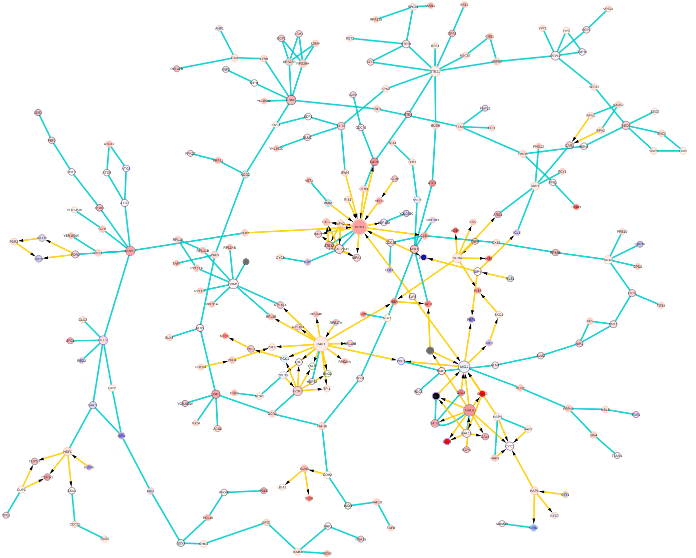
Pathway & Network Analysis of Omics Data: Network-Based Pathway Enrichment Analysis

Ali Shojaie
Department of Biostatistics
University of Washington
`faculty.washington.edu/ashojaie`

Summer Institute for Statistical Genetics – Australia, 2017

Yeast GAL Pathway

Ideker et al, 2001



Issues of Interest

- Incorporate the network information
- Consider changes in the gene (protein, metabolite) expressions
- Consider changes in the network structure
- Test the "effect" of pre-specified subnetwork/pathway, sharing common biological function, chromosomal location etc
- A general framework for inference in complex experiments

Recap: Gene Set Enrichment Analysis

Subramanian et al. (2005) proposed gene set enrichment analysis (GSEA); *Efron & Tibshirani* (2007) formalized the GSEA approach, and proposed a more efficient test statistic

- Test the significance of *a priori* defined gene sets
- Preserve the correlation among genes in the gene set
- Based on a competitive null hypothesis, where activity of each pathway is compared with other pathways, often using a permutation test
- Competitive tests of enrichment assume that a small number of genes have differential activity, and are very sensitive to the choice of gene sets, they also problem with
- Self-contained tests address these issues, but may be less efficient or sensitive to model assumptions (*Goemen & Buhlmann* (2007), *Ackermann & Strimmer* (2009))

Signaling Pathway Impact Analysis (SPIA)

- Combines classical overrepresentation analysis (ORA) with measure of perturbation of a given pathway
- A permutation procedure is used to assess the significance of the observed pathway perturbation (difficult to extend to comparison of > 2 conditions)
- Currently not applicable to all pathways (more later)
- Models each pathway separately (i.e., ignores connections between pathways)
- Implemented in the Bioconductor package `SPIA`

The SPIA Methodology

SPIA combines two types of evidence

(i) the over-representation of DE genes in a given pathway

- measured by the p-value for the given number of DE genes

$$P_{NDE} = P(X \geq N_{DE} \mid H_0)$$

The SPIA Methodology

SPIA combines two types of evidence

(ii) the abnormal perturbation of the pathway

- the perturbation for each gene in the pathway is defined as

$$PF(g_i) = \Delta E(g_i) + \sum_{j=1}^{p} \beta_{ij} \frac{PF(g_j)}{N_{DS}(g_j)}$$

  - $PF(g_i)$ is the perturbation factor of gene i (not known)
  - β_{ij} measures the effect of gene j on i; currently, $\beta_{ij} = 1$ if $j \to i$
  - $\Delta E(g_i)$ is the fold change in expression of gene i
  - $N_{DS}(g_j)$ is the number of genes downstream of gene j

