Composite Likelihood Approach To Gene Network Construction

Peter Song

Department of Biostatistics

University of Michigan

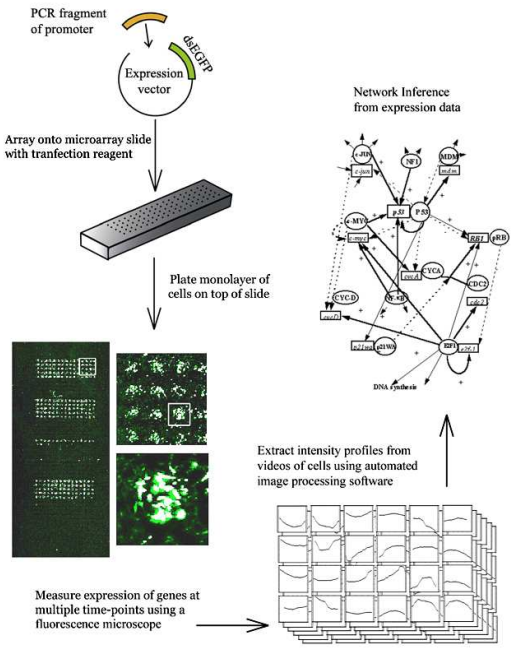
pxsong@umich.edu

Joint work with Xin Gao and Daniel Q. Pu.

Gene Network For Time-Course Microarray Data

- Technological advances allow biologists to collect gene expression data at multiple times within a relatively short period of time.
- Time series expression data are essential to understand individual cellular behaviors such as mobility, division and differentiation.
- Gene regulatory network is important knowledge of biological pathways.

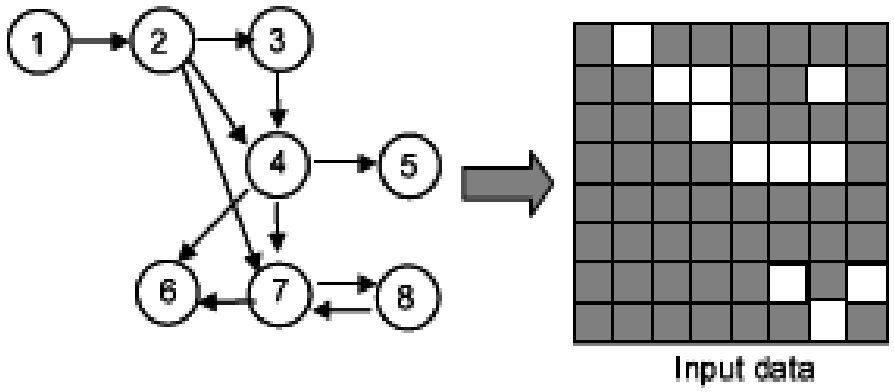
Figure 1: Reverse Transfection Reporter Microarray (Ziauddin and Sabatini, Nature, 2001)



Primary Tasks in Gene Network Reconstruction

- Who are related?
- How are they related, if related?

Figure 2: If related? A network defined by 0-1 connectivity.



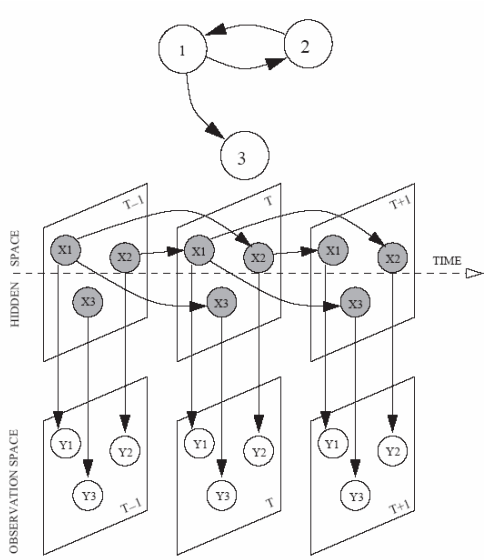
- The graphic representation describes pairwise relationships, which may be summarized by a square matrix of dependencies (correlation or partial correlation) or connectivity (0/1).
- Entries of the matrix are parameters of interest in gene network construction.
- How are they related?–**The meaning of entries in the relationship matrix.**
- **Statistical inference in gene network pertains to the existence and/or the direction of an edge.**

