NETWORK-BASED GENE SET ANALYSIS WITH INCOMPLETE NETWORK INFORMATION

ALEXEJ GOSSMANN

TULANE UNIVERSITY

DEPT. OF BIOSTATISTICS AND BIOINFORMATICS JOURNAL CLUB

2015/9/4

THIS PRESENTATION IS BASED ON:

1. A. Shojaie and G. Michailidis, *Analysis of Gene Sets Based on the Underlying Regulatory Network*. J of Computational Biol. 2009.
2. A. Shojaie and G. Michailidis, *Network Enrichment Analysis in Complex Experiments*. Stat Appl Genet Mol Biol. 2010.
3. J. Ma, A. Shojaie and G. Michailidis, *Network-Based Pathway Enrichment Analysis with Incomplete Network Information*. arXiv e-prints. 2014.

BACKGROUND

MOTIVATION

- Test the significance of a pre-specified subnetwork (e.g. pathway).
- Incorporate the network structure.
- Consider changes in the network structure between different experimental conditions (e.g. case — control).
- Consider changes in the gene (protein, metabolite) expression.

BACKGROUND

GENE SET ENRICHMENT ANALYSIS (SUBRAMANIAN ET. AL., 2005, EFRON AND TIBSHIRANI, 2007).

- Tests for the *joint* effect of biologically related groups of genes.
- Higher power and better interpretability than single gene analysis.
- Association measures (e.g. p-values) are computed for each gene separately. They are then combined into an *enrichment score* for each gene set without direct incorporation of correlations between genes.
- These methods do not incorporate network information.

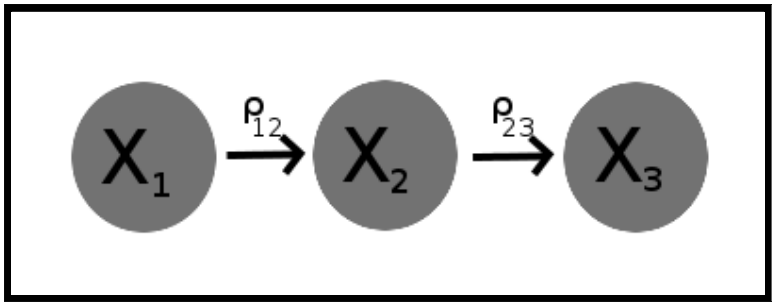
NETWORK-BASED GENE SET ANALYSIS (NETGSA)

- Combines the ideas of gene set analysis methods, network-based single gene analysis, and linear mixed models.
- Assesses changes in gene expression as well as network structure of arbitrary subnetworks (e.g. pathways) between different experimental conditions (e.g. case — control).
- Provides a general framework for inference in complex experiments using the linear mixed models theory.

MAIN IDEA

LINEAR MIXED MODEL REPRESENTATION

Very simple graph:



$$
\begin{aligned}
X_1 &= \gamma_1, \\
X_2 &= \rho_{12} X_1 + \gamma_2 = \rho_{12}\gamma_1 + \gamma_2, \\
X_3 &= \rho_{23} X_2 + \gamma_3 = \rho_{23}\rho_{12}\gamma_1 + \rho_{23}\gamma_2 + \gamma_3.
\end{aligned}
$$

where $\gamma_i \sim N(\mu_i, \sigma_\gamma^2)$.

$$
\Rightarrow X = \Lambda\gamma, \qquad \gamma \sim N(\mu, \sigma_\gamma^2 I).
$$

MAIN IDEA

LINEAR MIXED MODEL REPRESENTATION

- Let Y be the ith sample in the expression data.
- Assume that $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N(0, \sigma_\varepsilon^2 I)$ the noise.
- It follows that

$$Y = \Lambda\gamma + \varepsilon,$$
$$Y \sim N(\Lambda\mu, \sigma_\gamma^2 \Lambda\Lambda^T + \sigma_\varepsilon^2 I).$$

MAIN IDEA

LINEAR MIXED MODEL REPRESENTATION

- Multiple experimental conditions, e.g. case — control:

$$\begin{aligned} Y^C &= \Lambda^C \gamma^C + \varepsilon, \quad & \gamma^C &\sim N(\mu^C, \sigma_\gamma^2 I), \\ Y^T &= \Lambda^T \gamma^T + \varepsilon, \quad & \gamma^T &\sim N(\mu^T, \sigma_\gamma^2 I). \end{aligned}$$

