Creating In Silico Interactomes

- Tony Chiang
- Denise Scholtens
- Robert Gentleman
Objectives

- Define interactomes
  - Biological and in silico
- Describe the process of construction
- Relate the data structure
  - How this structure is comprehensive to detailing the data
  - Why this structure is good for some statistical modeling
- Simple examples in using the interactome
- Future Work
Introduction and Background

● Basic Terminology
  – Protein Complex
    ● Group of 2 or more associated proteins
    ● Conduct some biological process
  – Protein Complex Interactome
    ● Coordinated set of protein complexes
    ● Specific to each cell or tissue type
    ● Variable over environmental conditions
Graph Theoretic Representation

- **Hyper-graph**
  - Generalization of ordinary graph
    - Vertex set, \( V \), is the collection of unique proteins
      - Let \( |V| = n \)
    - Hyper-edge, \( E \), is the collection of unique protein complexes
      - Then \( |E| \leq 2^n - (n+1) \)

- **Interactome ↔ Hyper-graph**
  - Most protein complex identification experiments occur in some biological interactome
In Silico Interactome

- Collection of estimated protein complexes representing an in silico model organism
  - The ISI is a simulated organism with which we can conduct computational experiments
- ISI is modeled after biological interactomes
- Storage of the ISI
  - Incidence Matrix Representation of the Hyper-Graph
    - Rows indexed by the vertices (expressed proteins)
    - Columns indexed by the hyper-edges (complexes)
    - Incidence is equivalent to membership
Interactome to Incidence Matrix

<table>
<thead>
<tr>
<th>Complex1</th>
<th>Complex2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein1</td>
<td>1</td>
</tr>
<tr>
<td>Protein2</td>
<td>1</td>
</tr>
<tr>
<td>Protein3</td>
<td>1</td>
</tr>
<tr>
<td>Protein4</td>
<td>0</td>
</tr>
</tbody>
</table>
Why hyper-graph representation

The hyper-graph representation encapsulates more information than a graph representation.

We look at the example of PP2A I, II, III

By example, we show why protein-protein interaction graphs and co-membership graphs cannot incorporate protein membership information
Neither graph can determine Protein Complex Membership
A Hyper-Graph (Forgive me) details protein membership, co-membership, but not interaction data.
Constructing the ISI

- Presently, the simulated model organism is based on Saccharomyces cerevisiae
- Constructing the in silico interactome
  - Collecting protein complex composition data
    - Gene Ontology
    - MIPS
    - High Through-Put Affinity Purification - Mass Spectrometric Experimentation
      - Protein Complex Estimation via apComplex
ISI - Limitations

- Comprehensive
  - It does not contain an exhaustive list of all protein complexes since it reflects known biology

- Definitive
  - It contains mostly estimated protein complexes via both low and high through-put technologies

- Meant to replace experimental de novo research
  - It cannot give insight to unknown biological complexes and interactomes
ISI - Benefits

- **Dynamic**
  - It can be updated and modified as new data is discovered and old data is revised
- **Simplified**
  - Redundancies from different data sources can be eliminated as well as irrelevant protein complexes
- **Versatile**
  - An ISI can be modeled after any organism from yeast to mice to men
Why build in silico interactomes

- Reasons to build valid in silico interactomes:
  - Provides one single data structure with which to conduct in silico experiments
  - Provides tool with which simulated wet-lab experiments can be conducted
  - Use in the generation of multiple data sets
  - Develop tools and strategy for small scale experiments
  - Study of perturbation in networks
  - Effects of varying sampling paradigms on large, non-random networks
Integrating Data and Deriving Statistics

In Silico Interactome

Computational Statistics
In Silico Interactome for Yeast - ScISI

- Computational parsing data from GO and MIPS
  - Term mining
    - [Cc]omplex
    - Suffix “-ase” (e.g. RNA polymerase II)
    - Suffix “-some” (e.g. ribosome)
- Manual parsing resultant protein complexes
- Collecting estimates from apComplex
  - Experiments
    - Gavin et al. (2002, 2006*)
    - Ho et al. (2002)
    - Krogan et al. (2004)
ScISI - a model example

- In silico S. cerevisiae
  - 1661 unique expressed proteins
  - 734 distinct protein complexes

- Basic statistical profile
  - Complex
    - Cardinality range = [2,57]
    - Median cardinality = 4
    - Mean cardinality = 5.98
  - Protein
    - Membership range = [1,31]
    - Median membership = 1
    - Mean membership = 2.64
In Silico experiments on ScISI

- Determining protein complex structures
  - Let $A$ be the incidence matrix of ScISI
    - Then $[A^T]_{ij}$ counts the number of complexes to which protein $i$ and protein $j$ belong, that is how many complexes these two proteins share co-membership
  - Transformation gives a measure of protein affiliation but not direct binary interaction
Graphical representation of in silico experiments

- We make use of the equivalence of hyper-graphs to bi-partite graph
  - Equivalence is determined by letting the set of hyper-edges be the second set of nodes.

- The operation $A A^T$ is a contraction on the protein complex nodes of the bi-partite graph
  - This process takes us from protein complex membership to protein-protein complex co-membership
Bi-partite Graph: Protein Complex Membership

Ordinary Graph: Protein-Protein Complex Co-Membership
Where to from here?

- Let’s re-iterate the 5 reasons to build valid in silico interactomes:
  - Provides tool with which simulated wet-lab experiments can be conducted
  - Use in the generation of multiple data sets
  - Develop tools and strategy for small scale experiments
  - Study of perturbation in networks
  - Effects of varying sampling paradigms on large, non-random networks

- All 5 of which are still open ended…
Future Direction

● An interesting question…
  – Many of the protein complexes are estimates obtained from Affinity Purification - Mass Spectrometry experiments
  – Can we validate these estimates?
    ● Each interactome built needs to be validated before conducting computational experiments
  – We present two different methods to validate the interactomes.
Validating ISI

- Using direct binary interaction data to verify protein complex composition
  - Necessary and sufficient condition is that induced interaction graph be connected on the sub-set of proteins in each protein complex

- Hard to verify
  - Binary interaction data is sparse
  - Error Rates are extremely high
  - There is a need to decipher between true negative interactions between two proteins and un-tested interactions between two proteins
  - Induced interaction graph is almost always disconnected
Validating ISI

- Simulation Models
  - Simulate the AP-MS technology and derive data-sets on which we can apply estimation algorithm.
  - Determine how effective estimation algorithm based on statistical significance
  - Compare with other estimation algorithms