An Overview of Weighted Gene Co-Expression Network Analysis

Steve Horvath
University of California, Los Angeles
Contents

• How to construct a weighted gene co-expression network?
• Why use soft thresholding?
• How to detect network modules?
• How to relate modules to an external clinical trait?
• What is intramodular connectivity?
• How to use networks for gene screening?
• How to integrate networks with genetic marker data?
• What is weighted gene co-expression network analysis (WGCNA)?
• What is neighborhood analysis?
Philosophy of Weighted Gene Co-Expression Network Analysis

• Understand the “system” instead of reporting a list of individual parts
  – Describe the functioning of the engine instead of enumerating individual nuts and bolts
• Focus on modules as opposed to individual genes
  – this greatly alleviates multiple testing problem
• Network terminology is intuitive to biologists
How to construct a weighted gene co-expression 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
Steps for constructing a co-expression network

A) Microarray gene expression data
B) Measure concordance of gene expression with a Pearson correlation
C) The Pearson correlation matrix is either dichotomized to arrive at an adjacency matrix → unweighted network
   Or transformed continuously with the power adjacency function → weighted network
Power adjacency function results in a weighted gene network

\[ a_{ij} = |\text{cor}(x_i, x_j)\|^\beta \]

Often choosing beta=6 works well but in general we use the "scale free topology criterion" described in Zhang and Horvath 2005.
Comparing adjacency functions

Power Adjacency vs Step Function
Comparing the power adjacency function to the step function

- While the network analysis results are usually highly robust with respect to the network construction method there are several reasons for preferring the power adjacency function.
  - Empirical finding: Network results are highly robust with respect to the choice of the power beta.
  - Theoretical finding: Network Concepts make more sense in terms of the module eigengene.
How to detect network modules?
Module Definition

• Numerous methods have been developed
• Here, we use average linkage hierarchical clustering coupled with the topological overlap dissimilarity measure.
• Once a dendrogram is obtained from a hierarchical clustering method, we choose a height cutoff to arrive at a clustering.
• Modules correspond to branches of the dendrogram
The topological overlap dissimilarity is used as input of hierarchical clustering.

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

\[ DistTOM_{ij} = 1 - TOM_{ij} \]

- Generalized in Zhang and Horvath (2005) to the case of weighted networks
- Generalized in Yip and Horvath (2006) to higher order interactions
Using the topological overlap matrix (TOM) to cluster genes

- Here modules correspond to branches of the dendrogram

TOM plot

Genes correspond to rows and columns

Hierarchical clustering dendrogram

TOM matrix

Module: Correspond to branches
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

Multi Dimensional Scaling

Idea:
Use network distance in MDS
Heatmap view of module

Columns = tissue samples

Rows = Genes
Color band indicates module membership

Message: characteristic vertical bands indicate tight co-expression of module genes
Module Eigengene = measure of over-expression = average redness

Rows, = genes, Columns = microarray

The brown module eigengenes across samples
Module eigengenes can be used to determine whether 2 modules are correlated. If correlation of MEs is high-> consider merging.
How to relate modules to external data?
Clinical trait (e.g. case-control status) gives rise to a gene significance measure

- Abstract definition of a gene significance measure
  - GS(i) is non-negative,
  - the bigger, the more *biologically* significant for the i-th gene

Equivalent definitions

- GS.ClinicalTrait(i) = \(|\text{cor}(x(i),\text{ClinicalTrait})|\)
  where \(x(i)\) is the gene expression profile of the i-th gene
- GS(i) = \(|T\text{-test}(i)|\) of differential expression between groups defined by the trait
- GS(i) = \(-\log(p\text{-value})\)
A SNP marker naturally gives rise to a measure of gene significance

\[ GS\_SNP(i) = |\text{cor}(x(i), SNP)|. \]

- Additive SNP marker coding: AA->2, AB->1, BB->0
- Absolute value of the correlation ensures that this is equivalent to AA->0, AB->1, BB->2
  - Dominant or recessive coding may be more appropriate in some situations
  - Conceptually related to a LOD score at the SNP marker for the i-th gene expression trait
A gene significance naturally gives rise to a module significance measure

- Define module significance as mean gene significance
- Often highly related to the correlation between module eigengene and trait
Important Task in Many Genomic Applications: Given a network (pathway) of interacting genes how to find the central players?
Does this map tell you which cities are important? This one does!