The SPIA Methodology

- The accumulated activity of each gene is defined as

$$ACC(g_i) = B \cdot (I - B)^{-1} \Delta E$$

  - B is the normalized matrix of β's: $B_{ij} = \beta_{ij} / N_{DS}(g_j)$
  - ΔE is the vector of fold changes
  - Requires B to be invertible — would not work otherwise

- The total accumulated pathway perturbation is given by

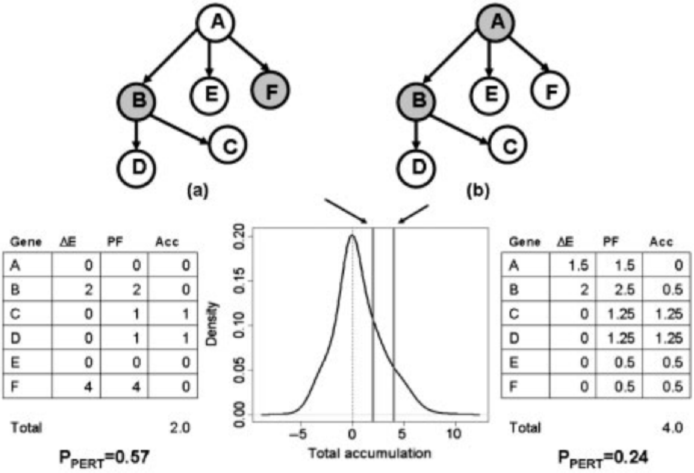
$$t_A = \sum_i ACC(g_i)$$

- The p-value for pathway perturbation is given by

$$P_{PERT} = P(T_A \geq t_A \mid H_0),$$

which is calculated by permutation

The SPIA Methodology



Gene	ΔE	PF	Acc
A	0	0	0
B	2	2	0
C	0	1	1
D	0	1	1
E	0	0	0
F	4	4	0

Total 2.0

P_{PERT}=0.57

Gene	ΔE	PF	Acc
A	1.5	1.5	0
B	2	2.5	0.5
C	0	1.25	1.25
D	0	1.25	1.25
E	0	0.5	0.5
F	0	0.5	0.5

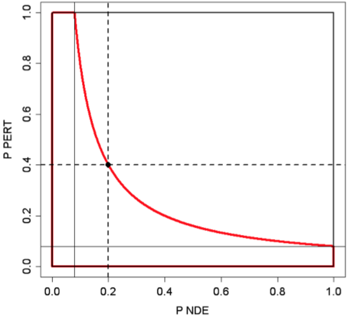
Total 4.0

P_{PERT}=0.24

The SPIA Methodology

SPIA combines two types of evidence

- The final p-value for each pathway is calculated based on the p-values from parts (i) and (ii):
  - $P_G(k) = c_k - c_k \ln(c_k)$
  - $c_k = P_{NDE}(k) P_{PERT}(k)$



An Example in R: Data on Colorectal Cancer

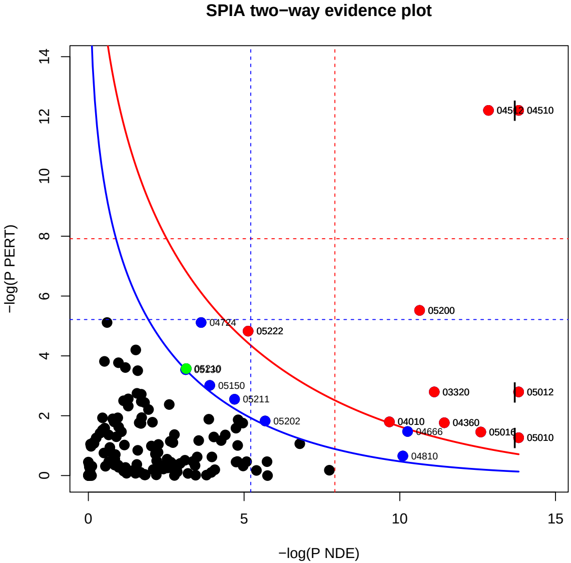
```
data(colorectalcancer)

#pathway analysis using SPIA
#use nB=2000 or higher for more accurate results
#uses older version of KEGG signaling pathways graphs
res <- spia(de=DE_Colorectal, all=ALL_Colorectal, organism="hsa", beta=NULL,
    nB=2000, plots=FALSE, verbose=TRUE, combine="fisher")

#now combine pNDE and pPERT using the normal inversion method without
#running spia function again
res$pG=combfunc(res$pNDE,res$pPERT,combine="norminv")
res$pGFdr=p.adjust(res$pG,"fdr")
res$pGFWER=p.adjust(res$pG,"bonferroni")
plotP(res,threshold=0.05)

#highlight the colorectal cancer pathway in green
points(I(-log(pPERT))~I(-log(pNDE)),data=res[res$ID=="05210",],col="green",
    pch=19,cex=1.5)
```

