership, then one can choose m so that the groups have high within group interconnectedness and low between group interconnectedness. To make this more specific, we assume that there is evidence that the nodes of group 1 are highly interconnected and that they are well separated from nodes of group 0. For example, in our

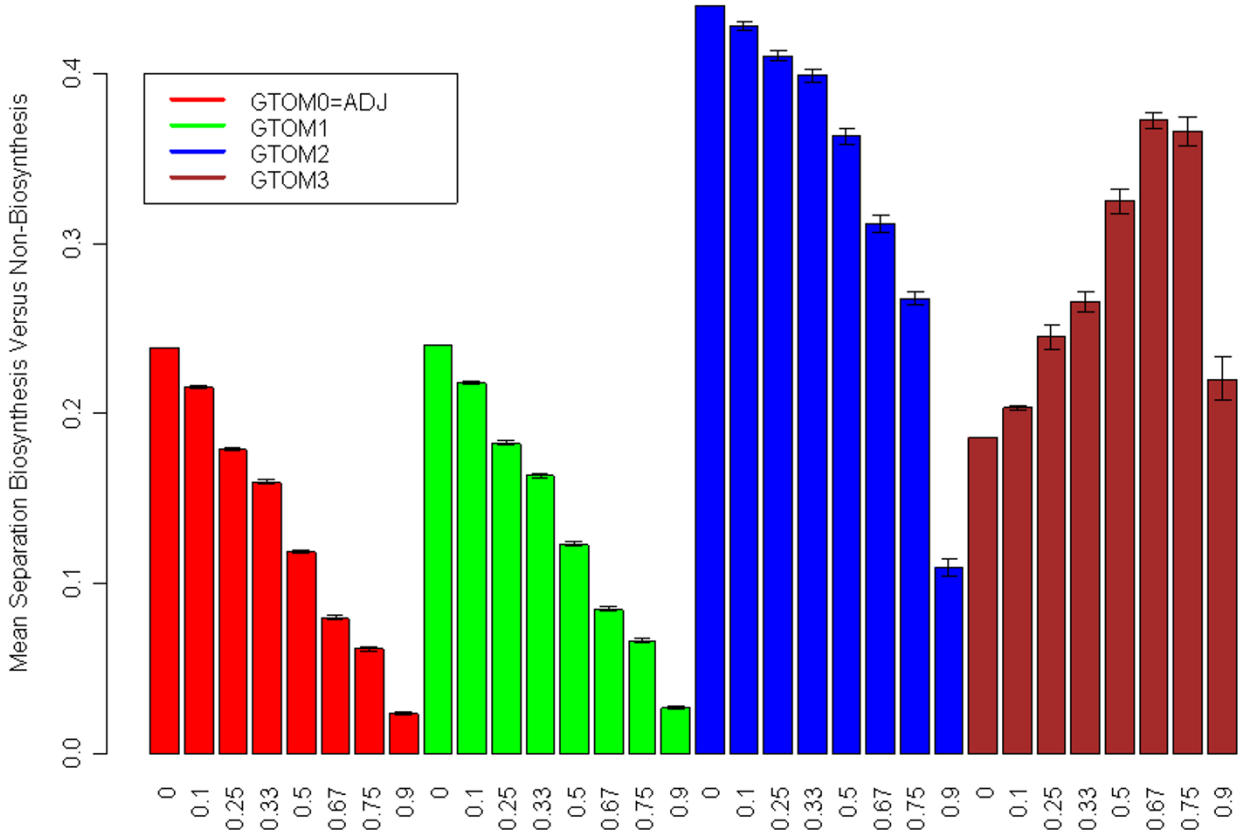

Figure 6
Separation of protein biosynthesis genes from non-protein biosynthesis genes in perturbed versions of the yeast network. The average separation (c.f. Eq. 6) is reported for GTOM0 (red), GTOM1 (green), GTOM2 (blue) and GTOM3 (brown). To assess the robustness of the GTOM measures to random deletions, we randomly deleted a proportion p of connections (adjacencies) and averaged the results across 20 draws. Note that GTOM2 outperforms the other measures if $p < 67\%$. GTOM3 outperforms GTOM2 if more than 67% of adjacencies are deleted. This illustrates that high values of m can counter the effect of misspecified (unknown or missing) adjacencies.

yeast gene co-expression network, we have shown above that the average GTOMm measure between protein biosynthesis genes (group 1) is larger than the average GTOMm measure between protein biosynthesis genes and non-protein biosynthesis genes (Figure 6).

