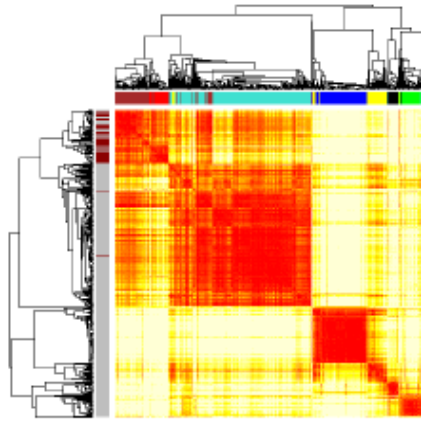


The Generalized Topological Overlap Matrix in Biological Network Analysis

Andy Yip, Steve Horvath
Email: shorvath@mednet.ucla.edu
Depts Human Genetics and Biostatistics,
University of California, Los Angeles

Contents

- Dissimilarity measures in undirected networks
- Dissimilarities based on shared neighbors
- Generalized topological overlap matrix
- Applications
- Simulation

Network Terminology

- Unweighted Network=adjacency matrix A=[a_{ij}], that encodes whether a pair of nodes is connected.
  - A is a symmetric matrix with entries in [0,1]
  - a_{ij}=1 nodes i and j are connected else 0
- HERE WE CONSIDER AN UNWEIGHTED NETWORKS
- Gene connectivity K= row sum of the adjacency matrix=number of direct neighbors

$$k_i = \sum_j a_{ij}$$

- Network Module=Subset of highly interconnected nodes

Basic Steps in Many Biological Network Analyses

Define an Adjacency Matrix

↓

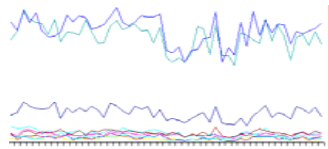
Measure of Node Dissimilarity

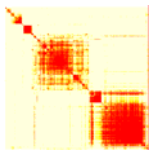
↓

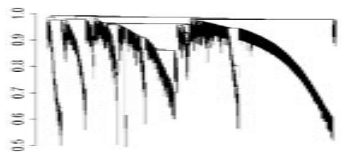
Identify Network Modules (Clustering)

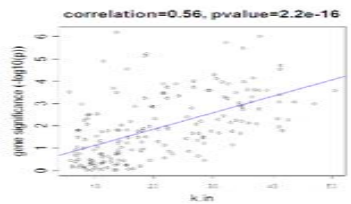
↓

Understand the biological meaning of modules and network concepts









What is a node dissimilarity? And why do we need it?

Mathematical Definition of a Dissimilarity measure

1) Symmetry: G(u,v)=G(v,u)
2) Non-negative G(u,v)>=0
3) G(u,u)=0

Major application: **module detection**

Module=cluster of “similar” nodes

Implementation: use the dissimilarity measure as input of a clustering procedure,

- e.g. average linkage hierarchical clustering,
- or partitioning around medoid clustering

Aside: node dissimilarities have many other uses, e.g. to study how a node dissimilarity between 2 interacting genes changes across conditions…

Possible measures of node dissimilarity

1. Simply use 1 minus the adjacency matrix
2. Length of shortest path connecting 2 nodes
3. Our focus: measures based on number of shared neighbors
   - Intuition: if 2 people share the same friends they are close in a social network

Similarity based on number of shared neighbors

Number neighbors shared by nodes i and j

$$\sum_{u \neq i,j} a_{iu} a_{uj}$$

Numerator of topological overlap measure GTOM1

$$\sum_{u \neq i,j} a_{iu} a_{uj} + a_{ij}$$

Idea: define the denominator so that the following requirements are satisified

i) numerator $\leq$ denominatory, i.e.

$0 \leq \text{GTOM(i,j)} \leq 1$

ii) denominator TOM(i,j)>0

Standard Topological Overlap measure (Ravasz et al 2002)

$$\mathrm{GTOM1}(i,j) = \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}}$$

$$\mathrm{dissGTOM1}(i,j) = 1 - GTOM1(i,j)$$

- Generalization to unweighted networks discussed in Zhang and Horvath (2005).
- Generalization to multiple nodes defined in Ai Li, S Horvath (2006) Multinode topological overlap matrix.

