

Generalizations of the Topological Overlap Matrix for Module Detection in Gene and Protein Networks

Ai Li and Steve Horvath

Email: shorvath@mednet.ucla.edu

Depts Human Genetics and Biostatistics,
University of California, Los Angeles

Mathematical Definition of an Undirected Network

Network=Adjacency Matrix

- A network can be represented by an adjacency matrix, $A=[a_{ij}]$, that encodes whether/how a pair of nodes is connected.
 - A is a symmetric matrix with entries in $[0,1]$
 - For unweighted network, entries are 1 or 0 depending on whether or not 2 nodes are adjacent (connected)
 - For weighted networks, the adjacency matrix reports the connection strength between gene pairs

Generalized Connectivity

- Gene connectivity = row sum of the adjacency matrix
 - For unweighted networks=number of direct neighbors
 - For weighted networks= sum of connection strengths to other nodes

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

Network Construction

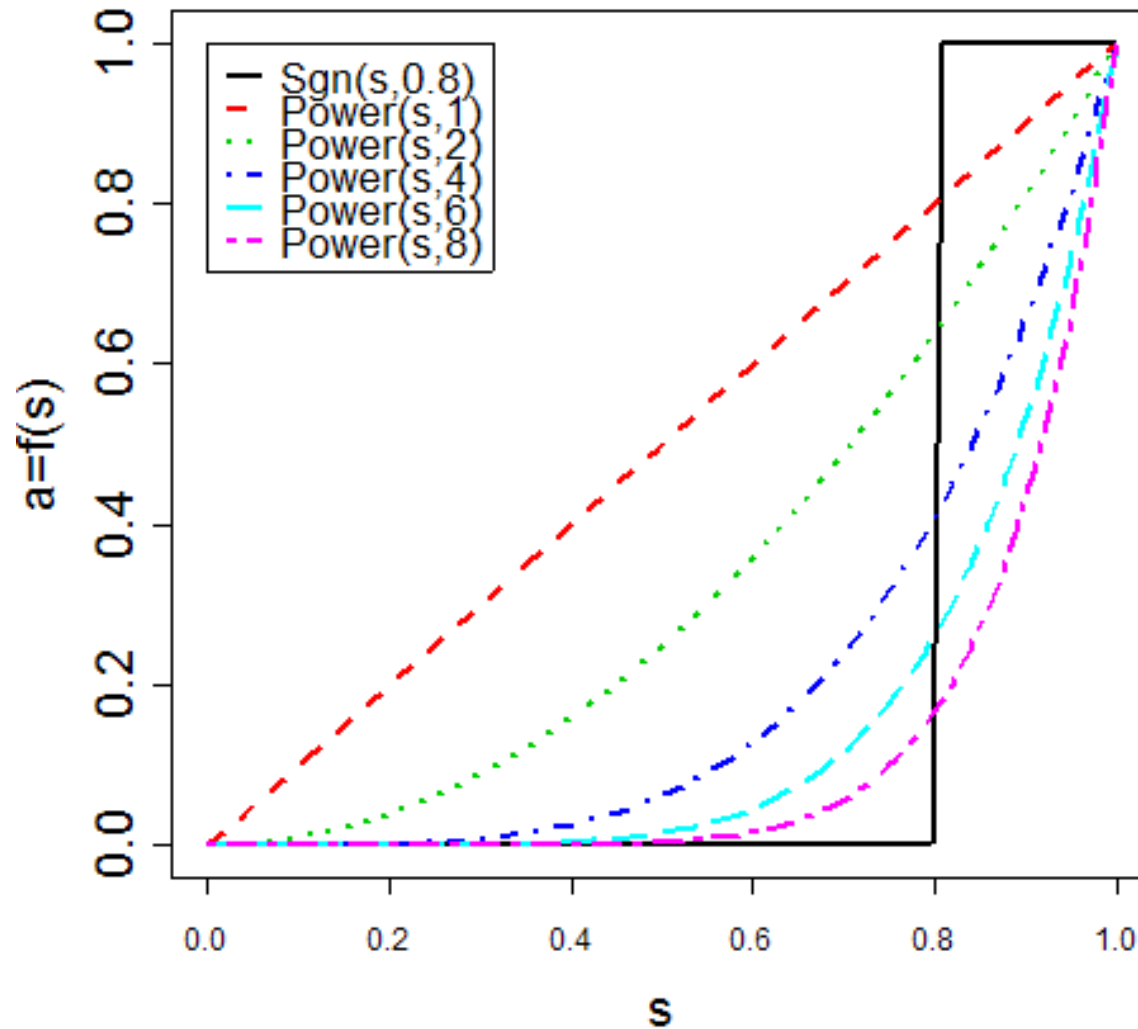
Bin Zhang and Steve Horvath (2005) "A General Framework for Weighted Gene Co-Expression Network Analysis", Statistical Applications in Genetics and Molecular Biology: Vol. 4: No. 1, Article 17.

An adjacency function is used to turn co-expression information into a network

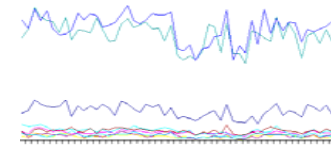
- Measure co-expression by the absolute value of the Pearson correlation
- Define an adjacency matrix by using an adjacency function $A(i,j) = AF(\text{cor}(x[i], x[j]))$
- The adjacency function AF is a monotonic function that maps $[0,1]$ onto $[0,1]$
- We consider 2 classes of AF
 - Hard Thresholding:
 - step function $AF(s) = I(s > \tau)$ with parameter τ
 - Soft thresholding:
 - Power Adjacency function $AF(s) = s^b$ with parameter b
- The choice of the AF parameters determines the properties of the network.

