Inferring Gene-Gene Interactions and Functional Modules Beyond Standard Models

Haiyan Huang
Department of Statistics, UC Berkeley

Feb 7, 2018
Background

- **Gene regulatory networks**
  - Directed, genes as nodes, a directed edge from gene A to gene B if A encodes a transcription factor that regulates the expression of B.
  - Need time-course data at small time intervals.
  - Study undirected networks as a simplification (when inferring gene networks using expression data).

- **Task of our interest**: To identify gene functional groups, many of which encode biological pathways, using gene expression data.
  - A network of relations needs to be learned using gene expression levels as covariates with functionally related genes having denser connections (stage 1).
  - Detect tightly knit sub-structures (stage 2).
• **High dimensionality** \((p >> n)\) often results in a high rate of false positives, severely limiting a genomic scale network exploration.

• Measures based on **pairwise marginal relationships** are often inappropriate for detecting higher level interactions.
  - A gene may interact with a group of genes despite having weak marginal correlations with the individual genes.

• **Partial correlation** takes the influence of other genes into account, e.g., Gaussian Graphical Models (GGM).
  - Inclusion of non-pathway genes in the conditional set can lead to spurious dependencies.
  - “Seed genes” (Kim et. al. (2012)), i.e. known pathway genes are useful to define an appropriate conditional set, but they are not always available.
  - Under GGM, the elements in the precision matrix are proportional to partial correlations, i.e., correlation between two genes conditioned on the rest genes.
Problem Formulation

Our formulation was motivated by considering GGM in a linear regression setting:

The partial correlation between gene $i$ and gene $j$ conditioned on a set of genes $Z$ is simply the correlation $\text{cor}(\varepsilon_1,\varepsilon_2)$ of the residuals $\varepsilon_1$ and $\varepsilon_2$ resulting from linearly regressing gene $i$ and gene $j$ against the genes in $Z$, respectively.

For instance, assume that genes A, B and C are interacting as a group, and that $G_A = \beta_A G_C + \varepsilon_1$ and $G_B = \beta_B G_C + \varepsilon_2$, where $G_A, G_B, G_C$ denote the expression profile for genes A, B and C, respectively.

If $\text{cor}(\varepsilon_1,\varepsilon_2) \approx 1$, then $G_A + \Gamma \, G_B + \Gamma \, G_C + \Gamma \approx 0$
Problem Formulation

Our formulation was motivated by considering GGM in a linear regression setting:

The partial correlation between gene $i$ and gene $j$ conditioned on a set of genes $Z$ is simply the correlation $\text{cor}(\varepsilon_1, \varepsilon_2)$ of the residuals $\varepsilon_1$ and $\varepsilon_2$ resulting from linearly regressing gene $i$ and gene $j$ against the genes in $Z$, respectively.

Therefore, under a GGM, it is reasonable to assume that functionally related genes form strong linear relationships, with one linear equation for one gene module (a set of functionally related genes), suggesting a new formulation for finding gene modules:

Finding strong linear models that involve a very small set of genes among the thousands of candidate genes.
An attempted solution

• To find strong linear models that involve a very small set of genes among the thousands of candidate genes, we introduce a procedure based on a novel application of sparse canonical correlation analysis (SCCA) to the gene expression data matrix.
Sparse Canonical Correlation Analysis (SCCA)

- SCCA finds vectors $\mathbf{a}$ and $\mathbf{b}$ that solve

$$\max_{\mathbf{a, b}} \mathbf{a}^T \mathbf{Y}^T \mathbf{X} \mathbf{b} \quad \text{subject to} \quad \mathbf{a}^T \mathbf{a} \leq 1, \mathbf{b}^T \mathbf{b} \leq 1, p_1(\mathbf{a}) \leq c_1, p_2(\mathbf{b}) \leq c_2,$$

where $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{Y} \in \mathbb{R}^{n \times q}$ have centered and scaled columns, $p_1$ and $p_2$ are convex functions. (Witten et al. (2009))

- For high-dimensional biological data, sparsity is necessary and we impose this using $L_1$ penalty, $p_1(\cdot) = p_2(\cdot) = \| \cdot \|_1$.

