Function Prediction

CMSC 858L
Predicting Protein Function from Networks

• Ultimately, we want to know how various processes in the cell work.

• A first step: figure out which proteins are involved in which biological role.

• What do we mean by a “biological role”?
  – Several different schemes:
    • Gene Ontology (largest, most widely used)
    • MIPS (good collection of known protein complexes)
    • KEGG (manually curated pathways)
Gene Ontology (GO)

- Node = manually defined function
- Directed, acyclic graph
- Main edges are either “is a” or “part of”

Curated collection of biological functions
Gene Ontology has 3 Sub-ontologies

- **Cellular component**: a part of the cell (a location, or organelle, or other structure)

- **Biological process**: a collection of steps that the cell carries out to achieve some purpose. E.g. cell division.

- **Molecular function**: a specific mechanism that a protein performs. E.g.
  - a kinase would have molecular function “phosphorylation”;
  - a transcription factor would have molecular function “DNA binding”

- Each protein may be *annotated* with several terms from each sub-ontology.
Edge Types

• **is_a**: like a C++ or Java subclass relationship.
  - A is_a B means A is a more specific version of B
  - E.g. “nuclear chromosome” is_a “chromosome”.

• **part_of**: A is some part of B
  - A piston is part_of an engine (but a piston is not an specific kind of engine)

• **Transitivity**:
  - If a protein is annotated with term A, it is implicitly annotated with all the ancestors of A (following every path to the root).
  - GO is explicitly designed so this is always true.
Gene Ontology is Complicated
1. Metabolism

1.1 Carbohydrate Metabolism
Glycolysis / Gluconeogenesis
Citrate cycle (TCA cycle)
Pentose phosphate pathway
Pentose and glucuronic acid interconversions
Fructose and mannose metabolism
Galactose metabolism
Ascorbate and aldarate metabolism
Starch and sucrose metabolism
Aminosugars metabolism
Nucleotide sugars metabolism
Pyruvate metabolism
Glyoxylic acid and dicarboxylate metabolism
Propanoate metabolism
Butanoate metabolism
C5-Branch dibasic acid metabolism
Inositol metabolism
Inositol phosphate metabolism

1.2 Energy Metabolism
Oxidative phosphorylation
Photosynthesis
Photosynthesis - antenna proteins
Carbon fixation in photosynthetic organisms
Reducive carboxylate cycle (CO2 fixation)
Methane metabolism
Nitrogen metabolism
Sulfur metabolism

1.3 Lipid Metabolism
Fatty acid biosynthesis
Fatty acid elongation in mitochondria
Fatty acid metabolism
Synthesis and degradation of ketone bodies
Biosynthesis of steroids
Bile acid biosynthesis
C21-Steroid hormone metabolism
Androgen and estrogen metabolism
Glycerolipid metabolism
Glycerophospholipid metabolism
Ether lipid metabolism
Sphingolipid metabolism
Arachidonic acid metabolism
Linoleic acid metabolism
alpha-Linolenic acid metabolism
Biosynthesis of unsaturated fatty acids

1.4 Nucleotide Metabolism
Purine metabolism
Pyrimidine metabolism

1.5 Amino Acid Metabolism
Glutamate metabolism

1.5 Amino Acid Metabolism

1.5.1 Glutamate metabolism
- Alanine metabolism
- Aspartate metabolism
- Glycine, serine and threonine metabolism
- Methionine metabolism
- Cysteine metabolism
- Valine, leucine and isoleucine degradation
- Valine, leucine and isoleucine biosynthesis
- Lysine biosynthesis
- Lysine degradation
- Arginine and proline metabolism
- Histidine metabolism
- Tyrosine metabolism
- Phenylalanine metabolism
- Tryptophan metabolism
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Urea cycle and metabolism of amino acids

1.6 Metabolism of Other Amino Acids
- beta-Alanine metabolism
- Taurine and hypouracil metabolism
- Aminophosphonate metabolism
- Selenoamino acid metabolism
- Cyanobacterial acid metabolism
- D-Glutamine and D-glutamate metabolism
- D-Arginine and D-ornithine metabolism
- D-Alanine metabolism
- Glutathione metabolism