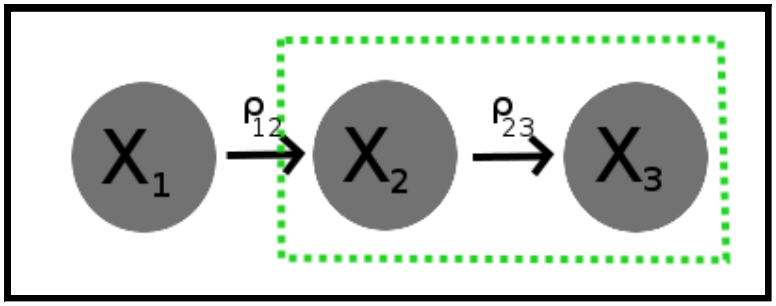
- Let $\beta := ((\mu^C)', (\mu^T)')'$, and rearrange Y, γ and ε into $np \times 1$ vectors, and redefine γ such that $E(\gamma) = 0$. Then we obtain a linear mixed model:

$$Y = \Psi\beta + \Pi\gamma + \varepsilon,$$

where $\gamma \sim N(0, \sigma_\gamma^2 I), \Pi = I \otimes \Lambda \in \mathbb{R}^{np \times np}$, $\Psi = \chi \otimes \Lambda \in \mathbb{R}^{np \times 2p}$ ($\chi \in \mathbb{R}^{n \times 2}$ decodes membership in treatment or control group).

MAIN IDEA

INFERENCE



- Use the LMM framework to test whether a subnetwork differs w.r.t. the network structure or gene expression levels between cases and controls.
- Any hypothesis test of the following form can be performed:

$$H_0 : l'\beta = 0 \quad \text{vs.} \quad H_1 : l'\beta \neq 0.$$

- Choose a contrast vector l that includes the interaction between genes 2 and 3 and excludes gene 1.

LMM GENERAL FORM

- $Y = \Psi\beta + \Pi\gamma + \varepsilon$, where β and γ are the fixed and respectively the random effects coefficients.
- $\gamma \sim N(0, \sigma_\gamma^2 I), \varepsilon \sim N(0, R(\theta_\varepsilon))$, where *temporal correlations* may be incorporated through R.
- Estimates of σ_γ^2 and θ_ε^2 are obtained based on the REML criterion.
- MLE of β:

$$\hat{\beta} = (\Psi' \hat{W}^{-1} \Psi)^{-1} \Psi' \hat{W}^{-1} Y,$$

where $W = \sigma_\gamma^2 \Pi\Pi' + R$.

LMM GENERAL FORM

INFLUENCE MATRIX

- The fixed and random effects design matrices Ψ and Π are determined from the *influence matrix* Λ.
- The influence matrix represents the effect of each gene on all of the other genes in the network.
- In general, $\Lambda = \sum_{r=0}^{\infty} A^r$, where A is the adjacency matrix of the network.
- In practice, one can compute

$$\Lambda = \lim_{\zeta \to 0} (I - \mathcal{L}(\zeta))^{-1} = (I - \mathcal{L}(0))^{+},$$

where $(\cdot)^{+}$ denotes the Moore-Penrose pseudo-inverse, and

$$\mathcal{L}(\zeta)_{ij} = \frac{A_{ij}}{(\sum_{j=1}^{p} |A_{ij}|) + \zeta}.$$

INFERENCE

INFERENCE

CHOICE OF THE CONTRAST VECTOR

- The choice of l in $H_0 : l'\beta = 0$ is crucial.
- For any indicator vector b determining a specific subnetwork, the vector $(b'\Lambda \cdot b)\beta$ includes the effects of all nodes in b, but excludes the effects of nodes not in b (here $\cdot$ denotes the Hadamard product).

NETWORK ESTIMATION UNDER EXTERNAL INFORMATION CONSTRAINS

- The correlation structure underlying a molecular network can be represented by a graph $G = (V, E)$.
- The edge set E is represented by the adjacency matrix A, where $A_{ij} \in (-1, 1)$ represents the strength of association between the respective nodes.
- Let E_1 contain known edges, and E_0 contain pairs of nodes with prior knowledge of no interaction.
- The objective is to estimate A subject to the external information in E_0 and E_1.

NETWORK ESTIMATION UNDER EXTERNAL INFORMATION CONSTRAINS

Using the framework of **Gaussian graphical models** the maximum likelihood estimate of A is given by

$$\min_{A \succ 0} (\mathrm{tr}(A\hat{\Sigma}) - \log\det A),$$

subject to

$$\sum_{i \neq j, (i,j) \notin E_0 \cup E_1} |A_{ij}| \leq t,$$

$$A_{ij} = 0, (i,j) \in E_0,$$

$$A_{ij} \neq 0, (i,j) \in E_1,$$

where $\hat{\Sigma}$ denotes the empirical covariance matrix.

