Metabolomics Pathway analysis

Anatoly Sorokin
Metabolomics

• Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind", the study of their small-molecule metabolite profiles. Daviss, Bennett (2005) The Scientist 19 (8): 25–28

• Younger sister?:
Metabolic network

• Pathway is a series of reactions converting set of substrate into set of products
• Pathway definition is subjective and non-standard
• Pathways are overlapping
• Easier to talk about whole network
  – FBA
  – Extreme pathway etc
Network representations

Stoichiometry matrix

Connectivity matrix

\[
\begin{pmatrix}
F6P & FDP & T3P1 & T3P2 & 13PG & ATP & ADP & NADH & NAD & Pi & H2O \\
-1 & 1 & 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \\
1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\
0 & -1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -1 & 0 & 1 & 0 & 0 & 1 & -1 & -1 & 0
\end{pmatrix}
\]

\[
\begin{pmatrix}
F6P & FDP & T3P1 & T3P2 & 13PG \\
0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 1 & 0 \\
0 & 1 & 0 & 1 & 1 \\
0 & 1 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0
\end{pmatrix}
\]
Matrix to the network

\[
\begin{pmatrix}
F6P & 0 & 1 & 0 & 0 & 0 \\
FDP & 1 & 0 & 1 & 1 & 0 \\
T3P1 & 0 & 1 & 0 & 1 & 1 \\
T3P2 & 0 & 1 & 1 & 0 & 0 \\
13PG & 0 & 0 & 1 & 0 & 0 \\
\end{pmatrix}
\]

Metabolite graph

Connectivity (Adjacency) matrix

\[
\begin{pmatrix}
r_1 & 0 & 1 & 1 & 0 & 0 \\
r_2 & 1 & 0 & 0 & 0 & 0 \\
r_3 & 0 & 1 & 0 & 1 & 1 \\
r_4 & 0 & 0 & 1 & 0 & 1 \\
r_5 & 0 & 0 & 1 & 1 & 0 \\
\end{pmatrix}
\]

Reaction graph
Currency metabolites

From glucose to pyruvate, ADP can not be used as a link.

Otherwise path length will be 2 instead of 9
(Jeong et al. 2000 Nature 407:651)
With or without currency metabolites

Metabolic network of *S. pneumonia* (616 reactions)
Network metrics for metabolic network

• A typical genome scale metabolic network contains one thousand reactions/metabolites.
• We need to characterize importance of nodes and edges in the network
Neighbours and degree

**Neighbours**: directly linked nodes

**K-neighbours**: nodes linked with a node in k steps.

**Degree**: the number of links to its neighbours from a node (may not equal to the number of neighbours).

For directed network: *input and output degree*.

For r2, neighbours are 2, 2-neighbours are 4

Degree is 2, input degree is 2 and output degree is 1.
Connection degree distribution

How node degrees distributed in a network.

$P(k) = ak^{-\gamma}$

$P(k)$: Percentage of nodes with a degree $k$ or not less than $k$ (Cumulative distribution).

**Power law degree distribution** indicates a **scale free network**: A few nodes (hubs) have very high degree while most nodes have very low degree.
Many real networks are scale free networks.

Robust on random failure but vulnerable under aimed attack at the highly connected nodes (hubs). Scale free feature is the result of evolution (rich get richer generative model, like web)
Hub metabolites

E. Coli metabolic network

Glycerate-3-phosphate, D-Ribose-5-phosphate, Acetyl-CoA, Pyruvate, D-Xylulose 5-phosphate

Most hubs are in central pathways. However, if currency metabolites are included in the network, Most hubs would be currency metabolites
Node Centrality

Closeness centrality of node $x$:

$$C(x) = \frac{n - 1}{\sum_{y \in U, y \neq x} d(x, y)} = \frac{1}{\bar{d}}$$

- $d(x,y)$: the path length between node $x$ and node $y$
- $U$: the set of all nodes
- $\bar{d}$: average path length between $x$ and the other nodes

The central nodes have short path lengths to other nodes in the network
The most central metabolites in the metabolic network of *E. coli*

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Centrality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyruvate</td>
<td>0.225</td>
</tr>
<tr>
<td>Actyl-CoA</td>
<td>0.210</td>
</tr>
<tr>
<td>Malate</td>
<td>0.204</td>
</tr>
<tr>
<td>2KD6PG</td>
<td>0.203</td>
</tr>
<tr>
<td>Acetate</td>
<td>0.201</td>
</tr>
<tr>
<td>Acetaldehyde</td>
<td>0.199</td>
</tr>
<tr>
<td>G3P</td>
<td>0.198</td>
</tr>
</tbody>
</table>

most central nodes ≠ highly connected nodes
Betweenness Centrality

- the fraction of shortest paths between pairs of nodes that pass through a given node.