The SPIA Methodology



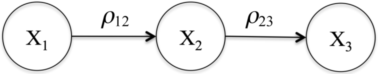
Network-Based Gene Set Analysis (NetGSA)

- Combines the ideas of gene set analysis methods, and network-based single gene analysis
- Generalizes SPIA, to allow for more complex experiments & incorporate interactions among pathways
- Assesses the overall behavior of arbitrary subnetworks (pathways): changes in gene expression & network structure
- Uses latent variables to model the interaction between genes defined by the network
- Uses mixed linear models for inference in complex data
- Computationally challenging for large networks (OK up to 3-4K nodes)

Problem Setup

- Gene (protein/metabolite) expression data for K experimental conditions and J_k time points
- Network information (partially) available in the form of a directed weighted graph $G = (V, E)$, with vertex set V corresponding to the genes/proteins/metabolites and edge set E capturing their associations
- Networks with directed $j \rightarrow k$ and/or undirected $j \leftrightarrow k$ edges
- Edges capture effects of nodes on their neighbors; the weight associated with each edge corresponds to partial correlations
- Represent the network by its adjacency matrix A: $A_{jk} \neq 0$ iff $k \rightarrow j$ and for undirected edges, $A_{jk} = A_{kj}$
- Pathways defined *a priori* based on common biological functions, etc

The Latent Variable Model: Main Idea



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

Thus $X = \Lambda\gamma$ where

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \rho_{12} & 1 & 0 \\ \rho_{12}\rho_{23} & \rho_{23} & 1 \end{pmatrix}$$

The Latent Variable Model

- Let Y be the ith sample in the expression data
- Let $Y = X + \varepsilon$, with X the signal and $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ the noise
- The influence matrix Λ measures the propagated effect of genes on each other through the network, and can be calculated based on the adjacency matrix A
- Using $X = \Lambda\gamma$, we get

$$Y = \Lambda\gamma + \varepsilon, \quad \Rightarrow \quad Y \sim N_p(\Lambda\mu, \sigma_\gamma^2 \Lambda\Lambda' + \sigma_\varepsilon^2 I_p)$$

where $\gamma \sim N_p(\mu, \sigma_\gamma^2 I_p)$ are latent variables

Mixed Linear Model Representation

Rearranging the expression matrix into *np*-vector **Y**, we can write

$$\mathbf{Y} = \mathbf{\Psi}\boldsymbol{\beta} + \mathbf{\Pi}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are fixed and random effect parameters and

$$\boldsymbol{\varepsilon} \sim N_{np}(\mathbf{0}, R(\theta_\varepsilon)), \qquad \boldsymbol{\gamma} \sim N_{np}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I_{np}})$$

- Can accommodate e.g. temporal Correlation through R

In general, the design matrices, $\mathbf{\Psi}$ and $\mathbf{\Pi}$ depend on the experimental settings (similar to ANOVA), and are functions of Λ

