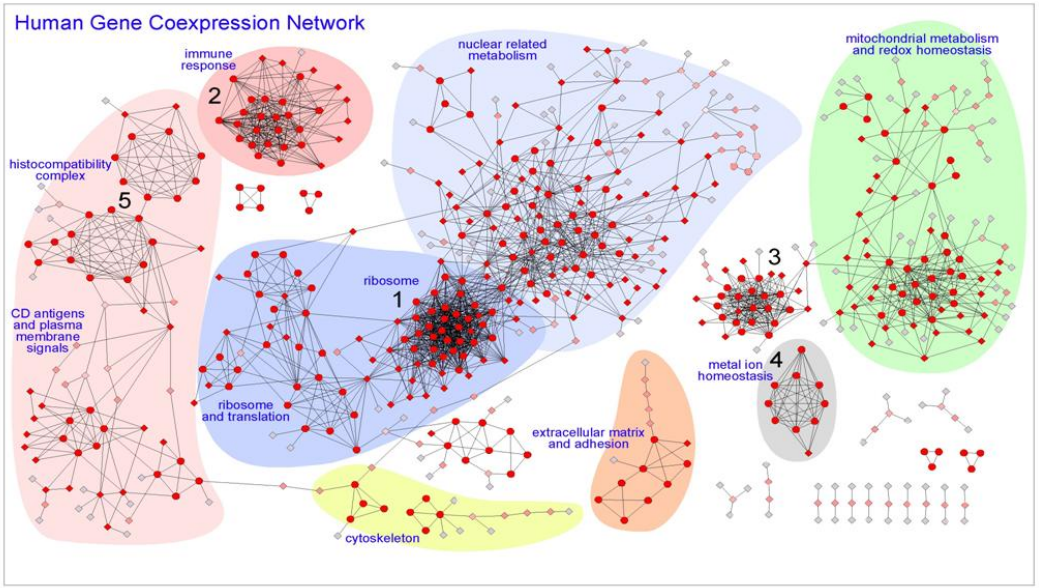
Network and Database for Omics Data Analysis

Agus Salim, Ph.D

Big Data Analytics: Application to Modern Genetics
Universitas Jember, 23 Dec 2015

Genes Network

- When analyzing Omics data, we tend to test each gene separately for association with outcome (disease)
- Treating each gene as 'singleton' is convenient, but is it appropriate?
- After all, experiments have shown that genes work in network with a collection of genes performing one major task



Network Database

- Over the years, the network has been mapped, refined and verified (experimentally or computationally)
- There are databases that contain network information, e.g.,
  - Gene Ontology Database (`http://geneontology.org/page/go-database`)
  - BioGRID (`http://thebiogrid.org/`: databases of interactions (mostly protein-protein and DNA-protein)
  - KEGG (`http://www.genome.jp/kegg/`: pathway and interaction databases
  - Reactome (`http://www.reactome.org/`): a carefully-curated pathway database

Incorporating Network into Data Analysis

One of the following approaches:

1. Test each gene separately, followed by testing for over/under-representation of genes from certain pathways among the statistically-significant genes (Example: GO Test, GSEA)
2. Incorporate the network from the start; the DE status of a gene depends on the status of its 'neighbor' (Markov Random Field-Based Algorithm)

Gene Ontology (GO) Test

Is the collection of significant genes over (under) represent genes from certain pathways?

- Uses the list of statistically-significant genes (e.g., from t-test) as input
- Perform the test pathway-by-pathway
- Uses Fisher's exact test for 2×2 table as the underlying stats tool.

Gene Ontology (GO) Test

- Uses the list of statistically-significant genes (e.g., from t-test) as input
- Perform the test pathway-by-pathway
- Uses Fisher's exact test for 2×2 table as the underlying stats tool.

Gene Ontology (GO) Test

Status	In pathway	Not in Pathway	Total
S	a	b	q
NS	c	d	$p - q$
Total	m_1	m_2	p

We can perform Fisher-exact test on this table.

Gene Ontology (GO) Test

If we use the following analogy with fair experiment involving marbles, Statistically significant (S) genes = drawn marbles, genes in the pathway = 'red' marbles, genes not in the pathway = 'blue' marbles.

If we draw q marbles from a bag containing m_1 'red' and $m_2 = p - m_1$ 'blue' marbles, then the chance of drawing exactly a 'red' ones is

$$\frac{C(m_1, a) \times C(p - m_1, b)}{C(p, q)}$$

- The number of genes from the pathway (A) follows Hypergeometric distribution, with

$$E(A) = m_1 \frac{q}{p}$$

Gene Ontology (GO) Test

The p-value can be calculated as follows (using p-value definition)

- For over-representation test: $P(\geq a)$
- For under-representation test: $P(\leq a)$

Gene-set Enrichment Analysis

- The GO test categorizes gene into 'S' and 'NS' categories, but within each category there is no differentiation based on the weight of evidence
- Some 'S' genes may be 'just' significant, likewise some 'NS' genes may just fail to meet the significance criterion.
- Gene-set Enrichment Analysis (GSEA) weights genes differently based on the strength of evidence