Comparing adjacency functions

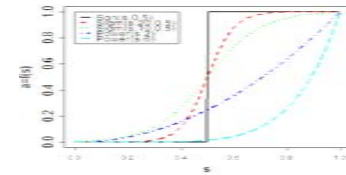
Power Adjacency vs Step Function



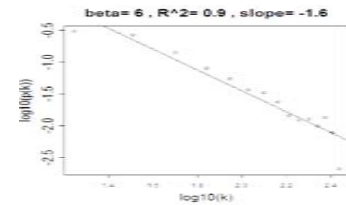
Define a Gene Co-expression Similarity



Define a Family of Adjacency Functions

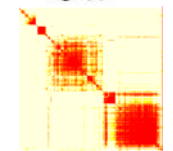


Determine the AF Parameters

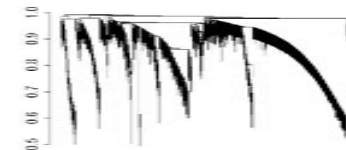


Our focus:

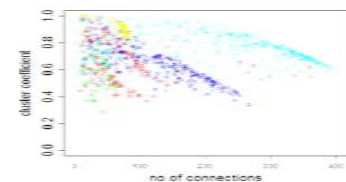
Define a Measure of Node Dissimilarity



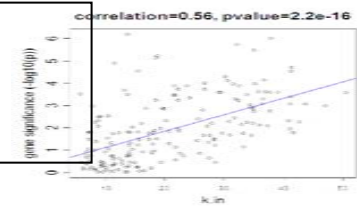
Identify Network Modules (Clustering)



Relate Network Concepts to Each Other



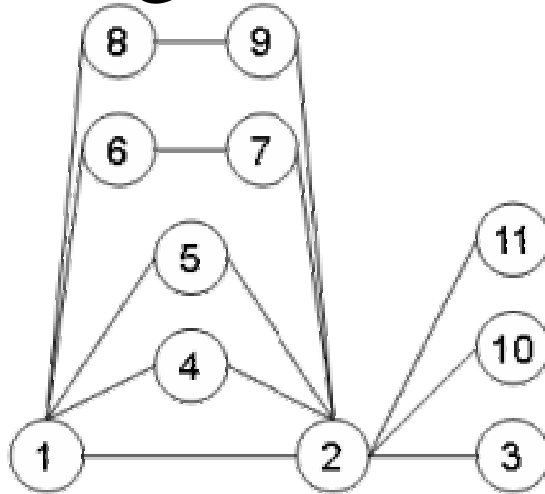
Relate the Network Concepts to External Gene or Sample Information



How to measure distance in a network?

- Graph theoretic answer: Geodesics
 - length of shortest path connecting 2 nodes
- More robust answer: look at shared neighbors
 - Intuition: if 2 people share the same friends they are close in a social network
 - Use the topological overlap measure based distance proposed by Ravasz et al 2002 (Science)

Set interpretation of the topological overlap matrix



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

$N_1(i)$ denotes the set of 1-step (i.e. direct) neighbors of node i

$||$ measures the cardinality

Adding $1-a(i,j)$ to the denominator prevents it from becoming 0.

Generalizing the topological overlap matrix to 2 step neighborhoods etc

- *Andy M. Yip and SH (2006) The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks.*
www.genetics.ucla.edu/labs/horvath/GTOM
- 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

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.

Topological Overlap leads to a network distance measure (Ravasz et al 2002)

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

$$DistTOM_{ij} = 1 - TOM_{ij}$$

- Generalization to unweighted networks is discussed in Zhang and Horvath (2005). Trivial since the formula is mathematically meaningful even if the adjacencies are real numbers in $[0,1]$

Defining Gene Modules
=sets of tightly co-regulated genes

Module Identification based on the notion of topological overlap

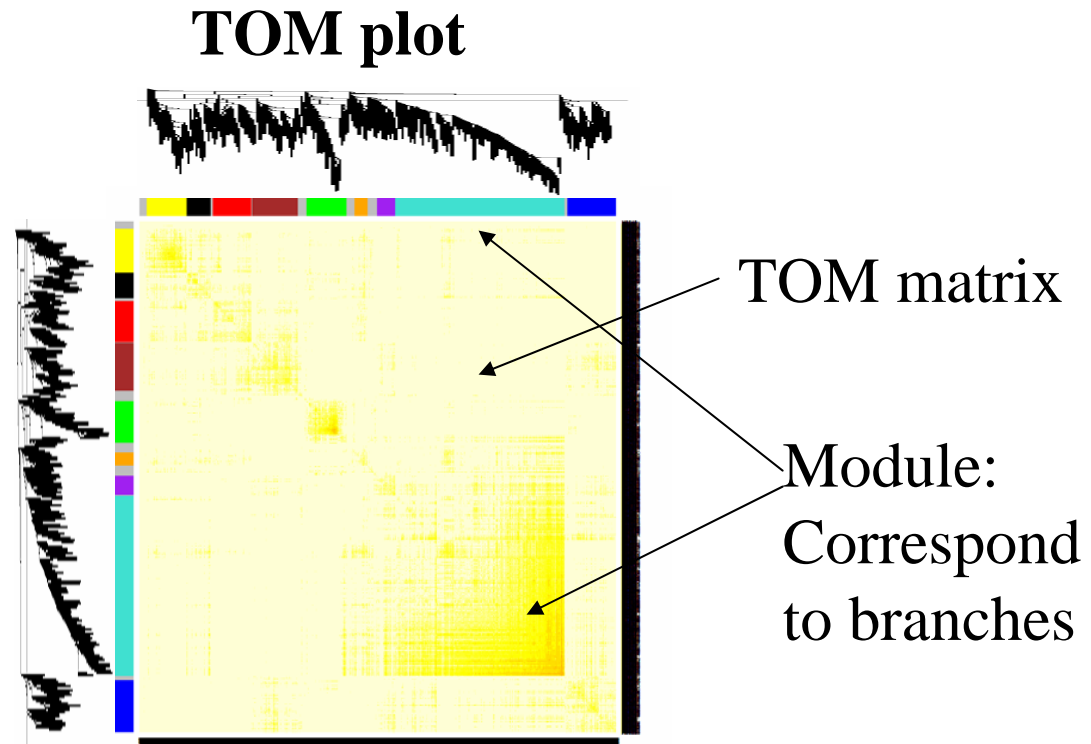
- An important aim of metabolic network analysis is to detect subsets (modules) of nodes that are tightly connected to each other.
- We adopt the definition of Ravasz et al (2002): modules are groups of nodes that have high topological overlap.