Inference using MLM

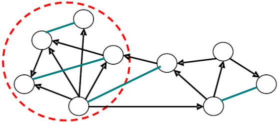
- For any contrast vector ℓ (a linear combination of fixed effects), can test:

$$H_0 : \ell\beta = 0 \quad vs. \quad H_1 : \ell\beta \neq 0$$

using the test statistic

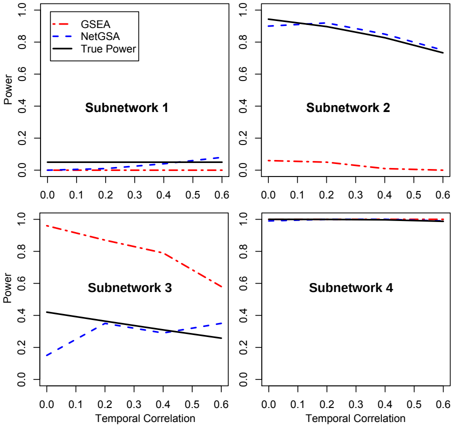
$$T = \frac{\ell\hat{\beta}}{\sqrt{\ell\hat{C}\ell'}} \qquad \text{with} \qquad C = \left(\Psi' W^{-1} \Psi\right)^{-1}$$

- Under the null, T has approximately *t*-distribution with degrees of freedom that needs to be estimated.
- ℓ should *de-couple the effects in each pathway* from others



Comparison in Simulated Data

Subnetwork	Mean	Network Influence
1	$\mu_1 = \mu_2 = 1$	$\rho_1 = \rho_2 = 0.2$
2	$\mu_1 = 1, \mu_2 = 2$	$\rho_1 = \rho_2 = 0.2$
3	$\mu_1 = \mu_2 = 1$	$\rho_1 = 0.2, \rho_2 = 0.7$
4	$\mu_1 = 1, \mu_2 = 2$	$\rho_1 = 0.2, \rho_2 = 0.7$



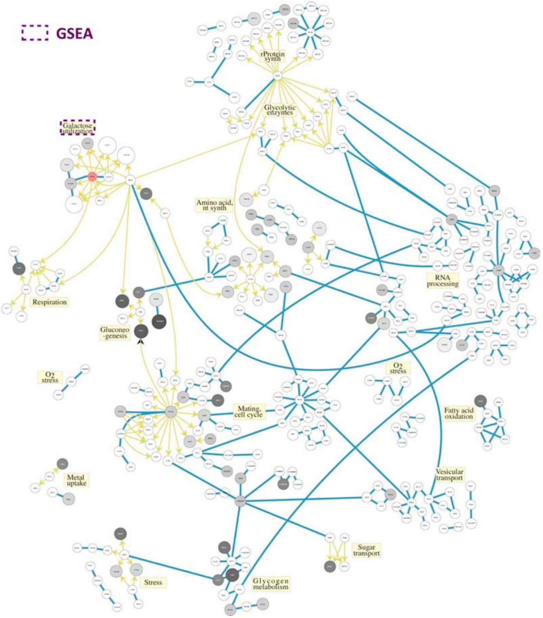
Yeast Galactose Utilization Pathway

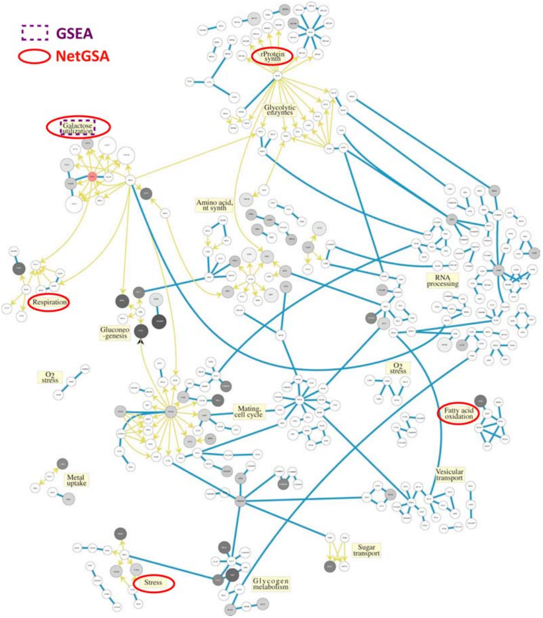
Ideker et al (2001) data on yeast Galactose Utilization Pathway

- Gene expression data for 2 experimental conditions: (gal+) and (gal–)
- Gene-gene and protein-gene interactions as well as association weights found from previous studies
- Q: which pathways respond to the change in growth medium?

