Visualization of Omics Data for Systems Biology: A Review in the Context of Network Biology

Laura Stevens August 24, 2017
Background and Significance

• High throughput studies continue to yield extensive and various amounts of biological data.
• Visualization is a critical component of both the analysis and the interpretation of these data.
• Users now have a variety of ways to visualize these various types of omics-scale data.
• The challenge is to create clear visualizations which can integrate multiple data types and ideally be integrated across multiple biological systems to provide biological insight without being inundated by the mass amount of information and complexity of the data.
Data Types

• Interaction Data
  • Interactions between Genes and Gene Products
  • Protein-Protein Interaction- (set of Proteins identified in a Sample)
  • Drugs and Proteins

• Profile Data
  • Set of Metabolites identified in a sample

• Expression
  • Gene Expression
  • Proteins Expression
  • Metabolite Concentrations
Data Types and Key Experimental Methods

Interaction Data-

- **Mass Spectrometry-**
  - compounds present in a sample are identified through the accurate measurements of their mass to charge ratios. This type of data is generated in many fields- Proteomics, metabolomics, and interactome mapping
  - typical datasets consists of lists of proteolytic peptides/metabolomic features characterized by their mass to charge ratios which can be used to deduce their sequences
  - These sequences can then be input into search engines to identify a given protein, peptide, or metabolite
  - Goal is to identify groups of molecules associated with a protein complex/metabolic pathway
  - Example protein complexes for gene regulation, metabolites for biomarkers, diseased phenotypes, drug targets

<table>
<thead>
<tr>
<th>Protein</th>
<th>Interacts</th>
<th>Protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAPK3</td>
<td>interacts SCRIB</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SEC61B</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SET</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SF3A3</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SF3B1</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SFTPC</td>
<td></td>
</tr>
<tr>
<td>MAPK3</td>
<td>interacts SGPL1</td>
<td></td>
</tr>
</tbody>
</table>
Data Types and Key Experimental Methods

Profile Data –

• Mass Spectrometry
  - Various Chromatography methods are used prior to MS/NMR due to chemical diversity of metabolites
  - Metabolites are identified based on fragmentation patterns
  - Data is in similar format to proteomics data

• NMR- Metabolites
  - A common method in metabolomics to provide detailed information on the molecular structure of compounds found in complex mixtures.
  - Unlike Mass Spectrometry, does not require analyte separation

• Goal is to gain insight into molecular mechanisms of cellular metabolic pathways and identify reliable biomarkers for disease and drug treatment
Data Types and Key Experimental Methods

• **Expression/Concentrations-**
  • measuring expression levels of large numbers of genes/proteins simultaneously
  • Normalize experimental and batch differences between samples and then identify up and downregulated genes/proteins/metabolites based on fold change level when comparing across samples (ex- healthy and non-healthy tissue).
  • Statistical Approaches are used to assess reliability of fold change measurements
  • Goal is to identify a set of genes/proteins/metabolites that share a related pattern of expression/concentration
    • Genes that are up/down regulated in a certain genotype, disease phenotype, or in response to a drug treatment
Data Types- Size and Variety Can be Daunting

• Can be a variety of different interactions: Metabolites, Proteins, Genes, SNPs
• Rolland et al tested potential interactions between proteins of 13000 genes, and reported 13,944 high quality human binary protein interactions between 4303 proteins with the possible interactions being much larger
• Metabolite often are associated with a variety of reactions with hundreds to thousands of associated compounds
Data Types - Size and Variety Can be Daunting

Data

• Can be measured over a range of time points
• Can be measured over a range of experimental conditions
• Requires extensive search through noisy, multivariate data
Visualizing Protein-Protein Interactions

- Iteratively dissect the data into smaller subsets
  - Clustering and ML techniques to identify higher order complexes
  - Same complex
  - Same subcellular locations
  - Specific functional category
Visualizing Protein-Protein Interactions-Systems

- Collapsing nodes to meta-nodes (a)
- Use of Layering and/or Node Coloring/Regions (b,c)
- Overlaying temporal information (c)
Visualizing Metabolic Pathways

- Highly connected
- Involved in multiple systems
Network Enrichment

- Filtering of Genes/Proteins with a particular expression profile
- Enrich networks to identify pathways/networks where proteins/genes are significantly under/over represented
Mapping of Expression Data On to Networks

• After identifying potentially relevant genes from genetic expression
• Data is mapped onto interaction networks, or onto identified pathways
• Small multiples/Animations/Sequential visualizations can be used for different conditions
• Expression levels as color gradient is most common
• Expression profiles with many conditions often link a separate visualization (heat map or profile plot)
Discussion

• Ideally Visualizations will not only integrate multiple types of data but can also integrate multiple types of networks to get an entire systems overview.

• If this were the case what would the meta-nodes be and what would the meta edges be?