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

Genes correspond to rows and columns

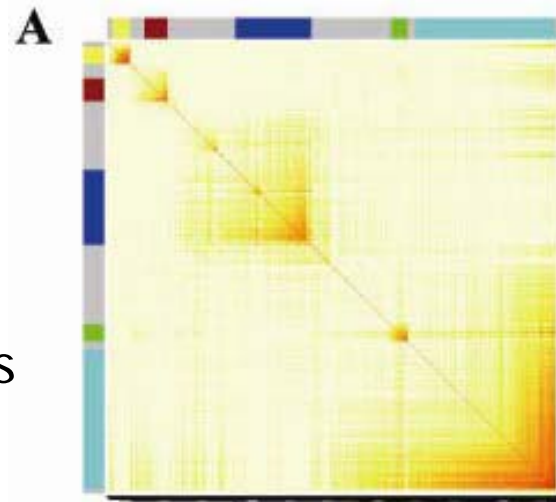
Hierarchical clustering dendrogram



Different Ways of Depicting Gene Modules

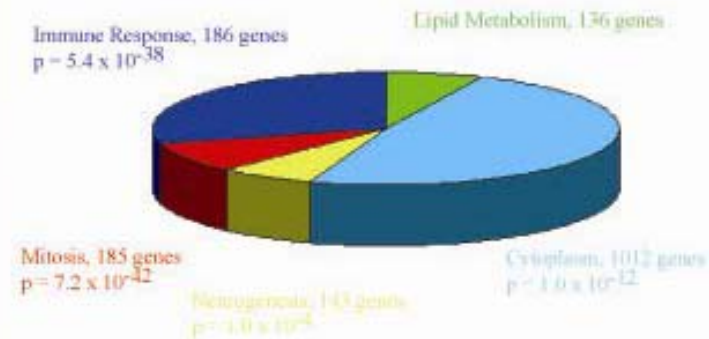
Topological Overlap Plot

- 1) Rows and columns correspond to genes
- 2) Red boxes along diagonal are modules
- 3) Color bands=modules