Gene-set Enrichment Analysis

Example: two-sample t-test with n_1 and n_2 subjects

- Apply two-sample t-test to each gene and let t_k be the test-statistic for gene k
- Assuming that data is Normally-distributed and equal variance, under H_0, $t_k \sim t_{n_1+n_2-2}$
- $E(t_k) = 0$, $Var(t_k) = \frac{n_1+n_2-2}{n_1+n_2-4} \approx 1$ if $n_1 + n_2$ is large enough
- Suppose we are testing a particular pathway and Ω set represent R number of genes in that pathway
- We form a new test-statistic

$$
\begin{aligned}
Z &= \frac{1}{\sqrt{R}} \sum_{k \in \Omega} t_k \\
&\approx \frac{1}{\sqrt{\sum_{k \in \Omega} var(t_k)}} \sum_{k \in \Omega} (t_k - E(t_k))
\end{aligned}
$$

GSEA

$$Z = \frac{1}{\sqrt{R}} \sum_{k \in \Omega} t_k$$

- Genes with stronger evidence (larger magnitude of t_k) will contribute more to the test-statistic
- By using Lindeberg's Central Limit Theorem, we can show that $Z \xrightarrow{d} N(0, 1)$
- Which means, we can obtain p-value by comparing Z to standard normal distribution

Markov Random Field (MRF) Approach

- Incorporating network information from the beginning
- The basic idea: the state of each gene (normal/supressed/enhanced) depends the status of its neighbor
- Neighbors are genes that are directly linked by edge(s)
- The probability of a normal gene is higher in the neighborhood of normal genes, and lower in the neighborhood full of suppressed/enhanced genes

Markov Random Field (MRF) Approach

- The state of each gene ONLY depends on the state of its direct neighbors
- There is probability distribution for gene under each state
- For each gene k, there are three states $x_k = -1, 0, 1$

$$p(x_k = d \mid \textit{neighbors}) \propto \exp\{\gamma_d + \beta_d u_k(d) - \beta_0 u_k(0)\}$$

where $(\gamma_d, \beta_d, \beta_0)$ are MRF-model parameters to be estimated from data, $u_k(d)$ is the proportion of gene k's neighbors with state $= d$.

Markov Random Field (MRF) Approach

- The state of each gene ONLY depends on the state of its direct neighbors
- There is probability distribution for gene under each state
- For each gene k, there are three states $x_k = -1, 0, 1$

$$p(x_k = d \mid \textit{neighbors}) \propto \exp\{\gamma_d + \beta_d u_k(d) - \beta_0 u_k(0)\}$$

where $(\gamma_d, \beta_d, \beta_0)$ are MRF-model parameters to be estimated from data, $u_k(d)$ is the proportion of gene k's neighbors with state $= d$.

- We have on-going work where we use Poisson-Gamma-Beta (PGB) distribution to model gene expression data under each state

Markov Random Field (MRF) Approach

- Computationally challenging, so we use a hybrid approach: Methods of Moments (MoM) and Maximum-Likelihood to estimate parameters.

$$f(y_i|x_i=0) = \frac{\alpha^{\kappa}\Gamma(y_{i..}+\kappa)}{\Gamma(\kappa)\prod_{j=1}^{m+n} y_{ij}!\,(m+n+\alpha)^{y_{i..}+\kappa}},$$

$$f(y_i|x_i=1) = \int\int \frac{e^{-m\lambda_i}\lambda_i^{y_{i.m}}}{\prod_{j=1}^{m} y_{ij}!} \cdot \frac{e^{-n\frac{\lambda_i}{\delta_i}}\frac{\lambda_i}{\delta_i}^{y_{i.n}}}{\prod_{j=m+1}^{m+n} y_{ij}!} g(\lambda_i;\kappa,\alpha)b(\delta_i;\sigma,\rho)d\lambda_i d\delta_i,$$

$$f(y_i|x_i=-1) = \int\int \frac{e^{-m\lambda_i}\lambda_i^{y_{i.m}}}{\prod_{j=1}^{m} y_{ij}!} \cdot \frac{e^{-n\lambda_i\delta_i}\lambda_i\delta_i^{y_{i.n}}}{\prod_{j=m+1}^{m+n} y_{ij}!} g(\lambda_i;\kappa,\alpha)b(\delta_i;\sigma,\rho)d\lambda_i d\delta_i$$

- Preliminary results: very encouraging (beat GO and GSEA!). We are hoping to submit for publication soon.

Some Useful References

- Falcon S, Gentleman R. (2007). Using GOstats to test gene lists for GO term association. *Bioinformatics* **23:** 257-8.
- Subramanian A et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA* **102:** 15545-50.
- Sajeewani IDM, Prendergast L, Salim A. PathDESeq: a powerful pathway-based differential expression analysis for sequencing data. To be submitted.