Integrative network based analysis of metabolomic and transcriptomic data for understanding biological mechanism of lung cancer

Christopher Wilson, PhD

Research Fellow, Moffitt Cancer Center

Moffitt Cancer Center

February 28, 2020

Lung Cancer

- Lung cancer is the leading cause of cancer deaths worldwide.
- Common therapy for lung cancer includes surgery, radiation, chemotherapy, and immunotherapy, which are debilitating to the patient can be harmful to healthy surrounding tissue.
- Moreover, immunotherapy has been found to improve the outcome of a small subset of lung cancers.
- There is a need for further understanding of the basic biology to develop targeted therapies for lung cancers.

Pre-clinical Data

- There were 18 mice which comprised 5 knockout groups (p53, DN-p63, TA-p63, DN-p73, TA-p73) and wild type (WT).
- Gene expression and metabolite abundance was measured for each of these mice.

- The following quality control was performed on the gene expression data:
  1. Poorly expressed genes were removed resulting 15192 genes.
  2. VOOM transformation

- The following quality control was performed on the metabolomic abundance data (366 metabolites):
  1. Log2 transform
  2. Averaged technical replicates
  3. Imputed missing data with KNN
  4. Pareto scaling

Multi-omics Data

Trends in Biotechnology
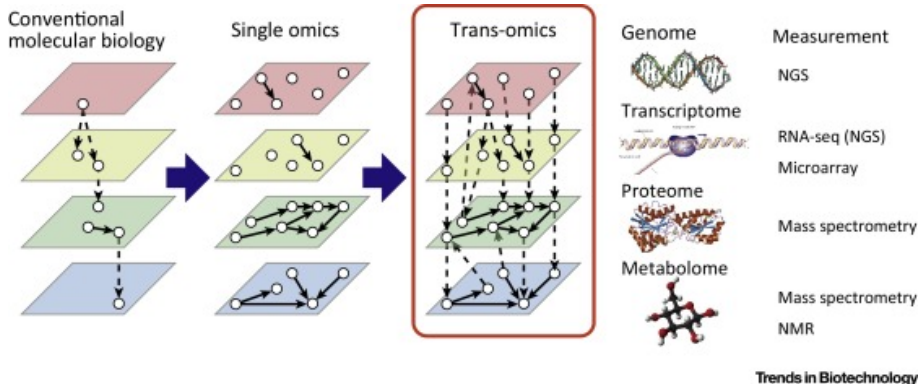
Our Integrative Approach
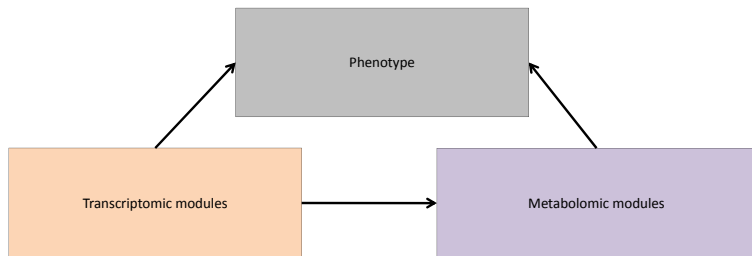
Proposed Method

- Modules are groups of genes or metabolites with similar expression profiles.
- Network based module detection using sparse Gaussian Graphical Models (GGMs).
- Hierarchical clustering on the Topographical Overlap Measure (TOM).
- Use first principal component to summarize each module.
- Association analysis is conducted between module and phenotype, and data types.
- Benefits to Weighted Gene Coexpression Network Analysis (WGCNA):
  - Direct association between elements
  - Sparse network representation
  - Reduces amount of multiple testing (General benefit of clustering)
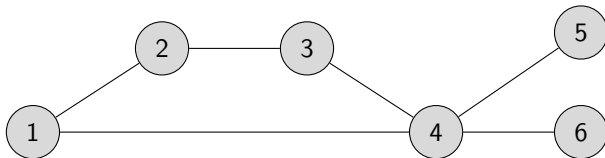
Topographical Overlap

- Topographical overlap (TOM_{ij}) is computed by the following formula:

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

where

  - a_{ij} strength of the connection between node i and j
  - k_i is the connectivity of node i, $k_i = \sum_u a_{iu}$
  - RHS is a special case for unweighted adjacency matrix
  - $N(i)$ is the neighbor set of node i

Example Network

- Consider nodes 1 and 2. $N(1) = \{2, 4\}$ and $N(2) = \{1, 3\}$, thus $k_1 = 2 = k_2$ and $|N(1) \cap N(2)| = 0$.

$$TOM_{12} = \frac{|N(1) \cap N(2)| + a_{12}}{\min(|N(1)|, |N(2)|) + 1 - a_{12}} = \frac{0 + 1}{2 + 1 - 1} = \frac{1}{2}$$
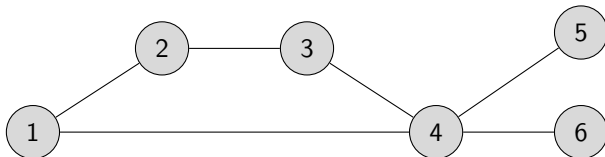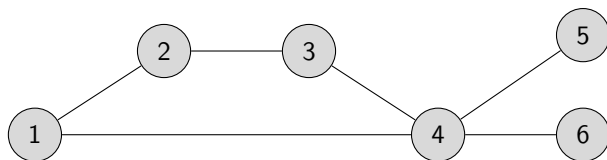
Example Network

