CS4220: Knowledge Discovery Methods for Bioinformatics
Unit 5: Biological Network

Limsoon Wong
Lecture Outline

- Overview of biological networks
- Use of biological networks in enhancing bioinformatics analysis
- Consistency, comprehensiveness, and compatibility of biological pathway databases
- Integrating pathway databases
- Reliability of PPIN
- Identifying noise edges in PPIN
- Identifying missing edges in PPIN
- An advanced example on assessment of PPIN
Overview of Biological Networks
Why Biological Networks?

- Complete genomes are now available
- Knowing the **genes** is not enough to understand how biology **functions**
- **Proteins**, not genes, are responsible for many cellular activities
- Proteins function by interacting w/ other proteins and biomolecules

**GENOME**  
**PROTEOME**  
**"INTERACTOME"**

Slide credit: See-Kiong Ng
Types of Biological Networks

• Natural biological pathways
  – Metabolic pathway
  – Gene regulation network
  – Cell signaling network

• Protein-protein interaction networks
Metabolic Pathway

- A series of biochem reactions in a cell
  - Catalyzed by enzymes
  - Step-by-step modification of an initial molecule to form another product that can
    - be used /store in the cell
    - initiate another metabolic pathway
Gene Regulation Network

- Gene regulation is the process that turns info from genes into gene products
- Gives a cell control over its structure & function
  - Cell differentiation
  - Morphogenesis
  - Adaptability, …
Cell Signaling Network

• It is the entire set of changes induced by receptor activation
  – Governs basic cellular activities and coordinates cell actions

• Cells communicate with each other
  – Direct contact (juxtacrine signaling)
  – Short distances (paracrine signaling)
  – Large distances (endocrine signaling)

• Errors result in cancer, diabetes, ...

Protein Interaction Network (PPIN)

- PPI usual refers to physical binding between proteins
  - Stable interaction
    - Protein complex
    - ~70% of PPIs
  - Transient interaction, modifying a protein for further actions
    - Phosphorylation
    - Transportation
    - ~30% of PPIs

- PPIN is usually a set of PPIs; it is not put into biological context

<table>
<thead>
<tr>
<th>Database</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG</td>
<td>KEGG (<a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a>) is one of the best known pathway databases (Kanehisa et al., 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.</td>
</tr>
<tr>
<td>WikiPathways</td>
<td>WikiPathways (<a href="http://www.wikipathways.org">http://www.wikipathways.org</a>) is a Wikipedia-based collaborative effort among various labs (Kelder et al., 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.</td>
</tr>
<tr>
<td>Reactome</td>
<td>Reactome (<a href="http://www.reactome.org">http://www.reactome.org</a>) is also a collaborative effort like WikiPathways (Vastrik et al., 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.</td>
</tr>
<tr>
<td>Pathway Commons</td>
<td>Pathway Commons (<a href="http://www.pathwaycommons.com">http://www.pathwaycommons.com</a>) collects information from various databases but does not unify the data (Cerami et al., 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.</td>
</tr>
<tr>
<td>PathwayAPI</td>
<td>PathwayAPI (<a href="http://www.pathwayapi.com">http://www.pathwayapi.com</a>) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh et al., 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.</td>
</tr>
</tbody>
</table>

Sources of Protein Interactions

<table>
<thead>
<tr>
<th>Database</th>
<th># nodes, # edges</th>
<th>URL</th>
<th>Build Focus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioGRID</td>
<td>10k, 40k</td>
<td><a href="http://thebiogrid.org">http://thebiogrid.org</a></td>
<td>Literature</td>
<td>(Stark et al., 2006)</td>
</tr>
<tr>
<td>DIP</td>
<td>2.6k, 3.3k</td>
<td><a href="http://dip.doe-mbi.ucla.edu">http://dip.doe-mbi.ucla.edu</a></td>
<td>Literature</td>
<td>(Xenarios et al., 2002)</td>
</tr>
<tr>
<td>HPRD</td>
<td>30k, 40k</td>
<td><a href="http://www.hprd.org">http://www.hprd.org</a></td>
<td>Literature</td>
<td>(Prasad et al., 2009)</td>
</tr>
<tr>
<td>IntAct</td>
<td>56k, 267k</td>
<td><a href="http://www.ebi.ac.uk/intact">http://www.ebi.ac.uk/intact</a></td>
<td>Literature</td>
<td>(Aranda et al., 2010)</td>
</tr>
<tr>
<td>MINT</td>
<td>30k, 90k</td>
<td><a href="http://mint.bio.uniroma2.it/mint">http://mint.bio.uniroma2.it/mint</a></td>
<td>Literature</td>
<td>(Chatr-aryamontri et al., 2007)</td>
</tr>
<tr>
<td>STRING</td>
<td>5200k, ?</td>
<td><a href="http://string-db.org">http://string-db.org</a></td>
<td>Literature, Prediction</td>
<td>(Szklarczyk et al., 2011)</td>
</tr>
</tbody>
</table>


and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010
Use of Biological Networks in Enhancing Bioinformatics Analysis
Recall from Unit 2 of the course…

Gene Expression Profile Analysis

- Each disease phenotype has some underlying cause
- There is some unifying biological theme for genes that are truly associated with a disease subtype
- Uncertainty in selected genes can be reduced by considering biological processes of the genes
- The unifying biological theme is basis for inferring the underlying cause of disease subtype

Contextualization!

Taming false positives by considering pathways instead of all possible groups

Group of Genes

- Suppose
  - Each gene has 50% chance to be high
  - You have 3 disease and 3 normal samples
- What is the chance of a group of 5 genes being perfectly correlated to these samples?

\[
\text{Prob(group of genes correlated) = } (1/2)^5
\]
- Good, \(< 1/2^6
\)
- # of groups = \(1000^5\)
- E(# of groups of genes correlated) = \(1000^5 \times (1/2^6) \approx 9.3 \times 10^{-7}\)

\(\Rightarrow\) Even more false positives?
- Perhaps no need to consider every group
Recall from Unit 2 of the course…

**Proteomic Profile Analysis**

- Suppose the failure to form a protein complex causes a disease
  - If any component protein is missing, the complex can’t form

⇒ Diff patients suffering from the disease can have a diff protein component missing
  - Construct a profile based on complexes?

