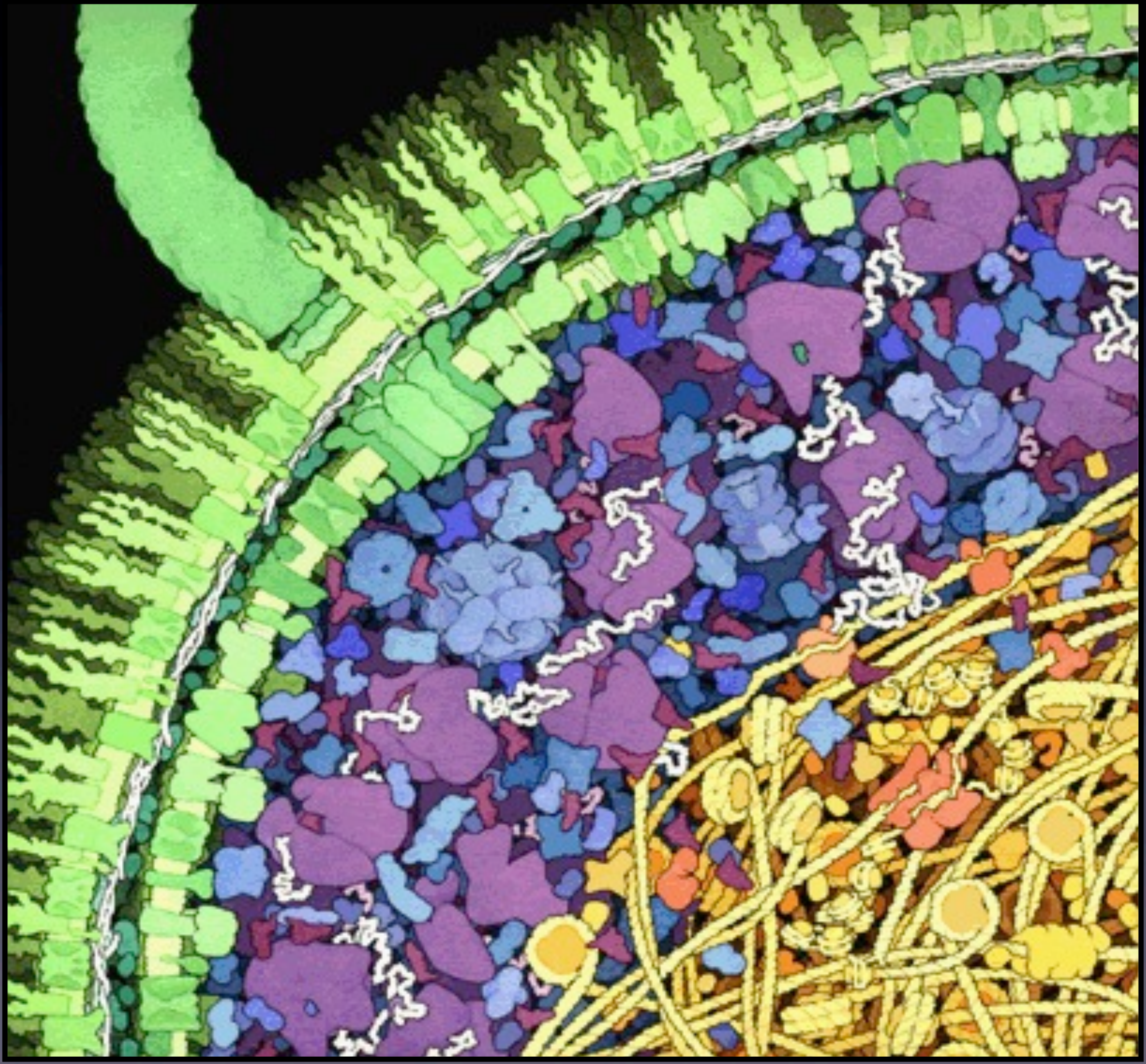
Identifying The Components of Cellular Pathways and Protein Complexes using Co-evolution