- Modified NIPALS algorithm (Li et al. (2011)) with tuning parameters $\lambda = (\lambda_1, \lambda_2)$. 
SCCA with random partition and subsampling

- Average \( a, b \) over a large number of subsampling and random partitions.
  - Subsample a fixed fraction of the genes
  - Random split the matrix (by genes) into \( x \) and \( y \) and run SCCA

**Proposition**

Assuming only one pathway \( K \) and only genes in \( K \) are correlated. For every partition \( t \), let \( c_t \) be the list of the absolute values \( |a_t| \) and \( |b_t| \) ordered according to the gene list. Let \( \bar{c} = \sum_{t=1}^{N} c_t / N \), where \( N \) is the number of partitions, then \( \exists B > 0 \)

\[
\lim_{N \to \infty} \lim_{n \to \infty} \left( \min_{i \in K} \bar{c}_i - \max_{j \notin K} \bar{c}_j \right) = B.
\]
1. SCCA implemented by a modified NIPALS algorithm involving iterative penalized regression with L1 penalty.

2. If some genes are known to operate in the same pathway, the penalty parameter corresponding to those genes could be lowered.
Subsampling helps the identification of weaker functional group

- We simulated the expressions of 20 genes.
- 6 of them form two disjoint functional gene groups.
- One group has 3 genes that are perfectly linearly related (see (a)).
- The other group has 3 genes that have relatively weaker linear relationships (see (b)).
- The left 14 genes are independent from each other.

Figure 2: Asymptotic values of $\bar{A}$ with (a) no subsampling and (b) 50% subsampling.
Subsampling helps the identification of overlapping functional groups

We simulated the expressions of 150 genes.

There are two overlapping functional gene groups, with each functional group having 15 genes and 5 genes in common.

Figure 1: Heatmap of $\tilde{A}$ with 50% subsampling.
Choose the amount of regularization

- $p = 150$, $n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 3$, $\lambda_b = 3$

- $p = 150$, $n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 9$, $\lambda_b = 9$

- $p = 150$, $n = 30$
- functional group: 1-15
- subsample 70%, $\lambda_a = 18$, $\lambda_b = 18$
The **entropy** of an edge weight matrix $\tilde{A}$ is defined as

$$H(A) = - \sum_{i<j, A_{ij} > 0} (A_{ij} / S_A) \log(A_{ij} / S_A),$$

where $S_A = \sum_{i<j} A_{ij}$.

$(\lambda_1, \lambda_2)$ leading to small entropy values are desirable.
Identify community structures in $\tilde{A}$

- Gene functional groups as communities in a network.
- Community detection methods (many are available; we implemented the below two)
  - Greedy algorithms — hierarchical clustering (HC)
  - Probabilistic models — stochastic block models (SBM) fitted using the pseudo-likelihood algorithm Chen et al. (2012)
Performance comparison for simulated data

- Classification rates: \( \text{precision} = \frac{TP}{TP + FP} \), \( \text{recall} = \frac{TP}{TP + FN} \).
- Cross-mix the following to allow comparisons
  - Stage 1: building edge weight matrices \( \mathbf{\bar{A}} \)
    - \textit{scca}
    - \textit{pearson}
    - \textit{module}: Transformed Pearson’s correlation matrix used in Langfelder et al. (2007).
  - Stage 2: detecting block structures from \( \mathbf{\bar{A}} \)
    - \textit{sbm}
    - \textit{hc}: Cutting the dendrogram when clusters of size less than 25 start to appear.
    - \textit{dynamic, hybrid}: HC with dendrogram cutting methods in dynamicTreeCut (Langfelder et al. (2008)).
- Six methods for comparison: \textit{scca.hc, pearson.hc, scca.sbm, pearson.sbm, module.dynamic, module.hybrid}. 
**Table:** Classification performance of different methods using datasets with $p = 500$, two functional groups, subsampling level 70%, and various levels (0%, 33% and 67%) of experiment dependency.