Epistatic Interaction Mining

- GWAS have linked many SNPs to diseases, but many genetic risk factors still unaccounted for
- Proteins coded by genes interact in cell
  ⇒ Some SNPs affect the phenotype in combination with other SNPs; i.e., epistasis

- Exhaustive search for epistatic effects has to test many combinations (>100,000²) of SNPs
  - Hard to get statistical significance
  - Take long time to run on computers
  ⇒ Use biological networks to narrow the search for two-locus epistasis
Disease Causal Gene Prioritization

- Genes causing the same or similar diseases tend to lie close to one another in PPIN

- Given disease Q. Look for proteins in PPIN interacting with many causal genes of diseases similar to Q

Protein Complex Prediction

• Nature of high-throughput PPI expts
  – Proteins are taken out of their natural context!

• A big “hub” and its “spokes” should probably be decomposed into subclusters
  – Each subcluster is a set proteins that interact in the same space & time; viz., a protein complex

• Can a protein interact with so many proteins simultaneously?

• Many complexes have highly connected cores in PPIN ➔ Find complexes by clustering

• Issue: How to identify low edge density complexes?
Protein Function Prediction

- Proteins with similar function are topologically close in PPIN
  - Direct functional association
  - Indirect functional association

A pair of proteins that participate in the same cellular processes or localize to the same cellular compartment are many times more likely to interact than a random pair of proteins

- Proteins with similar function have interaction neighborhoods that are similar

When proteins in the neighborhood of a protein X have similar functions to proteins in the neighborhood of a protein Y, then proteins X & Y likely operate in similar environment

Consistency, Comprehensiveness, and Compatibility of Biological Pathway Databases
<table>
<thead>
<tr>
<th>Database</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEGG</td>
<td>KEGG (<a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a>) is one of the best known pathway databases (Kanehisa <em>et al.</em>, 2010). It consists of 16 main databases, comprising different levels of biological information such as systems, genomic, etc. The data files are downloadable in XML format. At time of writing it has 392 pathways.</td>
</tr>
<tr>
<td>WikiPathways</td>
<td>WikiPathways (<a href="http://www.wikipathways.org">http://www.wikipathways.org</a>) is a Wikipedia-based collaborative effort among various labs (Kelder <em>et al.</em>, 2009). It has 1,627 pathways of which 369 are human. The content is downloadable in GPML format.</td>
</tr>
<tr>
<td>Reactome</td>
<td>Reactome (<a href="http://www.reactome.org">http://www.reactome.org</a>) is also a collaborative effort like WikiPathways (Vastrik <em>et al.</em>, 2007). It is one of the largest datasets, with over 4,166 human reactions organized into 1,131 pathways by December 2010. Reactome can be downloaded in BioPax and SBML among other formats.</td>
</tr>
<tr>
<td>Pathway Commons</td>
<td>Pathway Commons (<a href="http://www.pathwaycommons.com">http://www.pathwaycommons.com</a>) collects information from various databases but does not unify the data (Cerami <em>et al.</em>, 2006). It contains 1,573 pathways across 564 organisms. The data is returned in BioPax format.</td>
</tr>
<tr>
<td>PathwayAPI</td>
<td>PathwayAPI (<a href="http://www.pathwayapi.com">http://www.pathwayapi.com</a>) contains over 450 unified human pathways obtained from a merge of KEGG, WikiPathways and Ingenuity® Knowledge Base (Soh <em>et al.</em>, 2010). Data is downloadable as a SQL dump or as a csv file, and is also interfaceable in JSON format.</td>
</tr>
</tbody>
</table>
Low Comprehensiveness of Human Pathway Sources

Human pathways in Wikipathways, KEGG, & Ingenuity

Low Consistency of Human Pathway Sources

Gene Pair Overlap

Wiki vs KEGG

Wiki vs Ingenuity

KEGG vs Ingenuity

Gene Overlap

Wiki vs KEGG

Wiki vs Ingenuity

KEGG vs Ingenuity

Example: Human Apoptosis Pathway

<table>
<thead>
<tr>
<th>Apoptosis Pathway</th>
<th>Wiki x KEGG</th>
<th>Wiki x Ingenuity</th>
<th>KEGG x Ingenuity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Pair Count:</td>
<td>144 vs 172</td>
<td>144 vs 3557</td>
<td>172 vs 3557</td>
</tr>
<tr>
<td>Gene Count:</td>
<td>85 vs 80</td>
<td>85 vs 176</td>
<td>80 vs 176</td>
</tr>
<tr>
<td>Gene Overlap:</td>
<td>38</td>
<td>28</td>
<td>30</td>
</tr>
<tr>
<td>Gene % Overlap:</td>
<td>48%</td>
<td>33%</td>
<td>38%</td>
</tr>
<tr>
<td>Gene Pair Overlap:</td>
<td>23</td>
<td>14</td>
<td>24</td>
</tr>
<tr>
<td>Gene Pair % Overlap:</td>
<td>16%</td>
<td>10%</td>
<td>14%</td>
</tr>
</tbody>
</table>

The same low inter-database consistency (in gene overlap) is observed in pathways of other organisms

<table>
<thead>
<tr>
<th></th>
<th>M. musculus</th>
<th>S. cerevisiae</th>
<th>M. tuberculosis H37Rv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KEGG vs WikiPathways</td>
<td>WikiPathways vs MouseCyc</td>
<td>MouseCyc vs KEGG</td>
</tr>
<tr>
<td>Overlap Genes</td>
<td>2,611</td>
<td>532</td>
<td>919</td>
</tr>
<tr>
<td>Unique Genes</td>
<td>5,168</td>
<td>4,214</td>
<td>5,662</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.336</td>
<td>0.112</td>
<td>0.140</td>
</tr>
<tr>
<td></td>
<td>KEGG vs WikiPathways</td>
<td>WikiPathways vs YeastCyc</td>
<td>YeastCyc vs KEGG</td>
</tr>
<tr>
<td>Overlap Genes</td>
<td>801</td>
<td>402</td>
<td>480</td>
</tr>
<tr>
<td>Unique Genes</td>
<td>996</td>
<td>601</td>
<td>1,317</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.446</td>
<td>0.400</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>KEGG vs WikiPathways</td>
<td>WikiPathways vs MTBRvCyc</td>
<td>MTBRvCyc vs KEGG</td>
</tr>
<tr>
<td>Overlap Genes</td>
<td>141</td>
<td>60</td>
<td>432</td>
</tr>
<tr>
<td>Unique Genes</td>
<td>948</td>
<td>525</td>
<td>707</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.129</td>
<td>0.103</td>
<td>0.379</td>
</tr>
</tbody>
</table>

The same low inter-database consistency (in gene pair overlap) is observed in pathways of other organisms.

