The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks

Andy M. Yip  
Department of Mathematics  
National University of Singapore  
2, Science Drive 2  
Singapore 117543, Singapore  
Email: matymha@nus.edu.sg

Steve Horvath  
Department of Human Genetics and  
Department of Biostatistics  
University of California  
Los Angeles, CA 90095  
Email: shorvath@mednet.ucla.edu

Abstract— Systems biologic studies of gene and protein interaction networks have found that these networks are comprised of ‘modules’ (groups of tightly interconnected nodes). Module identification is an essential step towards understanding the whole network architecture. Here we will focus on module identification methods that are based on using a node dissimilarity measure in conjunction with a clustering method. More specifically, we introduce a general class of node dissimilarity measures based on the notion of ‘topological’ overlap, which has been found to be biologically meaningful in several applications.

The resulting generalized topological overlap measure (GTOM) generalizes the standard topological overlap measure (TOM) introduced by Ravasz et al. [1]. Specifically, the $m$-th order version of this family 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 main use of the GTOM measures is the identification of network modules (sets of tightly connected nodes). We discuss the properties of the GTOM measures and provide empirical evidence that they are useful in the context of gene co-expression network analysis. We show that the original TOM proposed by Ravasz et al. [1] favors the discovery of smaller modules whereas the higher-order GTOM's favor larger modules. Moreover, in our examples on yeast co-expression networks, we show that genes essential for the yeast's survival tend to have high GTOM similarity with each other.

I. INTRODUCTION

Genes and their protein products carry out cellular processes in the context of functional modules [2]. Thus it is natural to ask whether such modular organization can be revealed by gene- or protein interaction networks. Multiple approaches for defining gene modules have been proposed in the literature, e.g. [3]–[11]. Recent efforts devoted to module detection include a repeated edge-removal algorithm by Newman and Girvan [12] where at each step the edge with the highest centrality is removed, and a hierarchical clustering method using the topological overlap matrix (TOM) as input similarities by Ravasz et al. [1]. Here we aim to generalize the topological overlap dissimilarity measure since we and others have found that it results in biologically meaningful modules [1], [13], [14]. Further we study the properties of the resulting class of node distance measures.

The Topological Overlap Measure: The approach taken by Ravasz et al. on uncovering network modules essentially boils down to the problem of measuring how close pairs of nodes are in a network. Ravasz et al. demonstrate the potential of using the set of immediate neighbors (first order interactions) of a pair of nodes as a basis for measuring pair-wise similarity. Moreover, in their study of a E. coli metabolic network, they show that two substrates having a higher overlap are more likely to belong to the same functional class than substrates having a lower topological overlap. Such an interesting finding prompts the question whether a measure of topological similarity based on higher-order interactions would be useful to define more stable and larger modules. In this paper, we generalize Ravasz et al.'s approach by incorporating information from higher-order interactions into the similarity measure.

The organization of the rest of the paper is as follows. We start off in §I-A with two examples showing the superiority of GTOM measures in relating functional classes with modules and in predicting lethal gene deletions in yeast. The notion of GTOM is formally introduced in §II. Methods for module detection, together with illustrations in yeast microarray data are presented in §III. Finally, we give a conclusion in §IV.

A. Motivational Example

An important use of DNA microarray data is to annotate genes by clustering them on the basis of their gene expression profiles across several microarrays. 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 [15], [16]. It is standard to use the (Pearson) correlation coefficient as a co-expression measure.

To group genes with coherent expression profiles into modules, one typically uses hierarchical clustering [17] coupled with an appropriate distance measure. Here our emphasis is the distance measure and not the clustering procedure. Comparing different clustering procedures is beyond the scope of this paper. Once a dendrogram is obtained from a hierarchical clustering method, we need to choose a height cutoff in order arrive at a clustering. It is a judgement call where to cut the tree branches. When apply our proposed GTOM measures for clustering, we use a GTOM plot (see §III-B) to visually
aid the choice of the cutoff. 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.

Applications to a Yeast Dataset: To demonstrate how the GTOM measures are useful and biologically meaningful, we apply the proposed measures to gene co-expression networks constructed from microarray data. A dataset recording gene expression levels during different stages of cell cycles in yeasts [18] is employed throughout the paper.

