

Lecture 10: Network Analysis

BIOINF3005/7160: Transcriptomics Applications

Zhipeng Qu

School of Biological Sciences,
The University of Adelaide

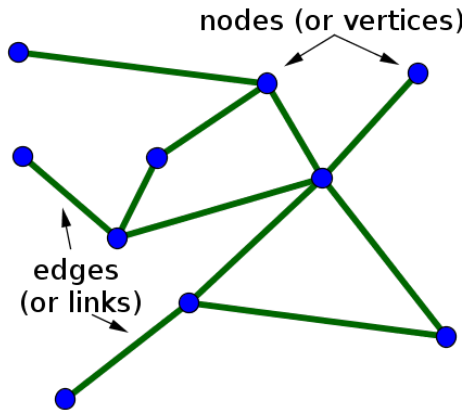
May 18th, 2020

- Introduction of network theory
- Types of biological networks
- Co-expression network analysis (WGCNA)

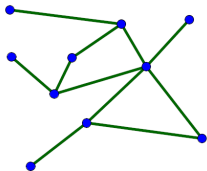
Part 1, Introduction of network theory

What is network?

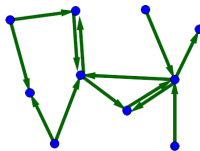
- Network theory is part of graph theory
- It has two basic components
 - ① Nodes (N): or vertices (genes, people, cities, PCs, etc.)
 - ② Edges (E): or links/connections/interactions connecting nodes



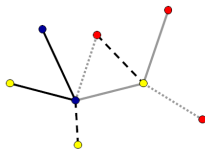
Types of networks



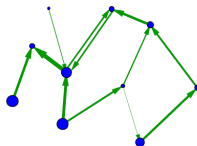
(A) Undirected network



(B) Directed network



(C) An undirected network where the nodes and edges have different types, as indicated by their colors and line styles.

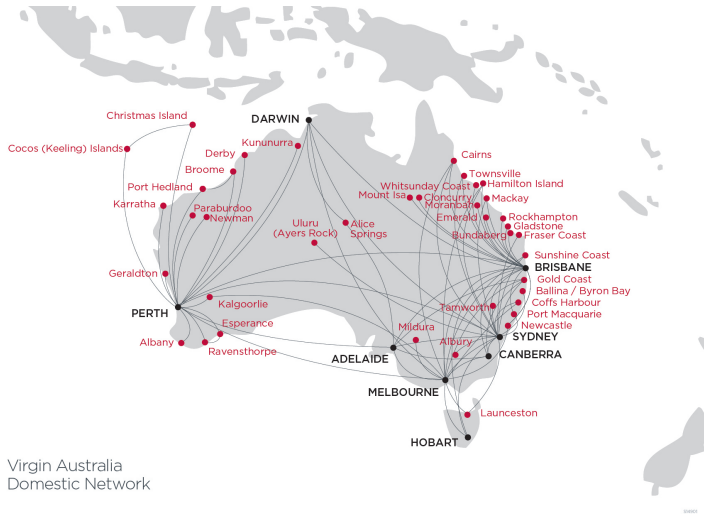


(D) A directed network where the edges and nodes have different weights, as indicated by their sizes.

Example 1 – Social network



Example 2 – Flight network

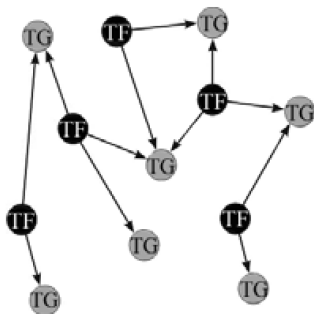


Part 2, Biological networks

Four types of biological networks

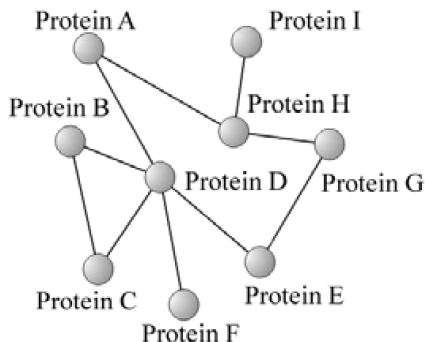
- Transcriptional regulatory network
- Protein-protein interaction (PPI) network
- Metabolic network
- Gene co-expression network