<table>
<thead>
<tr>
<th></th>
<th>M. musculus</th>
<th>S. cerevisiae</th>
<th>M. tuberculosis H37Rv</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KEGG vs WikiPathways</td>
<td>WikiPathways vs YeastCyc</td>
<td>MouseCyc vs KEGG</td>
</tr>
<tr>
<td>Overlap Gene Pairs</td>
<td>875</td>
<td>1,242</td>
<td>2,068</td>
</tr>
<tr>
<td>Unique Gene Pairs</td>
<td>55,489</td>
<td>33,312</td>
<td>38,891</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.016</td>
<td>0.036</td>
<td>0.050</td>
</tr>
<tr>
<td></td>
<td>KEGG vs WikiPathways</td>
<td>WikiPathways vs MTBRvCyc</td>
<td>MTBRvCyc vs KEGG</td>
</tr>
<tr>
<td>Overlap Gene Pairs</td>
<td>35</td>
<td>9</td>
<td>419</td>
</tr>
<tr>
<td>Unique Gene Pairs</td>
<td>2,909</td>
<td>1,479</td>
<td>3,524</td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.012</td>
<td>0.006</td>
<td>0.106</td>
</tr>
</tbody>
</table>

Example: TCA Cycle Pathway

<table>
<thead>
<tr>
<th>M. musculus</th>
<th>TCA cycle pathway</th>
<th>KEGG vs WikiPathways</th>
<th>KEGG vs MouseCyc</th>
<th>MouseCyc vs WikiPathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>31 vs 30</td>
<td>31 vs 13</td>
<td>13 vs 30</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>24</td>
<td>13</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.65</td>
<td>0.42</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>Gene Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>100 vs 30</td>
<td>100 vs 24</td>
<td>24 vs 30</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>10</td>
<td>9</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.083</td>
<td>0.078</td>
<td>0.149</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>H. sapiens</th>
<th>Fatty Acid Biosynthesis</th>
<th>KEGG vs WikiPathways</th>
<th>KEGG vs HumanCyc</th>
<th>HumanCyc vs WikiPathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>6 vs 22</td>
<td>6 vs 2</td>
<td>2 vs 22</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.12</td>
<td>0.33</td>
<td>0.04</td>
<td></td>
</tr>
<tr>
<td>Gene Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>12 vs 29</td>
<td>12 vs 2</td>
<td>2 vs 29</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.025</td>
<td>0.077</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>M. tuberculosis H37Rv</th>
<th>TCA cycle pathway</th>
<th>KEGG vs WikiPathways</th>
<th>KEGG vs MTBRvCyc</th>
<th>MTBRvCyc vs WikiPathways</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>35 vs 34</td>
<td>35 vs 10</td>
<td>10 vs 34</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>34</td>
<td>10</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.97</td>
<td>0.29</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Gene Pair</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>107 vs 37</td>
<td>107 vs 19</td>
<td>19 vs 37</td>
<td></td>
</tr>
<tr>
<td>Overlap</td>
<td>3</td>
<td>9</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>Jaccard Coefficient</td>
<td>0.021</td>
<td>0.077</td>
<td>0.098</td>
<td></td>
</tr>
</tbody>
</table>

Pathway sources are curated. They are incomplete; but they have few errors. Makes sense to combine them. But…

Incompatibility Issues

- Data extraction method variations
- Format variations
- Data differences
- Gene/GeneID name differences
- Pathway name differences

Data Format Variations

- KEGG
  - API Call
  - SOAP Data Format

- Wikipathway
  - Parse GPML
  - GPML Data Format

- Ingenuity
  - Manual Extraction
  - Graphical Format

Image credit: Donny Soh’s PhD dissertation, 2009
Integrating Pathway Databases
Things to deal with

- Any integration of incompatible pathway databases must deal with
  - Data extraction method variations
  - Format variations
  - Data differences
  - Gene name / gene id differences
  - Pathway name differences

- We discuss only pathway name differences

- For other issues, consult
The same pathways in the different sources are often given different names.

So how do we even know two pathways are the same and should be compared/merged?
Example of Pathway Name Differences

<table>
<thead>
<tr>
<th>IntPath</th>
<th>KEGG</th>
<th>WikiPathways</th>
<th>MouseCyc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatty Acid</td>
<td>Fatty acid</td>
<td>Fatty Acid</td>
<td>1. fatty acid biosynthesis initiation II</td>
</tr>
<tr>
<td>Biosynthesis</td>
<td>biosynthesis</td>
<td>Biosynthesis</td>
<td>2. very long chain fatty acid biosynthesis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. fatty acid biosynthesis initiation III</td>
</tr>
<tr>
<td>Cholesterol</td>
<td>Cholesterol</td>
<td></td>
<td>1. cholesterol biosynthesis III (via desmosterol)</td>
</tr>
<tr>
<td>Biosynthesis</td>
<td></td>
<td>Biosynthesis</td>
<td>2. cholesterol biosynthesis II (via 24,25-dihydrolanosterol)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. cholesterol biosynthesis I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. superpathway of cholesterol biosynthesis</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>Citrate cycle</td>
<td>TCA cycle</td>
<td>TCA Cycle</td>
</tr>
<tr>
<td>Glycolysis and</td>
<td>Glycolysis/</td>
<td>Glycolysis and</td>
<td>1. glycolysis I 2. glycolysis II</td>
</tr>
<tr>
<td>Gluconeogenesis</td>
<td>Gluconeogenesis</td>
<td>Gluconeogenesis</td>
<td></td>
</tr>
</tbody>
</table>

The table shows several examples of the same pathways with inconsistent referrals to pathway names in different databases.

Possible Ways to Match Pathways

• **Match based on name (LCS)**
  – Pathways w/ similar name should be the same pathway
  – But annotations are very noisy
  ⇒ Likely to mismatch pathways?
  ⇒ Likely to match too many pathways?