Gene-Gene Dependency: Beyond Linear Correlation

- A measure for gene-gene relationship should reflect the fact that the genes' induction and inhibition routes are determined through specific molecular interactions.
- In literature, most research work on gene network construction utilizes 0-1 connectivity or traditional correlation (e.g. partial correlation) as a dependency measure between two genes.
- Partial correlation, obtained from the inverse of a covariance matrix, is no longer available for measuring dependency between two time series.
- Cross-correlation function (CCF, Haugh, 1976), dynamic correlation (DC, Dubin and Muller, 2005), sample dynamic correlation (SDC, Opgen-Rheinard and Strimmer, 2006) are linear dependency measures.

- Correlation-based measures essentially reflect the strength of concordance and discordance between genes' expression processes.
- A single number is not sufficient to describe complicated molecular interactions between genes.
- We propose to use the mechanism of transitions to characterize gene-gene interactions, as by monitoring transitions of expression levels over time, we hope to tell a more detailed story about gene-gene dependency/interaction.
- This motivates us to consider a dynamical (stochastic) framework rather than a static framework to construct gene networks.

Figure 3: Three-gene networks: Static versus Stochastic.



Hidden Markov Model

- Assume X_{gtjs} denote the sth replicate (e.g. plate) of the gene expression value of gene g at time t, $t = 1, \ldots, T$, treatment level j, $j = 1, 2$.
- To investigate whether or not gene g is differentially expressed, a summary two-sample t-statistic may be considered:

$$Y_{g,t} = \frac{\overline{X}_{gt1} - \overline{X}_{gt2}}{\sqrt{\frac{s^2_{gt1}}{n_1} + \frac{s^2_{gt2}}{n_2}}}$$

- At time t, assume there is a hidden state

$$S_{g,t} = \begin{cases} 1, \text{gene } g \text{ is differentially expressed (DE)}, \\ 0, \text{gene } g \text{ is not differentially expressed (NDE)}. \end{cases}$$

- Given the hidden state $S_{g,t} = 0$, the conditional distribution of $Y_{g,t}$ is denoted as f_{t0}, whereas the conditional distribution of $Y_{g,t}$ when $S_{g,t} = 1$, is denoted as f_{t1}.
- The two densities are estimated by Efron's (2001) nonparametric empirical Bayesian method and held fixed in the remaining analysis.
- For a pair of genes, gene g_1, and gene g_2, with the joint hidden state vector $\boldsymbol{S}_t = (S_{g_1,t}, S_{g_2,t})' \in \{(0,0),(0,1),(1,0),(1,1)\}$.
- Assume that the joint state vector at time $t+1$, $\mathbf{S}_{t+1}$, obeys the Markovian property. Denote the joint transition matrix between time t and time $t+1$ by $\Lambda(\boldsymbol{S}_{t+1}|\boldsymbol{S}_t)$.

Pairwise Transition Dependency

- A measure of gene-gene dependency between two genes is defined by the following transition dependency matrix:

$$\boldsymbol{D}_{g_1,g_2} = \Lambda(S_{g_1,t+1}|S_{g_1,t}) \otimes \Lambda(S_{g_2,t+1}|S_{g_2,t}) - \Lambda(\boldsymbol{S}_{t+1}|\boldsymbol{S}_t).$$

- If the transitions of the two genes independently occur, $D_{g_1,g_2} = \mathbf{0}$.
- Deviations from zero indicate not only the magnitude of the dependency but also the nature of dependency effects.
- Therefore, testing hypothesis of $D_{g_1,g_2} = \mathbf{0}$ is useful to assess the relationship between the two genes.
- Note a matrix, rather than a number, is used to describe the gene-gene dependency.