Analysis of Yeast GAL Data

- Data:
  - gene expression data for 343 genes
  - 419 interactions found from previous studies and integration with protein expression (association among genes also available)
- Results:
  - GSEA finds *Galactose Utilization Pathway* significant
  - NetGSA finds several other pathways with biologically meaningful functions related to survival of yeast cells in gal–


GSEA
rProtein synth
Glycolytic enzymes
Galactose utilization
Amino acid, nt synth
Respiration
Gluconeo -genesis
O2 stress
RNA processing
O2 stress
Mating, cell cycle
Fatty acid oxidation
Metal uptake
Vesicular transport
Sugar transport
Stress
Glycogen metabolism


GSEA
NetGSA
rProtein synth
Glycolytic enzymes
Galactose utilization
Amino acid, nt synth
RNA processing
Respiration
Gluconeo -genesis
O2 stress
O2 stress
Fatty acid oxidation
Mating, cell cycle
Metal uptake
Vesicular transport
Sugar transport
Stress
Glycogen metabolism

Environmental Stress Response in Yeast

Gene expression data on Yeast Environmental Stress Response (ESR) (*Gasch et al.*, 2000)

- 3 combinations of experimental factor, heat shock and osmotic changes (sorbitol), over 3 time points
- Temporal correlation
- Network correlation
- Q: Which pathways indicate response to environmental stress
    - in different experimental conditions
    - over time

Yeast ESR Data

Gasch et al (2000)

- Gene Expression Data

Experiment	Obs. Time (after 33C)
Mild heat shock (*29C to 33C)*, no sorbitol	5, 15, 30 min
Mild Heat Shock, 1M sorbitol at 29C & 33C	5, 15, 30 min
Mild Heat Shock, 1M sorbitol at 29C	5, 15, 30 min

- Network Data
  - Use YeastNet (*Lee et al.*, 2007) for gene-gene interactions (102,000 interactions among 5,900 yeast genes)
  - Use independent experiments of *Gasch et al.* to estimate weights
  - Pathways are defined using GO functions