• **Are the followings good alternative approaches?**
  – Match based on overlap of genes
  – Match based on overlap of gene pairs
LCS vs Gene-Agreement Matching

- **Accuracy**
  - 94% of LCS matches are in top 3 gene agreement matches
  - 6% of LCS matches not in top 3 of gene agreement matches; but their gene-pair agreement levels are higher

- **Completeness**
  - Let Pi be pathway in db A that LCS cannot find match in db B
  - Let Qi be pathway in db B with highest gene agreement to Pi
  - Gene-pair agreement of Pi-Qi is much lower than pathway pairs matched by LCS

LCS is better than gene-agreement based matching!
LCS vs Gene-Agreement Matching

Gene-pair overlap percentage

- LCS consistently has higher gene-pair agreement
  ⇒ LCS is better than gene-agreement based matching!

LCS vs Gene-Pair Agreement Matching

The 8 pathway pairs singled out by LCS

The 24 pathway pairs singled out by maximal gene-pair overlap

Note: We consider only pathway pairs that have at least 20 reaction overlap.

The 16 pathway pairs depicted in LCS overlap.

LCS vs Gene-Pair Agreement Matching

- Gene-pair agreement match will miss when
  - Pathway P in db A has few overlap with pathway P in db B due to incompleteness of db, even if pathway name matches perfectly!
  - Example: wnt signaling pathway, VEGF signaling pathway, MAPK signaling pathway, etc. in KEGG don’t have largest gene-pair overlap w/ corresponding pathways in Wikipathways & Ingenuity

⇒ Bad for getting a more complete unified pathway P
LCS vs Gene-Pair Agreement Matching

• Pathways having large gene-pair overlap are not necessarily the same pathways

• Examples
  – “Synaptic Long Term Potentiation” in Ingenuity vs “calcium signalling” in KEGG
  – “PPAR-alpha/RXR-alpha Signaling” in Ingenuity vs “TGF-beta signaling pathway” in KEGG

⇒ Difficult to set correct gene-pair overlap threshold to balance against false positive matches
Further Improvement to LCS

• Please read the reference below (esp. page 10) for some of the improvements made to LCS
An Interesting Question

If two pathways are merged, how do you choose the name of the resulting merged pathway?

- Pick the longer of the two original names?
- Pick the shorter?
- Pick randomly?


<table>
<thead>
<tr>
<th>IntPath</th>
<th>KEGG</th>
<th>WikiPathways</th>
<th>MouseCyc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fatty Acid</td>
<td>Fatty acid</td>
<td>Fatty Acid</td>
<td>1. fatty acid biosynthesis initiation II</td>
</tr>
<tr>
<td>Biosynthesis</td>
<td>biosynthesis</td>
<td>Biosynthesis</td>
<td>2. very long chain fatty acid biosynthesis</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. fatty acid biosynthesis initiation III</td>
</tr>
<tr>
<td>Cholesterol</td>
<td></td>
<td>Cholesterol</td>
<td>1. cholesterol biosynthesis III (via desmosterol)</td>
</tr>
<tr>
<td>Biosynthesis</td>
<td></td>
<td>Biosynthesis</td>
<td>2. cholesterol biosynthesis II (via 24,25-dihydrolanosterol)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3. cholesterol biosynthesis I</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4. superpathway of cholesterol biosynthesis</td>
</tr>
<tr>
<td>TCA cycle</td>
<td>Citrate cycle (TCA cycle)</td>
<td>TCA cycle</td>
<td>TCA Cycle</td>
</tr>
<tr>
<td>Glycolysis and</td>
<td>Glycolysis/</td>
<td>Glycolysis and</td>
<td>1. glycolysis I 2. glycolysis II</td>
</tr>
<tr>
<td>Gluconeogenesis</td>
<td>Gluconeogenesis</td>
<td>Gluconeogenesis</td>
<td></td>
</tr>
</tbody>
</table>

The table shows several examples of the same pathways with inconsistent referrals to pathway names in different databases.
The Answer

• The general pathway name is chosen as the shortest pathway names from among the identified related pathways
  ⇒ This usually works well as the name of the integrated pathway
• But in some cases, the shortest name contains “suffix” or “prefix”---like “I”, “II”---that causes the integrated pathway name to give the wrong idea of describing only a specific aspect of the integrated pathway
  ⇒ Remove such suffixes and prefixes when generating integrated pathway names
• In a small number of cases, several similar pathways are included in one pathway name. In these cases, the shortest name is not appropriate as the name of the integrated pathway
  ⇒ Replace the keyword of the integrated pathway name to cover more pathway information

What have we learned?

• Significant lack of concordance betw db’s
  – Level of consistency for genes is 0% to 88%
  – Level of consistency for genes pairs is 0%-61%
  – Most db contains less than half of the pathways in other db’s

• Matching pathways by name is better than matching by gene overlap or gene-pair overlap
Reliability of PPIN
Sources of Protein Interactions

<table>
<thead>
<tr>
<th>Database</th>
<th># nodes, # edges</th>
<th>URL</th>
<th>Build Focus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioGRID</td>
<td>10k, 40k</td>
<td><a href="http://thebiogrid.org">http://thebiogrid.org</a></td>
<td>Literature</td>
<td>(Stark et al., 2006)</td>
</tr>
<tr>
<td>DIP</td>
<td>2.6k, 3.3k</td>
<td><a href="http://dip.doe-mbi.ucla.edu">http://dip.doe-mbi.ucla.edu</a></td>
<td>Literature</td>
<td>(Xenarios et al., 2002)</td>
</tr>
<tr>
<td>HPRD</td>
<td>30k, 40k</td>
<td><a href="http://www.hprd.org">http://www.hprd.org</a></td>
<td>Literature</td>
<td>(Prasad et al., 2009)</td>
</tr>
<tr>
<td>IntAct</td>
<td>56k, 267k</td>
<td><a href="http://www.ebi.ac.uk/intact">http://www.ebi.ac.uk/intact</a></td>
<td>Literature</td>
<td>(Aranda et al., 2010)</td>
</tr>
<tr>
<td>MINT</td>
<td>30k, 90k</td>
<td><a href="http://mint.bio.uniroma2.it/mint">http://mint.bio.uniroma2.it/mint</a></td>
<td>Literature</td>
<td>(Chatr-aryamontri et al., 2007)</td>
</tr>
<tr>
<td>STRING</td>
<td>5200k, ?</td>
<td><a href="http://string-db.org">http://string-db.org</a></td>
<td>Literature, Prediction</td>
<td>(Szklarczyk et al., 2011)</td>
</tr>
</tbody>
</table>


and Protein Complexes

- **CORUM**
  - http://mips.helmholtz-muenchen.de/genre/proj/corum
  - Ruepp et al, *NAR*, 2010
PPI Detection Assays

• Many high-throughput assays for PPIs
  – Y2H
  – TAP
  – Synthetic lethality

Generating *large amounts* of expt data on PPIs can be done with ease

• But ...