MCDB187

Proteins are Components of Molecular Machines

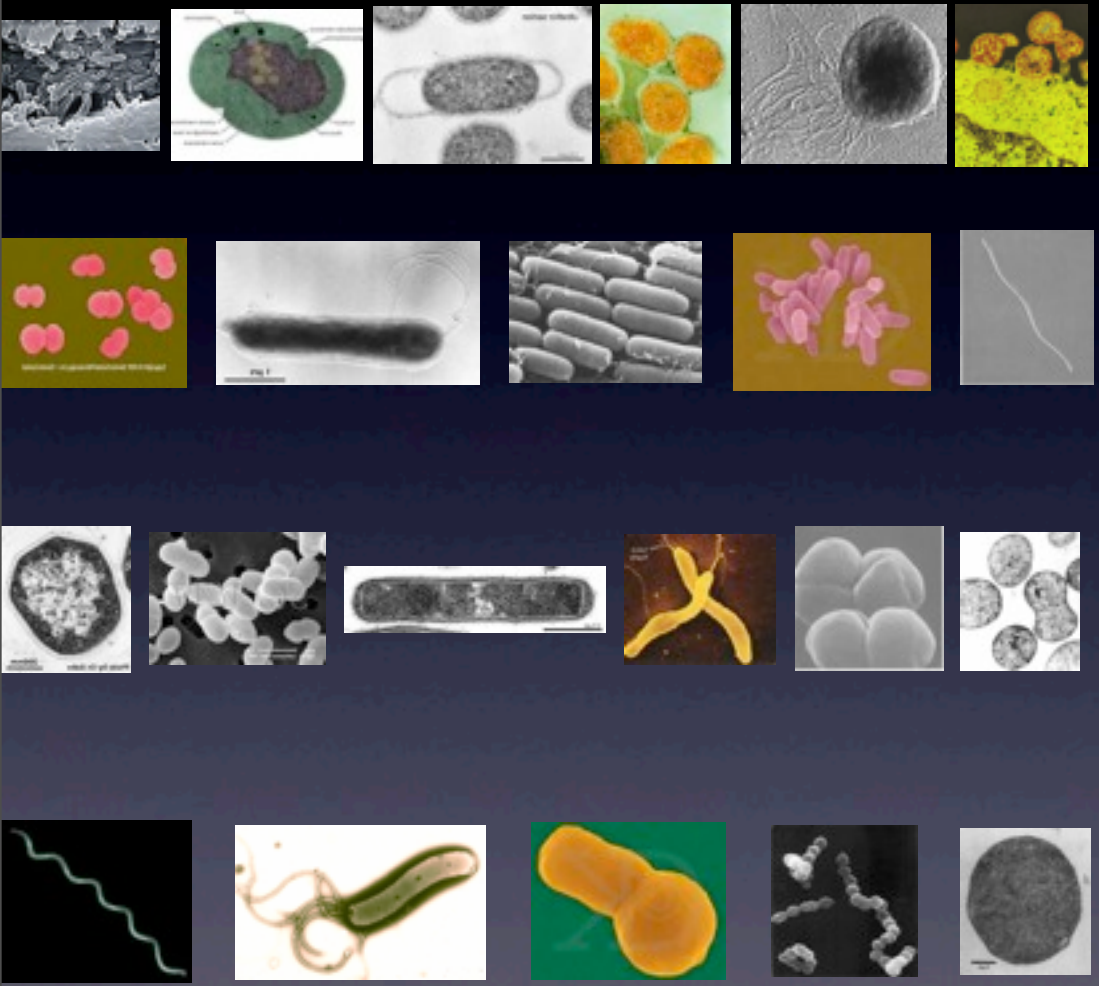


Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999 Dec 2;402(6761 Suppl):C47-52.