The topological overlap measures interconnectedness

- for an *unweighted* network, one can show that the topological overlap=1 only if the node with fewer links satisfies two conditions:
  - (a) all of its neighbors are also neighbors of the other node, i.e. it is connected to all of the neighbors of the other node and
  - (b) it is linked to the other node.
- In contrast, top. overlap=0 if i and j are unlinked and the two nodes don't have common neighbors.

Our set theory interpretation of the topological overlap matrix

m-step neighborhood

$$N_m(i) = \{ j \neq i \mid \text{minimum path length}(i, j) \leq m \}$$

Node Similarity based on number of shared 1-step neighbors

$$GTOM1(i, j) = \frac{| N_1(i) \cap N_1(j) | + a_{ij}}{\min(| N_1(i) |, | N_1(j) |) + 1 - a_{ij}}$$

Mathematically, identical to the topological overlap measure proposed in the supplement of Ravasz et al (2002)

Generalizing the topological overlap matrix to 2 step neighborhoods etc

- Simply replace the neighborhoods by 2 step neighborhoods in the following formula

$$GTOM\,2(i,j) = \frac{|N_2(i) \cap N_2(j)| + a_{ij}}{\min(|N_2(i)|, |N_2(j)|) + 1 - a_{ij}}$$

where $N_2(i)$ denotes the set of nodes within 2 steps of node i

Reference: Andy M. Yip and SH (2006) The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks.

www.genetics.ucla.edu/labs/horvath/GTOM

Computationally simple calculation of GTOMm

- GTOMm can be directly calculated from A+A*A+A*A*A+A…..A

  where * denotes matrix mutiplication
- Computation time driven by m matrix multiplications of A

Summary: dissimilarity measures based on an adjacency matrix A

Trivial dissimilarity for a network adjacency matrix $A = (a_{ij})$

$$\text{disGTOM}\,0(i, j) = 1 - a_{ij}$$

Standard topological overlap dissimilarity matrix based on 1 step neigbhorhood

$$\text{dissGTOM}\,1(i, j) = 1 - \frac{\sum_u a_{iu} a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} = 1 - \frac{|N_1(i) \cap N_1(j)| + a_{ij}}{\min(|N_1(i)|, |N_1(j)|) + 1 - a_{ij}}$$

Our generalization to m-step neighborhoods

$$\text{dissGTOM}\,m(i, j) = 1 - \frac{|N_m(i) \cap N_m(j)| + a_{ij}}{\min(|N_m(i)|, |N_m(j)|) + 1 - a_{ij}}$$

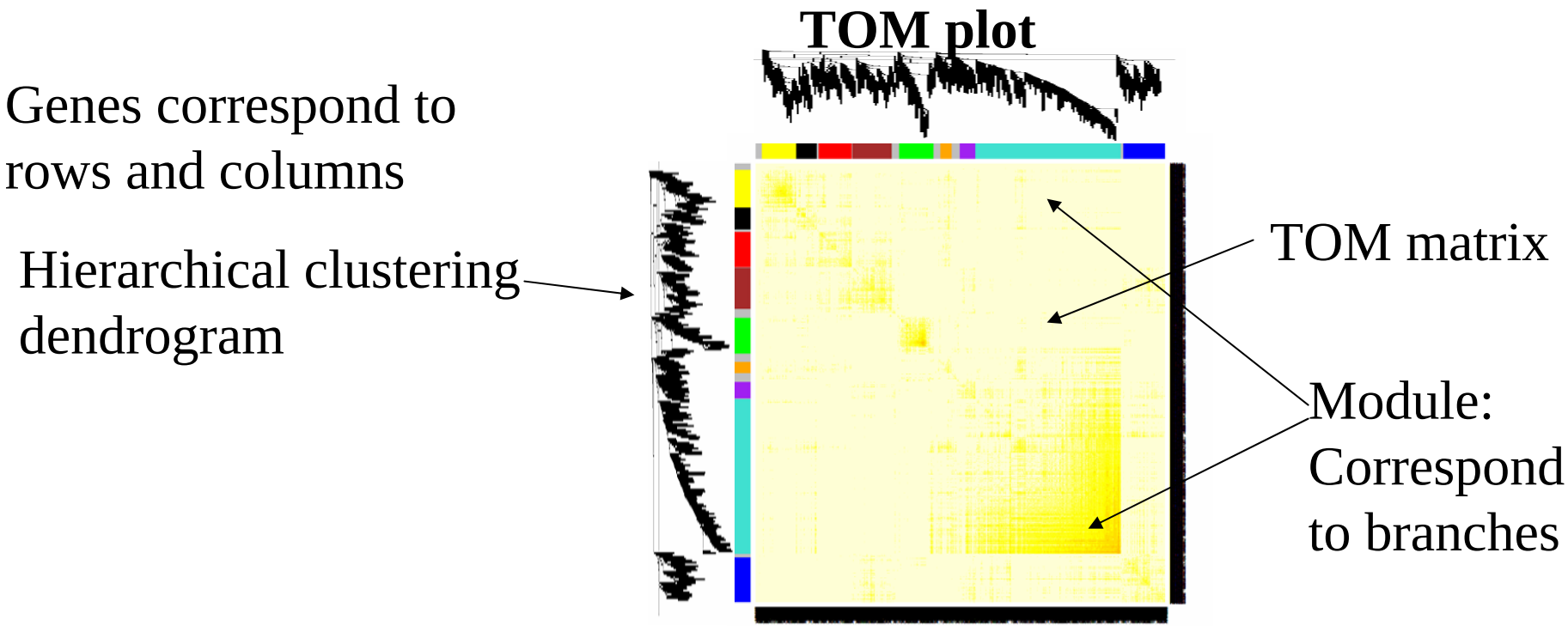
Defining Gene Modules
=sets of tightly co-regulated genes