High-throughput approaches sacrifice quality for *quantity*: (a) limited or biased coverage: *false negatives*, & (b) high error rates: *false positives*
Noise in PPI Networks

<table>
<thead>
<tr>
<th>Experimental method category</th>
<th>Number of interacting pairs</th>
<th>Co-localization (%)</th>
<th>Co-cellular-role (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>All: All methods</td>
<td>9347</td>
<td>64</td>
<td>49</td>
</tr>
<tr>
<td>A: Small scale Y2H</td>
<td>1861</td>
<td>73</td>
<td>62</td>
</tr>
<tr>
<td>A0: GY2H Uetz et al. (published results)</td>
<td>956</td>
<td>66</td>
<td>45</td>
</tr>
<tr>
<td>A1: GY2H Uetz et al. (unpublished results)</td>
<td>516</td>
<td>53</td>
<td>33</td>
</tr>
<tr>
<td>A2: GY2H Ito et al. (core)</td>
<td>798</td>
<td>64</td>
<td>40</td>
</tr>
<tr>
<td>A3: GY2H Ito et al. (all)</td>
<td>3655</td>
<td>41</td>
<td>15</td>
</tr>
<tr>
<td>B: Physical methods</td>
<td>71</td>
<td>98</td>
<td>95</td>
</tr>
<tr>
<td>C: Genetic methods</td>
<td>1052</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>D1: Biochemical, \textit{in vitro}</td>
<td>614</td>
<td>87</td>
<td>79</td>
</tr>
<tr>
<td>D2: Biochemical, chromatography</td>
<td>648</td>
<td>93</td>
<td>88</td>
</tr>
<tr>
<td>E1: Immunological, direct</td>
<td>1025</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>E2: Immunological, indirect</td>
<td>34</td>
<td>100</td>
<td>93</td>
</tr>
<tr>
<td>2M: Two different methods</td>
<td>2360</td>
<td>87</td>
<td>85</td>
</tr>
<tr>
<td>3M: Three different methods</td>
<td>1212</td>
<td>92</td>
<td>94</td>
</tr>
<tr>
<td>4M: Four different methods</td>
<td>570</td>
<td>95</td>
<td>93</td>
</tr>
</tbody>
</table>

Sprinzak et al., \textit{JMB}, 327:919-923, 2003

Large disagreement betw methods

- High level of noise
  \[\Rightarrow\text{Need to clean up before making inference on PPI networks}\]
Identifying Noise Edges in PPIN
Dealing with noise in PPIN using Reproducibility

- If a PPI is reported in a few independent expts, it is more reliable than those reported in only one expt

\[ r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i) \]

- \( r_i \) is reliability of expt source \( i \),
- \( E_{u,v} \) is the set of expt sources in which interaction between \( u \) and \( v \) is observed

Good idea. But you need to do more expts

\( \Rightarrow \) More time & more $ has to be spent
Dealing with noise in PPIN using Functional Homogeneity

Good idea. But the two proteins in the PPI you are looking at may not have functional annotation.

- If two proteins in a PPI participate in the same function or pathway, it is more reliable than those whose proteins not sharing function & pathway.

Exercise
- What fraction of yeast PPIs in BioGrid share function?
- What fraction of yeast protein pairs share function?
Dealing with noise in PPIN using Localization Coherence

- Two proteins should be in the same place to interact. Agree?

Good idea. But the two proteins in the PPI you are looking at may not have localization annotation.

Exercise
- What fraction of yeast PPIs in BioGrid are in the same cellular compartment?
- What fraction of yeast protein pairs are in the same cellular compartment?
Dealing with noise in PPIN using local topology around a PPI edge

- Two proteins participating in same biological process are more likely to interact
- Two proteins in the same cellular compartments are more likely to interact

CD-distance & FS-Weight: Based on concept that two proteins with many interaction partners in common are likely to be in same biological process & localize to the same compartment

Topological structure of the neighborhood of real PPIs

- Suppose 20% of putative PPIs are noise
  $\implies \geq 3$ purple proteins are real partners of both A and B
  $\implies A$ and B are likely localized to the same cellular compartment (Why?)

- Fact: Proteins in the same cellular compartment are 10x more likely to interact than other proteins
  $\implies A$ and B are likely to interact
Czekanowski-Dice Distance

- Given a pair of proteins \((u, v)\) in a PPI network
  - \(N_u\) = the set of neighbors of \(u\)
  - \(N_v\) = the set of neighbors of \(v\)

\[
CD(u, v) = \frac{2 |N_u \cap N_v|}{|N_u| + |N_v|}
\]

- Consider relative intersection size of the two neighbor sets, not absolute intersection size
  - Case 1: \(|N_u| = 1, |N_v| = 1, |N_u \cap N_v| = 1\), \(CD(u, v) = 1\)
  - Case 2: \(|N_u| = 10, |N_v| = 10, |N_u \cap N_v| = 10\), \(CD(u, v) = 1\)
Adjusted CD-Distance

• Variant of CD-distance that penalizes proteins with few neighbors

\[ wL(u,v) = \frac{2 \left| N_u \cap N_v \right|}{\left| N_u \right| + \lambda_u + \left| N_v \right| + \lambda_v} \]

\[ \lambda_u = \max\{0, \frac{\sum_{x \in G} \left| N_x \right|}{|V|} - \left| N_u \right| \}, \quad \lambda_v = \max\{0, \frac{\sum_{x \in G} \left| N_x \right|}{|V|} - \left| N_v \right| \} \]

