Identifying The Components of Cellular Pathways and Protein Complexes using Co-evolution

MCDB187
Proteins are Components of Molecular Machines

The Study of the Co-Evolution of Non-Homologous Proteins

- Because selection generally acts to maintain or delete entire complexes and pathways, pairs of proteins that are part of these will appear to co-evolve across bacteria.
- By studying the co-evolution of non-homologous proteins across these bacteria we attempt to reconstruct the components of complexes and pathways.
Bacterial Diversity

- 1000 fully sequenced genomes in Genbank
- 30,000 species represented in Genbank
- Sea may support 2,000,000*
- Soil may support 4,000,000*

## Methods to Infer Co-evolution

<table>
<thead>
<tr>
<th>Method</th>
<th>Basis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phylogenetic Profile</td>
<td>Pairs of genes that are always present or absent together</td>
</tr>
<tr>
<td>Rosetta Stone</td>
<td>Pairs of proteins that are fused in some organism</td>
</tr>
<tr>
<td>Gene Neighbor</td>
<td>Pairs of genes that are coded nearby in multiple organisms</td>
</tr>
<tr>
<td>Gene Cluster</td>
<td>Gene proximity within genome</td>
</tr>
</tbody>
</table>
Phylogenetic Profile

Phylogenetic Profiles of flagellar protein cluster together
Hypergeometric Distribution

How often do we observe an overlap of $k$ elements when we draw two lists of size $m$ and $n$ from a population of size $N$?

$$P(k \mid n, m, N) = \frac{\binom{n}{k} \binom{N-n}{m-k}}{\binom{N}{m}}$$

where

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$
Gene Neighbor Method

Linking Dihydrofolate reductase and Thymidilate synthase

<table>
<thead>
<tr>
<th>Genome (Contig)</th>
<th>Total Genes</th>
<th>Gene Separation</th>
<th>Contig Layout</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Escherichia coli K12</em> (Chromosome 1)</td>
<td>4289</td>
<td>1574</td>
<td></td>
</tr>
<tr>
<td><em>Agrobacterium tumefaciens</em> (Circular)</td>
<td>2721</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em> (Chromosome 2)</td>
<td>4036</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em> (Chromosome 4)</td>
<td>3816</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Bacillus halodurans</em> (Chromosome 1)</td>
<td>4066</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Bacillus subtilis</em> (Chromosome 1)</td>
<td>4100</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Bordetella pertussis</em> (Contig 104)</td>
<td>40</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Buchnera sp. APS</em> (Chromosome 1)</td>
<td>564</td>
<td>276</td>
<td></td>
</tr>
<tr>
<td><em>Caulobacter crescentus</em> (Chromosome 1)</td>
<td>3737</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td><em>Clostridium acetobutylicum</em> (Chromosome 1)</td>
<td>3672</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Corynebacterium diphtheriae</em> (Contig 26)</td>
<td>75</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td><em>Deinococcus radiodurans</em></td>
<td>2579</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>
Gene Neighbor Probability

The probability that a pair of genes $i, j$ in genome $k$ with $n_k$ genes would be separated a distance $d^*$ less than the observed distance $d$,

$$P(d_{ij}^* \leq d_{ij}) = \frac{2d_{ij}}{n_k - 1}$$

For a pair of genes $i, j$ across $m$ genomes

$$Q = \prod_{k=1}^{m} \frac{2d_{ij}}{n_k - 1}$$

The probability of observing a $Q^*$ less than the observed $Q$ is computed using the Gamma distribution.
Monomeric proteins that are found fused in another organism are likely to be functionally related and physically interacting.

Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D, Detecting protein function and protein-protein interactions from genome sequences. Science 285(5428):751-3, 1999
As in the case of Phylogenetic Profiles we can use the Hypergeometric distribution to estimate the statistical significance of the overlap.
If we model the start of genes as a random process, we can use the Poisson distribution to estimate the probability that two genes are separated by a distance greater than the observed one.
Here, a p-value threshold of 0.1 captures all but one of the genes for this operon.
Combining Inferences of Co-evolution from Previous

We combine the probabilities from the previous four methods to arrive at a single probability that two proteins co-evolve:

\[ P = \min(PP, RS, GN, OP) \]

This allows us to generate networks where proteins are linked if any one method generates a statistically significant link.
We test the network by asking how often we link together functionally related proteins. True and False Interactions are derived from Pathway Classification Schemes.
Benchmarking using Receiver Operator curves

- Find the P values associated with each protein pair
  - 1 2  \( P = .001 \)
  - 1 3  \( P = 0.1 \)
  - 1 4  \( P = 0.0001 \)
  - ....
  - 4000 3999  \( P = 0.5 \)
Benchmarking using Receiver Operator curves