- Consider nodes 1 and 2. $N(1) = \{2, 4\}$ and $N(2) = \{1, 3\}$, thus $k_1 = 2 = k_2$ and $|N(1) \cap N(2)| = 0$.

$$TOM_{12} = \frac{|N(1) \cap N(2)| + a_{12}}{\min(|N(1)|, |N(2)|) + 1 - a_{12}} = \frac{0 + 1}{2 + 1 - 1} = \frac{1}{2}$$

- Consider nodes 5 and 6. $N(5) = \{4\}$ and $N(6) = \{4\}$, thus $k_5 = 1 = k_6$ and $|N(5) \cap N(6)| = 1$.

$$TOM_{56} = \frac{|N(5) \cap N(6)| + a_{56}}{\min(|N(5)|, |N(6)|) + 1 - a_{56}} = \frac{1 + 0}{1 + 1 - 0} = \frac{1}{2}$$
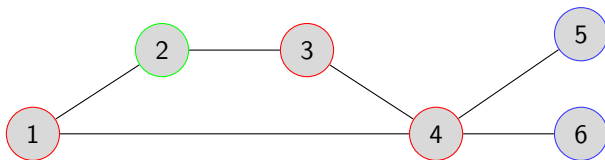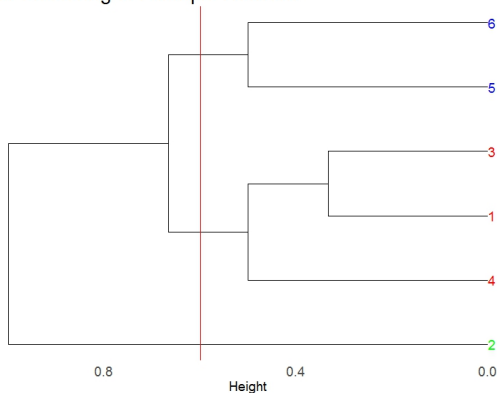
Example Network

- Consider nodes 4 and 5. $N(4) = \{1, 3, 5, 6\}$ and $N(5) = \{4\}$, thus $k_4 = 4$, $k_5 = 1$ and $|N(1) \cap N(2)| = 0$.

$$TOM_{45} = \frac{|N(4) \cap N(5)| + a_{45}}{\min(|N(4)|, |N(5)|) + 1 - a_{45}} = \frac{0 + 1}{1 + 1 - 1} = \frac{1}{1}$$

Example Network

Heirachical Clustering of Example Network

Gaussian Graphical Models

- Gaussian graphical models are networks constructed using conditional dependency.
- Edges between two elements are missing iff they conditionally independent.
- Partial correlation between two genes can be computed by
  - inverting the correlation matrix, or
  - calculating the correlation after adjusting each gene for all other genes

$$\hat{\beta}^{(i)} = \arg\min_{\beta \in \mathbb{R}^{p-1}} ||X^{(i)} - X^{(-i)}\beta||^2, \text{ and}$$

$$\hat{\rho}_{ij} = \text{sign}(\beta_j^{(i)})\sqrt{\hat{\beta}_i^{(j)}\hat{\beta}_j^{(i)}},$$

parcor package in R

- The parcor package in R is a tool that allows you to estimate partial correlation
- Additionally, different penalty terms can be used, specifically the ℓ_1 can be enforced to induce a sparse GGM.
- parcor allows for cross validation to select the shrinkage parameter.

Network Properties

	Mean Connectivity	Median Connectivity	Max Connectivity	Average Module Size
WGCNA (Genes)	7.35	4.88	30.91	56.90
Proposed Method (Genes)	3.068	3	10	42.80
WGCNA (Metabolites)	167.18	125.56	771.86	27.77
Proposed Method (Metabolites)	3.45	3	8	21.24

Table: Summary of network properties

Differential Analysis of Modules

	Number of Modules	Number of Significant (p<0.1)	TAp63 v. WT
WGCNA	265	23	7
Proposed Method	355	31	10

Table: Summary of gene module analysis

	Number of Modules	Number of Significant (p<0.1)	TAp63 v. WT
WGCNA	13	1	1
Proposed Method	17	2	2

Table: Summary of metabolite module analysis

Integrative Analysis of Modules

	Number of Genes Modules	Number of Metabolite Modules	Number of Significant Interactions
WGCNA	7	1	(0,0)
Proposed Method	17	2	(0,0)

Table: Summary of integrative analysis, red indicates gene and blue indicates metabolites.

Discussion & Future Work

- Discussion
  - Module detection can greatly reduce the burden of multiple testing and can allow for a potentially better biological understanding.
  - Using partial correlation to compute TOM provides more and smaller modules than using direct and indirect association.
  - Integrating multiple data types can lead to a much broader understanding of systems biology.
- Future Work
  - Study the underlying network properties more. (Does scale-free matter?)
  - Conduct simulation study to observe the network properties of WGCNA and our proposed method.
  - Develop method to allow elements to be assigned to multiple modules.
  - Investigate how to identify hubs

Acknowledgements
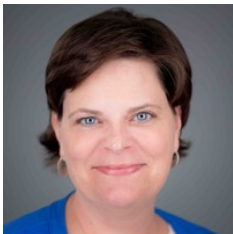
Brooke Fridley, PhD

Paul Stewart, PhD

- I would also like to thank Elsa Flores for allowing me to use her data for this presentation.

Thank you!