The most effective target to break down the network (Robustness of network)
Network Global Connectivity

Degree distribution tells nothing about global connectivity

The right network can have short average path length though not connected at all
Strongly or weakly connected components

A connected component is a maximal connected subgraph. Two nodes are defined to be in the same connected component if there exists a path between them. If link direction is considered it is strongly connected, otherwise weakly connected.
SC distribution in a metabolic network

One big SC and many small SCs

GSC: Giant strong component
Connectivity structure of MN

**Giant strong component (GSC)**
metabolites fully converted and convertible to each other

274 out of total 811 metabolites

**Substrate subset** (93)
converted to metabolite in GSC

**Product subset** (161)
produced from metabolites in GSC

**Isolated subset** (283)
Bow-tie: a general structure of biological and physical networks

- Metabolic network
- Signal transduction
- Web pages network
- Material processing and other tech. systems
Tools for network analysis

• KNEVA http://csb.inf.ed.ac.uk/kneva
• Cytoscape http://www.cytoscape.org/ (for Biological networks, mapping data), many plugins
• Bioconductor and R (SNA)
• Java and Python packages (NetworkX)
Network databases

• KEGG
• Metacyc
• Yeast (http://www.comp-sys-bio.org/yeastnet/)
• Human-specific networks
  – Recon1 (Palsson group, 1496 ORFs, 2004 proteins, 2766 metabolites and 3311 reactions)
  – EHMN (Edinburgh group, 2671 compounds, 2322 genes, 2823 reactions 66 pathways)
Metabolomics

• Chemometric
  – Spectral method based
  – Compounds are not defined
  – Feature extraction
  – Qualitative

• Quantitative
  – MS method based
  – Compounds are defined
  – Quantitative
Pathway analysis in metabolomics

• Quantitative metabolomics data is similar to microarray data
• Can be processed and understood in similar way
• MetaboAnalyst (www.metaboanalyst.ca) online tool for data analysis in metabolomics
Pathway analysis in metabolomics

- We have data in “standard” format similar to transcriptomics and proteomics
- We have networks and pathways
- We can apply standard pathway analysis
  - Pure metabolomics
  - Metabolomics/transcriptomics
- MetPA (http://metpa.metabolomics.ca) online tool for metabolic pathway analysis
## Compound mapping

<table>
<thead>
<tr>
<th>Query</th>
<th>Match</th>
<th>KEGG</th>
<th>HMDB</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1,6-Anhydro-beta-D-glucose</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1-Methyl nicotinamide</td>
<td>1-Methyl nicotinamide</td>
<td>C02918</td>
<td>HMDB00699</td>
<td></td>
</tr>
<tr>
<td>2-Aminobutyrate</td>
<td>L-Alpha-aminobutyric acid</td>
<td>C02356</td>
<td>HMDB00452</td>
<td></td>
</tr>
<tr>
<td><strong>2-Hydroxyisobutyrate</strong></td>
<td>(S)-3-Hydroxyisobutyric acid</td>
<td>C01188</td>
<td>HMDB00023</td>
<td>View</td>
</tr>
<tr>
<td>2-Oxoglutarate</td>
<td>Oxoglutaric acid</td>
<td>C00026</td>
<td>HMDB00208</td>
<td>View</td>
</tr>
<tr>
<td>3-Aminoisobutyrate</td>
<td>3-Aminoisobutanoic acid</td>
<td>C05145</td>
<td>HMDB03911</td>
<td></td>
</tr>
<tr>
<td>3-Hydroxyisobutyrate</td>
<td>3-Hydroxyisobutyric acid</td>
<td>C01089</td>
<td>HMDB00357</td>
<td></td>
</tr>
<tr>
<td><strong>3-Hydroxyisovalerate</strong></td>
<td>3-Hydroxy-3-methyl-2-oxobutanoic acid</td>
<td>C04181</td>
<td></td>
<td>View</td>
</tr>
<tr>
<td><strong>2-Indoxyl sulfate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4-Hydroxyphenylacetate</td>
<td>p-Hydroxyphenylacetic acid</td>
<td>C00642</td>
<td>HMDB00020</td>
<td></td>
</tr>
<tr>
<td>Acetate</td>
<td>Acetic acid</td>
<td>C00033</td>
<td>HMDB00042</td>
<td></td>
</tr>
<tr>
<td>Acetone</td>
<td>Acetone</td>
<td>C00207</td>
<td>HMDB01659</td>
<td></td>
</tr>
<tr>
<td>Adipate</td>
<td>Adipic acid</td>
<td>C06104</td>
<td>HMDB00448</td>
<td></td>
</tr>
<tr>
<td>Alanine</td>
<td>Alanine</td>
<td>C01401</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asparagine</td>
<td>1-Asparagine</td>
<td>C00152</td>
<td>HMDB00168</td>
<td></td>
</tr>
</tbody>
</table>
Pathway impact

• Calculate importance of metabolites, found in the pathway
  – Degree
  – Betweenness

\[
P_I = \frac{\sum_{\text{found}} \text{Imp}_i}{\sum_{\text{all}} \text{Imp}_i}
\]
Pathway Impact

- P is calculated from GSE analysis
- Most significant pathways has low impact
Metabolome/Transcriptome