Analogous to Eq. 6, we define the following measure of mean GTOM difference between the 2 groups

$$\mathrm{GTOMdiff}(m) := \frac{\sum_{i\neq j} t_{ij}^{[m]} I(\gamma_i = 1, \gamma_j = 1)}{\sum_{i\neq j} I(\gamma_i = 1, \gamma_j = 1)} - \frac{\sum_{i\neq j} t_{ij}^{[m]} I(\gamma_i = 1, \gamma_j = 0)}{\sum_{i\neq j} I(\gamma_i = 1, \gamma_j = 0)}$$

where the indicator function $I(\cdot)$ equals 1 if the condition is satisfied and 0 otherwise. Note that $\frac{\sum_{i\neq j} t_{ij}^{[m]} I(\gamma_i = 1, \gamma_j = 1)}{\sum_{i\neq j} I(\gamma_i = 1, \gamma_j = 1)}$ equals the mean interconnectedness of group 1 nodes and $\frac{\sum_{i\neq j} t_{ij}^{[m]} I(\gamma_i = 1, \gamma_j = 0)}{\sum_{i\neq j} I(\gamma_i = 1, \gamma_j = 0)}$ equals the mean interconnectedness between group 1 and group 0 nodes. Since high values of GTOMdiff(m) indicate a good separation between the 2 groups, it is natural to choose m as the value that maximizes GTOMdiff(m). Obviously, this criterion for choosing m only works if prior data allow one to define the external label γ.

