Network Neighborhood Analysis With The Multi-Node Topological Overlap Measure

Ai Li a, Steve Horvath a,b,*

aDepartment of Biostatistics, School of Public Health, University of California, Los Angeles, CA 90095-1772, USA,
bDepartment of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA 90095-7088, USA

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 neighborhood analysis 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 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/

1.1 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 \( \sum_{u \notin \{i,j\}} a_{iu} \cdot a_{ju} \). The topological overlap \( T_{ij} \) 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):

\[
T_{ij} = \left\{ \begin{array}{ll}
\frac{\min\{\sum_{u \notin \{i,j\}} a_{iu} \cdot a_{ju} + a_{ij}\}}{1 + \sum_{u \notin \{i,j\}} a_{iu} - a_{ij}} & \text{if } i \neq j \\
1 & \text{if } i = j.
\end{array} \right.
\]

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 \( \sum_{u \notin \{i,j\}} a_{iu} - a_{ij} = 0 \). In the following, we use 0 \( \leq a_{ij} \leq 1 \) to prove that 0 \( \leq T_{ij} \leq 1 \). Since \( \sum_{u \notin \{i,j\}} a_{iu} - a_{ij} \) and \( \sum_{u \notin \{i,j\}} a_{ju} - a_{ij} \) which implies \( \sum_{u \notin \{i,j\}} a_{iu} - a_{ij} \leq \min\{\sum_{u \notin \{i,j\}} a_{iu} - a_{ij}, \sum_{u \notin \{i,j\}} a_{ju} - a_{ij}\} \). Along

© 2006 The Author(s)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Associate Editor: Golan Yona
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

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

(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

$$|N(i, j)| = \sum_{u \neq i, j} a_{iu} a_{ju},$$

(3)

$$|N(-i, j)| = \sum_{u \neq i} a_{iu} - a_{ij},$$

$$|N(-i, -j)| = \sum_{u \neq i, j} a_{iu} a_{ju} - a_{ij} a_{jk}.$$  

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

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

(4)

where

$$|N(i, j, k)| = \sum_{u \neq i, j, k} a_{iu} a_{ju} a_{ku},$$

(5)

$$|N(-i, j, k)| = \sum_{u \neq i, j} a_{iu} a_{ju} - a_{ik} a_{jk},$$

$$|N(-i, -j, k)| = \sum_{u \neq i, j} a_{iu} a_{ju} - a_{ik} a_{jk}.$$  

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.

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 $S_0$ such that if more than $S_0$ 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 $S_0$, the neighborhood is more interconnected than 95 percent of the locally permuted neighborhoods, we chose a neighborhood size close to $S_0$ 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.

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. $a_{ij} = \left(\text{corr}(x_i, x_j)\right)^6$, 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.
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 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.