Module Identification based on the notion of topological overlap

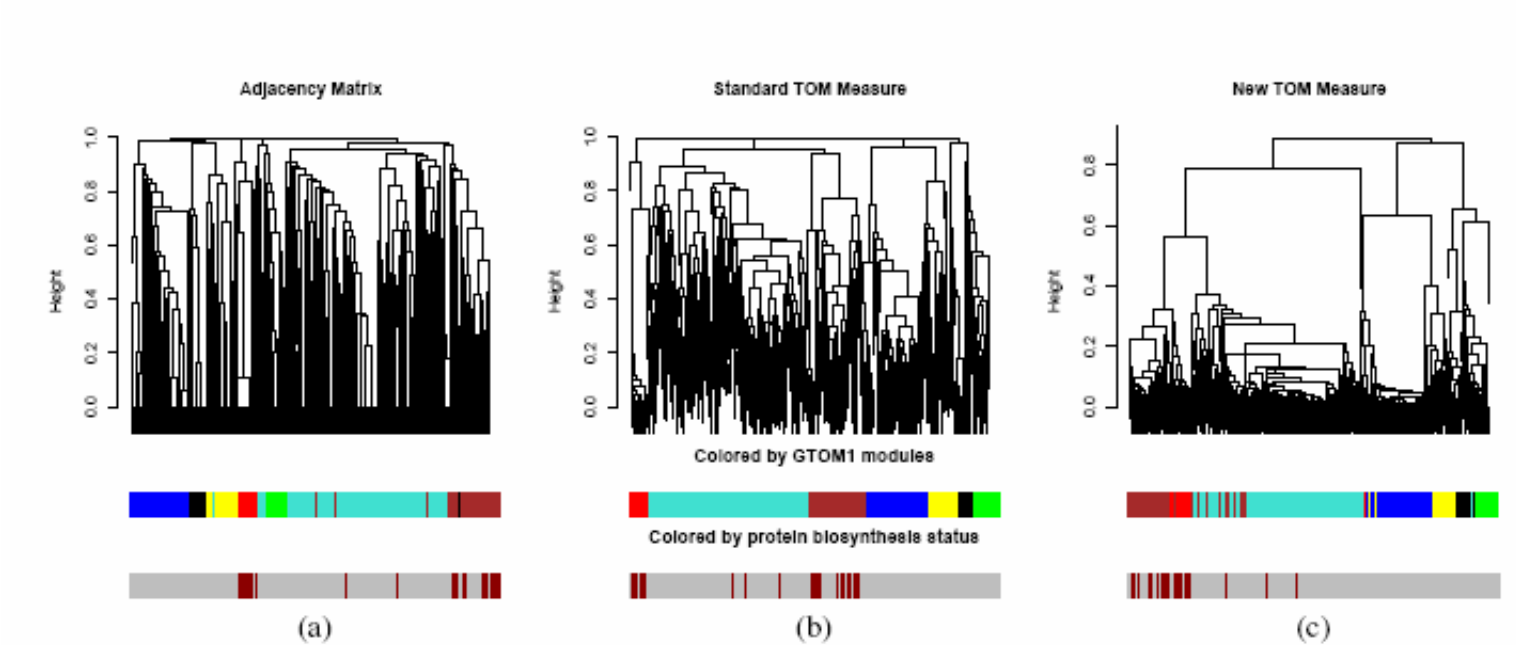
- An important aim of metabolic network analysis is to detect subsets (modules) of nodes that are tightly connected to each other.
- We adopt the definition of Ravasz et al (2002): modules are groups of nodes that have high topological overlap.