Interpretation of Transition Dependency Matrix

- Inhibition Effect:

$$P(S_{g_1,t+1}=1|S_{g_1,t}=0)P(S_{g_2,t+1}=1|S_{g_2,t}=1) \\ -P(\mathbf{S}_{t+1}=(1,1)'|\mathbf{S}_t=(0,1)') \quad >0$$

- DE state of gene g_2 reduces the probability of gene g_1 changing from NDE state to DE state.
- Induction Effect:

$$P(S_{g_1,t+1}=1|S_{g_1,t}=0)P(S_{g_2,t+1}=1|S_{g_2,t}=1) \\ -P(\mathbf{S}_{t+1}=(1,1)'|\mathbf{S}_t=(0,1)') \quad <0$$

- DE state of gene g_2 increases the probability of gene g_1 changing from NDE state to DE state.

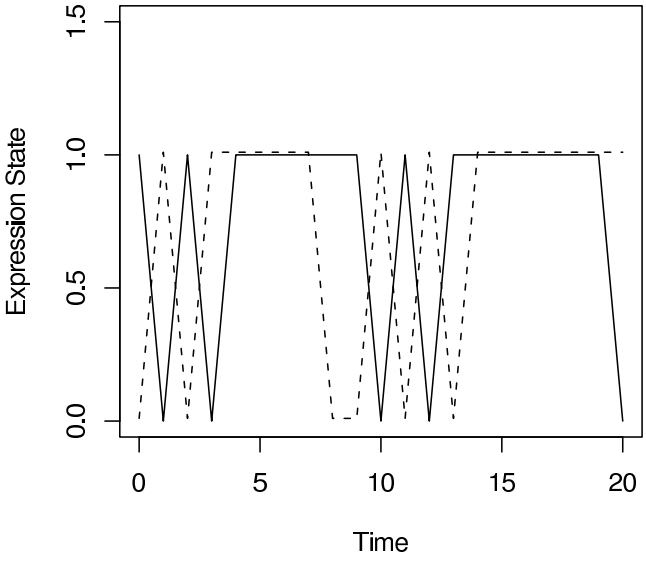
Example: Dependency Matrix vs Traditional Correlation

- The joint transition matrix is

$$\Lambda = \begin{bmatrix} 0.80 & 0.10 & 0.10 & 0.00 \\ 0.10 & 0.10 & 0.70 & 0.10 \\ 0.10 & 0.70 & 0.10 & 0.10 \\ 0.00 & 0.10 & 0.10 & 0.80 \end{bmatrix}.$$

- The resulting stationary distribution is $\pi = (0.25, 0.25, 0.25, 0.25)$.
- Thus, the two marginal processes have zero correlation (the same prob to have concordant and discordant pairs).
- But these two genes are indeed a conjugate pair; that is, they reconcile to reach an equilibrium state.

- When two genes are at the same hidden states, the two series undergo equilibrium periods.
- When the equilibrium is broken, the two series oscillate rapidly to reach the equilibrium.



- The sample correlation is -0.06, little evidence on such dependency.
- However, the dependency matrix $\mathbf{D}$ is the difference of the two:

$$\begin{bmatrix} 0.80 & 0.10 & 0.10 & 0.00 \\ 0.10 & 0.10 & 0.70 & 0.10 \\ 0.10 & 0.70 & 0.10 & 0.10 \\ 0.00 & 0.10 & 0.10 & 0.80 \end{bmatrix} - \begin{bmatrix} 0.3025 & 0.2475 & 0.2475 & 0.2025 \\ 0.2475 & 0.3025 & 0.2025 & 0.2475 \\ 0.2475 & 0.2025 & 0.3025 & 0.2475 \\ 0.2025 & 0.2475 & 0.2475 & 0.3025 \end{bmatrix}.$$

Statistical Inference: Composite Likelihood Approach

- For a given network of N genes, the grand joint transition matrix is of $2^N \times 2^N$ dimension.
- The high dimensionality of the parameter space makes infeasible the computation of the full likelihood.
- Composite likelihood method helps us to carry out "dimension reduction" (e.g. based on pairs) and perform a valid inference.
- It needs an EM type algorithm in the composite likelihood context.