NETWORK ESTIMATION UNDER EXTERNAL INFORMATION CONSTRAINS

Denote Z the $m \times p$ data matrix.

1. For every node i:

$$\hat{\theta}^i = \text{argmin}_\theta \frac{1}{m} ||Z_i - Z_{-i}\theta||_2^2 + 2\lambda \sum_{j \neq i} t_j |\theta_j|,$$

   where t_j is 0 if j is known to be a neighbor of i, ∞ if j and i are known to be disconnected, and 1 otherwise.

2. Get the network structure $\hat{E}$, where an edge (i, j) is estimated if $\hat{\theta}_i^j \neq 0$ or $\hat{\theta}_j^i \neq 0$.

3. $\hat{A} = \min_{A \in S_{\hat{E}}^p} (\text{tr}(A\hat{\Sigma}) - \log \det A)$, where S_E^p denotes the set of all $p \times p$ positive definite matrices such that $A_{ij} = 0$ for all $(i, j) \notin E$ with $i \neq j$.

NETWORK ESTIMATION UNDER EXTERNAL INFORMATION CONSTRAINS

THEOREM 2.3 IN MA ET. AL. (2014)

Let A_0 be the adjacency matrix of the true model. Under certain assumptions, with high probability it holds that

$$\|\hat{A} - A_0\|_2^2 \leq \|\hat{A} - A_0\|_F = O_P\left[(S\log(p - rp)/m)^{1/2}\right],$$

where S is the total number of true edges and r is the percentage of external information.

THEOREM 2.1 IN SHOJAIE ET. AL. (2010) AND COROLLARY 3.1 IN MA ET. AL. (2014)

Assume that $S = o(m/\log p)$ then the proposed test statistic (shown previously) based on the estimated network is an asymptotically most powerful unbiased test.

SIMULATION RESULTS

- First experiment:
  - $m = 40$ and $p = 64$.
  - 8 subnetworks, each with 8 members.
  - There is a 20% probability for subnetworks to connect to each other.
  - Under the null, all subnetworks have the same topology and all nodes have mean expression 1.
  - Under the alternative, the proportion of nodes that have mean changes of magnitude 1 is 0%, 40%, 40%, 50%, 0%, 40%, 40%, 50% for subnetworks 1—8, and subnetworks 5—8 differ in network structure to the null equivalent by 10%.
- Second experiment has a similar design except:
  - $m = 100$ and $p = 160$ with 20 members in each subnetwork.
  - Mean changes of magnitude 0.3 for 0%, 40%, 60%, 80%, 0%, 40%, 60%, 80% of nodes.

SIMULATION RESULTS

DEVIANCE MEASURES FOR NETWORK ESTIMATION

Let r denote the percentage of prior information about the network structure.

		$p = 64$				$p = 160$			
	r	FPR(%) ♯	FNR(%) ♭	MCC †	Fnorm ‡	FPR(%)	FNR(%)	MCC	Fnorm
	0.0	7.99	9.04	0.48	0.57	2.90	1.16	0.54	0.30
Null	0.2	6.68	9.76	0.52	0.55	2.29	1.53	0.59	0.28
	0.8	1.74	3.91	0.79	0.38	0.55	1.07	0.83	0.19
	0.0	8.27	8.48	0.48	0.54	2.44	0.72	0.58	0.31
Alternative	0.2	6.87	8.93	0.51	0.51	1.96	0.93	0.62	0.29
	0.8	1.82	0.91	0.80	0.35	0.44	1.43	0.85	0.18

♯ FPR(%), false positive rate in percentage;
♭ FNR(%), false negative rate in percentage;
† MCC, Matthews correlation coefficient;
‡ Fnorm, Frobenius norm loss.

SIMULATION RESULTS

ESTIMATED POWERS FOR EACH PATHWAY

NetGSA is compared to **Gene Set Analysis** (Efron & Tibshirani, 2007). Powers were calculated based on the FDR controlling procedure of Benjamini & Hochberg (1995) with $q = 0.05$.