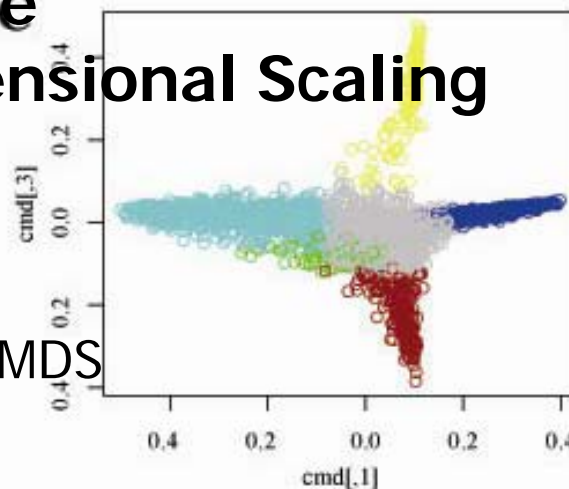
Gene Functions

B



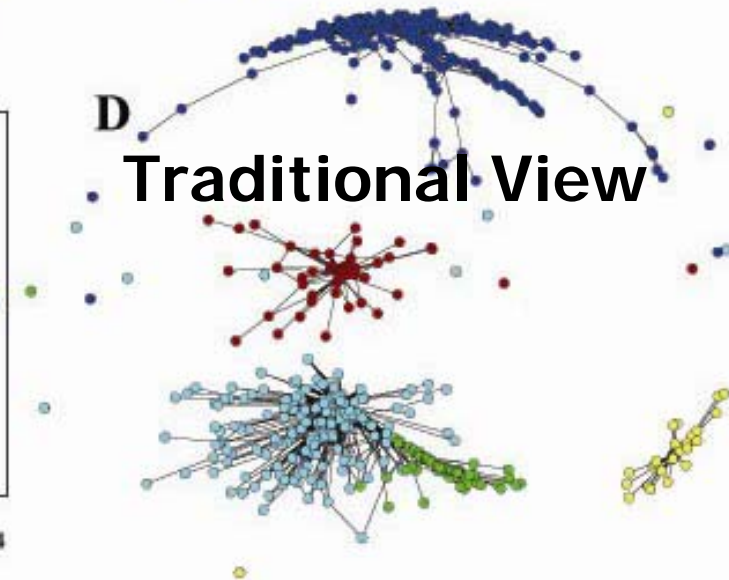
We propose Multi Dimensional Scaling

Idea:
Use network distance in MDS



D

Traditional View

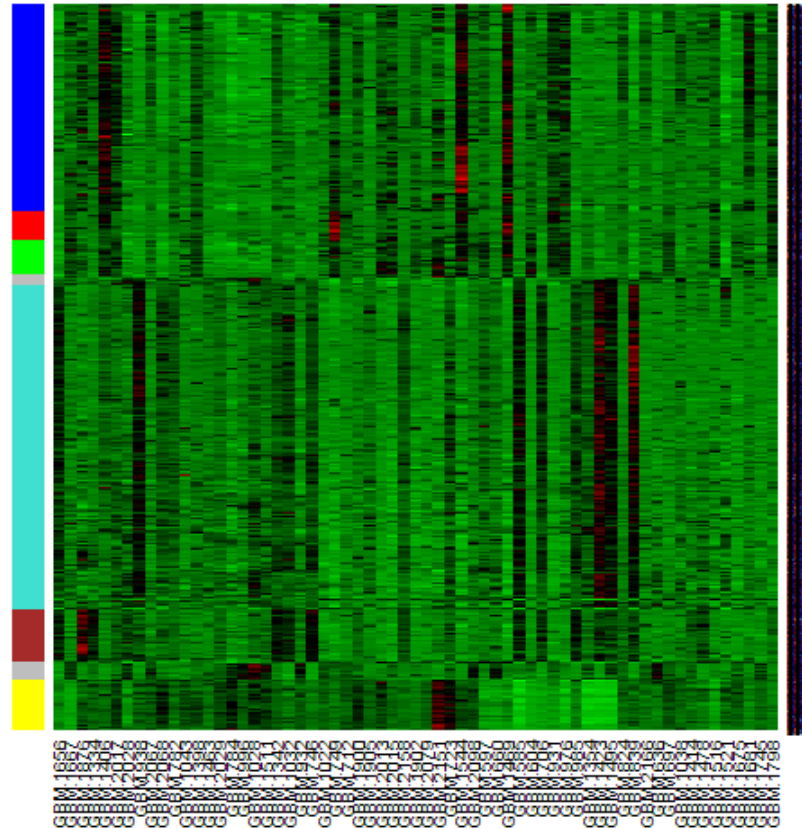


More traditional view of gene co-expression modules

Columns=Brain tissue samples

Rows=Genes

Color band indicates
module membership

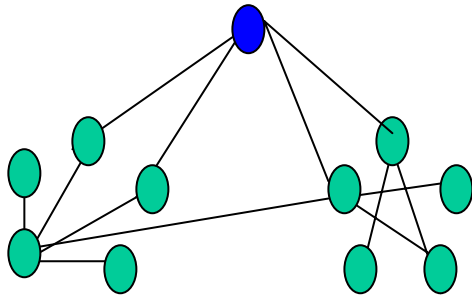


Message: characteristic vertical bands indicate tight co-expression of module genes

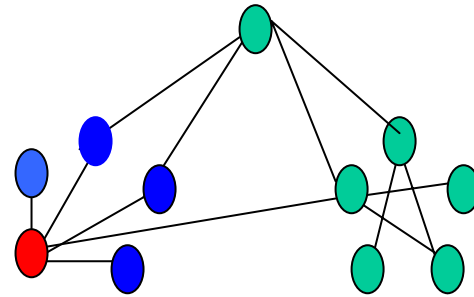
Module-Centric View of Networks

Intra-modular connectivity is biologically and mathematically more meaningful than whole network connectivity

- Whole network connectivity



- Intramodular connectivity



Example: our yeast network example (next slide) that relates intramodular connectivity to knock-out essentiality.

Message: module definition is an important first step towards defining the concept of *intramodular* connectivity.

Yeast network application

M Carlson, B Zhang, Z Fang, PS Mischel, S Horvath, SF Nelson (2006) Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks", BMC Genomics 2006, 7:40

<http://www.biomedcentral.com/1471-2164/7/40/>

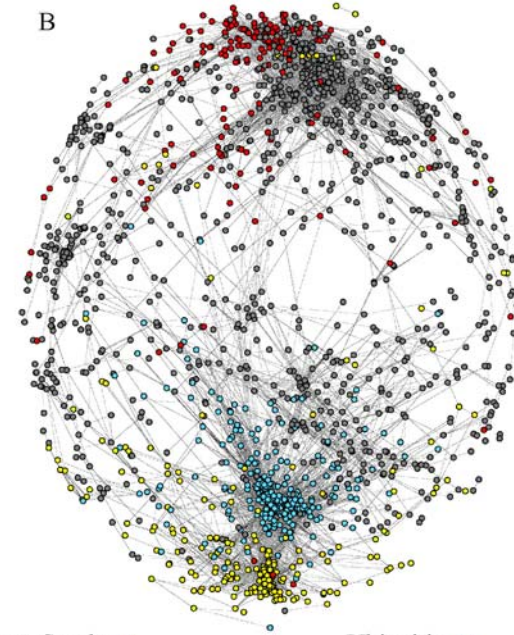
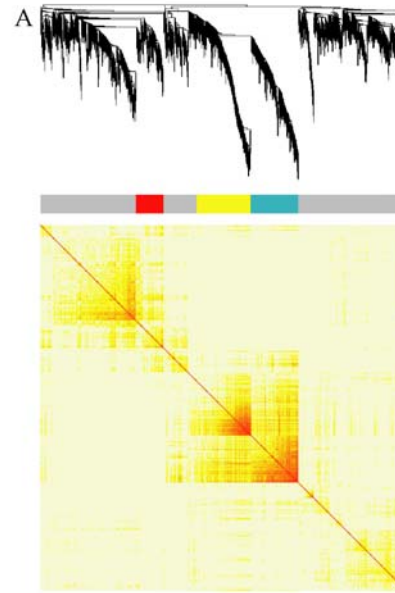
Yeast Data Analysis

Marc Carlson et al (2006)

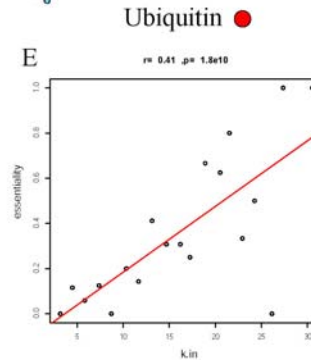
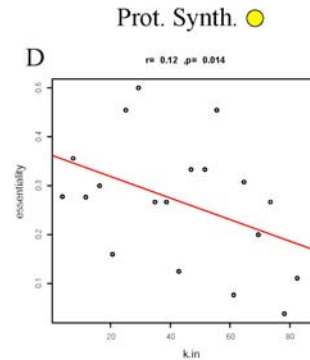
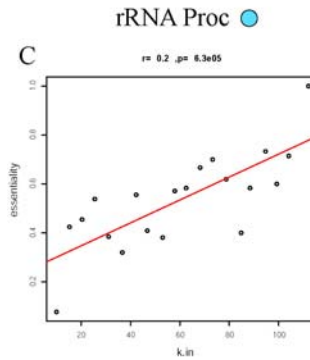
Within Module Analysis