Transcriptional regulatory network



A transcriptional regulatory network has two components: transcription factor (TF) and target genes (TG), where TF regulates the transcription of TGs

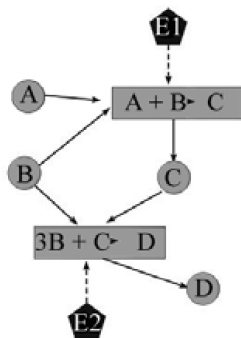
Protein-protein interaction network



Protein-protein interaction networks: two proteins are connected if there is a docking between them

Costa, Luciano da F., Rodrigues, Francisco A., & Cristino, Alexandre S.. (2008). Complex networks: the key to systems biology. *Genetics and Molecular Biology*, 31(3), 591-601

Metabolic network



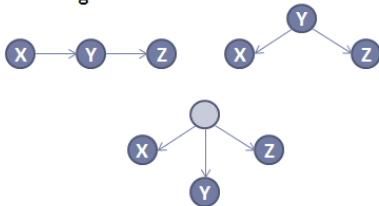
A metabolic network is constructed considering the reactants, chemical reactions and enzymes

Gene co-expression network

Gene Co-expression



Gene Regulation



Genes are found to be co-expressed across samples, it is not determined whether X activates Y and Y activates Z, or Y activates X and Z, or another gene activates three of them

Part 3, Co-expression network analysis

WGCNA

- Weighted Gene Co-expression Network Analysis
- R package
 - Can be applied on different types of high-throughput biological data
 - To understand the “global” expression patterns (modules) instead of “individual” genes
 - Biologically meaningful by linking to biological traits
 - Fast and robust

Overview of co-expression network analysis

Construct a gene co-expression network

Rationale: make use of interaction patterns among genes

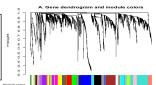
Tools: correlation as a measure of co-expression



Identify modules

Rationale: module (pathway) based analysis

Tools: hierarchical clustering, Dynamic Tree Cut

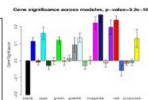


Relate modules to external information

Array Information: clinical data, SNPs, proteomics

Gene Information: ontology, functional enrichment

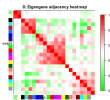
Rationale: find biologically interesting modules



Study module relationships

Rationale: biological data reduction, systems-level view

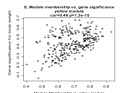
Tools: Eigengene Networks



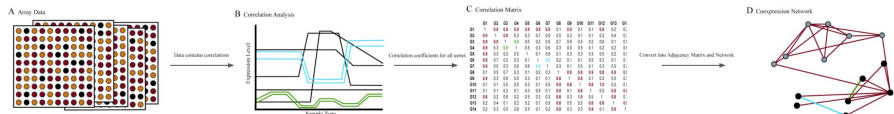
Find the key drivers in interesting modules

Rationale: experimental validation, biomarkers

Tools: intramodular connectivity, causality testing



Steps of constructing co-expression network



- A, get gene expression data (microarrays, RNA-Seq, etc.)
- B, measure co-expression between genes using a correlation coefficient (e.g. Pearson correlation)
- C, transform the correlation matrix to a dichotomized adjacency matrix (unweighted network) or transform to an continuous adjacency matrix (weighted network) using the power adjacency function

Gene expression data

- Microarrays, RNA-Seq and others
- Filter low-expressed genes
- All genes or subset of genes
- Normalization of gene expression data

Co-expression similarity

$$s_{ij} = |\text{cor}(x_i, x_j)|$$

- Co-expression similarity s_{ij} is denoted as the absolute value of the correlation coefficient between the node (gene) i and node (gene) j
- Different correlation measurement methods can be used, such as Pearson correlation, Spearman correlatin, Biweight midcorrelation, mutual information, etc.

Representation of a network -- Adjacency matrix

A network can be represented by an adjacency matrix $A[a_{ij}]$, including connection status/strength of node pairs.