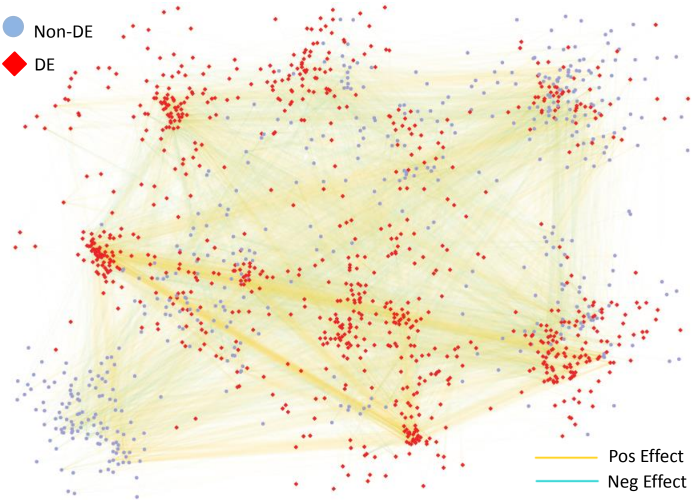
Model and Results

- Model: Let j and k be indices for time and levels of sorbitol

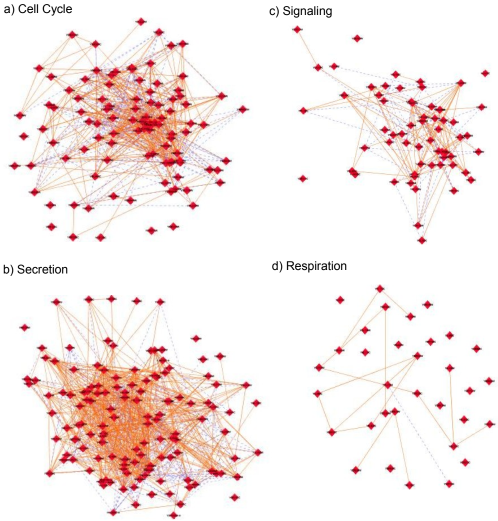
$$\mathbb{E}Y_{11} = \Lambda\mu, \qquad \mathbb{E}Y_{jk} = \Lambda(\mu + \alpha_j + \delta_k) \quad j, k = 2, 3$$

- Temporal correlation is modeled directly via R (as $AR(1)$ process)
- Results:
  - $\sim$ 3000 genes,
  - 47 pathways showed significant changes of expression
  - 24 pathways showed changes over time
  - 29 pathways showed changes in response to different sorbitol levels
  - 12 pathways showed both types of changes
  - Significant pathways overlap with the gene functions recognized by *Gasch et al.*

Yeast ESR Network

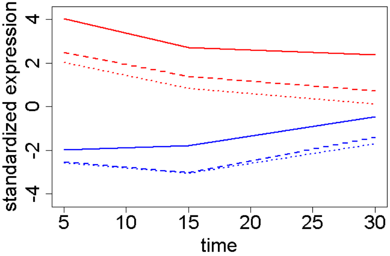


Significant subnetworks



Expression Profiles

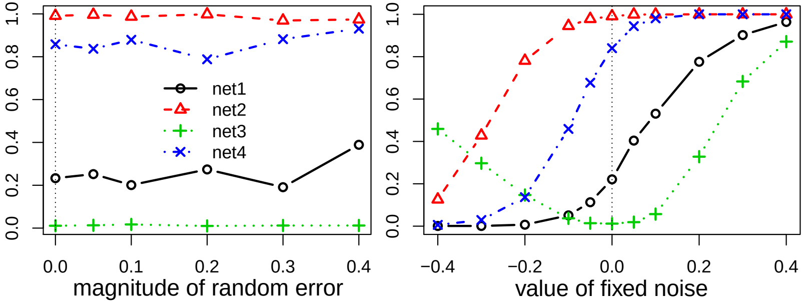
Average Standardized Expression Levels of Pathways



- Induced and Suppressed Pathways
- Can observe the transient patterns of expressions as predicted by *Gasch et al.*

Effect of Noise In Network Information

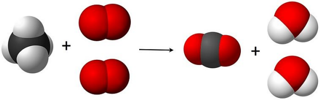
- Let $\tilde{A}$ be observed network information, and A be the truth.
- It can be shown that, if $\|\tilde{A} - A\|$ is small then, NetGSA still works (is *asymptotically most powerful unbiased test*)



Metabolic Profiling in Bladder Cancer

Metabolic profiling of bladder cancer (BCa) (*Putluri et al.*, 2012)

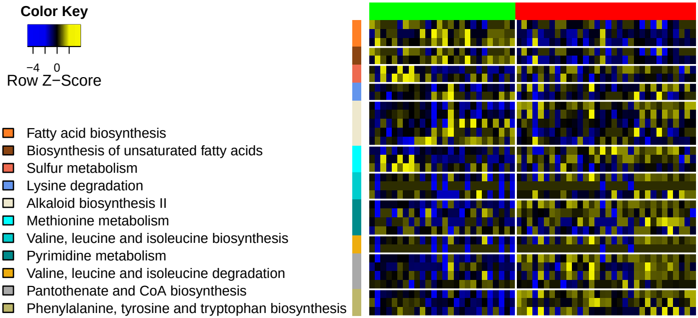
- 58 bladder cancer and adjacent benign samples
- Pathways information obtained from KEGG



- Varying number of identified metabolites per pathway (3-15)
- Q: Which pathways show differential activity in BCa?