Findings

- 1) The intramodular connectivities are related to gene essentiality:
 - in the turquoise module there is a positive relationship.
 - In the yellow module there may be an inverse relationship.
- 2) Modules are highly preserved across different data sets



Prob(Essential)



Connectivity k

Topological Overlap Matrix
for Multiple Nodes
Ai Li and SH

Topological Overlap Measure for 2 Nodes

The topological overlap matrix can be re - written as follows

$$TOM_{ij} = \frac{\sum_{u \neq i, j} a_{iu} a_{ju} + a_{ij}}{\min(\sum_{u \neq i, j} a_{iu}, \sum_{u \neq i, j} a_{ju}) + 1}$$

One can easily prove that if $0 \leq a_{ij} \leq 1$ then

$$0 \leq TOM_{ijk} \leq 1$$

The topological overlap of two nodes reflects their similarity in terms of the commonality of the nodes they connect to.

Topological Overlap Matrix for 3 nodes

$$MTOM_{ijk} = \frac{\sum_{u \neq i, j, k} a_{iu} a_{ju} a_{ku} + a_{ij} + a_{ik} + a_{kj}}{\min\left(\sum_{u \neq i, j, k} a_{iu} a_{ju}, \sum_{u \neq i, j, k} a_{iu} a_{ku}, \sum_{u \neq i, j, k} a_{ju} a_{ku}\right) + 3}$$

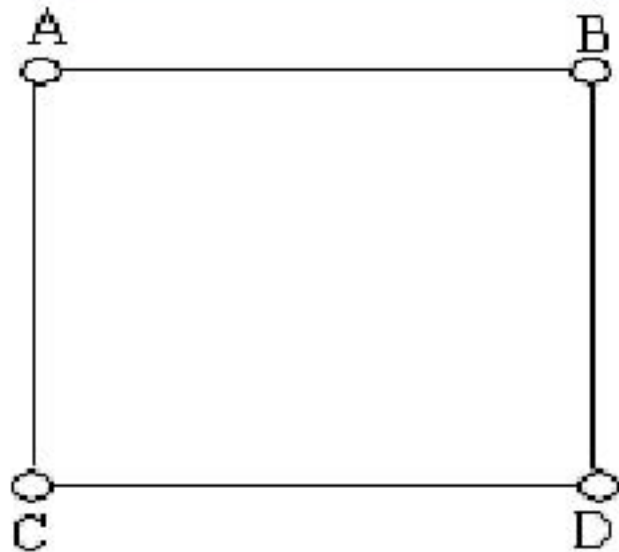
One can easily prove that if $0 \leq a_{ij} \leq 1$ then

$$0 \leq MTOM_{ijk} \leq 1$$

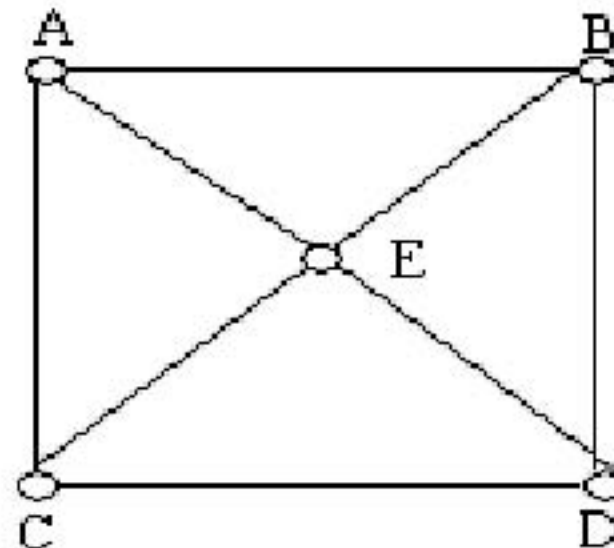
Topological Overlap Matrix for 4 nodes

$$MTOM_{ijkl} = \frac{\sum_{u \neq i, j, k, l} a_{iu} a_{ju} a_{ku} a_{lu} + a_{ij} + a_{ik} + a_{il} + a_{jk} + a_{jl} + a_{kl}}{\min\left(\sum_{u \neq i, j, k, l} a_{iu} a_{ju} a_{ku}, \sum_{u \neq i, j, k, l} a_{iu} a_{ju} a_{lu}, \sum_{u \neq i, j, k, l} a_{iu} a_{ku} a_{lu}, \sum_{u \neq i, j, k, l} a_{ju} a_{ku} a_{lu}\right) + \binom{4}{2}}$$

Topological Overlap for 4 nodes A,B,C,D



$$\frac{0+4}{0+6} = 0.667$$



All share the node E, MTOM increases

$$\frac{1+4}{1+6} = 0.714$$

Using the multinode topological
measure for neighborhood analysis

Many biological questions can be interpreted as neighborhood analysis

Abstract definition: Find the network neighborhood of an initial (seed) set of highly interconnected nodes.

- Examples

- A) Drosophila protein-protein interaction network: Find the neighborhood of a set of essential proteins. Hypothesis: it should be enriched with essential proteins as well → predicting knock out effect
- B) Consider survival time as an idealized gene expression profile. Find the neighborhood of genes in the corresponding gene co-expression network. Hypothesis: it should be enriched with prognostic genes that are associated with survival time → variable selection
- C) Yeast protein network: Find the neighborhood of a set of cell-cycle related genes. Hypothesis: it should be enriched with other cell-cycle related genes → useful for annotation

Recursive and non-recursive approaches for defining an MTOM neighborhood of size S

Recursive approach

- Input a seed of starting nodes and the neighborhood size S
- For each node outside of the current neighborhood compute its MTOM value with the current version of the neighborhood.
- Add the node with highest MTOM value to the neighborhood.
- Repeat b) and c) until the neighborhood size is reached.

Advantage: results in neighborhoods with high MTOM values

Disadvantage: computationally intensive.