• Suppose average degree is 4, then
  – Case 1: \(|N_u| = 1, |N_v| = 1, |N_u \cap N_v| = 1\), \(wL(u,v) = 0.25\)
  – Case 2: \(|N_u| = 10, |N_v| = 10, |N_u \cap N_v| = 10\), \(wL(u,v) = 1\)
A thought…

\[ wL(u, v) = \frac{2 | N_u \cap N_v |}{| N_u | + \lambda_u + | N_v | + \lambda_v} \]

- Weight of interaction reflects its reliability

⇒ Can we get better results if we use this weight to re-calculate the score of other interactions?
Iterated CD-Distance

- \( wL^0(u,v) = 1 \) if \((u,v)\in G\), otherwise \( wL^0(u,v)=0 \)

- \( wL^1(u,v) = \frac{|N_u \cap N_v| + |N_u \cap N_v|}{|N_u| + \lambda_u + |N_v| + \lambda_v} \)

- \( wL^k(u,v) = \frac{\sum_{x \in N_u \cap N_v} wL^{k-1}(u,x) + \sum_{x \in N_u \cap N_v} wL^{k-1}(v,x)}{\sum_{x \in N_u} wL^{k-1}(u,x) + \lambda_u + \sum_{x \in N_v} wL^{k-1}(v,x) + \lambda_v} \)

- \( \lambda^k_u = \max\{0, \frac{\sum_{x \in V} \sum_{y \in N_x} wL^{k-1}(x,y)}{|V|} - \sum_{x \in N_u} wL^{k-1}(u,x) \} \)

- \( \lambda^k_v = \max\{0, \frac{\sum_{x \in V} \sum_{y \in N_x} wL^{k-1}(x,y)}{|V|} - \sum_{x \in N_v} wL^{k-1}(v,x) \} \)
Validation

• DIP yeast dataset
  – Functional homogeneity is 32.6% for PPIs where both proteins have functional annotations and 3.4% over all possible PPIs
  – Localization coherence is 54.7% for PPIs where both proteins have localization annotations and 4.9% over all possible PPIs

• Let’s see how much better iterated CD-distance is over the baseline above, as well as over the original CD-distance/FS-weight
How many iteration is enough?

Cf. ave functional homogeneity of protein pairs in DIP < 4%
ave functional homogeneity of PPI in DIP < 33%

• Iterated CD-distance achieves best performance wrt functional homogeneity at k=2
• Ditto wrt localization coherence (not shown)
How many iteration is enough?

<table>
<thead>
<tr>
<th>noise level</th>
<th>k</th>
<th>#common PPIs</th>
<th>avg_rank_diff</th>
<th>avg_score_diff</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%</td>
<td>1</td>
<td>5669</td>
<td>540.21</td>
<td>0.10</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5870</td>
<td>144.86</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5849</td>
<td>67.00</td>
<td>0.01</td>
</tr>
<tr>
<td>300%</td>
<td>1</td>
<td>5322</td>
<td>881.77</td>
<td>0.18</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5664</td>
<td>367.45</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5007</td>
<td>249.85</td>
<td>0.02</td>
</tr>
<tr>
<td>500%</td>
<td>1</td>
<td>5081</td>
<td>1013.14</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>5502</td>
<td>625.46</td>
<td>0.12</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>5008</td>
<td>317.33</td>
<td>0.05</td>
</tr>
<tr>
<td>1000%</td>
<td>k=1</td>
<td>4472</td>
<td>1187.10</td>
<td>0.28</td>
</tr>
<tr>
<td></td>
<td>k=2</td>
<td>5101</td>
<td>1021.69</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>k=20</td>
<td>5264</td>
<td>614.66</td>
<td>0.13</td>
</tr>
</tbody>
</table>

- Iterative CD-distance at diff k values on noisy network
  ⇒ # of iterations depends on amt of noise
Identifying False Positive PPIs

- Iterated CD-distance is an improvement over previous measures for assessing PPI reliability

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

Liu et al. GIW2008, pp. 138-149
Identifying False Negative PPIs

Iterated CD-distance is an improvement over previous measures for predicting new PPIs

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

Liu et al. GIW2008, pp. 138-149
5-Fold Cross-Validation

• **DIP core dataset**
  – Ave # of proteins in 5 groups: 986
  – Ave # of interactions in 5 training datasets: 16723
  – Ave # of interactions in 5 testing datasets: 486591
  – Ave # of correct answer interactions: 307

• **Measures:**
  – sensitivity = TP/(TP + FN)
  – specificity = TN/(TN + FP)
  • #negatives >> #positives, specificity is always high
  • >97.8% for all scoring methods
  – precision = TP/(TP + FP)
5-Fold X-Validation

• Iterated CD-distance is an improvement over previous measures for identifying false positive & false negative PPIs

Liu et al. *GIW2008*, pp. 138-149
Combining multiple types of info to predict whether a PPI edge is real

- Sometimes you do have additional independent info available
  - Several PPI expts
  - Functional annotations
  - Localization information

- You can combine these pieces of info in the following standard way:

\[ r_{u,v} = 1 - \prod_{i \in E_{u,v}} (1 - r_i) \]

- \( r_i \) is reliability of expt source \( i \),
- \( E_{u,v} \) is the set of expt sources in which interaction betw u and v is observed
Another way to combine more types of info to predict if a PPI is real

- Overlay literature co-occurrence, gene co-expression, etc. on PPIN
- Machine learning to learn characteristic of real PPI

Identifying Missing Edges in PPIN
### PPI Prediction Methods

<table>
<thead>
<tr>
<th>Method Name</th>
<th>Protein/Domain Interaction</th>
<th>Physical Interaction/Functional Association</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene co-expression</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Synthetic lethality</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Gene cluster and gene neighbor</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Phylogenetic profile</td>
<td>P, D</td>
<td>F</td>
</tr>
<tr>
<td>Rosetta Stone</td>
<td>P</td>
<td>F</td>
</tr>
<tr>
<td>Sequence co-evolution</td>
<td>P, D</td>
<td>F</td>
</tr>
<tr>
<td>Classification</td>
<td>P, D</td>
<td>P</td>
</tr>
<tr>
<td>Integrative</td>
<td>P, D</td>
<td>P</td>
</tr>
<tr>
<td>Domain association</td>
<td>D</td>
<td>P</td>
</tr>
<tr>
<td>Bayesian networks</td>
<td>P, D</td>
<td>F, P</td>
</tr>
<tr>
<td>Domain pair exclusion</td>
<td>D</td>
<td>P</td>
</tr>
<tr>
<td>p-Value</td>
<td>D</td>
<td>P</td>
</tr>
</tbody>
</table>

Second column shows if method is designed to predict protein (P) or domain (D) interactions (note that predicted domains can also be used for verifying protein interactions). Third column shows if the method can be used to infer direct physical interaction (P) or indirect functional association (F).