Using the TOM matrix to cluster genes

- To group nodes with high topological overlap into modules (clusters), we typically use average linkage hierarchical clustering coupled with the TOM distance measure.
- Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering.
  - Here modules correspond to branches of the dendrogram



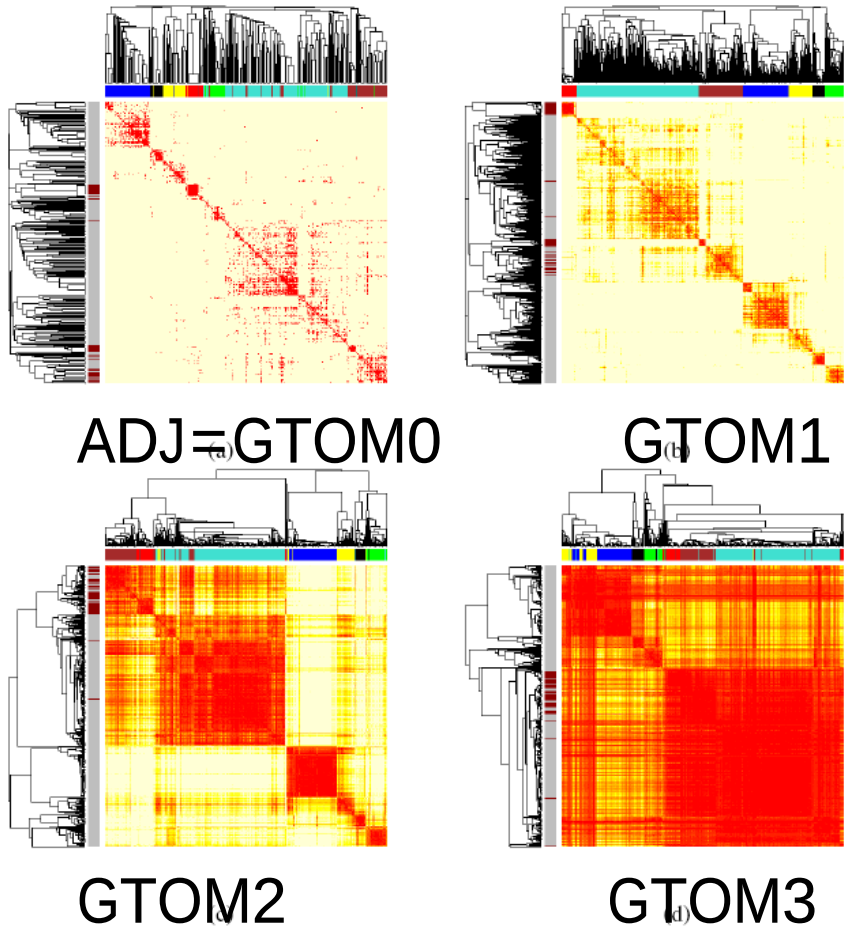
Comparison of 3 different similarities in capturing the functional class `protein biosynthesis'.