- Patil and Nielsen 2005
- Convert metabolic network into compound-enzyme

```
<table>
<thead>
<tr>
<th>Glucose</th>
<th>ATP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucose-6-phosphate</td>
<td>ADP</td>
</tr>
<tr>
<td>cytosol</td>
<td></td>
</tr>
</tbody>
</table>
```

```
Glucose  
ATP      
ADP      
Glucose-6-phosphate  
```

```
Hexokinase
```

```
Hexokinase
ATP
ADP
Glucose-6-phosphate
```
Identify reporter metabolite

- Calculate Z-score for each enzyme

\[ Z_{ni} = \theta^{-1}(1 - p_i) \]

- Calculate Z-score for metabolite

\[ Z_{metabolite} = \frac{1}{\sqrt{k}} \sum Z_{ni} \]
Tools and database

• Experiment repository: www.metabolome-express.org
• Metscape2 metscape.ncibi.org
• Vanted wanted.ipk-gatersleben.de/
• MetPA metpa.metabolomics.ca/
• MetaboAnalyst
SBGN

• To analyse
• To discuss
• To share
Can a Biologist Understand This Diagram?

Can a Biologist Understand This Diagram?

Do these arrows have the same meaning?

Can a Biologist Understand This Diagram?

Do these arrows have the same meaning?

Are these different types of entity?

What is this?

Can a Biologist Understand This Diagram?

Do these arrows have the same meaning?

What is this?

Are these different types of entity?

How about these?

What Happens if one Cannot Read the Blueprint
Graph Trinity: Three Languages in One

- Process Description maps
  - Unambiguous
  - Mechanistic
  - Sequential
  - Combinatorial explosion

- Entity Relationship maps
  - Unambiguous
  - Mechanistic
  - Non-Sequential

- Activity Flow maps
  - Ambiguous
  - Conceptual
  - Sequential
Three Orthogonal Projections of Biology
SBGN Process Description Language

- Inspired and based on Kitano’s Process Diagram Notation
- A Process Description (PD) Diagram represents all molecular processes and interactions occurring between various biochemical entities
- It depicts how entities transition forms as a result of biochemical reactions (including non-covalent modifications such as binding)
- Most of the classic metabolic pathways (e.g., glycolysis and TCA cycle) in biochemistry textbooks were drawn in this approach
- Though not the conventional approach for drawing signaling pathways, this approach captures the details of biochemical reactions within the pathway network and provides, in most cases, unambiguous interpretation of pathway mechanisms
Macromolecules: biochemical substances that are built up from the covalent linking of pseudo-identical units. Examples of macromolecules include proteins, nucleic acids (RNA, DNA), and polysaccharides (glycogen, cellulose, starch, etc.).
Macromolecular Pools: State Variables

- Pool is set of molecules somehow undistinguishable
- Molecules can be in different state
  - (Non)phosphorylated
  - Open/close channel
  - Modified at some state
Complex and Multimer

• Represents complexes of molecules held together by non-covalent bonds
• Multimer require cardinality
• Can have state variables
  – In multimer it means that all monomers have same state
  – Use complex if not the same states
Key Concept: Process

- Process: conversion of element of one pool to another
- Special cases:
  - Non-covalent binding
    - Association
    - Dissociation
  - Incompleteness
    - Uncertain process
    - Omitted process

Diagram:
- Association
- Dissociation
- Process
- Uncertain process
- Omitted process
Arcs

- Using pools by process
  - Consumption/production
  - Stoichiometry (optional)
- Regulating process rate
  - Stimulation
  - Inhibition
  - Catalysis
- Requirement for process
  - Necessary stimulation
Compartments

• Container to represent physical or logical structure
  – Free form
  – Visually thicker line
• The same entity pools in different compartments are different
• Compartments are independent
• Overlapping do not mean containment
Neuro-muscular Junction
Activity Flow: Abstraction

• Main concept is **Biological Activity**
  – Each node represents an activity, but not the entity
  – Multiple nodes can be used to represent activities from one entity (e.g., receptor protein kinase)
  – One node can be used to represent activities from a group of entities (e.g., a complex, generics etc.)
Material and Conceptual Types in AF

- Activity node is rectangular to emphasize similarity to reaction
- Unit of information has shape according to node type
- Unit of information can carry name of entity, which has the activity

Activity of ion
Activity of protein
Phenotype
Perturbation

mt:ion
mt:prot
Regulatory Arcs

- Operates on activities
- Shows influences
  - Positive
    - Catalysis
    - Stimulation
  - Negative
    - Inhibition
  - Required
    - Necessary stimulation

Positive influence
Negative influence
Necessary stimulation
Unknown influence
Logical Gates

• Three main logic operations
  – AND: all are required
  – OR: any combination is required
  – NOT: prevent influence

• Crucial for AF
  – No complex
  – No outcome
  – No modifications
Activity Flow Map is Ambiguous

- AF diagrams are ambiguous
- An AF diagram should be associated with either a PD or ER diagram, if possible
- Automatic conversion between PD and/or ER to AF is planned