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



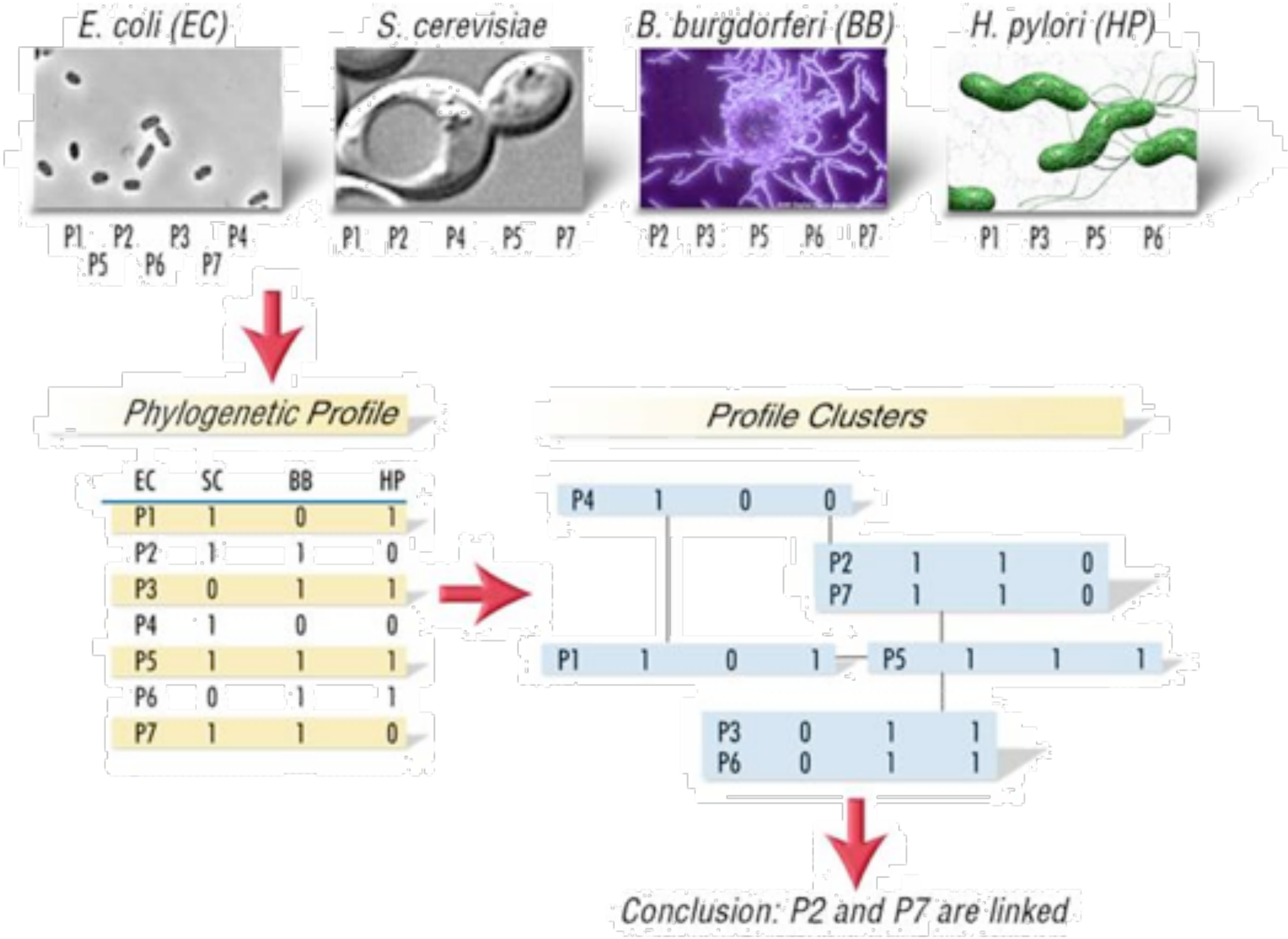
- ·1000 fully sequenced genomes in Genbank
- ·30,000 species represented in Genbank
- ·Sea may support 2,000,000*
- ·Soil may support 4,000,000*

*T.P. Curtis, W.T. Sloan, and J.W. Scannell. 2002. Estimating prokaryotic diversity and its limits Proc Natl Acad Sci USA 99: 10494-10499.

Methods to Infer Co-evolution

Method	Basis
Phylogenetic Profile	Pairs of genes that are always present or absent together
Rosetta Stone	Pairs of proteins that are fused in some organism
Gene Neighbor	Pairs of genes that are coded nearby in multiple organisms
Gene Cluster	Gene proximity within genome