Notation

- Let $\boldsymbol{Y}_g = (Y_{g,1}, \ldots, Y_{g,T})'$ denote the vector of summary statistics from the expression data of gene g. Simultaneously consider N genes expression profile together $\boldsymbol{Y} = (\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_N)$, corresponding to the hidden vectors of states $\boldsymbol{S} = (\boldsymbol{S}_1, \ldots, \boldsymbol{S}_N)$, whereas $\boldsymbol{S}_g = (S_{g,1}, \ldots, S_{g,T})'$.
- 2^{2N} transition probabilities: $\boldsymbol{\Lambda} = [P\{(S_{1,t}, \ldots, S_{N,t}) | (S_{1,t-1}, \ldots, S_{N,t-1})\}]$.
- To apply the CL method, we reparametrize these transition probabilities as follows: Univariate transitions, bivariate transitions given univariate transitions, trivariate transitions given bivariate transitions, and so on.

- The CL method considers the simultaneous inferences of all the pairwise gene-gene dependency and discards higher order transitions.
- Let $\mathbf{\Lambda}^{gg'}$ denote the bivariate transition matrix of the vector $(S_{g,t}, S_{g',t})$. Let $\mathbf{\Lambda}^{g}$ denote the marginal transition matrix of $S_{g,t}$, and $\mathbf{\Lambda}^{g'}$ denote the marginal transition matrix of $S_{g',t}$.
- The composite likelihood based on all the pairs takes the form:

$$\begin{aligned} L_c(\mathbf{Y}) &= \prod_{(g<g')} P(\mathbf{Y}_g, \mathbf{Y}_{g'}) \\ &= \prod_{(g<g')} E_{f(\mathbf{S}_g)} E_{f(\mathbf{S}_{g'})} P(\mathbf{S}_g, \mathbf{S}_{g'}) P(\mathbf{Y}_g|\mathbf{S}_g) P(\mathbf{Y}_{g'}|\mathbf{S}_{g'}). \end{aligned}$$

- Note that the dimensionality of this model is now only $O(N^2)$.

Composite Likelihood EM (CLEM) Algorithm

- As usual, the hidden Markov process is treated as missing data, so the EM algorithm in the contexts of composite likelihood is needed.
- Let $\{f(z;\theta), z \in \mathcal{Z}, \theta \in \Theta\}$ be a parametric statistical model, with $\mathcal{Z} \subseteq \mathcal{R}^n$, $\Theta \subseteq \mathcal{R}^d$, $n \geq 1$, and $d \geq 1$. Consider a set of measurable events $\{\mathcal{A}_i : i \in C \subseteq \mathcal{N}\}$. A composite likelihood is defined as

$$L_c(\theta; z) = \prod_{i \in C} L_i(\theta; z)^{w_i},$$

where $L_{c,i}(\theta; z) = f(z \in \mathcal{A}_i; \theta)$, with $\{w_i, i \in C\}$ being a set of suitable weights.

- There exists another sample space $\mathcal{Y}$ and a many-to-one (attrition) mapping $z \to y(z)$ from $\mathcal{Z}$ to $\mathcal{Y}$. Instead of observing the complete data z in $\mathcal{Z}$, we observe the incomplete data y in $\mathcal{Y}$.
- The observed composite likelihood is defined as $L_c(\theta; y) = \prod_{i \in C} L_{c,i}(\theta; y)^{w_i}$ with $L_{c,i}(\theta; y) = \int_{\mathcal{A}_i \cap \mathcal{Z}(y)} f(z; \theta) dz$, and $\mathcal{Z}(y)$ denoting the subset of $\mathcal{Z}$ determined by the equation $y = y(z)$.
- Varin et al. (2005) considered the pairwise EM algorithm. We considered a general setup for all the theorems given below.

- In the spirit of dimension reduction, instead of conditioning on the entire $\mathbf{y}$ which may be of large dimension and has complicated dependency structure, we consider an expected composite likelihood based on subsets:

$$Q_c(\theta|\theta_n) = \sum_{i \in C} w_i E\left\{\log f(z \in \mathcal{A}_i; \theta) | y_{\mathcal{A}_i}, \theta_n\right\},$$

  where $y_{\mathcal{A}_i} = \{y(z) : z \in \mathcal{A}_i \cap \mathcal{Z}(y)\}$.
- **E Step**: Given the current estimate θ_n, obtain the expected composite likelihood $Q_c(\theta|\theta_n)$.
- **M Step**: Maximize $Q_c(\theta|\theta_n)$ with respect to θ, and update the estimate θ_{n+1}. Repeat these two-step iterations until convergence.