Based on the expression data, the absolute pair-wise (Pearson) correlation coefficient between the expression profiles of each pair of genes is calculated. Then, a network with each node representing one gene is constructed. An edge between two nodes is present if their absolute correlation coefficient exceeds a threshold $\tau = 0.7$. Our methods work for any threshold. A discussion for how this threshold is chosen is beyond the scope of the article. Here, we obtain the threshold $\tau$ by using the scale-free criterion proposed by Zhang and Horvath [19]. Such a topology implies the existence of ‘hub genes’ [20] and the robustness to random perturbations [21] which are biologically relevant features. The yeast network we used also possesses a scale-free topology.

Module Detection: Figure 1 shows the modules (as branches of the dendrogram) detected by applying hierarchical clustering with 3 different similarity measures: The adjacency matrix (GTOM0), Ravasz et al.’s TOM (GTOM1) and a generalized TOM (GTOM2) presented in §II. Genes belong to the functional class ‘protein biosynthesis’ are grouped in the vicinity of each other by the GTOM2 measure.

Prediction of Lethal Gene Deletions: It has been reported independently by many research groups, which used different biological networks or constructed the same type of networks in different ways, genes with high nodal degree (i.e. adjacency matrix based connectivity) are more likely to be essential, i.e. they are lethal if knocked out. To further elaborate these findings, we pick two essential genes with highest nodal degrees in the network. Then, for each of the two hub genes, we find out the 20 genes in the network that are nearest to the hub gene where the distance is measured by the GTOM1, GTOM2 and one minus the absolute correlation (between expression profiles). Finally, we compute the proportion of essential genes among the top $k$ nearest neighbors of the hub genes for $k = 1, 2, \ldots, 20$. The rationale is that if a gene shares a lot of neighboring genes with the essential hub genes, then it is more likely to be essential as well. The proportions of essential genes among the nearest neighbors of the first and the second essential hub genes are shown in Figure 2(a) and Figure 2(b) respectively.

II. GENERALIZED TOPOLOGICAL OVERLAP MATRICES

In this section, we formally introduce the notion of generalized topological overlap for measuring pair-wise similarity. We start with a network encoded by its corresponding adjacency matrix $A = [a_{ij}]$ which is a symmetric with binary entries. By convention, the diagonal elements are assumed to be zero.

The topological overlap of two nodes reflects their similarity in terms of the commonality of the nodes they connect to. Ravasz et al. [1] define the topological overlap matrix $T = [t_{ij}]$ as follows

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

where, $l_{ij} = \sum_u a_{iu}a_{uj}$, $k_i = \sum_u a_{iu}$ and the index $u$ runs across all nodes of the network. $^1$ Basically, $t_{ij}$ is an indicator for the agreement between the sets of neighboring nodes of $i$ and $j$. The inclusion of the term $a_{ij}$ in the numerator makes $t_{ij}$ explicitly depends on whether there is a direct link between the two nodes in question. The purpose of the quantity $1 - a_{ij}$ in the denominator is to avoid double-counting $i$ as a neighbor of $j$ and vice versa.

Our generalization of the TOM is motivated by the observation that formula (1) can be expressed as

$$t_{ij} = \begin{cases} \frac{|N_i \cap N_j| + a_{ij}}{\min(|N_i|, |N_j|) + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (2)$$

where $N_i(i)$ denotes the set of neighbors of excluding itself and $\cdot \cdot$ denotes the number of elements (cardinality) in its argument. The quantity $|N_i \cap N_j|$ measures the number of common neighbors that nodes $i$ and $j$ share whereas $|N_i(i)|$ gives the number of neighbors of $i$.

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 $m$ (see Figure 3), i.e.,

$$N_m(i) := \{ j \neq i \mid \text{dist}(i,j) \leq m \} \quad (3)$$

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

$$t_{ij}^{[m]} = \begin{cases} \frac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min(|N_m(i)|, |N_m(j)|) + 1 - a_{ij}} & \text{if } i \neq j \\ 1 & \text{if } i = j. \end{cases} \quad (4)$$

We call the matrix $T^{[m]} = [t_{ij}^{[m]}]$ the $m$-th order generalized topological overlap matrix (GTOM$m$). This quantity simply measures the agreement between the nodes that are reachable from $i$ and from $j$ within $m$ steps. When $m = 1$, we obtain back the original TOM in formula (1).