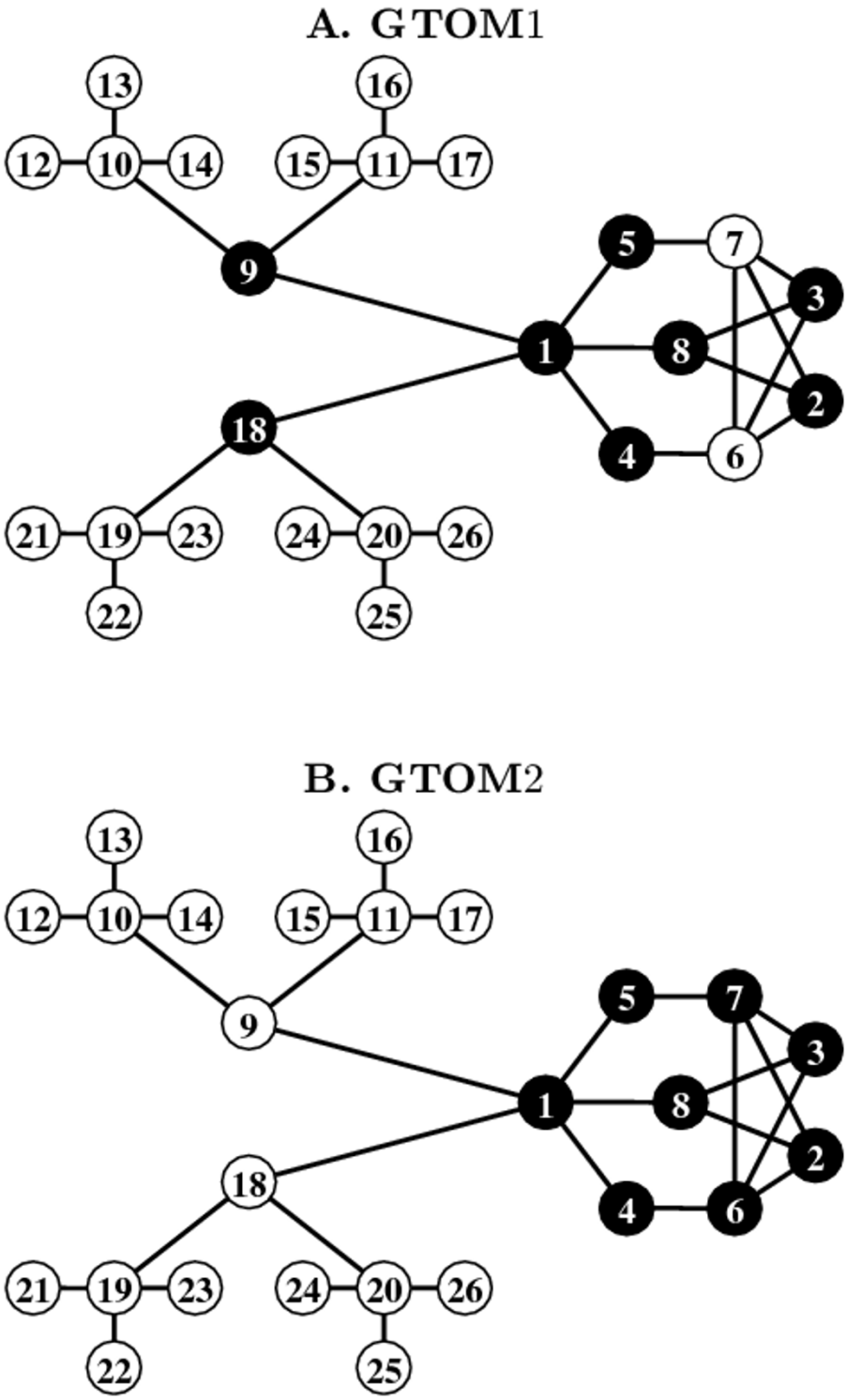

Figure 7
A simple example where GTOM2 is superior to GTOM1. GTOM neighborhood of size *S* = 7 around node 1. **A**. GTOM1 neighbors are colored in black. **B**. GTOM2 neighbors are colored in black. Note that GTOM2 detects the 'true' neighborhood (comprised on nodes 1 through 8) while GTOM1 misses nodes 6 and 7.

Discussion

Several measures that keep track of shared 1 step neighbors have been proposed in the literature, e.g. [28]. Here, we propose a natural generalization of the widely used topological overlap matrix. This class of new measures is constructed by keeping track of the number of *m*-step neighbors that a pair of nodes share. The GTOM similarity measure is normalized to take on values in the unit interval. A corresponding dissimilarity measure can be defined by subtracting the GTOM similarity from 1.

While we find it a worthwhile goal for future research to develop statistical or heuristic criteria for choosing *m*, we find that a main advantage of GTOM*m* is that it allows one to assess the robustness of network analysis results. In many applications (e.g. module definition, neighborhood analysis), it will be worthwhile to show that the results are relatively robust with respect to *m* since this indicates that the biological signal is strong. While we present applications where non-standard choices of *m* lead to superior results, we have found in several other (unreported) applications that the results are robust with respect to *m* = 0,1, 2. By randomly deleting network adjacencies in the yeast gene co-expression network application, we have shown that large values of *m* can counter the effect of misspecified (missing) adjacencies. GTOM*m* becomes uninformative if *m* is larger than the network diameter. Thus, GTOM*m* will be useful in networks with moderate or large 'degree of separation' (average path length between any pair of nodes). Since biological networks tend to have low diameters [29], we expect that low values of *m* will be preferable in most applications. But we have provided two real data applications where *m* = 2 is preferable over *m* = 1. In general, the GTOM measures with lower orders *m* will be useful for discovering small modules while those with higher orders favor the discovery of larger modules.