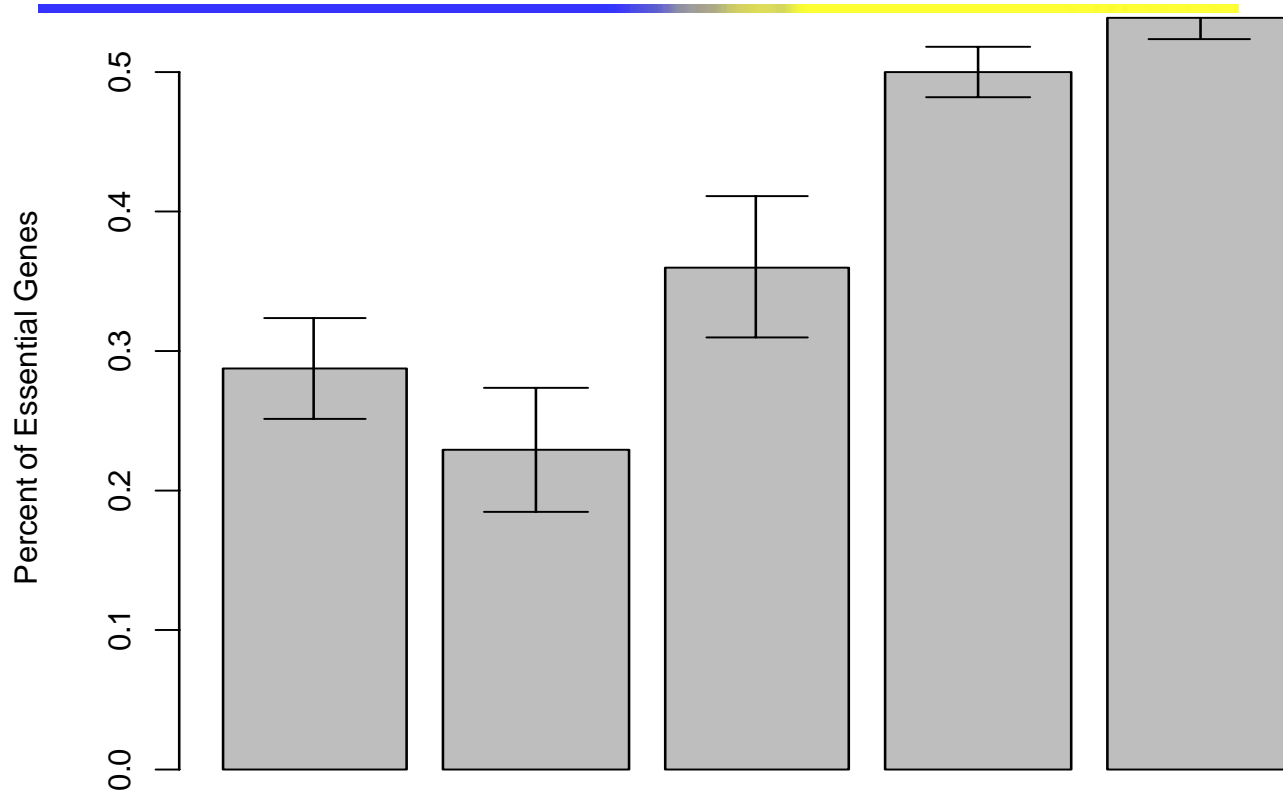
Non-recursive approach: carry out step b) and select a neighborhood based on the highest MTOM values with the seed

Fly protein-protein network analysis (BioGrid Data)

Goal: study the neighborhood of highly connected essential genes
Record the proportion of genes that remain essential as a function of
different seed genes

Drosophila, protein-protein network

Percent of Essential Genes , p-value = $2.0e-45$



Screening: Naïve, non-rec, recurs, recurs, recurs

Initial: 1 1 1 2 3

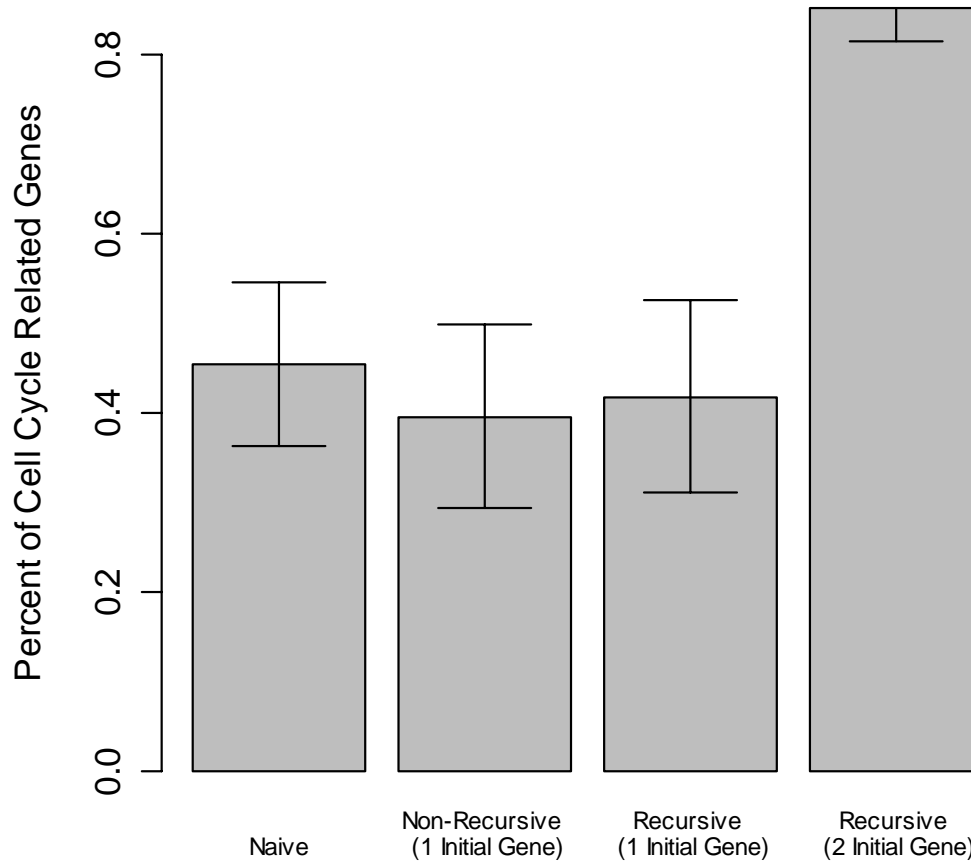
Using 3 genes as seed of recursive MTOM analysis leads to neighborhoods with the highest percentage of essential genes