<table>
<thead>
<tr>
<th>Method</th>
<th>Pathway 1 0%</th>
<th>Pathway 1 33%</th>
<th>Pathway 1 67%</th>
<th>Pathway 2 0%</th>
<th>Pathway 2 33%</th>
<th>Pathway 2 67%</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td><code>scca.hc</code></td>
<td>0.861</td>
<td>0.533</td>
<td>0.831</td>
<td>0.441</td>
<td>0.811</td>
<td>0.433</td>
</tr>
<tr>
<td><code>module.dynamic</code></td>
<td>0.718</td>
<td>0.3</td>
<td>0.742</td>
<td>0.333</td>
<td>0.764</td>
<td>0.38</td>
</tr>
<tr>
<td><code>module.hybrid</code></td>
<td>0.439</td>
<td>0.407</td>
<td>0.544</td>
<td>0.447</td>
<td>0.453</td>
<td>0.385</td>
</tr>
<tr>
<td><code>pearson.hc</code></td>
<td>0.238</td>
<td>0.233</td>
<td>0.497</td>
<td>0.427</td>
<td>0.471</td>
<td>0.393</td>
</tr>
<tr>
<td></td>
<td>Pathway 2 0%</td>
<td>Pathway 2 33%</td>
<td>Pathway 2 67%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td><code>scca.hc</code></td>
<td>0.808</td>
<td>0.487</td>
<td>0.890</td>
<td>0.489</td>
<td>0.833</td>
<td>0.420</td>
</tr>
<tr>
<td><code>module.dynamic</code></td>
<td>0.758</td>
<td>0.4</td>
<td>0.808</td>
<td>0.347</td>
<td>0.8</td>
<td>0.4</td>
</tr>
<tr>
<td><code>module.hybrid</code></td>
<td>0.565</td>
<td>0.473</td>
<td>0.529</td>
<td>0.387</td>
<td>0.455</td>
<td>0.46</td>
</tr>
<tr>
<td><code>pearson.hc</code></td>
<td>0.438</td>
<td>0.387</td>
<td>0.323</td>
<td>0.307</td>
<td>0.460</td>
<td>0.273</td>
</tr>
</tbody>
</table>
Arabidopsis data

- Shoot tissue dataset from *A. thaliana* subject to oxidation stress, composed of 22810 genes, 13 experiments and two replicates for each experiment.

- Select genes that have
  - a reasonable large variance across the experiment. Also remove genes with a suspiciously high experiment variance.
  - small replicate discrepancy
  - reasonably large expression levels

\[ p = 2718. \]

- `scca.hc` vs. `pearson.hc`, `module.dynamic`, `module.hybrid`
Of the 13 groups found,

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Enriched GO term</th>
<th>Number of genes with enriched terms</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chloroplast organellar gene</td>
<td>10 out of 15</td>
<td>$1.10 \times 10^{-4}$</td>
</tr>
<tr>
<td>2</td>
<td>Phenylpropanoid-flavonoid biosynthesis</td>
<td>3 out of 4</td>
<td>$6.65 \times 10^{-7}$</td>
</tr>
<tr>
<td>3</td>
<td>Glucosinolate biosynthesis</td>
<td>7 out of 7</td>
<td>$1.95 \times 10^{-14}$</td>
</tr>
<tr>
<td>4</td>
<td>Chloroplast organellar gene</td>
<td>3 out of 3</td>
<td>$7.83 \times 10^{-3}$</td>
</tr>
<tr>
<td>5</td>
<td>Ribosome</td>
<td>10 out of 15</td>
<td>$7.20 \times 10^{-13}$</td>
</tr>
<tr>
<td>8</td>
<td>Ribosome</td>
<td>5 out of 6</td>
<td>$8.31 \times 10^{-8}$</td>
</tr>
<tr>
<td>10</td>
<td>Photosystem I or II</td>
<td>8 out of 10</td>
<td>$2.87 \times 10^{-14}$</td>
</tr>
<tr>
<td>12</td>
<td>Endomembrane system</td>
<td>3 out of 4</td>
<td>$2.35 \times 10^{-3}$</td>
</tr>
</tbody>
</table>

The Arabidopsis Information Resource
(http://www.arabidopsis.org/tools/bulk/index.jsp)
Results using *module.dynamic* and *module.hybrid*

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Enriched GO term</th>
<th>Number of genes with enriched terms</th>
<th>P-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>Cell wall</td>
<td>16 out of 81</td>
<td>$4.46 \times 10^{-6}$</td>
</tr>
<tr>
<td>10</td>
<td>Defense response</td>
<td>29 out of 78</td>
<td>$1.58 \times 10^{-2}$</td>
</tr>
<tr>
<td>11</td>
<td>Phenylpropanoid-flavonoid biosynthesis</td>
<td>11 out of 76</td>
<td>$5.42 \times 10^{-12}$</td>
</tr>
</tbody>
</table>