- A symmetric matrix with n columns and n rows (n is the number of nodes/genes)
- Diagonal entries of $A[a_{ij}]$ are all 1

Weighted and unweighted networks

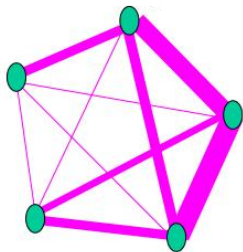
Weighted network and unweighted network

- Entries of $A[a_{ij}]$ in weighted network are continuous values in $[0, 1]$, representing the connection strength of two nodes/genes
- Entries of $A[a_{ij}]$ in unweighted network are either 1 or 0, representing two nodes are or aren't connected

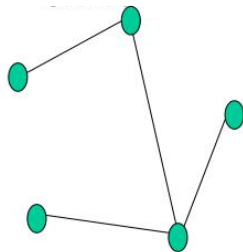
$$\begin{array}{c}
 \left| \begin{array}{cccc}
 1 & 0.8 & 0.1 & 0.2 \\
 0.8 & 1 & 0.3 & 0.6 \\
 0.1 & 0.3 & 1 & 0.1 \\
 0.2 & 0.6 & 0.1 & 1
 \end{array} \right|
 \qquad
 \left| \begin{array}{cccc}
 1 & 1 & 0 & 0 \\
 1 & 1 & 0 & 1 \\
 0 & 0 & 1 & 0 \\
 0 & 1 & 0 & 1
 \end{array} \right|
 \end{array}$$

Weighted and unweighted networks

Weighted network



Unweighted network



Unsigned and signed correlation networks

Correlation can be positive or negative $[-1, 1]$

There are two ways to handle correlation signs in weighted co-expression network

- **Unsigned network** uses absolute correlation values

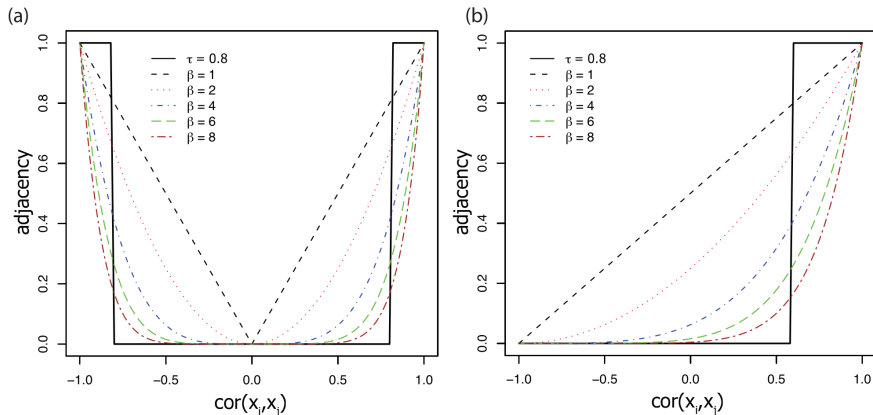
$$a_{ij} = |cor(x_i, x_j)|^\beta$$

- **Signed network** preserves sign info using a shift value

$$a_{ij} = \left| \frac{1 + cor(x_i, x_j)}{2} \right|^\beta$$

What is β ?

Comparison of adjacency and correlation in unsigned and signed network



Mason, M.J., Fan, G., Plath, K. et al. Signed weighted gene co-expression network analysis of transcriptional regulation in murine embryonic stem cells. *BMC Genomics* 10, 327 (2009)

What is β ?

β is the power in soft thresholding with a power function, and it has following advantages in co-expression network analysis:

- **Robustness:** results are highly robust with respect to the choice of the power β (Zhang et al 2005)
- **Calibration/normalization** of different networks becomes straightforward
- **Module preservation** statistics are particularly sensitive for measuring connectivity preservation in weighted networks
- **Math reason** (Horvath S and Dong J 2008)

How to choose β ?