A limitation of our approach is that it is only defined for unweighted networks, i.e. the entries of the adjacency matrix should be 0 or 1. Another limitation is that we only consider the topological overlap between 2 nodes. A multi-node extension of the GTOM1 measure is presented in [30].

Conclusion

The generalized topological overlap measure can serve as a filter for countering the effect of spurious or missing connections. The order *m* of the topological overlap measure can serve as a tuning parameter for interconnectedness that trades off sensitivity versus specificity. Since different orders of *m* probe different neighborhoods, adjusting *m* allows the user to consider network modules at different 'zoom' levels. We provide additional Materials and Methods as well as the statistical software code, a tutorial along with customized R functions, and the accompanying data files at the web page [31]. Thus, the reader should be able to reproduce all of our findings.

Methods

Topological overlap matrix

The topological overlap of two nodes reflects their similarity in terms of the commonality of the nodes they connect to. Note that in an unweighted network, the number of shared neighbors of nodes i and j is given by $\sum_{u\neq i,j} a_{iu} a_{uj}$. Ravasz *et al*. [1] define the topological overlap measure t_{ij} as follows

$$t_{ij} = \begin{cases} \dfrac{l_{ij} + a_{ij}}{\min\{k_i, k_j\} + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases} \qquad (7)$$

where $l_{ij} = \sum_{u\neq i,j} a_{iu} a_{uj}$, $k_i = \sum_{u\neq i} a_{iu}$. An advantage of the quantity 1 - a_{ij} in the denominator is that it prevents the denominator from becoming 0 when the connectivities (degrees) k_i and k_j are 0. Since $a_{ij} \leq 1$, one can easily show that $l_{ij} = \sum_{u\neq i,j} a_{iu} a_{uj} \leq \sum_{u\neq i} a_{iu} - a_{ij} = k_i - a_{ij}$. It follows that $l_{ij} \leq \min(k_i - a_{ij}, k_j - a_{ij})$ and that the numerator of t_{ij} is smaller than the denominator, i.e. $0 \leq t_{ij} \leq 1$.

We remark that the definition of TOM given in [1] is slightly different from Eq. (7):

$(l_{ij} + a_{ij})/\min\{k_i, k_j\}$. In a personal communication with E. Ravasz, the definition in Eq. (7) is preferred, which is also given in the online supporting material of [1]. The inclusion of the term a_{ij} in the numerator makes t_{ij} explicitly depend on the existence of a direct link between the two nodes in question.

An algorithm for computing GTOM

In this subsection, we present computational formulas for $T^{[m]}$. In this subsection, we assume that the diagonal of A has been set to 0. Then the ij-th entry of the matrix power A^m counts the number of paths of length m connecting nodes i and j [32]. But the paths are not necessarily geodesic and may contain cycles. Then the matrix $S^{[m]} \equiv [\, s_{ij}^{[m]} \,] := A + A^2 + \ldots + A^m$ counts how many distinct paths of length smaller than or equal to m connect each pair of nodes. Thus, we have $N_m(i) = \{j \neq i \mid s_{ij}^{[m]} > 0\}$. If we define a binary matrix $B^{[m]}$ to be

$$b_{ij}^{[m]} = \begin{cases} 1 & \text{if } s_{ij}^{[m]} > 0 \text{ and } i \neq j, \\ 0 & \text{otherwise,} \end{cases}$$

then $N_m(i) \equiv \{j \neq i | b_{ij}^{[m]} = 1\}$. To obtain the number of shared m-step neighbors, $|N_m(i) \cap N_m(j)|$, we simply take the inner product of the i-th and the j-th columns of $B^{[m]}$ which can be obtained from the matrix $(B^{[m]})^2 = [|N_m(i) \cap N_m(j)|]$ because of the symmetry of $B^{[m]}$. In particular, $|N_m(i)|$ is given by the i-th diagonal entry of $(B^{[m]})^2$. These values can then be used to compute $T^{[m]}$ using formula (3). It is worth repeating that the formulas in this subsection assume that the diagonal of the adjacency matrix is 0. Since matrix multiplication is computationally expensive, the computation of $S^{[m]}$ may be sped up using the formula $A(S^{[m-1]} + I)$.