- Sort pairs by P value
- 101 234  \( P = 0.000001 \)
- 1000 300  \( P = 0.00002 \)
- 3456 423  \( P = 0.00004 \)
- ....
- 57 399  \( P = 1 \)
Benchmarking using Receiver Operator curves

- Determine whether each pair is a TP or FP association (based on pathways)

- 101 234  \( P = 0.000001 \)  TP
- 1000 300  \( P = 0.00002 \)  TP
- 3456 423  \( P = 0.00004 \)  FP
- ....
- 57 399  \( P = 1 \)
Benchmarking using Receiver Operator curves

- Compute fraction of TP and FP pairs as a function of rank
  - 101 234  $P = 0.000001$  TP  $1/1000,0/5000$
  - 1000 300  $P = 0.00002$  TP  $2/1000,0/5000$
  - 3456 423  $P = 0.00004$  FP  $2/1000,1/5000$
  - ....
  - 57 399  $P = 1$  FP  $1,1$
Receiver Operator Characteristic Curve

Pathway Recovery in E. Coli
(KEGG pathways)

TP = same pathway
FP = different pathways
Networks of Co-evolving Proteins

We can generate networks of co-evolution by selecting only pairs of proteins whose probability of co-evolution is above a threshold.
Bacterial Flagella Network Using Combined Methods

Order: 3rd
Proteins: 48

Method (# links)
- PP (32)
- RS (0)
- GN (111)
- GC (0)
- Multiple (15)

KEGG Pathways
- Bacterial chemotaxis
- Flagellar Assembly
- Type III protein secretion
- Oxidative phosphorylation
- ATP Synthesis
- Photosynthesis
- Aminosugars metabolism
- Glutamate metabolism

Saturday, April 27, 13
Bacterial Flagella Network Using Combined Methods
Bacterial Flagella Network Using Combined Methods
Bacterial Flagella Network Using Combined Methods

Order: 3rd
Proteins: 48

Method (# links)
- PP (32)
- RS (0)
- GN (111)
- GC (0)
- Multiple (15)

KEGG Pathways
- Bacterial chemotaxis
- Flagellar Assembly
- Type III protein secret...
- Oxidative phosphoryl...
- ATP Synthesis
- Photosynthesis
- Aminosugars metabol...
- Glutamate metabolism

Saturday, April 27, 13
Bacterial Flagella Network Using Combined Methods
Bacterial Flagella Network Using Combined Methods
Bacterial Flagella Network Using Combined Methods

Order: 3rd
Proteins: 48

Method (# links)
- PP (32)
- RS (0)
- GN (111)
- GC (0)
- Multiple (15)

KEGG Pathways
- Bacterial chemotaxis
- Flagellar Assembly
- Type III protein secretion
- Oxidative phosphorylation
- ATP Synthesis
- Photosynthesis
- Aminosugars metabolism
- Glutamate metabolism
Alternative Representations of Network

Classical Network

Genome-Wide Functional Linkage Map

Strong M, Graeber TG, Beeby M, Pelligrini M, Thompson MJ, Yeates TO, Eisenberg D. Inference and Visualization of Protein Networks in Mycobacterium tuberculosis Based on Hierarchical Clustering of Whole Genome Functional Linkage Maps. Submitted to Nucleic Acids Research

Saturday, April 27, 13
Hierarchical Clustering Reveals Modular Evolution
Clusters are Enriched for Pathways and Complexes
Examples of Clusters that Contain Components of Biochemical Pathways
Cluster Reveals Additional ORFs Involved in Lipopolysaccharide Biosynthesis
Clusters are also Enriched for Subunits of Protein Complexes

True positive interactions are between subunits of known complexes and false positive ones are between subunits of different complexes.

For high confidence links, we recover one third of true interactions and only one thousandth of the false positive ones.
Clusters Containing Subunits of Protein Complexes

Cytochrome c oxidase controls the last step of food oxidation

ATP Synthase
Identification of an Uncharacterized Protein Complex
Conclusions

• Protein modules appear to co-evolve across bacterial species

• Modules are enriched for proteins that participate in the same pathway or complex
We have constructed a database that contains co-evolution links between the genes of 150 fully sequenced genomes.

The Prolinks database may be accessed through the Proteome Navigator web browser interface at:

prolinks.mbi.ucla.edu/
Proteome Navigator Access Page

Proteome Navigator

Search by Database Identifier

GenBank: [Input]

Show Protein

OR

Search by Protein Characteristic

Genome: Escherichia coli K12

Gene Name: [Input]
Annotation: [Input]
COG Description: [Input]
InterPro Domain: [Input]
KEGG Pathway: [Input]
PSORT Location: [Input]
EC Number: [Input]
Amino Acid Sequence: [Input]

Number of Criteria to Display: [Dropdown]

contains: [Input]
contains: [Input]
contains: [Input]

Reset Criteria | Search Proteins
Proteome Navigator