13 groups in total with sizes from 60 to 293

<table>
<thead>
<tr>
<th>Group ID</th>
<th>Enriched GO term</th>
<th>Number of genes with enriched terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>62</td>
<td>NA</td>
<td>0 out of 6</td>
</tr>
<tr>
<td>63</td>
<td>Chloroplast</td>
<td>4 out of 6</td>
</tr>
<tr>
<td>64</td>
<td>Located in plasma membrane</td>
<td>2 out of 5</td>
</tr>
<tr>
<td>65</td>
<td>Located in plasma membrane</td>
<td>3 out of 5</td>
</tr>
<tr>
<td>66</td>
<td>Pyridoxine biosynthetic process</td>
<td>2 out of 5</td>
</tr>
</tbody>
</table>

66 groups in total with sizes from 5 to 81
Conclusion on the SCCA method

• High precision
• More conceptually appealing.
  • Under a regression setting, it captures group interactions by providing an aggregated measure of gene partial correlations when the correct conditional set is unknown
• Flexible enough to incorporate prior knowledge when available
Gene-Fishing:
A semi-supervised, non-parametric clustering procedure with bagging
Background

Biomedical questions of interest
1. Identifying novel cholesterol-metabolism regulators that may contribute (or be related) to cardiovascular disease risk.
2. Characterizing the diversity of cholesterol metabolism across different tissues.

Datasets:
1. CAP LCLs Dataset (n=426): RNA-seq and expression array of subject-derived lymphoblastoid cell lines (LCLs) were established from CAP subjects; demographic information and plasma LDLC are also available.
2. GEUVADIS Dataset (n=465; publicly available): the transcriptome of 465 LCL samples (from the 1000 Genome Project) were deeply sequenced on Illumina HiSeq2000.
3. GTex Dataset (publicly available): 7051 samples for 43 different tissues were collected for RNA-Seq analysis.
Our Initial Analysis:

Spectral analysis of all 14028 genes (Cap LCL expression dataset)

Spectral analysis of 80 (GO-annotated) cholesterol metabolism associated genes

Visualization of gene-gene correlation matrix

21 genes
Interesting patterns become hidden as noise increases

- N=80
- N=1000
- N=3000
- N=5000
- N=8000
- N=10000

The 21 cholesterol synthetic pathway genes
Random genes
Workflow of gene fishing

Repeat for 1000 times

Gene name | Capture Frequency Rate
--- | ---
G₁ | CFR₁
G₂ | CFR₂
...
Gₙ | CFRₙ

Reducing search space
Making search better focused
Aggregating clustering results from different rounds to make the discovery robust

A ranked gene list

Histogram of Capture Frequency Rate (CAP-LCL)
Comparison of Reproducibility between the CAP LCL dataset and the GEUVADIS dataset

- Gene Fishing
- Guilt-By-Association (median)
- Guilt-By-Association (mean)
- Guilt-By-Association (max)
- WGCNA
Application to GTex dataset
Application to GTex dataset

- lipid metabolic process
- long-chain fatty-acyl-CoA biosynthetic process
- oxidation-reduction process
- response to nutrient
- metabolic process
- cholesterol homeostasis
- carbohydrate metabolic process
- gluconeogenesis
- tricarboxylic acid cycle
- mitochondrial electron transport, NADH to ubiquinone
- protein glycosylation
- intracellular protein transport
- ER to Golgi vesicle-mediated transport
- retrograde vesicle-mediated transport, Golgi to ER
- protein N-linked glycosylation via asparagine
- mitochondrial respiratory chain complex I assembly
- tumor necrosis factor-mediated signaling pathway
- proteasome-mediated ubiquitin-dependent protein catabolic process
- T cell receptor signaling pathway
- cytoskeleton organization
- microtubule-based process
- cell-cell adhesion
- negative regulation of transcription from RNA polymerase II
- cell adhesion
- cell surface receptor signaling pathway
- transmembrane receptor protein tyrosine kinase signaling pathway
- T cell costimulation
- T cell activation
- negative regulation of apoptotic process
- regulation of immune response
Acknowledgements

- SCCA project
  Inferring gene-gene interactions and functional modules using sparse
  canonical correlation analysis.

- Gene-Fishing project
  - UC Berkeley: Dr. Ke Liu (Postdoc); Yun Zhou (Student in
    Biostatistics); Tal Ashuach (Student in Computational Biology);
    Peter Bickel (Professor, Statistics)
  - Children’s Hospital at Oakland Research Institute: Dr. Elizabeth
    Theusch (Postdoc) Dr. Marisa Wong Medina (Scientist)