Algorithmic improvement of public cellular pathway and process definitions

Enrico Glaab
Luxembourg Centre for Systems Biomedicine
Motivation for pathway analyses

- How do the changes in omics data relate to known cellular functions?
- Are there specific cellular pathways / molecular networks which display an over-representation of changes in my data?
Motivation (2): Complex diseases as pathway perturbations

Alterations in different biomolecules of a cellular pathway or network can cause similar disruptions downstream

Example: Colorectal carcinoma

- Mutation deactivating APC has the same overall effect as mutations preventing degradation of β-catenin (Segditsas et al., 2006)

→ Strategy: Analyze alterations at the level of molecular networks and pathways to complement single gene/protein level analyses

Wnt/β-catenin signaling pathway
(⚡ = affected by disease-related mutations)
Motivation (3): The “curse of dimensionality“

When analyzing increasing numbers of genes (features):

- the space spanned by these features grows exponentially (no. of features = no. of dimensions)
  → the available data tends to become sparse

→ discrimination between different sample groups (e.g. patients vs. controls) becomes more difficult

→ **Strategy**: Use pathway activity representations of the data to reduce the number of dimensions
Pathway / gene set resources

- Many public databases on functional gene sets and pathways available
- Both generic, multi-organism pathway collections and specialized collections (e.g. disease pathways such as the PD map)
- Format standardization efforts underway (BioPax, SBGN/SBML)
Representations of pathways / functional gene groups

**GENE SETS**

- Mitochondrion
  - Gene.1
  - Gene.2
  - Gene.3

- P53 signaling
  - Gene.2
  - Gene.4
  - Gene.5

→ Find gene sets whose members are enriched among the differentially expressed genes
→ pure statistical scoring

**NETWORKS**

→ Identify network regions enriched in expression alterations
→ scoring topological + expression criteria

**DIAGRAMS**

→ Score pathways with regulatory consistent expression alterations
→ scoring topology + expression changes + consistency criteria
Inconsistencies between pathway definitions

• Pathways are usually manually curated → subjective decisions on members & boundaries

• A pathway defined for the same cellular process may look entirely different in two separate databases, e.g. “p53 signaling“:

  **BioCarta** (p53 signaling)
  **Invitrogen iPath** (p53 signaling)
  **KEGG** (p53 signaling)
Improving pathway definitions using networks

- **Questions**: Can we make pathway definitions more objective? Can we improve existing pathways according to quantitative criteria (compactness, connectivity, density)?

- **Strategy**: Use genome-scale networks to redefine pathways:
  - protein-protein interactions
  - genetic interactions
  - gene co-expression relations

  → large-scale, higher coverage, less biased
  → can also reveal communication between pathways (“cross-talk“)
PathExpand: Network-based pathway extension

- **Idea:** Extend pathways by adding genes that are "strongly connected" to the pathway-nodes and increase the pathway-"compactness" in a network.

**Pathway extension criteria:** Add a node $v$ to set $P$ if:

- $v$ has a pathway-neighbour and $\text{degree}(v) > 1$; and
- $\frac{\text{#pathway-links}(v,p)}{\text{#outside-links}(v,p)} > T_1$; or
- $\frac{\text{#triangle-links}(v,p)}{\text{#possible_triangles}(v,p)} > T_2$; or
- $\frac{\text{#pathway-links}(v,p)}{\text{#pathway-nodes}(p)} > T_3$; and
- avg. shortest path distance in $\{P,v\}$ smaller than in $P$

**black** = pathway-nodes
**red** **blue** **green** = nodes added based on different criteria
**Known cancer pathway:** “BTG family proteins and cell cycle regulation” (BioCarta)

- Disconnected nodes become connected
- Increased pathway-compactness

Added known cancer gene

original pathway
added nodes
PathExpand: Cross-validation

**Question**: Can randomly deleted genes in the original pathways be recovered by the expansion?

→ 3-step cross-validation procedure:

1. Randomly remove 10% of the pathway members (among proteins with at least one partner in the pathway)

2. Apply the proposed extension procedure as well as 100 random extensions (random sampling among candidates)

3. Estimate p-value-like significance scores:

\[
\sum_{i \in P} \left( \frac{\sum_{i=1}^{100} I(recovery_{\text{random}}_i \geq recovery_{\text{proposed}})}{100} \right) / |P|
\]
PathExpand: Semantic similarity analysis

• **Goal**: Quantify pairwise similarities between protein annotations

  **Method**: Jiang & Conrath's semantic GO term similarity measure

• Compute avg. GO-term similarity between pathway-proteins and added proteins

  → compare to random extension model
Biological applications (1): Alzheimer’s disease

- More than 20 proteins annotated in our PPI network
- 5 proteins added by the extension process (circled)
- 3 known to be associated with the disease
- 2 novel candidates: METTL2B, TMED10* (*putative early-onset AD mutations reported)
Biological applications (2): Interleukin signaling

- Complex system of intracellular signaling cascades
- New putative pathway regulators identified
- New “cross-talk proteins” identified (associated with multiple pathways)

Two functions: pathway-regulation & pathway-communication?
Biological applications (3): Enrichment analysis

**Classical approach:** Test enrichment of experimentally derived gene sets in cellular pathway members (one-sided Fisher exact test)

→ **Idea:** replace original pathways by extended versions

**Example:** Enrichment analysis for pancreatic cancer mutated genes:

<table>
<thead>
<tr>
<th>Cellular Process database</th>
<th>Cellular process</th>
<th>Pathway size</th>
<th>Number of pathway mutated genes</th>
<th>Number of mutated genes among added proteins</th>
<th>Mutated genes among added proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biocarta</td>
<td>Agrin Postsynaptic Differentiation</td>
<td>38</td>
<td>5</td>
<td>2</td>
<td>PGM5, PLEKHG2</td>
</tr>
<tr>
<td>Kegg</td>
<td>Fc epsilon RI signaling pathway</td>
<td>112</td>
<td>10</td>
<td>5</td>
<td>DOCK2, MAPKBPI, DUSP19, ATF2, RASGRP3</td>
</tr>
<tr>
<td>Kegg</td>
<td>ErbB signaling pathway</td>
<td>190</td>
<td>13</td>
<td>7</td>
<td>VPS13A, MAPKBPI, NEK8, LIG3, DUSP19, AFF2, GLTSCR1</td>
</tr>
</tbody>
</table>
Biological applications (4): Pancreatic cancer

- “Cell cycle G1/S check point process” - extension procedure adds 7 proteins
- 6 of the added proteins are involved in cell cycle regulation
- the 7\textsuperscript{th} (TGIF2) is known to be mutated in pancreatic cancer
- points to functional role of added proteins
PathExpand: Conclusion & Summary

- The method integrates two sources of information, extending **canonical pathways** using large-scale **protein interaction data**

- Three **evaluated methods**: cross-validation, GO-term semantic similarity and enrichment analysis

- Extended pathways are more compact and provide insights on **pathway regulators**, the **cross-talk** between pathways and gene set **functional enrichment**
References

4. N. Vlassis, E. Glaab, GenePEN: analysis of network activity alterations in complex diseases via the pairwise elastic net, Statistical Applications in Genetics and Molecular Biology (2015), 14(2), 221