When is a pathway changed?

Sorin Draghici

Robert J. Sokol, MD Endowed Chair

Department of Computer Science
Department of Obstetrics and Gynecology
Wayne State University
A gene set is not a
What is a pathway?

- A sub-system with some properties
  - Has some components (e.g. genes, biochemical compounds, etc)
  - Which are linked by some interactions (eg. activation/repression, reactions, etc.)
  - Interacts with the rest of the system through well defined inputs and outputs
  - There are more shared properties between the components of the sub-system than between these and the rest of the system (all genes related to insulin signaling, all reactions involved in the Krebs cycle, etc.)

- Can be modeled well by a graph
  - Components are represented by nodes
  - Interactions are represented by directed edges
  - The edge direction represents signals, reactions, etc.
Problem definition

• Input:
  • A set of variables (e.g., genes, proteins, etc.) for which there are significant measured differences between the given phenotype and control - measured gene expression changes.
  • A set of pathways describing sub-systems involving the given variables (e.g., signaling pathways, metabolic pathways) - pathway database (e.g., KEGG, Reactome).

• Output:
  • Rank the sub-systems in the decreasing order of the amount of disruption suffered
  • If possible, identify those sub-systems for which the disruption is significant (i.e. unlikely to be due to chance or noise alone)
  • Identify potential mechanisms that led to the measured expression changes.
Overview of existing

• Over-representation analysis (ORA)
  - Relative enrichment (GenMapp, Dahlquist, 2002)
  - Model based statistical approach (Onto-Express, Draghici, 2003)
  - Resampling based approaches (SAFE, Barry, 2005)

• Functional class-scoring (FCS)
Classical approaches

- Relative enrichment
  - $N$ – genes on the array used
  - $K$ – differentially regulated genes
  - $NP$ – genes on the pathway
  - $NRP$ – differentially regulated genes on the pathway

$$RE = \frac{\frac{N_{RP}}{NP}}{\frac{K}{N}}$$
Classical approaches

Model based statistical approach (hypergeometric)

- $N$ – genes on the array used
- $K$ – differentially regulated genes
- $N_P$ – genes on the pathway
- $N_{RP}$ – differentially regulated genes on the pathway

The probability of having exactly $N_{RP}$ genes on the given pathway:

$$P(X = N_{RP}|N, M, K) = \binom{N_P}{N_{RP}} \binom{N - N_P}{K - N_{RP}} \binom{N}{K}$$

The probability of having more than or equal to $N_{RP}$ genes on the given pathway just by chance:

$$p = 1 - \sum_{i=0}^{N_{RP}} \binom{N_P}{i} \binom{N - N_P}{K - i} \binom{N}{K}$$
Classical approaches

- Resampling based statistical approach
  - SAFE - Significance Analysis of Function and Expression
  - t-test at gene level
  - Wilcoxon rank sum as default global statistic
  - Permutation based method
  - Uses FDR for multiple testing correction
Gene Set Enrichment Analysis (GSEA)

All the genes are ranked based on the correlation to the phenotype.

Going down in the ranked list, the Enrichment Score (ES) will be increased if a gene exists in the gene set ($P_{\text{hit}}$), and decreased if not ($P_{\text{miss}}$).

The final Gene Set Enrichment Score is the maximum deviation from zero.

$$P_{\text{hit}}(S, i) = \sum_{g_i \in S} \frac{|r_j|^p}{N'_R}, \quad \text{where} \quad N_R = \sum_{g_i \in S} |r_j|^p$$

$$P_{\text{miss}}(S, i) = \sum_{g_i \not\in S} \frac{1}{(N - N_H)}.$$
Limitations of gene set

- Classical techniques only consider the number of genes on a given pathway ignoring other crucial aspects
  - in the classical approach all genes are the same
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their identity
    (not good because a small change in a transcription factors could trigger large downstream effects, for instance)