The nodes with the largest number of links (connections) are most important!

**Slide courtesy of A Barabasi**
What is intramodular connectivity?
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

\[ k_i = \sum_j a_{ij} \]
Gene significance versus intramodular connectivity kIN
How to use networks for gene screening?
Intramodular connectivity $kIN$ versus gene significance $GS$

• Note the relatively high correlation between gene significance and intramodular connectivity in some modules

• In general, $kIN$ is a more reliable measure than $GS$

• In practice, a combination of $GS$ and $k$ should be used

• Module eigengene turns out to be the most highly connected gene (under mild assumptions)
What is weighted gene co-expression network analysis?
**Construct a network**
Rationale: make use of interaction patterns between genes

**Identify modules**
Rationale: module (pathway) based analysis

**Relate modules to external information**
Array Information: Clinical data, SNPs, proteomics
Gene Information: gene ontology, EASE, IPA
Rationale: find biologically interesting modules

**Study Module Preservation across different data**
Rationale:
- Same data: to check robustness of module definition
- Different data: to find interesting modules

**Find the key drivers in interesting modules**
Tools: intramodular connectivity, causality testing
Rationale: experimental validation, therapeutics, biomarkers
What is different from other analyses?

- **Emphasis on modules (pathways) instead of individual genes**
  - Greatly alleviates the problem of multiple comparisons
    - Less than 20 comparisons versus 20000 comparisons
- **Use of intramodular connectivity to find key drivers**
  - Quantifies module membership (centrality)
  - Highly connected genes have an increased chance of validation
- **Module definition is based on gene expression data**
  - No prior pathway information is used for module definition
  - Two module (eigengenes) can be highly correlated
- **Emphasis on a unified approach for relating variables**
  - Default: power of a correlation
  - Rationale:
    - puts different data sets on the same mathematical footing
    - Considers effect size estimates (cor) and significance level
    - p-values are highly affected by sample sizes (cor=0.01 is highly significant when dealing with 100000 observations)
- **Technical Details: soft thresholding with the power adjacency function, topological overlap matrix to measure interconnectedness**
Case Study 1:
Finding brain cancer genes
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**

**Multi Dimensional Scaling**

**Traditional View**
Comparing the Module Structure in Cancer and Normal tissues

55 Brain Tumors

VALIDATION DATA: 65 Brain Tumors

Messages:
1) Cancer modules can be independently validated
2) Modules in brain cancer tissue can also be found in normal, non-brain tissue.

--> Insights into the biology of cancer
Mean Prognostic Significance of Module Genes

Message: Focus the attention on the brown module genes
Module hub genes predict cancer survival

1. Cox model to regress survival on gene expression levels

2. Defined prognostic significance as $-\log_{10}(\text{Cox-p-value})$ the survival association between each gene and glioblastoma patient survival

3. A module-based measure of gene connectivity significantly and reproducibly identifies the genes that most strongly predict patient survival
The fact that genes with high intramodular connectivity are more likely to be prognostically significant facilitates a novel screening strategy for finding prognostic genes

- **Focus on those genes with significant Cox regression p-value AND high intramodular connectivity.**
  - It is essential to take a module centric view: focus on intramodular connectivity of disease related module

- Validation success rate= proportion of genes with independent test set Cox regression p-value<0.05.
- Validation success rate of network based screening approach (68%)
- Standard approach involving top 300 most significant genes: 26%
Validation success rate of gene expressions in independent data

300 most significant genes (Cox p-value<1.3*10^{-3}) — 26%

Network based screening p<0.05 and high intramodular connectivity — 67%
The network-based approach uncovers novel therapeutic targets

Five of the top six hub genes in the mitosis module are already known cancer targets: topoisomerase II, Rac1, TPX2, EZH2 and KIF14. We hypothesized that the 6-th gene ASPM gene is novel therapeutic target. ASPM encodes the human ortholog of a drosophila mitotic spindle protein.