1.7 Glycan Biosynthesis and Metabolism
- N-Glycan biosynthesis
- High-mannose type N-glycan biosynthesis
- N-Glycan degradation
- O-Glycan biosynthesis
- Chondroitin sulfate biosynthesis
- Heparan sulfate biosynthesis
- Keratan sulfate biosynthesis
- Glycosaminoglycan degradation
- Lipopolysaccharide biosynthesis
- Peptidoglycan biosynthesis
- Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
- Glycosphingolipid biosynthesis - lactoseries
- Glycosphingolipid biosynthesis - neo-lactoseries
- Glycosphingolipid biosynthesis - globoseries
- Glycosphingolipid biosynthesis - ganglioseries
- Glycan structures - biosynthesis 1
- Glycan structures - biosynthesis 2
- Glycan structures - degradation

1.8 Biosynthesis of Polyketides and Nonribosomal Peptides
- Type I polyketide structures
- Biosynthesis of 12-, 14- and 16-membered macrolides
- Biosynthesis of ansamycins
- Biosynthesis of type II polyketide backbone
- Biosynthesis of type II polyketide products
MIPS has annotation terms organized in trees

- Function Catalog (FunCat): a collection of functions and biological processes organized as a tree.

- Manually annotated protein complexes (e.g. at left)
  - also organized as a tree

<table>
<thead>
<tr>
<th>Complex</th>
<th>Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>2-oxoglutarate dehydrogenase</td>
</tr>
<tr>
<td>40</td>
<td>Alpha-agglutinin anchor</td>
</tr>
<tr>
<td>60</td>
<td>Anaphase promoting complex (APC)</td>
</tr>
<tr>
<td>70</td>
<td>Anthranilate synthase</td>
</tr>
<tr>
<td>75</td>
<td>Arginase</td>
</tr>
<tr>
<td>80</td>
<td>Arginine-specific carbamoylphosphate synthase</td>
</tr>
<tr>
<td>90</td>
<td>Assembly complexes</td>
</tr>
<tr>
<td>100</td>
<td>Calcineurin B</td>
</tr>
<tr>
<td>110</td>
<td>cAMP-dependent protein kinase</td>
</tr>
<tr>
<td>120</td>
<td>Casein kinase</td>
</tr>
<tr>
<td>123</td>
<td>Catalase</td>
</tr>
<tr>
<td>125</td>
<td>Cell cycle checkpoint complexes</td>
</tr>
<tr>
<td>130</td>
<td>Chaperone containing T-complex TRIC (TCP RING Complex)</td>
</tr>
<tr>
<td>132</td>
<td>CTP synthetase</td>
</tr>
<tr>
<td>133</td>
<td>Cyclin-CDK (Cyclin-dependent kinases) complexes</td>
</tr>
<tr>
<td>140</td>
<td>Cytoskeleton</td>
</tr>
<tr>
<td>143</td>
<td>D-arabinose dehydrogenase</td>
</tr>
<tr>
<td>145</td>
<td>delta3-cis-delta2-trans-enoyl-CoA isomerase</td>
</tr>
<tr>
<td>150</td>
<td>Endonuclease Scel, mitochondrial</td>
</tr>
<tr>
<td>160</td>
<td>Exocyst complex</td>
</tr>
<tr>
<td>170</td>
<td>Fatty acid synthetase, cytoplasmic</td>
</tr>
<tr>
<td>172</td>
<td>Fatty acid synthetase, mitochondrial</td>
</tr>
<tr>
<td>177</td>
<td>Gim complexes</td>
</tr>
<tr>
<td>180</td>
<td>Prenyltransferases</td>
</tr>
<tr>
<td>190</td>
<td>Glucan synthases</td>
</tr>
<tr>
<td>200</td>
<td>Glycine decarboxylase</td>
</tr>
<tr>
<td>210</td>
<td>H+-ATPase, plasma membrane</td>
</tr>
<tr>
<td>220</td>
<td>H+-transporting ATPase, vacuolar</td>
</tr>
<tr>
<td>225</td>
<td>Hexokinase 2</td>
</tr>
<tr>
<td>230</td>
<td>Histone acetyltransferase complexes</td>
</tr>
<tr>
<td>240</td>
<td>Histone deacetylase complexes</td>
</tr>
</tbody>
</table>
Basic Methods for Predicting Function

• Majority Rule
• Neighborhood enrichment
• Minimum Multiway Cut
• “Functional Flow”
Neighboring Proteins More Likely to Share Function

• (Yu et al., 2008)
Majority Rule

- Proteins with known function + network topology $\rightarrow$ function assignment for unknown proteins.
- Guilt by association
- Majority Rule:
  - Can weight contribution by edge weight.
  - Doesn’t take into account connections between neighbors
  - Or annotations at distance $> 1$
Neighborhood Approaches, e.g.:

- Let \( N(u,r) \) be all the proteins within distance \( r \) to \( u \).

\[
f(u, r, a) = |\{u \in N(u, r) : u \text{ has function } a\}|
\]

= # of proteins in neighborhood with function \( a \)

\[
e(u, r, a) = |N(u, r)| \cdot \frac{|\{u \in V : u \text{ has function } a\}|}{|V|}
\]

= Expected # of proteins in neighborhood with function \( a \)

\[
Score(u, r, a) = \frac{(f(u, r, a) - e(u, r, a))^2}{e(u, r, a)}
\]

\( \approx \chi^2 \) statistic measures how surprising it is to see the observed # of proteins annotated with \( a \) in the neighborhood

- Protein \( u \) is assigned function \( \arg\max_a \text{Score}(u,r,a) \)
Problems with neighborhood

- Neighborhood with radius 2 gives the same scores for black and gray functions to nodes u and v:

(Nabieva, Singh, 2008)
**Minimum Multiway $k$-Cut:** Partition the nodes so that each of $k$ (sets of) terminal nodes is in a different partition & the number of edges cut is minimized.

- One “terminal node set” for each function, containing proteins known to have that function.
- NP-hard: simulated annealing; integer programming
Integer Programming

• General optimization framework:
  – Describe system by set of variables

\[
\text{IP := }
\begin{align*}
\text{- Minimize a linear function.} \\
\text{- Subject to linear constraints (\(\leq\) or \(\geq\)).} \\
\text{- While requiring the variables to be \(\{0,1\}\).}
\end{align*}
\]

• Computationally hard, but many advanced solver packages:
  – CPLEX, COIN-OR, ABACUS, FortMP, LINGO, …
Integer Programming (IP) Formulation for Multiway Cut

Introduce 0/1 variables associated with each node and edge:

Intuition: \( x_{u,a} \) is 1 if node \( u \) is assigned to annotation \( a \); 0 otherwise.

Intuition: \( x_{u,v,a} \) is 1 if both \( u \) and \( v \) are assigned to annotation \( a \); 0 otherwise.

Intuition: \( x_{u,v,b} \) is 1 if both \( u \) and \( v \) are assigned to annotation \( b \); 0 otherwise.
IP for Min Multiway Cut

\[
\text{maximize } \sum_{\{u,v\} \in E, a} x_{u,v,a} \quad \text{Maximize \# of "monochromatic edges"}
\]
\[
\text{Equivalent to minimizing the number of cut edges.}
\]

Subject to:
\[
x_{u,x} \quad \text{and} \quad x_{u,v,a} \in \{0, 1\}
\]
\[
\sum_{a} x_{u,a} = 1 \quad \text{Each node gets exactly 1 annotation}
\]
\[
x_{u,v,a} \leq x_{u,a} \quad \text{Can set } x_{u,v,a} \text{ to 1 iff both its endpoints are 1}
\]
\[
x_{u,v,a} \leq x_{v,a}
\]
\[
x_{u,a} = 1 \text{ if } a \in \text{annot}(u)
\]
\[
x_{u,a} = 0 \text{ if } a \not\in \text{annot}(u) \neq \emptyset \quad \text{Fix variables for nodes with known annotations.}
\]
Problem with Simple Cut Approaches

- Every cut is equally likely:

  but this node is more likely to be grey than black

(Nabieva, Singh, 2008)
Each node $u$ has a "reservoir" at each time step $t$.

At every time step, water flows "downhill" from the more filled reservoir to the more empty reservoir, up to the capacity of the edge.

If there isn't enough water to fill the downhill pipes, it is distributed proportionally to the capacity of the edge.

Every function $f$ is considered separately. Score($u, f$) is the total water that passed through $u$ when considering $f$. Predicted function for $u$ is the function with the highest score.
Performance of These Predictions on Yeast

(Nabieva et al, 2005)
Summary

• Guilt-by-association = proteins near one another in the network are more likely to have the same function.

• Neighborhood 1 does better than larger neighborhoods Perhaps because the structure of the neighborhood is not taken into account.

• Integer programming NP-hard, but often practical. Can obtain multiple solutions in 2 ways:
  – Random perturbation of weights
  – Solving successive problems with additional constraints.

• “Functional flow” is an embodiment of a general technique: “information” being passed along the network.