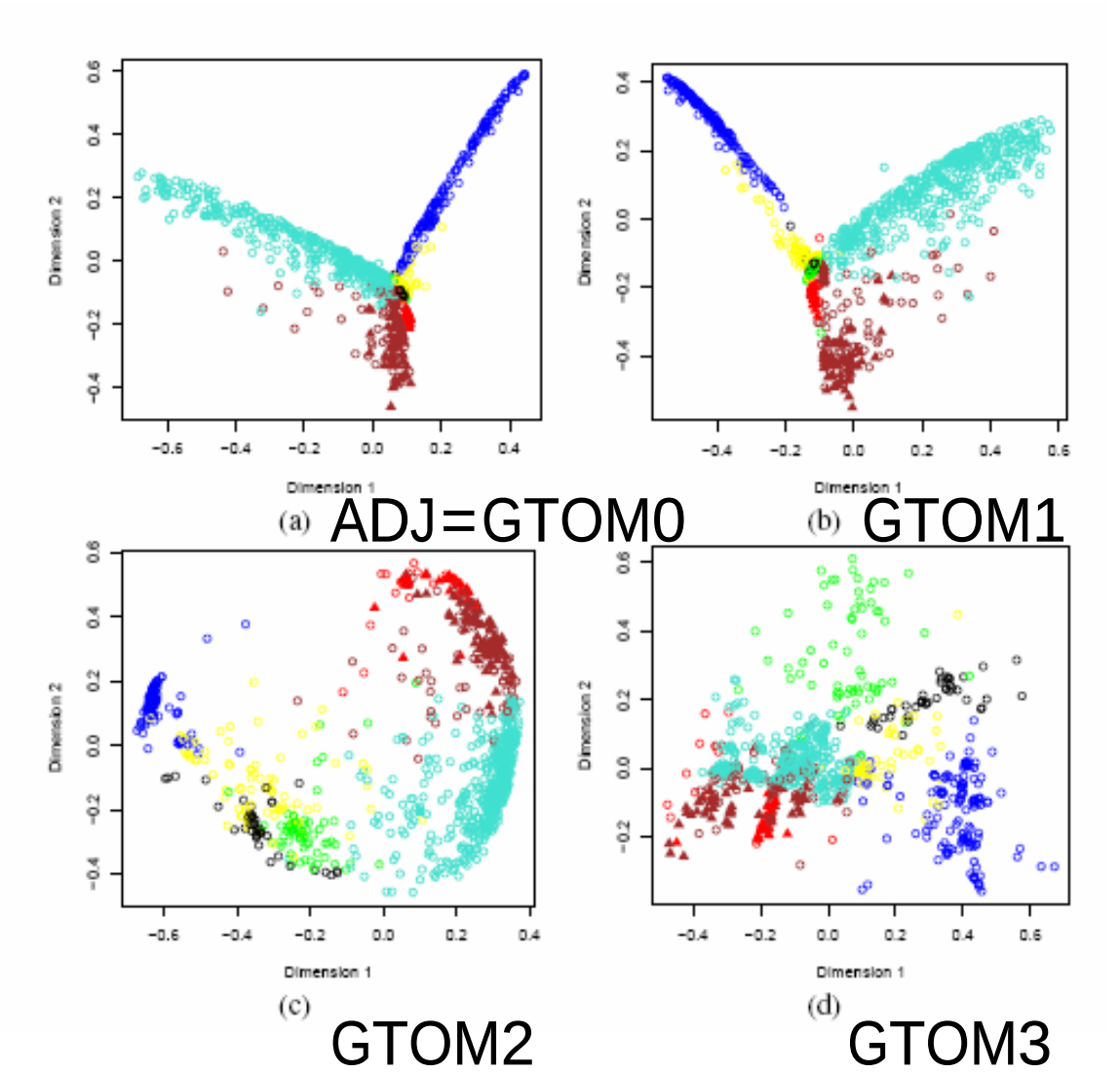
- (a) ADJ=GTOM0 (b) GTOM1 (c) GTOM2
- The middle row shows the color bar ordered by the corresponding dendrogram but colored by the module assignment with respect to the TOM measure in (b), the bottom shows the color bar ordered by the corresponding dendrogram where genes belong to the class `protein biosynthesis' are colored in dark red.
- Almost all protein biosynthesis genes are grouped together by the GTOM2 measure whereas the other two measures tend to distribute the class over two modules.

Topological Overlap Matrix Plots for different GTOM measures, yeast

- Overall, modules are quite robust with respect to the GTOM measure.
- Smaller modules are more visible in GTOM0 and GTOM1 plots
- Larger modules are more pronounced in GTOM2 and GTOM3 plots