Biological validation: siRNA mediated inhibition of ASPM
Case Study 2

What changed?

• Despite pronounced phenotypic differences, genomic similarity is ~96% (including single-base substitutions and indels)\(^1\)
  – Similarity is even higher in protein-coding regions

Assessing the contribution of regulatory changes to human evolution

• Hypothesis: Changes in the regulation of gene expression were critical during recent human evolution (King & Wilson, 1975)

• Microarrays are ideally suited to test this hypothesis by comparing expression levels for thousands of genes simultaneously
Gene expression is more strongly preserved than gene connectivity

Hypothesis: molecular wiring makes us human

Raw data from Khaitovich et al., 2004

Mike Oldham
$p = 1.33 \times 10^{-4}$  

$p = 8.93 \times 10^{-4}$  

$p = 1.35 \times 10^{-6}$  

$p = 1.33 \times 10^{-4}$
Connectivity diverges across brain regions whereas expression does not
Conclusions: chimp/human

- Gene expression is highly preserved across species brains
- Gene co-expression is less preserved
- Some modules are highly preserved
- Gene modules correspond roughly to brain architecture
- Species-specific hubs can be validated in silico using sequence comparisons
What is neighborhood analysis?
MTOM software

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 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 topological overlap 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.
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

Using 3 genes as seed of recursive MTOM analysis leads to neighborhoods with the highest percentage of essential genes.
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.
Software and Data Availability

• Sample data and R software tutorials can be found at the following webpage

• http://www.genetics.ucla.edu/labs/horvath/CoexpressionNetwork
Acknowledgement

Jason Aten, Jun Dong, Sud Doss, Tom Drake, Tova Fuller, Anatole Ghazalpour, Dan Geschwind, Peter Langfelder, Ai Li, Wen Lin, Jake Lusis, Michael Mason, Paul Mischel, Nicole MacLennan, Ed McCabe, Atila Van Nas, Stan Nelson, Mike Oldham, Roel Ophoff, Anja Presson, Lin Wang, Wei Zhao
A short methodological summary of the publications.

- How to construct a gene co-expression network using the scale free topology criterion? Robustness of network results. Relating a gene significance measure and the clustering coefficient to intramodular connectivity:

- Theory of module networks (both co-expression and protein-protein interaction modules):

- What is the topological overlap measure? Empirical studies of the robustness of the topological overlap measure:

- Software for carrying out neighborhood analysis based on topological overlap. The paper shows that an initial seed neighborhood comprised of 2 or more highly interconnected genes (high TOM, high connectivity) yields superior results. It also shows that topological overlap is superior to correlation when dealing with expression data:

- Gene screening based on intramodular connectivity identifies brain cancer genes that validate. This paper shows that WGCNA greatly alleviates the multiple comparison problem and leads to reproducible findings:

- The relationship between connectivity and knock-out essentiality is dependent on the module under consideration. Hub genes in some modules may be non-essential. This study shows that intramodular connectivity is much more meaningful than whole network connectivity:

- How to integrate SNP markers into weighted gene co-expression network analysis? The following 2 papers outline how SNP markers and co-expression networks can be used to screen for gene expressions underlying a complex trait. They also illustrate the use of the module eigengene based connectivity measure kME:

- The following application presents a 'supervised' gene co-expression network analysis. In general, we prefer to construct a co-expression network and associated modules without regard to an external microarray sample trait (unsupervised WGCNA). But if thousands of genes are differentially expressed, one can construct a network on the basis of differentially expressed genes (supervised WGCNA):

- The following paper presents a differential co-expression network analysis. It studies module preservation between two networks. By screening for genes with differential topological overlap, we identify biologically interesting genes. The paper also shows the value of summarizing a module by its module eigengene:
THE END