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*, 4(1)

Horvath S, Dong J (2008) Geometric Interpretation of Gene Coexpression Network Analysis. *PLOS Computational Biology* 4(8): e1000117

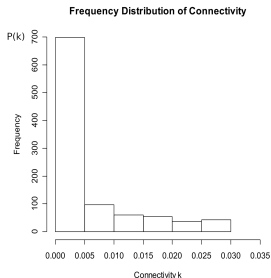
Connectivity in co-expression network

$$k_i = \sum_{j \neq i} a_{ij}$$

Row sum of the adjacency matrix $A[a_{ij}]$ as the gene connectivity

- Number of direct neighbors in unweighted networks
- Sum of connection strengths to other nodes in weighted networks

Scale free topology

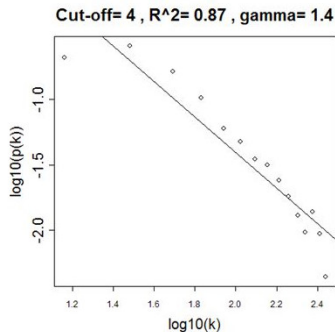


- Scale free topology refers to the frequency distribution of the connectivity k
- Gene/node connectivity k follows power law
- $p(k)$ is the proportion of nodes that have connectivity k
- When the network grows, the underlying structure remains the same

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*, 4(1)

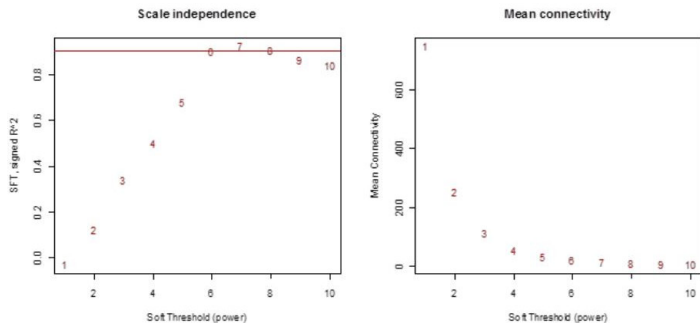
How to check scale free topology

Idea: Log transformation $p(k)$ and k and look at scatter plot



Linear model fitting R^2 index can be used to quantify goodness of fit

Zhang, B., & Horvath, S. (2005). A General Framework for Weighted Gene Co-Expression Network Analysis, *Statistical Applications in Genetics and Molecular Biology*, 4(1)

Choose power β to satisfy scale free topology

- In practice, we use the lowest value where the curve starts to **saturate**
- How to measure interconnectedness in a network?

Topological overlap matrix and dissimilarity matrix

$$TOM_{ij} = \frac{\sum_u a_{iu}a_{uj} + a_{ij}}{\min(k_i, k_j) + 1 - a_{ij}} \subseteq [0, 1]$$

$\sum_u a_{iu}a_{uj}$ denotes the number of shared neighbors of node i and j in an unweighted network, and k denotes the gene connectivity which is the row sum of adjacencies

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

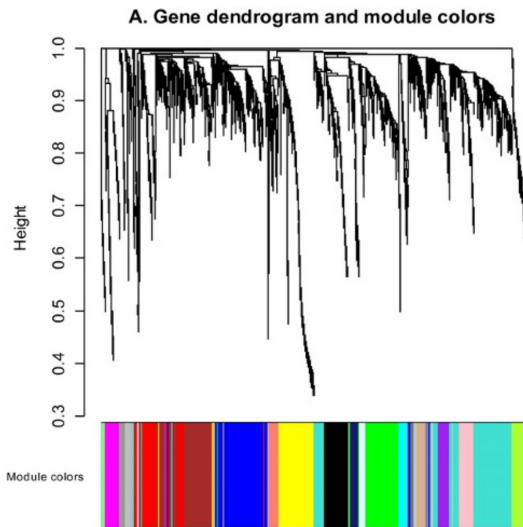
Topological overlap matrix TOM_{ij} is transformed to dissimilarity matrix $DistTOM_{ij}$ for clustering

How to detect co-expression network modules (clusters)

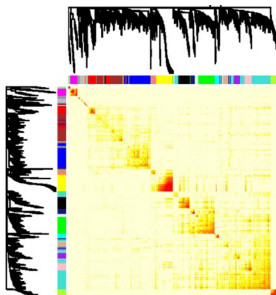
Definition of module in co-expression network analysis

- Perform hierarchical clustering based on the topological overlap dissimilarity matrix $DistTOM_{ij}$
- Define modules as branches of the resulting cluster tree
- Modules are either labeled by integers (1, 2, 3 ...) or equivalently by colors (turquoise, blue, brown, etc)

Gene dendrogram and modules



Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008)

Using the topological overlap matrix TOM_{ij} to cluster genes

- Use heatmap to visualize TOM_{ij} in this TOM plot
- Genes correspond to rows and columns
- Modules correspond to red squares along the diagonal

How to summarize the expression profiles in a module?