<table>
<thead>
<tr>
<th>Probe Name</th>
<th>Gene Name</th>
<th>Neuron</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>209642_at</td>
<td>BUB1</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>218355_at</td>
<td>KIF4A</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>222077_at</td>
<td>RACGAP1</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>219918_at</td>
<td>ASPM</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>207828_at</td>
<td>CENPF</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>202580_at</td>
<td>FOXM1</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>202870_at</td>
<td>CDC20</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>202095_at</td>
<td>BIRC5</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>221591_at</td>
<td>FLJ10156</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>218009_at</td>
<td>PRC1</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>204641_at</td>
<td>NEK2</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>209172_at</td>
<td>CENPF</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>209464_at</td>
<td>AURKB</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>212020_at</td>
<td>MKI67</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>212023_at</td>
<td>MKI67</td>
<td>unknown</td>
<td>yes</td>
</tr>
<tr>
<td>204444_at</td>
<td>KIF11</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>212494_at</td>
<td>BRRN1</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>204026_at</td>
<td>ZWINT</td>
<td>unknown</td>
<td>unknown</td>
</tr>
<tr>
<td>203213_at</td>
<td>CDC2</td>
<td>unknown</td>
<td>yes</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Probe Name</th>
<th>Gene Name</th>
<th>Neuron</th>
<th>Cancer</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>208464_at</td>
<td>GRIA4</td>
<td>yes</td>
<td>unknown</td>
<td>0.62</td>
</tr>
<tr>
<td>221623_at</td>
<td>BCAN</td>
<td>yes</td>
<td>unknown</td>
<td>0.406</td>
</tr>
<tr>
<td>91920_at</td>
<td>BCAN</td>
<td>yes</td>
<td>unknown</td>
<td>0.226</td>
</tr>
<tr>
<td>219107_at</td>
<td>BCAN</td>
<td>yes</td>
<td>unknown</td>
<td>0.212</td>
</tr>
<tr>
<td>216476_at</td>
<td>LOC115131</td>
<td>unknown</td>
<td>unknown</td>
<td>0.142</td>
</tr>
<tr>
<td>222301_at</td>
<td>CROC4</td>
<td>yes</td>
<td>unknown</td>
<td>0.093</td>
</tr>
<tr>
<td>212655_at</td>
<td>BDGL29</td>
<td>unknown</td>
<td>unknown</td>
<td>0.297</td>
</tr>
<tr>
<td>213768_s_at</td>
<td>ASCL1</td>
<td>yes</td>
<td>yes</td>
<td>0.233</td>
</tr>
<tr>
<td>209988_s_at</td>
<td>ASCL1</td>
<td>yes</td>
<td>yes</td>
<td>0.191</td>
</tr>
<tr>
<td>209987_s_at</td>
<td>ASCL1</td>
<td>yes</td>
<td>yes</td>
<td>0.129</td>
</tr>
<tr>
<td>212265_s_at</td>
<td>QKI</td>
<td>yes</td>
<td>unknown</td>
<td>0.163</td>
</tr>
<tr>
<td>218902_at</td>
<td>NOTCH1</td>
<td>yes</td>
<td>yes</td>
<td>-0.023</td>
</tr>
<tr>
<td>202981_s_at</td>
<td>SIAH1</td>
<td>unknown</td>
<td>yes</td>
<td>0.085</td>
</tr>
<tr>
<td>221776_s_at</td>
<td>BRD7</td>
<td>unknown</td>
<td>yes</td>
<td>0.068</td>
</tr>
<tr>
<td>212615_s_at</td>
<td>FLJ12178</td>
<td>unknown</td>
<td>unknown</td>
<td>0.123</td>
</tr>
<tr>
<td>212616_s_at</td>
<td>FLJ12178</td>
<td>unknown</td>
<td>unknown</td>
<td>0.126</td>
</tr>
<tr>
<td>213891_s_at</td>
<td>TCF4</td>
<td>unknown</td>
<td>yes</td>
<td>0.091</td>
</tr>
<tr>
<td>201310_s_at</td>
<td>CSorf13</td>
<td>yes</td>
<td>yes</td>
<td>0.072</td>
</tr>
<tr>
<td>214239_s_at</td>
<td>RNF110</td>
<td>unknown</td>
<td>yes</td>
<td>0.076</td>
</tr>
<tr>
<td>213551_s_at</td>
<td>RNF110</td>
<td>unknown</td>
<td>yes</td>
<td>0.075</td>
</tr>
</tbody>
</table>
with the network connectivity $k$ of the initial protein (Spearman correlation $r = 0.36$, p-value $\leq 0.001$) 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 hubs 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., 2005). 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