- The position and role of the DE genes on the pathways
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their position and role on the pathway (eg INSR on insulin and adherens junction pathways)
Insulin signaling pathway

Insulin receptor (INSR) is the crucial gene on this pathway
Adherens junction

INSR is one of many tyrosine kinase receptors
Limitations of gene set

• Classical techniques only consider the number of genes on a given pathway ignoring other crucial aspects
  - in the classical approach all genes are the same
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their identity (not good because a small change in a transcription factors could trigger large downstream effects, for instance)

• The position and role of the DE genes on the pathways
  - all current methods yield the same results for any subset of genes on the pathway, irrespective of their position and role on the pathway (eg INSR on insulin and adherens junction pathways)
  - genes upstream on the pathway could have a much greater impact than genes downstream

• The topology of the pathway
  - all current methods yield the same results even if the topology of the pathway is drastically changed
  - genes with many outgoing regulatory connections could perturb the pathway more than genes with few such connections
  - are the DE genes scattered randomly across the pathway or grouped together in a clique that might suggest a coherent perturbation propagating in the pathway?

• The specific expression values measured are completely ignored by all existing techniques
Impact analysis

For each input gene \( g \), a gene perturbation factor \( PF(g) \) depends on:

- User-provided normalized fold change of the gene
- Number and amount of perturbation of genes upstream
- The interactions on the pathway (its topology)

\[
PF(g) = \Delta E(g) + \sum_{u \in US_g} \beta_{ug} \frac{PF(u)}{N_{ds}(u)}
\]

\( PF(g) \) – perturbation factor of gene \( g \)
\( \Delta E(g) \) – change in expression level of gene \( g \)
\( US_g \) – set of genes directly upstream of \( g \)
\( N_{ds}(u) \) - number of genes directly downstream of \( u \)
\( \beta_{ug} \) - efficiency of the connection between \( u \) and \( g \)
Pathway perturbation

For each pathway, the impact of topology, gene interactions, and gene fold changes come into play and are captured through a pathway perturbation factor:

$$PF(P_i) = \frac{\sum_{g \in P_i} |PF(g)|}{|\Delta E| \cdot N_{\text{de}}(P_i)}$$

$$|\Delta E| = \sum_{k=1}^{N} \frac{|\Delta E|}{N}$$

mean fold change over all DE genes

$N_{\text{de}}(P_i)$ is the number of DE genes on the given pathway $P_i$

$PF(g)$ is the perturbation of the gene $g$
Pathway impact factors

The impact factor of a pathway depends on:

1. A probabilistic term - depends on the proportion of differentially regulated genes on the given pathway
2. A perturbation term - depends on the specific genes that are differentially regulated, the measured amount of expression change and the interactions on the pathway

\[ IF(P_i) = \log \left( \frac{1}{p_i} \right) + PF(P_i) \]

\[ P_i \] - pathway i
p-value calculated based
What the impact

Hypergeometric and perturbation analyses

\[-\log(P(\text{PF, pf}|H_0)) \leq \alpha \leq -\log(P(X, N_{de}|H_0)) - \log(P(\text{PF, pf}|H_0)) = ct\]
How does one validate pathway analysis methods?

How do we assess the quality of the results obtained with each pathway analysis method? How do we compare different methods?

Possible methods:

1. Validation against known pathways associated to a specific disease in “some” datasets

2. Use “target” pathways
### Lung adenocarcinoma

- 97 genes identified as good markers of survival

<table>
<thead>
<tr>
<th>Gene</th>
<th>Log2 Fold Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAG1</td>
<td>-1.19047619</td>
</tr>
<tr>
<td>CASP4</td>
<td>-1.063829787</td>
</tr>
<tr>
<td>FADD</td>
<td>1.57</td>
</tr>
<tr>
<td>P63</td>
<td>1.37</td>
</tr>
<tr>
<td>ST4</td>
<td>2.39</td>
</tr>
<tr>
<td>BAG1</td>
<td>-1.19047619</td>
</tr>
<tr>
<td>CASP4</td>
<td>-1.063829787</td>
</tr>
<tr>
<td>FADD</td>
<td>1.57</td>
</tr>
<tr>
<td>P63</td>
<td>1.37</td>
</tr>
<tr>
<td>ST4</td>
<td>2.39</td>
</tr>
</tbody>
</table>