Applications: Yeast Protein-Protein Network, cell cycle gene prediction

- This network is build up based on the yeast protein-protein interaction (PPI) from the Munich Information Center for Protein Sequence (MIPS)
- 3858 proteins with 7196 pair-wise physical interactions
- 101 cell cycle related genes found in Kyoto Encyclopedia of Genes and Genome (KEGG).
- We considered each of the 101 cell cycle genes as initial protein.
- Neighborhood size= 10
- MTOM based recursive approach and other approaches

Yeast Protein-Protein Network: Neighborhood analysis for predicting cell cycle related genes.

Percent of Cell Cycle Related Genes , p-value= 1.9e-08



Recursive MTOM analysis with 2 initial genes works best

Brain Cancer Network Application I:

Finding the neighborhood of 5 cancer genes in a brain cancer gene co-expression network

- A major advantage of the MTOM approach is that it allows one to input more than 1 probe set as initial neighborhood.
- In this application, we were interested in finding the neighborhood of five highly correlated cell mitosis related cancer genes: TOP2A, Rac1, TPX2, EZH2 and KIF14.
- Neighborhood size = 20
- Out of 20 probes 13 are cancer related.

Brain Cancer Network Application II: Finding the Neighborhood of a Clinical Outcome (Survival Time)

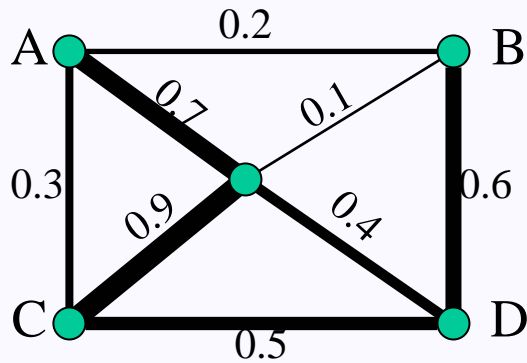
- Recursive MTOM based neighborhood of size $S=20$ of the patient survival time (TTS)
- Result: highly enriched in cancer- and neuron related genes:
 - **11** probe sets are related to neuron cells
 - **10** probe sets are related to cancers.
- A standard approach which simply selects a neighborhood on the basis of the absolute values of the correlations between gene expression profile and survival time, leads to a neighborhood with fewer cancer- and neuron related genes. only 4 probe sets are related to neuron cells and 6 probe sets are related to cancer.

Using local permutations to determine the neighborhood size

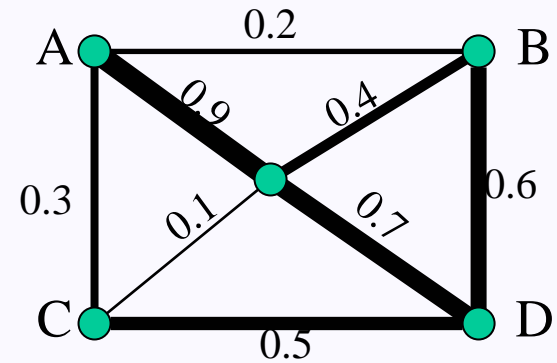
Local Permutation to Choose a Neighborhood Size

- The permutation test is based on comparing the MTOM values of the observed network neighborhood to those of permuted versions of the network.
- Global (whole network) permutations*: it can noise up the module structure of the network and thus leads to very large neighborhoods comprised of possibly thousands of nodes.
- Local Permutation*: for a given node in the initial neighborhood set, it permutes the adjacencies with all other nodes while keeping the remaining adjacencies intact.