Since $a_{ii} = 0$ and $t_{ii}^{[m]} = 1$, we can rewrite the definition above in a more compact way:

$$t_{ij}^{[m]} = \frac{|N_m(i) \cap N_m(j)| + a_{ij} + I_{i=j}}{\min(|N_m(i)|, |N_m(j)|) + 1 - a_{ij}}. \quad (5)$$

Here, $I_{i=j}$ equals to 1 if and only if $i = j$.

It is convenient and intuitive to define as well the GTOM0 as $T^{[0]} = A + I$ which only considers the direct link between the pair of nodes in question.

An example comparing modules detected by the GTOM1 and GTOM2 similarities is given in Figures 3 and 4.

$^1$The definition given in [1] is slightly different: $(l_{ij} + a_{ij})/\min(k_i, k_j)$. In a personal communication with E. Ravasz, the definition in Eq. (1) is preferred, which is also given in the online supporting material of [1].
Fig. 1. Comparisons of 3 different similarities in capturing the functional class ‘protein biosynthesis’. (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 hierarchical clustering to the corresponding similarity matrix, 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 where genes belong to the class ‘protein biosynthesis’ are colored in dark red. 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.

Fig. 2. (a) Proportions of essential genes among the $k$ nearest neighbors of the first essential hub gene for $k = 1, 2, \ldots, 20$. (b) Proportions of essential genes among the nearest neighbors of the second essential hub gene for $k = 1, 2, \ldots, 20$. The two hub genes are chosen so that they are essential to yeast’s survival and have the highest nodal degrees in the network. We observe from the figures that genes that are close to the hub genes with respect to the GTOM dissimilarity are more likely to be essential than genes that are close with respect to the GTOM1 measure. Moreover, both GTOM1 and GTOM2 outperform the correlation based dissimilarity.

A. A Simple Method for Computing GTOM

In this subsection, we present computational formulas for $T^{[m]}$.

We first note that the matrix $A^{m}$ counts the number of paths of length $m$ connecting the pair of nodes in place [22]. Here, the paths are not necessarily geodesic and may contain cycles. Hence, the matrix $S^{[m]} := s_{ij}^{[m]} = A + A^2 + \ldots + A^m$ gives exactly how many distinct paths with length smaller than or equal to $m$ connecting each pair of nodes. Thus, we have $N_m(i) \equiv \{ j \neq i \mid s_{ij}^{[m]} > 0 \}$. 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 $N_m(i) \equiv \{ j \neq i \mid b_{ij}^{[m]} = 1 \}$. Thus, to obtain the value $|N_m(i) \cap N_m(j)|$, we simply take the inner product of the $i$-th row and the $j$-th columns of $B^{[m]}$ which can be obtained from the matrix $B^{[m]} := [b_{ij}^{[m]}] = [N_m(i) \cap N_m(j)]$ because of the symmetry of $B^{[m]}$. In particular, $|N_m(i)|$ is given by the $i$ diagonal entry of $(B^{[m]})^2$. These values can then be used to compute $T^{[m]}$ using formula (4).

In the actually implementation, one may recursively update $S^{[m]}$ by the formula $A(S^{[m-1]} + I)$.

B. Properties of GTOM

Each entry of GTOM$_m$ lies between 0 (no common neighbors, no direct link) and 1 (one set of neighbors is a subset...
of the other, has a direct link). Moreover, it increases with the number of common \( m \)-step neighbors. This can be seen by the following observations.

Suppose two nodes \( i \) and \( j \) share \( c \) number of common \( m \)-step neighbors, not counting \( i \) and \( j \), i.e., \( |N_m(i) \cap N_m(j)| = c \), c.f. Eq. (3). Without loss of generality, assume that \( |N_m(i)| \leq |N_m(j)| \). If there is a direct link between \( i \) and \( j \) (implying \( i \neq j \)), then the \( m \)-th order generalized topological overlap between \( i \) and \( j \) is given by \( t_{ij}^{[m]} = (c + 1)/|N_m(i)| \) where \( 0 \leq c \leq |N_m(i)| - 1 \). On the other hand, if there is no direct link between \( i \) and \( j \), then we have \( t_{ij}^{[m]} = (c + I_{i=j})/(|N_m(i)| + 1) \) where \( 0 \leq c \leq |N_m(i)| \). Thus the results follow.