- Additional gene list continues with similar format.
Lung cancer

- Poor vs. good survival in lung adenocarcinoma (Beer, Draghici et al., *A systems biology approach for pathway level analysis.* Genome Research, 17:1537-1545, 2007

Implemented in the R-OntoTools and SPIA Bioconductor packages, iPathway-Guide
Datasets

- 24 datasets
- 12 different conditions

Target pathway

Comparing analysis

- 24 datasets
- 229 KEGG pathways
- 12 different conditions

Is there maybe a gold standard?

- Knock-out (KO) data sets
  - In a KO experiment, the precise cause of the phenotype changes is known
  - The target pathways (true positives) include all pathways that contain the KO gene
  - A good pathway analysis method would be able to retrieve as many TPs as possible and ideally, pinpoint
Data set: KO of gene Myd88

- The knockout gene of GSE19793 dataset is myd88 in mice

- Wild type mice of identical background were used as control groups.

- 10 biological repeats are performed for the wild type and myd88 samples.

- The platform that is used in this experiment is Affymetrix Mouse Genome 430 2.0 Array.

<table>
<thead>
<tr>
<th>Target Pathways</th>
<th>Entry</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. NF-kappa B signaling pathway</td>
<td>4064</td>
</tr>
<tr>
<td>2. Apoptosis</td>
<td>4210</td>
</tr>
<tr>
<td>3. Toll-like receptor signaling pathway</td>
<td>4620</td>
</tr>
<tr>
<td>4. Salmonella infection</td>
<td>5132</td>
</tr>
<tr>
<td>5. Pertussis</td>
<td>5133</td>
</tr>
<tr>
<td>6. Legionellosis</td>
<td>5134</td>
</tr>
<tr>
<td>7. Leishmaniasis</td>
<td>5140</td>
</tr>
<tr>
<td>8. Chagas disease (American trypanosomiasis)</td>
<td>5142</td>
</tr>
<tr>
<td>9. African trypanosomiasis</td>
<td>5143</td>
</tr>
<tr>
<td>10. Malaria</td>
<td>5144</td>
</tr>
<tr>
<td>11. Toxoplasmosis</td>
<td>5145</td>
</tr>
<tr>
<td>12. Tuberculosis</td>
<td>5152</td>
</tr>
<tr>
<td>13. Hepatitis B</td>
<td>5161</td>
</tr>
<tr>
<td>14. Measles</td>
<td>5162</td>
</tr>
<tr>
<td>15. Influenza A</td>
<td>5164</td>
</tr>
<tr>
<td>16. Herpes simplex infection</td>
<td>5168</td>
</tr>
</tbody>
</table>
Enrichment vs. Impact Analysis on the

Hypergeometric p-value (eg. DAVID, Ingenuity Pathway)

Perturbation p-value (iPathway-Guide, R-Onto-Tools, SPIA)
Impact Analysis results on the Myd88 KO
Coherent Change of perturbation propagation in Toxoplasmosis pathway.