Langfelder, P., Horvath, S. WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008)

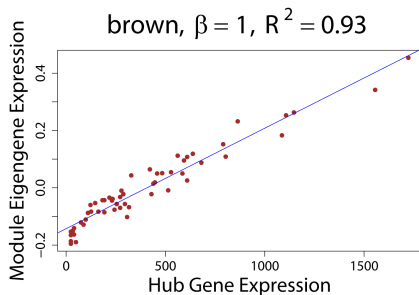
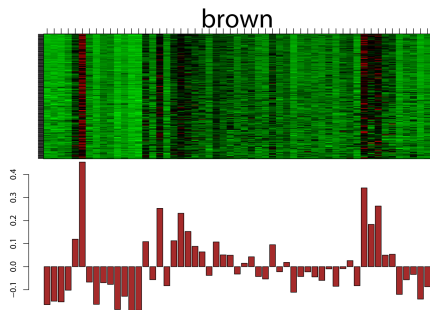
Two answers

Question: How to summarize the expression profiles in a module?

- **Math answer:** Module eigengene (first principal component)
- **Network answer:** The most highly connected intramodular hub gene

Module Eigengene

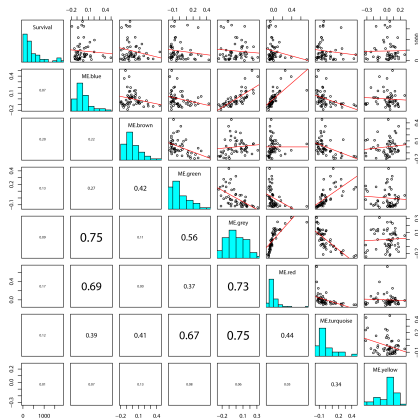
Module eigengene is the measurement of over expression, which equals the average redness in the heatmap



Eigengenes can be used for?

- Measure the correlation between modules, e.g. allows one to determine whether modules should be merged
- They allow one to relate modules to clinical traits and SNPs

Eigengenes can be used for?



The pairwise scatter plots among the module eigengenes of different modules and cancer survival time T

[How to find biologically meaningful module?](#)

Horvath S, Dong J (2008) Geometric Interpretation of Gene Coexpression Network Analysis. PLOS Computational Biology 4(8): e1000117

Trait-based gene significance measure in co-expression network

We define the trait-based gene significance measure by raising the correlation between the i th gene expression profile x_i and the clinical trait T to a power β

$$GS_i = |\text{cor}(x_i, T)|^\beta$$

- The higher GS_i the more *biologically* significant is gene i
- T denotes sample trait, which can be quantitative (e.g., body weight) or binary (e.g., disease status)

Module significance measure

We define a module significance measure as the average gene significance (GS_i) of the genes in a module

$$ModuleSignif = \frac{\sum_i GS_i}{n}$$

- Often highly related to the correlation between module eigengene and trait
- Can be used to identify biologically significant subnetworks or pathways

How to find central players (hubs) in an interesting module?

Eigengene-based intramodular connectivity measure (Module membership measure)

$$kME_i = ModuleMembership(i) = cor(x_i, ME)$$

- $kME(i)$ is simply the correlation between the i -th gene expression profile (x_i) and the module eigengene (ME)
- kME close to 1 means that the gene is a hub gene
- Can be used to find genes that are members of two or more modules (fuzzy clustering)
- Module eigengene can be interpreted as the most highly connected gene

Lecture summary -- how to perform gene co-expression network analysis with WGCNA

- Get gene expression data: microarrays, RNA-Seq, etc.
- Calculate similarity/correlation matrix s_{ij}
- Transform to adjacency matrix a_{ij}
 - Signed and unsigned
 - Scale free topology and power β
- Transform to topological overlap matrix TOM_{ij} and dissimilarity matrix $DistTOM_{ij}$
- Co-expression modules
 - how to summarize the expression profiles in a module? -- Eigengene
 - how to find biologically interesting module? -- Gene significance and module significance measure
 - how to find hub genes in a module? -- Eigengene-based intramodular measure

Thank you!