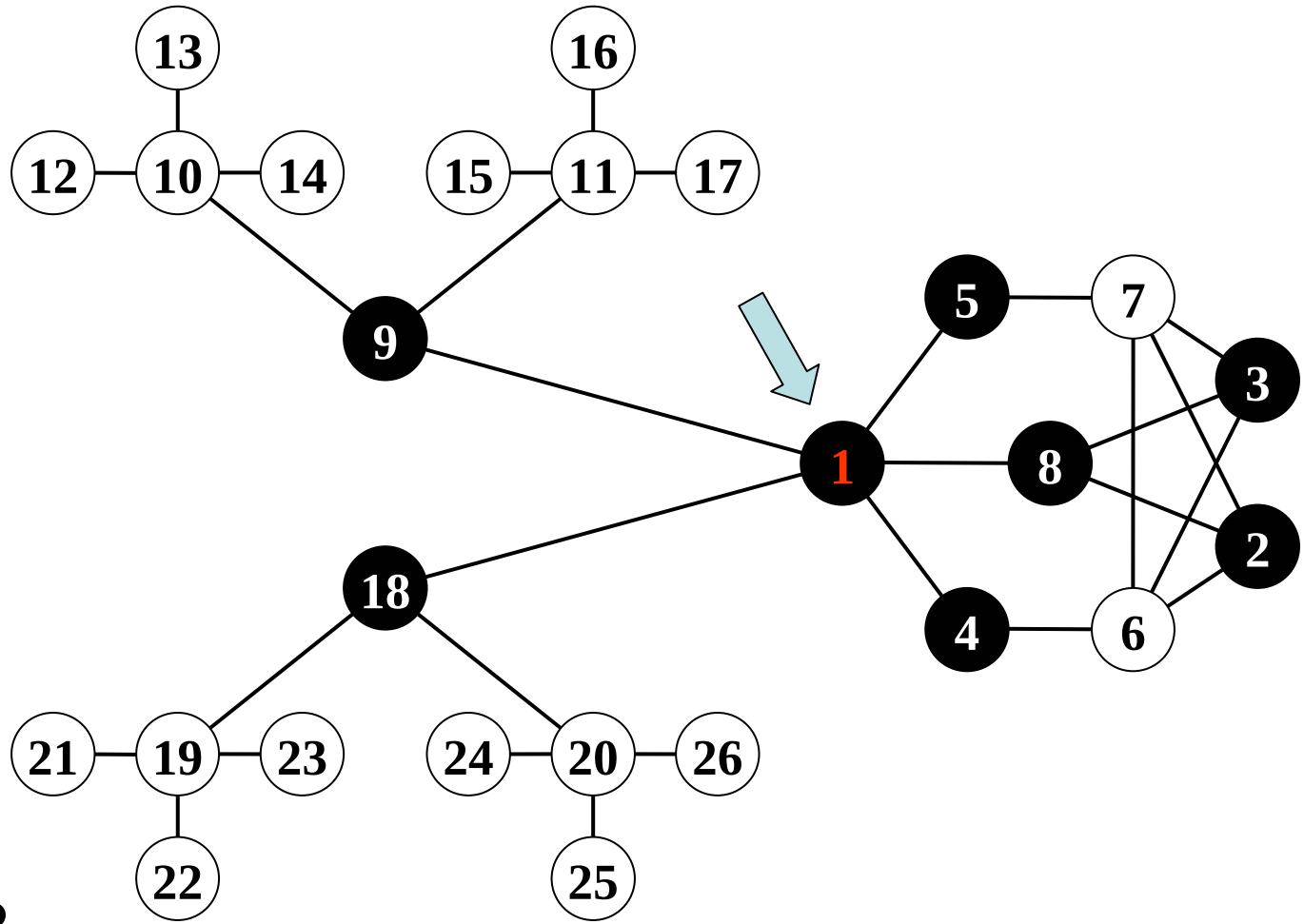
Multidimensional Scaling Plots involving different GTOMs



Simple simulated example where GTOM2 is better than GTOM1 and GTOM0

Example, when GTOM2 is superior to GTOM1 or GTOM0

- Top 8 GTOM2 neighbors of Node 1 are exactly Node 1 – Node 8.
- TOM1 neighbors of Node 1 are Node 1,4,5,8,9,18.