---

**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.

<table>
<thead>
<tr>
<th>Probe Name</th>
<th>Gene Name</th>
<th>Neuron</th>
<th>Cancer</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>208464_at</td>
<td>GRIA4</td>
<td>yes</td>
<td>unknown</td>
<td>0.62</td>
</tr>
<tr>
<td>204522_s_at</td>
<td>TOX</td>
<td>unknown</td>
<td>unknown</td>
<td>0.601</td>
</tr>
<tr>
<td>206170_at</td>
<td>ADRB2</td>
<td>unknown</td>
<td>unknown</td>
<td>0.539</td>
</tr>
<tr>
<td>216247_at</td>
<td>RPS20</td>
<td>unknown</td>
<td>unknown</td>
<td>0.537</td>
</tr>
<tr>
<td>214028_s_at</td>
<td>TDRD3</td>
<td>unknown</td>
<td>unknown</td>
<td>0.526</td>
</tr>
<tr>
<td>213447_at</td>
<td>IPW</td>
<td>unknown</td>
<td>unknown</td>
<td>0.522</td>
</tr>
<tr>
<td>207113_s_at</td>
<td>TNF</td>
<td>yes</td>
<td>yes</td>
<td>0.514</td>
</tr>
<tr>
<td>218036_s_at</td>
<td>CGI-07</td>
<td>unknown</td>
<td>unknown</td>
<td>0.504</td>
</tr>
<tr>
<td>209160_at</td>
<td>AKR1C3</td>
<td>unknown</td>
<td>yes</td>
<td>0.5</td>
</tr>
<tr>
<td>206107_at</td>
<td>RGS11</td>
<td>unknown</td>
<td>yes</td>
<td>0.496</td>
</tr>
<tr>
<td>209782_s_at</td>
<td>DBP</td>
<td>unknown</td>
<td>unknown</td>
<td>0.495</td>
</tr>
<tr>
<td>211653_s_at</td>
<td>AKR1C1</td>
<td>unknown</td>
<td>yes</td>
<td>0.494</td>
</tr>
<tr>
<td>213778_s_at</td>
<td>ZFP276</td>
<td>unknown</td>
<td>unknown</td>
<td>0.49</td>
</tr>
<tr>
<td>215119_at</td>
<td>MYR8</td>
<td>yes</td>
<td>unknown</td>
<td>0.486</td>
</tr>
<tr>
<td>202753_at</td>
<td>p44S10</td>
<td>unknown</td>
<td>yes</td>
<td>-0.481</td>
</tr>
<tr>
<td>209292_at</td>
<td>ID4</td>
<td>unknown</td>
<td>yes</td>
<td>0.481</td>
</tr>
<tr>
<td>204530_s_at</td>
<td>TOX</td>
<td>unknown</td>
<td>unknown</td>
<td>0.481</td>
</tr>
<tr>
<td>221974_at</td>
<td>SNRPN</td>
<td>unknown</td>
<td>unknown</td>
<td>0.481</td>
</tr>
<tr>
<td>219188_s_at</td>
<td>LRPI6</td>
<td>unknown</td>
<td>yes</td>
<td>0.48</td>
</tr>
<tr>
<td>205630_at</td>
<td>CRH</td>
<td>yes</td>
<td>unknown</td>
<td>0.478</td>
</tr>
</tbody>
</table>

**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.

**Figure 4.** The Multi-Node Topological Overlap Measure

---
Li and Horvath

neighborhood analysis methods. We argue that MTOM leads to more meaningful results than standard weighted gene co-expression networks. Here we use the model to illustrate the value of choosing multiple nodes as seeds for this article to discuss the relationship of this simple model to actual the structure of co-expression networks, it is beyond the scope of this simple model was motivated by our unpublished research on To evaluate out method, we simulated a network model motivated

in neighborhoods constructed using alternative methods. In this app-
llication, 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 our 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.

\[
x_{ij}^{(k)} = m_{ij}^{(k)} \times (\frac{i}{n_k})^{1/4} + \epsilon_{ij}^{(k)}
\]

where the stochastic noise \( \epsilon_{ij}^{(k)} \) was simulated to follow a normal distribution with mean 0 and variance \( 6 \). The vector \( m_{ij}^{(k)} \) 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).

\[
m_{ij}^{(1)} = 1.5 \times I_{30 < j \leq 45} + 1 \times I_{45 < j \leq 60},
\]

\[
m_{ij}^{(2)} = 1 \times I_{0 < j \leq 15} + 1 \times I_{15 < j \leq 30},
\]

\[
m_{ij}^{(3)} = 1 \times I_{0 < j \leq 15} + 1 \times I_{15 < j \leq 30} + 1 \times I_{30 < j \leq 45},
\]

\[
m_{ij}^{(4)} = 1 \times I_{0 < j \leq 15} + 1 \times I_{30 < j \leq 45},
\]

\[
m_{ij}^{(5)} = 1.5 \times I_{0 < j \leq 15} + 1.5 \times I_{30 < j \leq 45} + 0.5 \times I_{45 < j \leq 60},
\]

\[
m_{ij}^{(6)} = 0.
\]

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 \( 2000 \times 2000 \) 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. \( a_{ij} = |\text{cor}(x_i, x_j)|^\beta \), 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...
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 ave-
mage of the pairwise similarities between the nodes. Since the
average pairwise similarity measure is computationally much sim-
pler, 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 recur-
sively 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 biol-
gical signal in the resulting neighborhoods for the different applica-
tions. 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 topologi-
cal overlap measure. To illustrate the value of taking a multi-node
perspective when defining neighborhoods, we generalize the stan-
dard pairwise topological overlap measure (TOM) to measure the
topological overlap of multiple nodes (MTOM). MTOM is a natu-
ral 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 neighbor-
hoods. But it may be worth-while to explore the use of alternative,
more time consuming, construction methods. For example, step-
wise 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.
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.

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
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