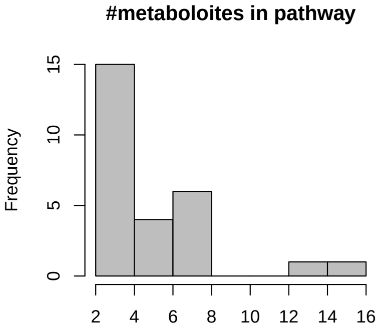
Metabolic Profiling in BCa

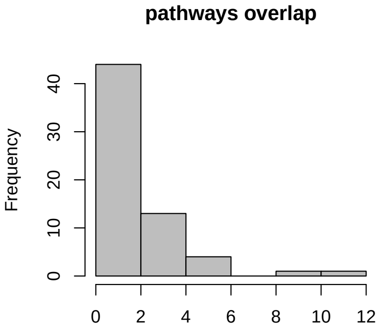
- 63 metabolites identified, mapped to 70 pathways
- 27 pathways with at least 3 members



Metabolic Profiling in BCa

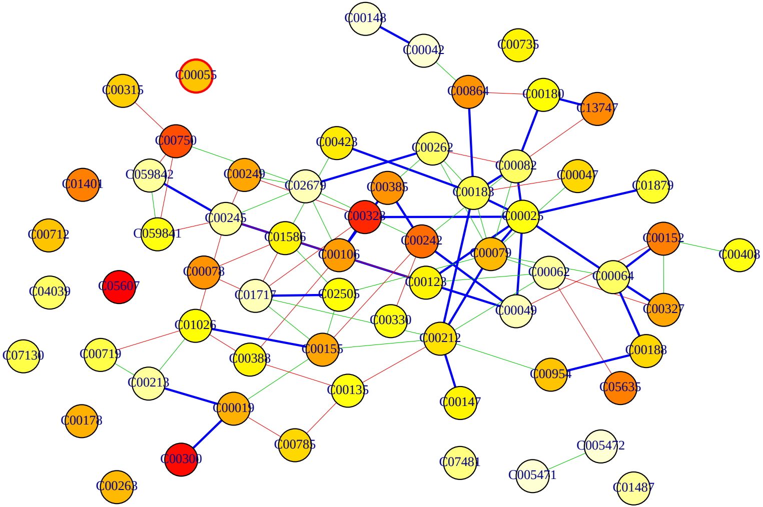
- Small pathway sizes & significant overlap among pathways


#metaboloites in pathway


pathways overlap

- Existing methods may not work well...

Metabolic Interaction Network



Significant Pathways

- GSEA does not identify any pathway as differential
- GSA identifies Fatty Acid Biosynthesis as differential
- NetGSA identifies another 7 pathways corresponding to Amino Acid Metabolism in BCa, also observed by *Putluri et al* (2012)

R package `netgsa`

- Basic usage:

```
NetGSA(A, x, y, B)
```

- `A`: list of m weighted adjacency matrices ($p \times p$) for conditions $1, \ldots, m$ (e.g. normal vs cancer), to capture network changes
- `B`: a $K \times P$ 0-1 matrix of pathway membership: $B_{k,j} = 1$ if gene/protein/metabolite j in pathway k
- Output: test statistics and p-values for each pathway
- `NetGSA` takes weighted `A`s as input. However, the package includes functions that allow you to enter a (partial) edge list as input, and estimate `A`s (only for undirected networks)

Summary

- Network-based enrichment analysis methods (SPIA, NetGSA) can be more powerful (if their assumptions are not violated!)
- Active area of research: a number of other methods have been recently proposed
- Focus is shifting towards estimating changes in the structure of networks: differential network biology[1]

[1] Ideker & Krogan (2012)