You can also use our earlier topology scores, e.g., CD-distance to predict novel PPIs.
PPI Prediction by Gene Clusters

• Gene clusters or operons encoding co-regulated genes are usually conserved, despite shuffling effects of evolution

⇒ Find conserved gene clusters

• Predict the genes to interact & form operons

PPI Prediction by Phylogenetic Profiling

- Components of complexes and pathways should be present simultaneously in order to perform their functions

- Functionally linked and interacting proteins co-evolve and have orthologs in the same subset of fully sequenced organisms

---

PPI Prediction by Rosetta Stone

- Some interacting proteins have homologs in other genomes that are fused into one protein chain, a so-called **Rosetta Stone protein**

- Gene fusion occurs to optimize co-expression of genes encoding for interacting proteins

---

PPI Prediction by Seq Co-Evolution

- Interacting proteins co-evolve
  - Changes in one protein leading to loss of function are compensated by correlated changes in another protein

- Co-evolution is quantified by correlation of distance matrices used to construct the trees

See [Juan et al, *PNAS*, 105(3):934-939, 2008] for an impt further development to this idea.
PPI Prediction by Iterated CD-Distance

Cf. ave localization coherence of protein pairs in DIP < 5%
ave localization coherence of PPI in DIP < 55%

$$wL^k(u,v) = \frac{\sum_{x \in Nu \cap Nv} w_{L}^{k-1}(u,x) + \sum_{x \in Nu \cap Nv} w_{L}^{k-1}(v,x)}{\sum_{x \in Nu} w_{L}^{k-1}(u,x) + \lambda^u_u + \sum_{x \in Nv} w_{L}^{k-1}(v,x) + \lambda^v_v}$$

- Predict \( (u,v) \) interact if \( wL^k(u,v) \) is large
What have we learned?

- It is possible to predict PPIs using a variety of information and methods
  - Gene cluster, gene fusion, phylogenetic profile, sequence co-evolution, …

For those who are interested to go further:
- How do you predict cross-species PPI’s between a host and a pathogen?
Must Read


Good to Read

• Soh et al. **Consistency, Comprehensiveness, and Compatibility of Pathway Databases.** *BMC Bioinformatics*, 11:449, 2010

• Sprinzak et al. **How reliable are experimental protein-protein interaction data?.** *JMB*, 327:919-923, 2003

• Liu et al. **Assessing and predicting protein interactions using both local and global network topological metrics.** *GIW 2008*, pp. 138-149

• Juan et al. **High-confidence prediction of global interactomes based on genome-wide coevolution networks.** *PNAS*, 105(3):934-939, 2008

• Enright et al. **Protein interaction maps for complete genomes based on gene fusion events.** *Nature*, 402:86–90, 1999

• Pellegrini et al. **Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles.** *PNAS*, 96:4285–4288, 1999

Comparative Analysis and Assessment of M. tuberculosis H37Rv Protein-Protein Interaction Datasets
Outline

• Low similarity betw two MTB PPI datasets

• Hypothesis: One or both of them are of very poor quality?

• Evaluating the quality of the two datasets
  – Informative GO assessment
  – PPI Functional Intensity matrix
  – Correlation w/ gene expression profiles
  – Interologs from diff organisms’ expt PPI’s
  – Integrated-pathway gene-pair relationships

This part of the lecture is based on

Low Similarity of H37Rv PPI Datasets

- **STRING v 8.3 M. tuberculosis H37Rv PPI datasets**
  
<table>
<thead>
<tr>
<th>STRING database prediction method</th>
<th>Number of PPIs</th>
<th>STRING database prediction method</th>
<th>Number of PPIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbourhood</td>
<td>12,706</td>
<td>Transferred neighborhood</td>
<td>78,376</td>
</tr>
<tr>
<td>Co-expression</td>
<td>0</td>
<td>Transferred co-expression</td>
<td>4,393</td>
</tr>
<tr>
<td>Experiments</td>
<td>4</td>
<td>Transferred experimental</td>
<td>4,129</td>
</tr>
<tr>
<td>Databases</td>
<td>7,030</td>
<td>Transferred databases</td>
<td>629</td>
</tr>
<tr>
<td>Text mining</td>
<td>2,715</td>
<td>Transferred text mining</td>
<td>11,074</td>
</tr>
<tr>
<td>Gene fusion</td>
<td>2,646</td>
<td>Co-occurrence</td>
<td>159,213</td>
</tr>
<tr>
<td>All STRING database PPIs</td>
<td>248,574</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Summary of various subsets of H37Rv PPIs in STRING and their sources.

- **High-throughput bacterial two-hybrid (B2H) dataset**: 8042 PPIs covering 2907 proteins

- **Overlap # of PPIs betw the two datasets**: 276
Why is the overlap so low?

• The STRING PPI dataset is too low quality?
  – MTB PPI’s in STRING come from predictions. Perhaps these predictions are all wrong

• The MTB B2H PPI dataset is too low quality?
  – B2H & Y2H technologies are very noisy. Perhaps these PPIs are produced by bad B2H expt

• Both of the datasets are too low quality?

• Let us try to find out which of the above is the case
Fraction of Overlap Expt PPIs wrt STRING Score

- STRING and MTB B2H PPI dataset have higher similarity when STRING score is > ~750
- Similar observation for many other species
  \[ \Rightarrow \text{Set threshold at } \sim 750 \text{ for high-quality STRING PPI's} \]
Functional Intensity Matrix

• STRING PPI (score >770) shows strong diagonal intensity
  ⇒ Many PPIs have interacting partners with same function

• MTB B2H PPI dataset contains lots of PPI that are functional uninformative or unknown
Correlated Gene Expression Profiles

- Large portion of STRING PPI (score >770) have correlated gene expression profiles
- Many MTB B2H PPI’s do not have correlated gene expression profiles in the two interacting partners
  - Note that PPIs with either of interaction partner w/o gene expression profile are assigned a Pearson's correlation coefficient to be -1.0. MTB B2H PPI dataset has large # of PPIs w/o gene expression profile
Informative GO Term Coherence

62% of STRING PPIs (score >770) have the same informative GO term in both proteins.