Using hierarchical clustering for module detection

By using the topological overlap measure as an input of the average linkage hierarchical clustering procedure, we define modules as discrete branches of the clustering tree (e.g. Figure 2). As in all hierarchical clustering analysis, it is a judgement call where to cut the tree branches. When detecting modules using hierarchical clustering, we use GTOM plots to aid the choice of the dendrogram's height cutoff (see the Results section). Thus the modules are found by inspection: a height cutoff value is chosen in the dendrogram such that some of the resulting branches correspond to the discrete diagonal blocks (modules) in the GTOM plot. The robustness of the module definition with respect to the height cut-off can be explored using our online R software tutorial.

Yeast gene co-expression network construction

Two genes in our co-expression network are linked if they are highly correlated across the samples. To construct the gene co-expression networks from the microarray data [22], we first select the 4000 yeast genes having the highest variance across the microarray samples. Then we calculated all possible pairwise Pearson correlations for the 4000 genes across the microarrays. Because microarray data can be noisy and the number of samples is often small, absolute values of the correlations were thresholded using a relatively large hard threshold of $\tau = 0.7$. This threshold corresponds to a significance level of $p = 8.7 \times 10^{-8}$ (Fisher's correlation test) and leads to an *approximate* scale-free topology as described in [33]. Such a topology implies the existence of 'hub genes' [34] and the robustness to random perturbations [35] which are biologically desirable properties. The topology of the yeast network data is further discussed in [3].

Authors' contributions

Both authors jointly developed the methods, implemented them, and wrote the article.

Acknowledgements

The authors would like to thank Marc Carlson, Jun Dong, Dan Geschwind, Peter Langfelder, Ai Li, Paul Mischel, Stan Nelson, Mike Oldham, Anja Presson for helpful comments. The work was supported in parts by grant 1U19AI063603 01.