- **Theorem 1** The proposed CLEM algorithm possesses the ascent property: the composite likelihood of the observed data is nondecreasing as we update the estimates: $\sum_{i \in C} w_i \log f(y_{\mathcal{A}_i}; \theta_{n+1}) \geq \sum_{i \in C} w_i \log f(y_{\mathcal{A}_i}; \theta_n)$.
- **Theorem 2** We assume that $\Theta_{\theta_0} = \{\theta \in \Theta : L_c(\theta; y) \geq L_c(\theta_0, y)\}$ is compact for any $L_c(\theta_0; y) > -\infty$ and L_c is continuous in Θ and differentiable in the interior of Θ. Under the smoothness assumption of the function $Q(\theta|\theta')$ in both θ and θ', the CLEM algorithm converges to the stationary point of the observed composite likelihood surface.

Application of CLEM Algorithm on Gene Network

- Sort out constraints in the transition probabilities.
- Invoke re-parametrization to facilitate the estimation.
- E-step is carried out by the Forward-Backward algorithm.
- M-step is carried out by quasi-Newton algorithm with the utility of Lagrange multipliers for the constraints.

Selection of Network Topology

- In practice, usually one starts with a gene of interest, and screen all possible pairwise dependencies with this target gene.
- At the completion of this screening analysis, a small pool of genes are identified for a joint analysis, say, under an FDR level.
- Because of involved falsely discovered genes in the previous screening analysis, there is a need to select the final network among a few candidate networks.
- Each candidate network corresponds to a specification of a composite likelihood and hence leads to the network-specific estimation for the vector of model parameters θ.

- To derive an AIC-type model selection criterion, minimize the expected Kullback-Leibler distance based on the composite likelihood, which is equivalent to maximizing the following form:

$$E_{f_0(y)}\left[\sum_{i\in C} E_{f_0(z)} w_i \left\{\log\, f(Z\in\mathcal{A}_i;\hat{\theta}_{MCL}(y))\right\}\right],$$

  where Z is the future observation and f_0 is the true model.
- In the composite likelihood context, the proposed first-order unbiased selection statistic (Varin and Vidoni, 2005) is

$$\ell_c(\hat{\theta}_{MCL};\mathbf{Y}) + tr\{\hat{V}(\mathbf{Y})\hat{H}(\mathbf{Y})^{-1}\}.$$

- For our application, we use the slightly modified statistic as follows:

$$\ell_c(\hat{\theta}_{MCL};\mathbf{Y}) + tr\{\hat{\Sigma}(\mathbf{Y})\hat{H}(\mathbf{Y})\}.$$

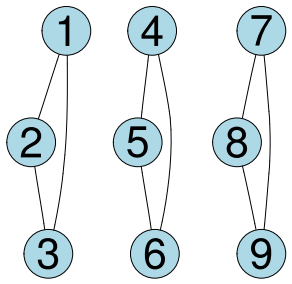
Simulation Study I: Performance of CLEM Algorithm

- Consider a three-gene network, including $2^3 \times 2^3 = 64$ transition probabilities.
- A simpler case of all pairwise transition probabilities ($3 * 3^2 = 27$).
- $M = 30, T = 40$ (not realistic but for asymptotics), and 1000 simulations.
- Λ^{ab} is shown, and the others are similar.

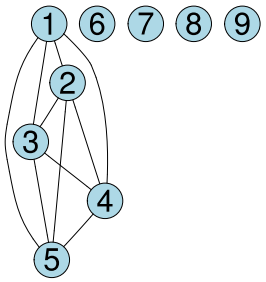

```
transProb=
        0.15      0.15      0.35      0.35
        0.15      0.15      0.35      0.35
        0.35      0.35      0.15      0.15
        0.35      0.35      0.15      0.15

MEANab=
        0.1498    0.1503    0.3510      0.3487
        0.1486    0.1496    0.3508      0.3507
        0.3502    0.3501    0.1499      0.1496
        0.3501    0.3496    0.1506      0.1495
SDab=
        0.0229    0.0223    0.0309      0.0284
        0.0232    0.0230    0.0302      0.0305
        0.0303    0.0297    0.0235      0.0234
        0.0296    0.0309    0.0238      0.0227
```