Local Permutation to Choose a Neighborhood Size



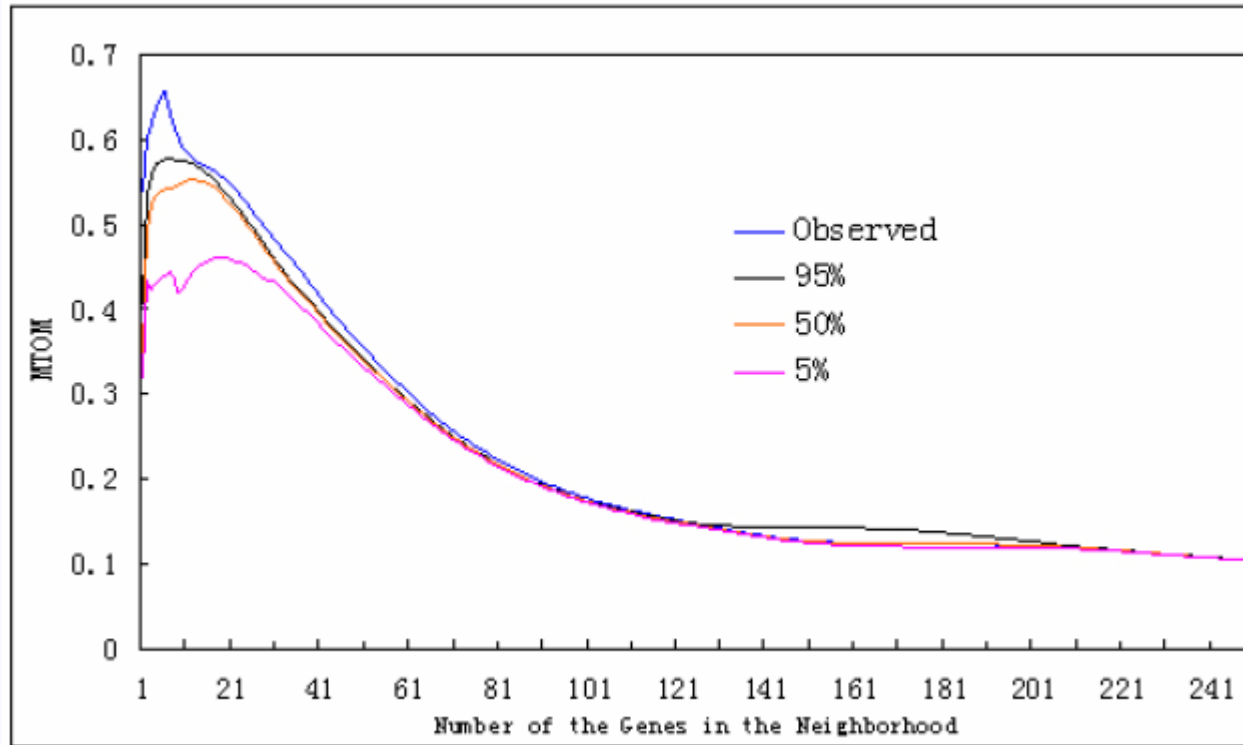
Original Weighted Network



One Possible Local Permutation

Local permutation also works for unweighted network

Local Permutation to Choose a Neighborhood Size



Comparing the MTOM values of the observed network with locally permuted versions to determine the neighborhood size S

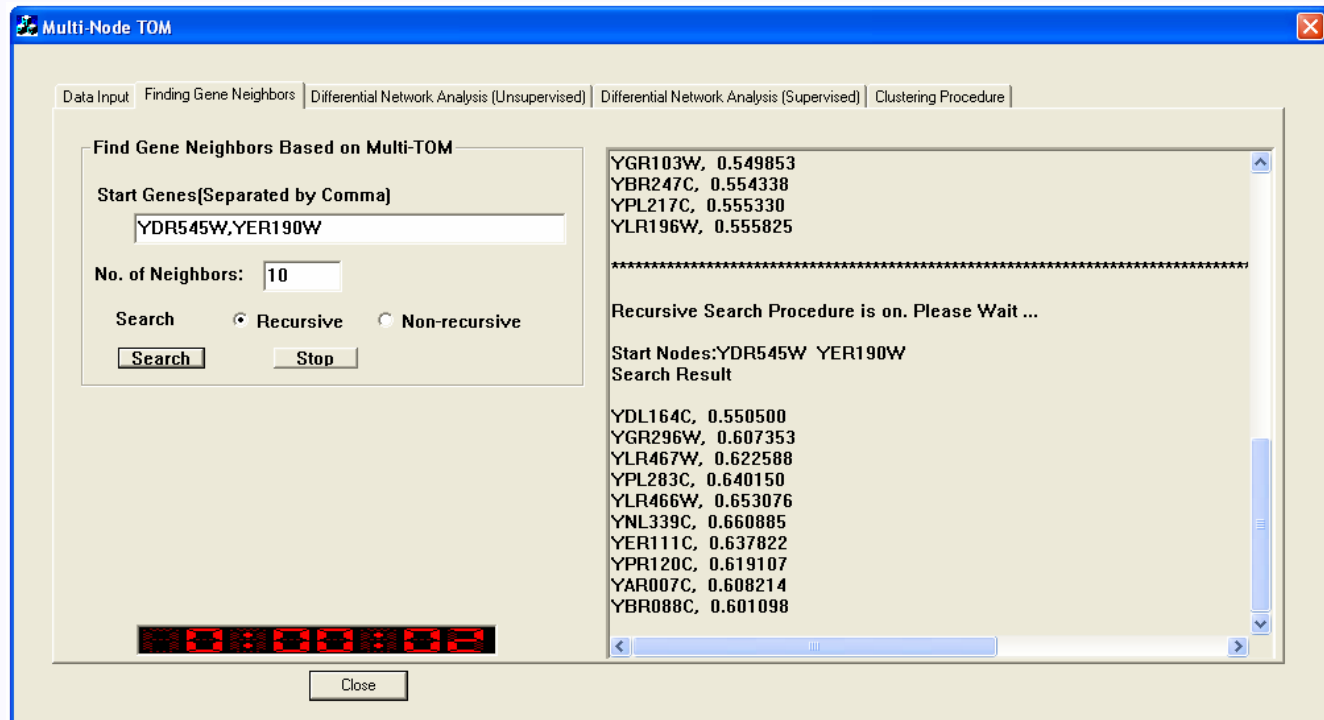
Discussion

- In several applications, we find that the topological overlap matrix leads to biologically meaningful results.
- Since it considers shared neighbors, it tends to be more robust to spurious connections.
- We described how to use the multinode TOM in network neighborhood analysis and provide empirical evidence that it produces biologically meaningful results.
- In our real data applications we find that the results are particularly good if
 - More than 1 node is used in the initial seed set
 - Highly connected hub nodes are used as seed
 - The nodes in the seed have high topological overlap themselves.
- MTOM can also be used in conjunction with a suitable clustering analysis to define modules.

MTOM Software Availability

Visual C++ implementation of the multinode TOM software can be found here

<http://www.genetics.ucla.edu/labs/horvath/MTOM/>



Acknowledgement

Biostatistics/Bioinformatics

- Ai Li, doctoral student UCLA
- Jun Dong, Postdoc UCLA
- Wei Zhao, Postdoc UCLA
- Andy Yip, Assistant Prof, Math, Singapore

Brain Cancer/Yeast

- Marc Carlson, Postdoc, UCLA
- Paul Mischel, Prof, UCLA
- Stan Nelson, Prof, UCLA

Webpages and References

- This talk and relevant R code

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/>

- Main Network Webpage

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

- *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

- *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/>