	$p = 64$						$p = 160$					
Pathway	0.2 ♯	0.8♯	E♭	T†	GSA-s‡	GSA-c ‡	0.2	0.8	E	T	GSA-s	GSA-c
1	0.07	0.09	0.02	0.05	0.06	0.00	0.07	0.07	0.03	0.05	0.06	0.11
2	0.14	0.13	0.09	0.14	0.18	0.00	0.32	0.29	0.14	0.22	0.32	0.02
3	0.47	0.59	0.66	0.78	0.62	0.02	0.62	0.64	0.63	0.73	0.78	0.01
4	0.93	0.92	0.94	0.98	0.68	0.22	0.75	0.88	0.90	0.94	0.95	0.08
5	0.26	0.18	0.09	0.13	0.10	0.00	0.48	0.38	0.25	0.32	0.16	0.02
6	0.39	0.28	0.35	0.47	0.26	0.00	0.73	0.66	0.54	0.66	0.42	0.01
7	0.61	0.65	0.82	0.92	0.63	0.02	0.93	0.94	0.99	0.99	0.90	0.02
8	0.78	0.79	0.82	0.91	0.51	0.04	0.96	0.97	0.97	1.00	0.97	0.09

♯ 0.2/0.8 refer to Network-based Gene Set Analysis with 20%/80% external information;

♭ E refers to Network-based Gene Set Analysis with the exact networks;

† T refers to the true power;

‡ GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 1000 permutations, respectively.

APPLICATION TO GENOMICS AND METABOLOMICS

APPLICATION 1

- The metabolomics data set (Putluri et al., 2011) examines changes in the metabolic profile between cancer and adjacent benign tissue specimens, with 31 samples from the cancer class and 28 from a benign class.
- The total number of metabolites detected is 63.
- The network of metabolic interactions is estimated subject to external information extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG).
- Tests of differential activity of biochemical pathways extracted from KEGG were performed using a false discovery rate correction with $q = 0.001$.

APPLICATION TO GENOMICS AND METABOLOMICS

APPLICATION 1

Pathway	NetGSA[♯]	GSA-s [†]	GSA-c [†]
Fatty acid biosynthesis	< 0.001	1.000	1.000
Aminoacyl-tRNA biosynthesis	< 0.001	< 0.001	0.458
Tryptophan metabolism	< 0.001	< 0.001	0.338
Pantothenate and CoA biosynthesis	< 0.001	< 0.001	0.395
Phenylalanine, tyrosine and tryptophan biosynthesis	< 0.001	1.000	1.000
beta-Alanine metabolism	< 0.001	< 0.001	0.338
Neuroactive ligand-receptor interaction	< 0.001	0.006	0.338
Phenylalanine metabolism	< 0.001	< 0.001	0.542
Pyrimidine metabolism	0.003	< 0.001	0.395
ABC transporters	0.005	< 0.001	0.624
Histidine metabolism	0.029	< 0.001	0.111
hsa00220	0.146	< 0.001	0.672

[♯] NetGSA refers to Network-based Gene Set Analysis;

[†] GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

APPLICATION TO GENOMICS AND METABOLOMICS

APPLICATION 2

- Gene expression profiles of 1416 genes for 62 normal and 24 lung cancer patients.
- Interaction networks for both normal and lung cancer conditions were estimated based on the external topology information from the BioGRID Database.
- Tests for enrichment of 61 pathways extracted from the KEGG data base were performed using a false discovery rate correction with $q = 0.001$.

APPLICATION TO GENOMICS AND METABOLOMICS

APPLICATION 2

Pathway	NetGSA[♯]	GSA-s [†]	GSA-c [†]
Glycerophospholipid metabolism	< 0.001	0.338	0.389
PPAR signaling pathway	< 0.001	0.338	1.000
Glycine, serine and threonine metabolism	< 0.001	0.137	0.357
Cysteine and methionine metabolism	< 0.001	0.229	0.357
Glycerolipid metabolism	< 0.001	0.404	0.520
TGF-beta signaling pathway	< 0.001	0.404	0.518
Fructose and mannose metabolism	< 0.001	0.308	0.408
Neurotrophin signaling pathway	< 0.001	0.308	0.588
Phosphatidylinositol signaling system	< 0.001	0.404	0.379
ErbB signaling pathway	< 0.001	0.035	0.249
mTOR signaling pathway	< 0.001	0.138	0.357

[♯] NetGSA refers to Network-based Gene Set Analysis;

[†] GSA-s/GSA-c refer to Gene Set Analysis with self-contained/competitive null hypothesis in 3000 permutations, respectively.

SUMMARY & IDEAS

- NetGSA can be more powerful than methods which are not network-based.
- NetGSA is computationally challenging for large networks.
- Generalized linear mixed model framework can be used to adapt the method to discrete data (e.g. SNP data).
- Other types of data can be integrated into the linear mixed model as additional fixed effects terms.
- Correlations among the phenotype samples can be incorporated into a regression model by extracting a random effects structure from a network among the subjects.