Only 13% of MTB B2H PPI's show informative GO term coherence. No diff from random PPI's.
## GO Term Coherence:
Comparison w/ Other Y2H PPI’s

<table>
<thead>
<tr>
<th></th>
<th>CC</th>
<th>BP</th>
<th>MF</th>
<th>GO</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. cerevisiae</em> Y2H PPI</td>
<td>49.74%</td>
<td>38.43%</td>
<td>35.90%</td>
<td>40.58%</td>
</tr>
<tr>
<td><em>S. cerevisiae</em> random PPI</td>
<td>5.03%</td>
<td>3.70%</td>
<td>7.72%</td>
<td>7.08%</td>
</tr>
<tr>
<td>Info GO ratio of <em>S. cerevisiae</em> (Y2H PPI / random)</td>
<td>9.88</td>
<td>10.38</td>
<td>4.65</td>
<td>5.73</td>
</tr>
<tr>
<td>Info GO term No.</td>
<td>57</td>
<td>173</td>
<td>78</td>
<td>365</td>
</tr>
<tr>
<td><em>C. jejuni</em> Y2H PPI</td>
<td>47.37%</td>
<td>17.01%</td>
<td>16.27%</td>
<td>17.91%</td>
</tr>
<tr>
<td><em>C. jejuni</em> random PPI</td>
<td>35.61%</td>
<td>11.04%</td>
<td>12.91%</td>
<td>13.97%</td>
</tr>
<tr>
<td>Info GO ratio of <em>C. jejuni</em> (Y2H PPI / random)</td>
<td>1.33</td>
<td>1.54</td>
<td>1.26</td>
<td>1.28</td>
</tr>
<tr>
<td>Info GO term No.</td>
<td>3</td>
<td>26</td>
<td>22</td>
<td>51</td>
</tr>
<tr>
<td><em>Synechocystis</em> Y2H PPI</td>
<td>80.77%</td>
<td>26.16%</td>
<td>25.30%</td>
<td>29.47%</td>
</tr>
<tr>
<td><em>Synechocystis</em> random PPI</td>
<td>44.94%</td>
<td>6.89%</td>
<td>8.59%</td>
<td>9.73%</td>
</tr>
<tr>
<td>Info GO ratio of <em>Synechocystis</em> (Y2H PPI / random)</td>
<td>1.80</td>
<td>3.79</td>
<td>2.95</td>
<td>3.03</td>
</tr>
<tr>
<td>Info GO term No.</td>
<td>3</td>
<td>30</td>
<td>30</td>
<td>63</td>
</tr>
<tr>
<td><em>M. tuberculosis</em> B2H PPI</td>
<td>38.46%</td>
<td>14.36%</td>
<td>9.57%</td>
<td>13.83%</td>
</tr>
<tr>
<td><em>M. tuberculosis</em> random PPI</td>
<td>38.30%</td>
<td>11.35%</td>
<td>8.35%</td>
<td>11.51%</td>
</tr>
<tr>
<td>Info GO ratio of <em>M. tuberculosis</em> (B2H PPI / random)</td>
<td>1.00</td>
<td>1.26</td>
<td>1.15</td>
<td>1.20</td>
</tr>
<tr>
<td>Info GO term No.</td>
<td>3</td>
<td>32</td>
<td>32</td>
<td>67</td>
</tr>
</tbody>
</table>

The table is a summary of the percentage of PPIs having coherent informative GO terms, number of informative GO terms and “Info GO ratio” in each of the representative two-hybrid PPI datasets in four organisms. Info GO ratio = percentage of PPIs in two-hybrid PPI dataset having coherent informative GO terms / percentage of PPIs in random PPI dataset having coherent informative GO terms.

- **MTB B2H PPI** has the lowest GO term coherence among all Y2H/B2H datasets, after normalization.
STRING PPI Overlap:
Comparison w/ Other Y2H PPI’s

- MTB B2H PPI has much lower overlap with STRING PPI compared to other Y2H PPI datasets
How well do MTB PPIs in STRING & B2H agree with interologs predicted from expt PPIs of other species

<table>
<thead>
<tr>
<th>Source PPIs for homology transfer</th>
<th>STRING database experimental PPIs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identify homology and transfer</td>
<td></td>
</tr>
<tr>
<td>PIDE &gt;30</td>
<td>HVAL &gt;20</td>
</tr>
<tr>
<td>Coverage &gt;0.2</td>
<td>E-VALUE &lt;1*e-6</td>
</tr>
<tr>
<td>E-VALUE &lt; 1*e-10</td>
<td>Identified by Inparanoid (default parameters)</td>
</tr>
<tr>
<td>Benchmarks</td>
<td></td>
</tr>
<tr>
<td>Jaccard coefficient</td>
<td>0.00354</td>
</tr>
<tr>
<td>Precision</td>
<td>0.00375</td>
</tr>
<tr>
<td>Recall</td>
<td>0.0588</td>
</tr>
<tr>
<td>Overlapping PPIs</td>
<td>5</td>
</tr>
</tbody>
</table>

**Table 3 Results of predicted interologs from STRING experimental PPIs**

- Interologs predicted from expt PPIs of related species agree better w/ MTB PPIs in STRING
Conclusion

Why is the overlap so low?

- The STRING PPI dataset is too low quality?
  - MTB PPI’s in STRING come from predictions. Perhaps these predictions are all wrong

- The MTB B2H PPI dataset is too low quality?
  - B2H & Y2H technologies are very noisy. Perhaps these PPIs are produced by bad B2H expt

- Both of the datasets are too low quality?

- Let us try to find out which of the above is the case
Acknowledgements

Lui Guimei  Donny Soh  Zhou Hufeng

• Lots of ideas on PPIN cleansing here came from my colleague, Dr Liu Guimei

• Lots of results on consistency & comprehensiveness of pathway databases, and their integration, came from my ex-students, Donny Soh & Zhou Hufeng

• The slides on MTB B2H PPI quality assessment are from the work of my ex-student, Zhou Hufeng