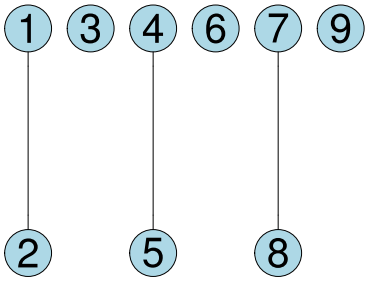
Simulation Study II: Selection of Network Topology



(a) Network N1



(b) Network N2



(c) Network N3

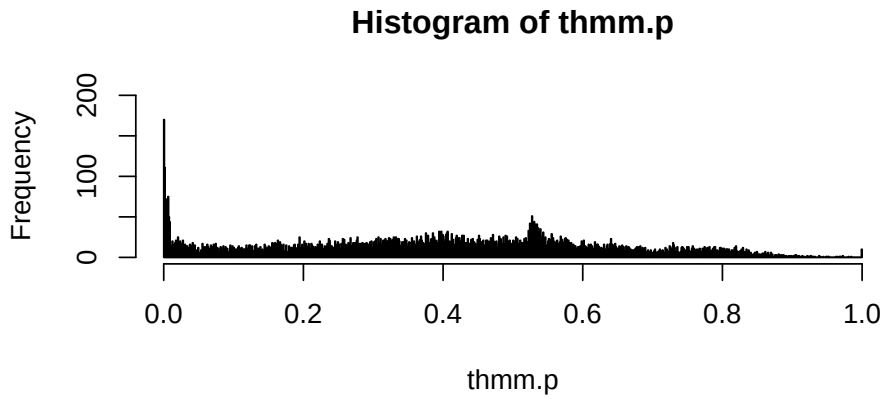
(d) Network N4

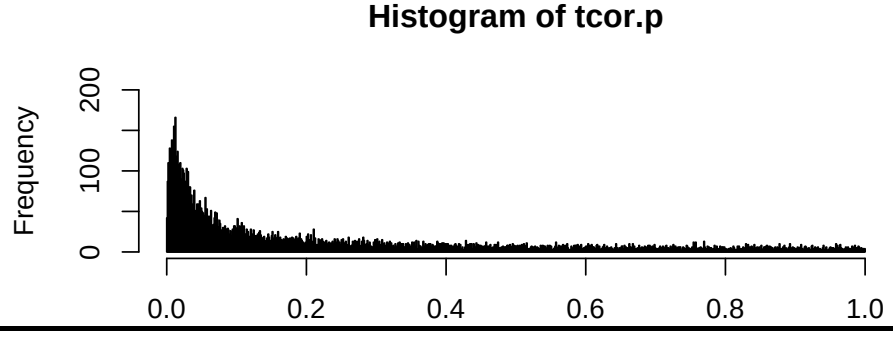
Empirical success rates of selecting the true network topology (N1) versus each of the candidate networks using the COMP-AIC method over 100 simulation data sets.

Model	# Par.	Rate ($M = 10, T = 10$)	Rate ($M = 10, T = 20$)
N1	108	0.77	0.96
N2	117	0.20	0.03
N3	54	0.00	0.01
N4	27	0.03	0.00

Data Analysis

- A set of 15 oligonucleotide microarrays in an experiment designed by Kobayashi et al. (2005, J. Leukocyte Biology) to study spontaneous neutrophil apoptosis and regulation of cell survival by granulocyte macrophage-colony stimulating factor (GM-CSF).
- The experiment was performed at 5 (T) time points (0, 6, 12, 18, 24 hrs), with 3 (M) blood donors.
- We considered a total of 12,624 genes, with the focus of gene 139 (CD44) which was found to be a key player in human immune system for host defense and transduction.
- Preliminary analysis concerned all two-gene networks, all the p-values are plotted in a histogram below.