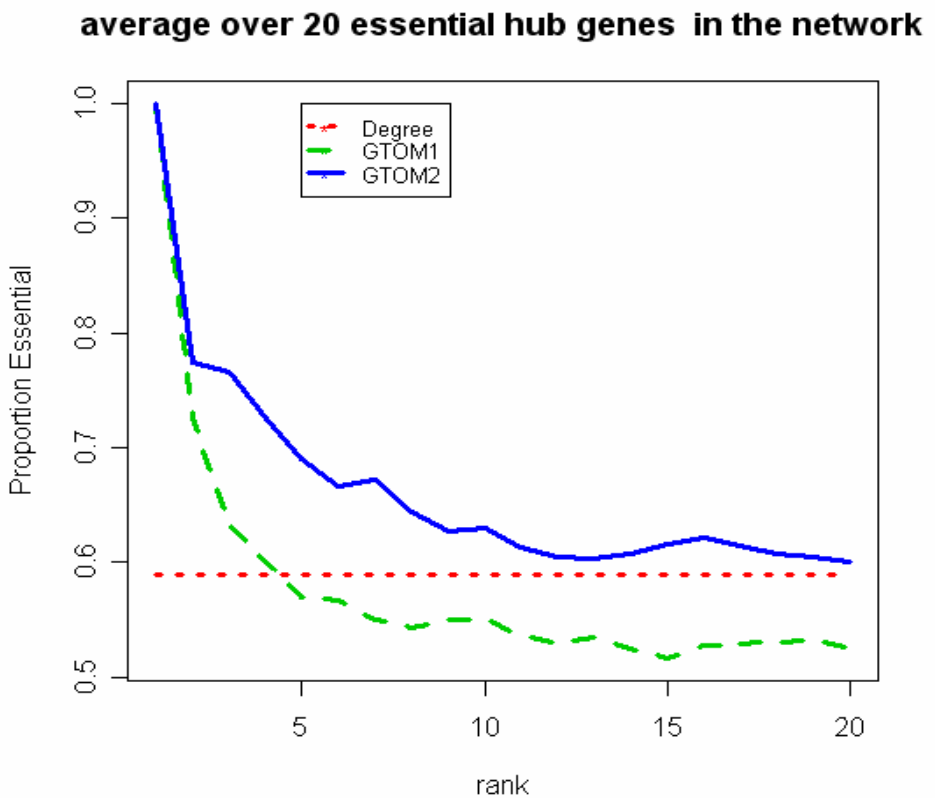
Black circles: 8 closest GTOM1 neighbors of node 1

Predicting essential proteins in a fly network

- Idea: start with a single highly connected essential protein and consider its closest neighbors based on a dissimilarity measure
- One would hope that the most similar (closest) neighbors are also essential since they may be part of the same pathway
- Data protein-protein interaction data from Biogrid
- Essentiality: determined by knock-out experiments

GTOM2 outperforms GTOM1 and GTOM0 in the fly protein-protein network

- Y-axis proportion of essential genes amony the closest neighbors
- X-axis size of closest neighborhood



Discussion

- Since the topological overlap matrix considers shared neighbors, it tends to be more robust to spurious connections.
- Limitation of GTOM: it rquires an unweighted network (binary adjacencies)
- GTOM is based on pairwise overlap.
  - In contrast, MTOM is based on multi-node overlap.
- Overall, GTOM0, GTOM1 and GTOM2 lead to similar clusters (modules).

Our experience

- In most applications, we find that GTOM1 is better than GTOM0
- Often GTOM1 performs better than GTOM2
- But in the fly network GTOM2 is better than GTOM1
- GTOMm with m>2 tends to lump nodes together→ loss of resolution

Acknowledgement

Biostatistics/Bioinformatics

- Ai Li, doctoral student UCLA (MTOM software)
- Jun Dong, Postdoc UCLA
- Wei Zhao, Postdoc UCLA
- Lin Wang
- Bin Zhang

Collaborators

Marc Carlson, Dan Geschwind, Paul Mischel, Stan Nelson, Mike Oldham, and many more

Webpages and References

- This talk and relevant R code
  - *Yip A, Horvath S (2006) The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks Proceedings Volume Gene Networks: Theory and Application Workshop at BIOCOMP'06, Las Vegas* http://www.genetics.ucla.edu/labs/horvath/GTOM/
  - *Ai Li, Steve Horvath (2006) The Multi-Point Topological Overlap Matrix for Gene Neighborhood Analysis. Proceedings Volume Gene Networks: Theory and Application Workshop at BIOCOMP'06, Las Vegas* http://www.genetics.ucla.edu/labs/horvath/MTOM/
  - *Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.* www.bepress.com/sagmb/vol4/iss1/art17
- Yeast Co-Expression Network

  MRJ Carlson, B Zhang, Z Fang, PS Mischel, S Horvath, SF Nelson, Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks", BMC Genomics 2006, 7:40 (3 March 2006). http://www.biomedcentral.com/1471-2164/7/40/

Appendix