Next, we consider the situation where every pair of nodes are reachable to each other within \( m \) steps. This requires \( m \) to be large enough and the network only has one component. In this case, we have \( |N_m(i)| = n - 1 \) and \( |N_m(i) \cap N_m(j)| = n - 2 \) for all \( i \) and \( j \) where \( n \) is the size of network. Thus,

\[
t_{ij}^{[m]} = \frac{n - 2 + a_{ij} + 2I_{i=j}}{n - a_{ij}} = 1 - \frac{2}{n - a_{ij}} + \frac{2(a_{ij} + I_{i=j})}{n - a_{ij}}.
\]

When \( n \) is large, we have \( T^{[m]} \approx (1 - 2n^{-1})I + 2n^{-1}(A + I) \), and therefore, GTOM\( m \) behaves like GTOM0 (\( \equiv A + I \)). Since most biological networks have an average path length of 2 to 6 [23], it is expected that GTOM\( m \) is useful for small values of \( m \), but may depend on the size and nature of the network. Ravasz et al. [1] provide empirical evidence that GTOM1 yields biologically meaningful results. In our motivational example, we provide evidence that GTOM2 can lead to biologically more meaningful results than GTOM1 when dealing with large modules.

III. MODULE DETECTION

A. GTOM-based Dissimilarity Measure

A rationale for considering the generalized topological overlap matrix is that nodes that are part of highly interconnected modules have high topological overlap with their neighbors. The GTOM\( m \) \( T^{[m]} \) is a similarity measure [17] since it is non-negative, symmetric and each entry increases with the relative overlapping between two corresponding sets of neighbors. To turn it into a dissimilarity measure, it is subtracted from one, i.e., the generalized topological overlap-based dissimilarity measure is defined by

\[
d_{ij}^{T,[m]} = 1 - t_{ij}^{[m]}.
\]

Next, we described two common dissimilarity measures which are used for comparison with our proposed measures. We have found that a power of the correlation coefficient can sometimes approximate the TOM measure quite well, which is why we consider it here. Specifically, we consider the following class of correlation based dissimilarities:

\[
d_{ij}^{C,[p]} = 1 - |\rho_{ij}|^p.
\]

Here, \( \rho_{ij} \) is the (Pearson) correlation coefficient between the expression profiles of \( i \) and \( j \). Setting \( p = 1 \) yields the absolute correlation coefficient which is in standard use for clustering genes. We also consider \( p = 6 \) since we find that the resulting distance is highly related to GTOM1 in the yeast dataset.

Our next example shows the relationship between these dissimilarities.

1) Comparing the Dissimilarity Measures: Figure 5 shows the correlation between six dissimilarity measures, GTOM-based \( d_{i,j}^{T,[m]} \) for \( m = 0,1,2,3 \) and correlation-based \( d_{i,j}^{C,[p]} \) for \( p = 1,6 \), when applied to the yeast dataset. For this dataset, we arrive at the following results. First, \( d_{i,j}^{C,[6]} \) is highly correlated (\( > 0.8 \)) with the lower-order GTOM dissimilarities, \( d_{i,j}^{T,[0]} \) and \( d_{i,j}^{T,[1]} \). This is in concordance with the results reported in [19] which uses a cancer dataset. The other correlation-based measure \( d_{i,j}^{C,[1]} \) is highly correlated (\( 0.79 \)) with \( q^{T,[2]} \). Second, the higher-order GTOM dissimilarities \( d_{i,j}^{T,[2]} \) and \( d_{i,j}^{T,[3]} \) 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 \( d_{i,j}^{T,[3]} \) is concentrated around 0 while that of the others are concentrated around 1. This shows that many pairs of nodes considered as dissimilar under lower-order GTOM measures become similar when measured with higher-order ones.