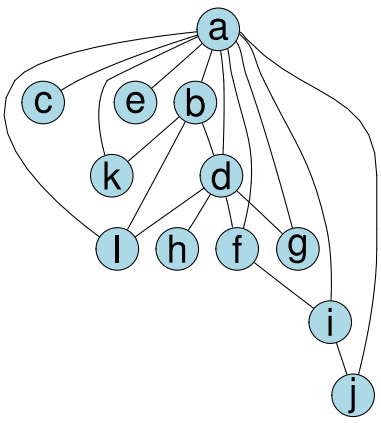
- Using FDR control at 0.1 level, we detected 302 significant genes in the pairwise analysis.
- The 15 most significant candidate genes dependent with CD44 are listed.
- Kobayashi et al. (2005) highlighted Caspase 8 as one of the most important genes in the apoptosis regulation. It's ranked at 5th in our HMM list and at 308th in the DC list. The estimated dependency matrix $\hat{D}$ is

$$\begin{bmatrix} 0.24 & -0.24 & -0.24 & 0.24 \\ 0.55 & -0.28 & -0.19 & -0.07 \\ 0.50 & -0.19 & -0.28 & -0.03 \\ 0.22 & -0.22 & -0.22 & 0.22 \end{bmatrix}.$$

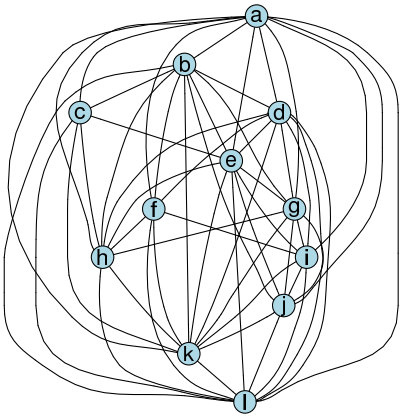
Probe	p_{DC}	p_{HMM}	Gene Title
38336_at	0.00110891	9.505e-05	FERM domain containing 4B
947_at	0.04692277	0.00011089	Gene function unknown
39237 _at	0.51548517	0.00017426	mitogen-activated protein ...
40968 _at	0.00367525	0.00017426	suppressor of cytokine signaling 3
31491_s _at	0.00107723	0.00019010	caspase 8 (CASP8)
36985 _at	0.02434851	0.00020594	isopentenyl-diphosphate delta ...
36344 _at	0.01527129	0.00022178	coagulation factor II receptor-like ...
1441_s _at	0.01954851	0.00023762	tumor necrosis factor receptor ...
31792 _at	0.00267723	0.00023762	annexin A3 (ANXA3)
33289_f _at	0.00365941	0.00023762	zinc finger protein 263 (ZNF263)
953_g _at	0.01698218	0.00023762	Gene function unknown
35799 _at	0.02151287	0.00025347	DnaJ (Hsp40) homolog, subfamily B
2035_s _at	0.00327921	0.00026931	enolase 1, (alpha) (ENO1)
31318 _at	0.03653069	0.00028515	Gene function unknown
296 _at	0.03504159	0.00030099	Gene function unknown

Network of CD44

- Build the network for CD44 (denoted by a) using the top 15 probes, where 4 probes with unknown biological functions are not considered, for the ease of interpretation.
- First, conducted a pairwise analysis for all paired edges, and found 8 more edges significant than the previous result (11 edges) at the threshold $p < 0.0003$. This formed one candidate network.
- Second, relaxing the threshold to $p < 0.005$, we found 55 edges significant. This formed another candidate network.
- Third, invoked COMP-AIC and found Network 1 is better supported by the data and strong biological evidences.



Network 1 (19 edges)



Network 2 (55 edges)

Concluding Remarks

- In the reconstruction of gene networks, the transition probability appears to be more appealing than the correlation-based dependency measures to characterize gene-gene dependency/interaction.
- The CL framework provides an efficient dimension reduction inference to deal with high-throughput gene expression data that have complicated dependency structures.
- Also, the CL framework allows us to estimate parameters in the pairwise transition probabilities when they are modeled by some common features, e.g environmental covariates.
- Assuming stationary transition is a limitation of the method. Extensions of this work would lead to better and faster machinery to the reconstruction of gene networks.

Thanks For Your Attention!