References

1. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL: **Hierarchical organization of modularity in metabolic networks.** *Science* 2002, **297(5586):**1551-1555.
2. Ye Y, Godzik A: **Comparative Analysis of Protein Domain Organization.** *Genome Biology* 2004, **14(3):**343-353.
3. Carlson MR, Zhang B, Fang Z, Horvath S, Mishel PS, Nelson SF: **Gene Connectivity, Function, and Sequence Conservation: Predictions from Modular Yeast Co-expression Networks.** *BMC Genomics* 2006, **7(40):**.
4. Horvath S, Zhang B, Carlson M, Lu K, Zhu S, Felciano R, Laurance M, Zhao W, Shu Q, Lee Y, Scheck A, Liau L, Wu H, Geschwind D, Febbo P, Kornblum H, Cloughesy T, Nelson S, Mischel P: **Analysis of Oncogenic Signaling Networks in Glioblastoma Identifies ASPM as a Novel Molecular Target.** *Proc Natl Acad Sci USA* 2006, **103(46):**17402-17407.
5. Oldham MC, Horvath S, Geschwind DH: **Conservation and evolution of gene coexpression networks in human and chimpanzee brains.** *Proc Natl Acad Sci U S A* 2006, **103(47):**17973-17978.
6. Kaufman L, Rousseeuw P: *Finding Groups in Data: An Introduction to Cluster Analysis* New York: John Wiley & Sons, Inc; 1990.
7. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411(6833):**41-42.
8. Jeong H, Oltvai Z, Barabási A: **Prediction of Protein Essetiality Based on Genome Data.** *ComPlexUs* 2003, **1:**19-28.
9. Hahn MW, Kern AD: **Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks.** *Molecular Biology and Evolution* 2005, **22(4):**803-806.
10. Stark C, Breitkreutz B, Reguly T, Boucher L, Breitkreutz A, Tyers M: **BioGRID: A General Repository for Interaction Datasets.** *Nucleic Acids Res* 2006, **34:**D535-9.
11. Hartwell L, Hopefield JSL, Murray A: **From Molecular to Modular Cell Biology.** *Nature* 1999, **402(6761 Suppl):**C47-52.
12. BarJoseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21(11):**1337-42.
13. Isaacs FJ, Hasty J, Cantor CR, Collins JJ: **Prediction and measurement of an autoregulatory genetic module.** *Proc Natl Acad Sci U S A* 2003, **100(13):**7714-9.
14. Lubovac Z, Olsson B, Gamalielsson J: **Combining topological characteristics and domain knowledge reveals functional modules in protein interaction networks.** In *Proc CompBioNets* Lyon, France: College Publications; 2005:93-106.
15. Prinz S, Avila-Campillo I, Aldridge C, Srinivasan A, Dimitrov K, Siegel AF, Galitski T: **Control of yeast filamentous-form growth by modules in an integrated molecular network.** *Genome Res* 2004, **14(3):**380-90.
16. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34(2):**166-76.
17. Thimm O, Blasing O, Gibon Y, Nagel A, Meyer S, Kruger P, Selbig J, Muller LA, Rhee SY, Stitt M: **MAPMAN: A user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes.** *Plant J* 2004, **37(6):**914-39.
18. Tornow S, Mewes HW: **Functional modules by relating protein interaction networks and gene expression.** *Nucleic Acids Res* 2003, **31(21):**6283-9.
19. Toyoda T, Konagaya A: **KnowledgeEditor: a new tool for interactive modeling and analyzing biological pathways based on microarray data.** *Bioinformatics* 2003, **19(3):**433-4.
20. Xu X, Wang L, Ding D: **Learning module networks from genome-wide location and expression data.** *FEBS Lett* 2004, **578(3):**297-304.
21. Newman M, Girvan M: **Finding and Evaluating Community Structure in Networks.** *Physical Review E* 2004, **69(2):**026113.
22. Spellman P, Sherlock G, Zhang M, Iyer V, Anders K: **Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast**

Saccharomyces Cerevisiae by Microarray Hybridization. *Mol Biol Cell* 1998, **9(12):**3273-3297.
23. Eisen M, Spellman P, Brown P, Botstein D: **Cluster Analysis and Display of Genome-wide Expression Patterns.** *Proc Natl Acad Sci USA* 1998, **95(25):**14863-14868.
24. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci U S A* 1999, **96:**2907-2912.
25. Ghazalpour A, Doss S, Zhang B, Plaisier C, Wang S, Schadt E, Thomas A, Drake T, Lusis A, Horvath S: **Integrating Genetics and Network Analysis to Characterize Genes Related to Mouse Weight.** *PloS Genetics* 2006, **2(8):**el30.
26. Zhou X, Kao MC, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data.** *Proc Natl Acad Sci U S A* 2002, **99(20):**12783-12788.
27. Cox T, Cox M: *Multidimensional Scaling* Boca Raton: Chapman and Hall/CR,C; 2001.
28. Goldberg DS, Roth FP: **Assessing experimentally derived interactions in a small world.** *Proc Natl Acad Sci U S A* 2003, **100:**4372-4376.
29. Newman M: **The Structure and Function of Complex Networks.** *SIAM Review* 2003, **45(2):**167-256.
30. Li A, Horvath S: **Network neighborhood analysis with the multi-node topological overlap measure.** *Bioinformatics* 2007, **23(2):**222-231.
31. **Weighted Gene Co-expression Network Page** [http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork/]
32. Wasserman S, Faust K: *Social Network Analysis: Methods and Applications. Structural Analysis in the Social Science* New York: Cambridge University Press; 1994.
33. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-expression Network Analysis.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4:**Article 17.
34. Barabási A, Albert R: **Emergence of Scaling in Random Networks.** *Science* 1999, **286:**509-512.
35. Albert R, Barabási A: **Error and Attack Tolerance of Complex Networks.** *Nature* 2000, **406(6794):**378-382.