We visualize the dissimilarity measures using classical multi-dimensional scaling plots [24], see §III-C. But to define modules, we use hierarchical clustering and topological overlap matrix plots [1] which are reviewed next.
B. 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. Towards this end, it has been suggested to use a topological overlap matrix plot (TOM or GTOM plot) which is a color-coded depiction of the values of \( d_{ij}^{T,m} \), i.e. the TOM based dissimilarity measure described above [1].

1) Comparing the GTOM Plots: When detecting modules using hierarchical clustering, we use GTOM plots to aid the choice of the dendrogram’s height cutoff. The four GTOM plots corresponding to the zeroth- to third-order GTOM are shown in Figure 6. The dataset used here is the same as the one in Figure 1. 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. As in all hierarchical clustering analysis, it is a judgement call where to cut the tree branches. Thus the modules are found by inspection: a height cutoff value is chosen in the dendrogram where to cut the tree branches. Thus the modules are found by the proposed GTOM measures with lower orders allow discovery of coarser-scale modules. We also show analytically that GTOM 1 is similar to the adjacency matrix (GTOM 0) if \( m \) is larger than the diameter of the network. Thus GTOM 1 will be useful in networks with moderate or large 'degree of separation' (average path length between any pair of nodes).

C. Multi-dimensional Scaling Plots

To visualize the GTOM-based dissimilarities, we use classical multi-dimensional scaling plots, which are shown in Figure 7. All the plots are color-coded according to the modules with respect to GTOM 1 depicted in Figure 1. The relative position of the points are well-preserved as we can see that points having the same color are almost always clustered together.

To make the plots more prolific, we encode some functional category information into the plots. 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 and they are colored in black, brown and gray. The plot using the GTOM 1 dissimilarity in Figure 7(b) shows a more clear separation between the black and brown modules.

IV. DISCUSSION

A class of natural generalizations of the widely used TOM similarity, called generalized topological overlap, is proposed. This class of new measures is constructed by counting the number of \( m \)-step neighbors that a pair of nodes share and then normalized suitably. The proposed GTOM measure is independent of the number of paths and the number of geodesic paths connecting a node and its \( m \)-step neighbor.

In the literature, several authors have found GTOM 1 useful in defining modules. However, we provide evidence that GTOM 2 is also capable of identifying cohesive and biologically plausible modules, especially when one is interested in larger modules such as the class of protein biosynthesis genes in our examples. Moreover, we show that genes that are close the essential hub genes with respect to GTOM 2 are more likely to be essential that genes that are close to the essential hub genes with respect the GTOM 1. In general, the GTOM measures with lower orders allow discovery of modules in finer scales while those with higher orders favor discovery of coarser-scale modules. We also show analytically that GTOM\( m \) is similar to the adjacency matrix (GTOM 0) if \( m \) is larger than the diameter of the network. Thus GTOM\( m \) will be useful in networks with moderate or large 'degree of separation' (average path length between any pair of nodes).
TOM plots are widely used to visualize modularity in networks. But we demonstrate that classical MDS plots are particularly handy for visualizing the relative distance between the nodes.

We also observe from the MDS plots that there is a tendency of consolidation as the order of the GTOM measure increases. This phenomenon can be seen in Figure 7(d) where GTOM3 is used and a few “sinks” (points of attraction) are formed. Intuitively speaking, each cluster becomes tighter whereas different clusters are more separated.

ACKNOWLEDGEMENT

The authors would like to thank Jun Dong, Ai Li, Bin Zhang, Marc Carlson, Stan Nelson, and Paul Mischel for their helpful discussions.

REFERENCES

Fig. 7. Multi-dimensional scaling plots of various GTOM-based dissimilarities. (a) GTOM0. (b) GTOM1. (c) GTOM2. (d) GTOM3. The coloring scheme is used to reflect the 7 modules shown in Figure 1(b) detected by using hierarchical clustering on the GTOM1-based dissimilarity. The symbol ‘▲’ denotes genes that belong to the functional category ‘protein biosynthesis’. Genes belong to other classes are denoted by a ‘◦’. The overall cluster structures are generally preserved across the different GTOM measures. But the spatial distributions of points vary to a large extent. Genes in the ‘protein biosynthesis’ class are clustered together.