Based on total perturbation:

Based on total accumulation:
Impact pathway analysis

  > 700 citations

  > 500 citations
Pathway cross-talk phenomena
Fat remodeling experiment

<table>
<thead>
<tr>
<th>title</th>
<th>p_adj.fdr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parkinson’s disease</td>
<td>2.05E-06</td>
</tr>
<tr>
<td>Alzheimer’s disease</td>
<td>3.57E-06</td>
</tr>
<tr>
<td>Huntington’s disease</td>
<td>3.36E-05</td>
</tr>
<tr>
<td>Leishmaniasis</td>
<td>0.000326241</td>
</tr>
<tr>
<td>Phagosome</td>
<td>0.000667716</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>0.001153498</td>
</tr>
<tr>
<td>Oocyte meiosis</td>
<td>0.001680805</td>
</tr>
<tr>
<td>Cardiac muscle contraction</td>
<td>0.001680805</td>
</tr>
<tr>
<td>Toll-like receptor signaling pathway</td>
<td>0.001875685</td>
</tr>
<tr>
<td>PPAR signaling pathway</td>
<td>0.001875685</td>
</tr>
<tr>
<td>Chemokine signaling pathway</td>
<td>0.015452048</td>
</tr>
<tr>
<td>Lysosome</td>
<td>0.021129303</td>
</tr>
<tr>
<td>B cell receptor signaling pathway</td>
<td>0.025295923</td>
</tr>
<tr>
<td>Systemic lupus erythematosus</td>
<td>0.029207716</td>
</tr>
<tr>
<td>Complement and coagulation cascades</td>
<td>0.0342643</td>
</tr>
<tr>
<td>Cytokine-cytokine receptor interaction</td>
<td>0.034618019</td>
</tr>
<tr>
<td>Chagas disease</td>
<td>0.046607705</td>
</tr>
<tr>
<td>Progesterone-mediated oocyte maturation</td>
<td>0.053020989</td>
</tr>
<tr>
<td>Fc epsilon RI signaling pathway</td>
<td>0.05480725</td>
</tr>
<tr>
<td>Leukocyte transendothelial migration</td>
<td>0.05480725</td>
</tr>
</tbody>
</table>

Before eliminating cross-talk
Case Study 1: Fat Remodeling Experiment in Obese Mice (day 3 vs. day 0)
Independent functional module present in the Parkinson's, Huntington’s, Alzheimer’s, and Cardiac Muscle Contraction pathways.
Eliminate pathway cross-talk

Before eliminating

- Donato et al, Genome Research 23 (11), 1885-1893, 2013

After eliminating

- Mitochondria $8.08E-10$
- Phagosome $9.33E-09$
- Cell Cycle+Oocyte Meiosis $5.76E-08$
- PPAR signaling pathway $0.001031617$
- Complement+Systemic pathway $0.002154387$
- * Toll-like receptor signaling pathway $0.051196718$
- MAPK signaling pathway $0.115839005$
- B cell receptor signaling pathway $0.145781479$
- Lysosome $0.187446378$
- Natural killer cell mediated cytotoxicity $0.187446378$
- * Cell cycle $0.229087797$
- Calcium signaling pathway $0.229087797$
- Cell adhesion molecules (CAMs) $0.2583244$
- NOD-like receptor signaling pathway $0.2583244$
- Vascular smooth muscle contraction $0.424980819$
- Dilated cardiomyopathy $0.424980819$
- * Oocyte meiosis $0.432407943$
- Type I diabetes mellitus $0.432407943$
Individual pathway bias
Null distributions (GSA)
GO term dependencies and overlap

![Diagram showing GO term dependencies and overlap]
GO term dependencies and overlap
GO term dependencies and overlap
The Gene Ontology true path rule
GO term dependencies and overlap

SY Rhee, V Wood, K Dolinski, S Draghici
Recommendations for the

- Use all knowledge available i.e. use pathways, not gene sets if possible (do consider the signals and interactions between genes - they are important!!)

- Use methods that can assess pathway impact based on the topology and calculate significance based on resampling (e.g. impact analysis), not simple enrichment
  - Draghici et al, Genome Research 17 (10), 1537-1545

- Use methods that can identify putative mechanisms based on known pathway topology (you have just seen one)

- Take into consideration and eliminate individual pathway bias
  - Nguyen et al, Proceedings of the IEEE 105 (3), 496-515
Thank you!

Sorin Draghici
sorin@wayne.edu