Phylogenetic Profile

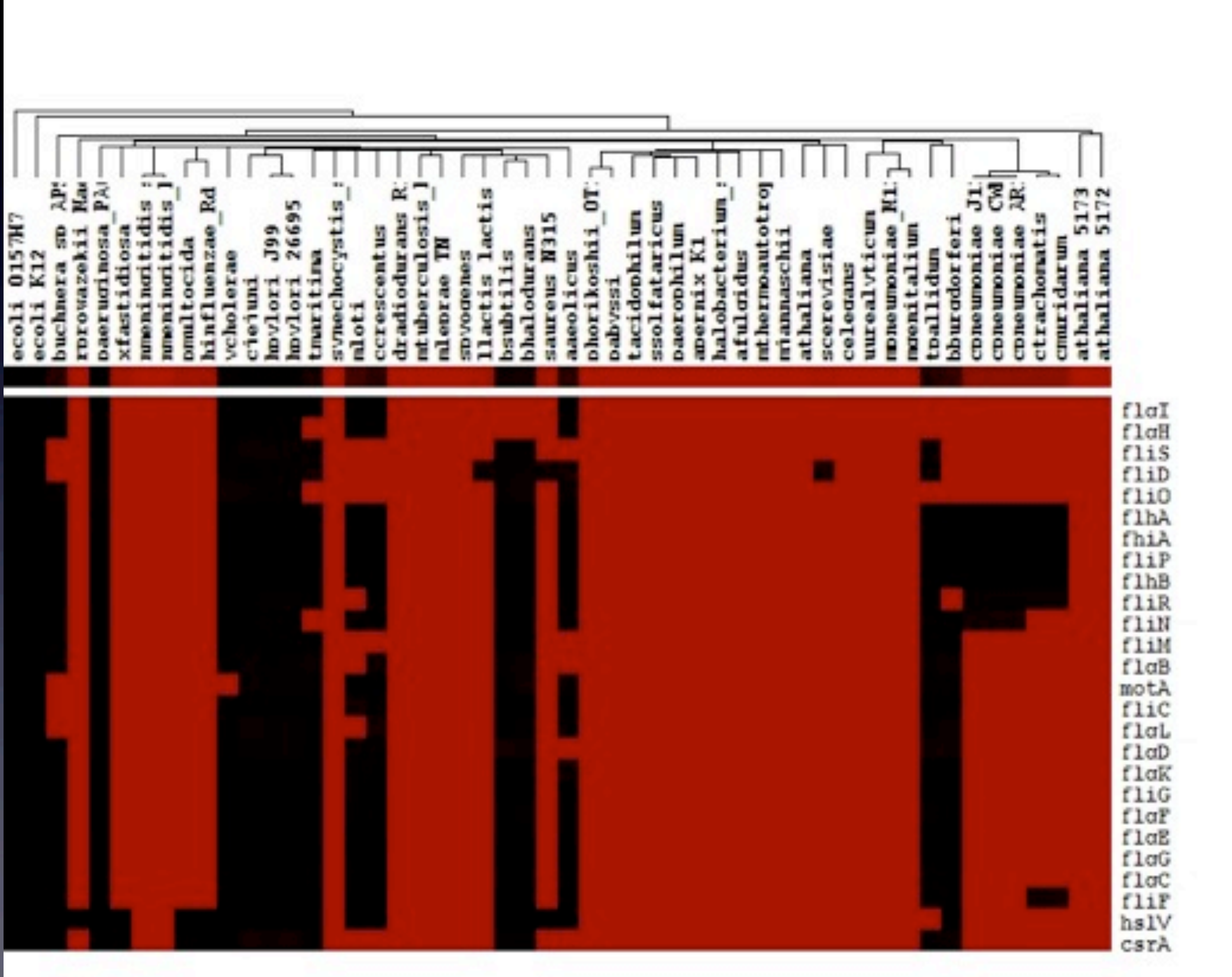


EC	SC	BB	HP
P1	1	0	1
P2	1	1	0
P3	0	1	1
P4	1	0	0
P5	1	1	1
P6	0	1	1
P7	1	1	0

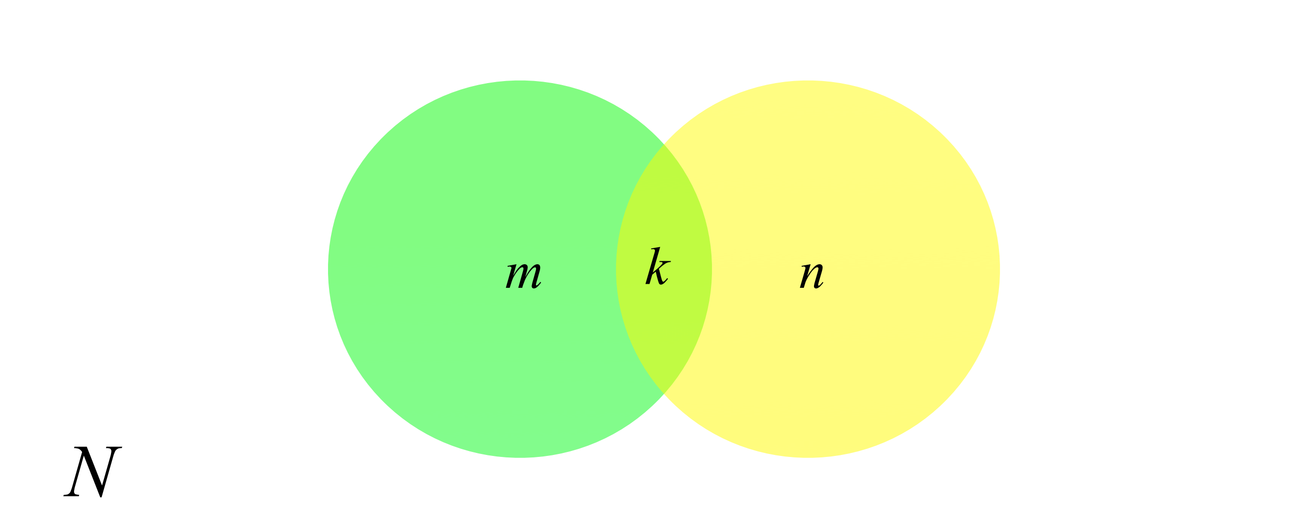
Conclusion: P2 and P7 are linked

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO, Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 96(8): 4285-8,. 1999

Phylogenetic Profiles of flagellar protein cluster together



Hypergeometric Distribution



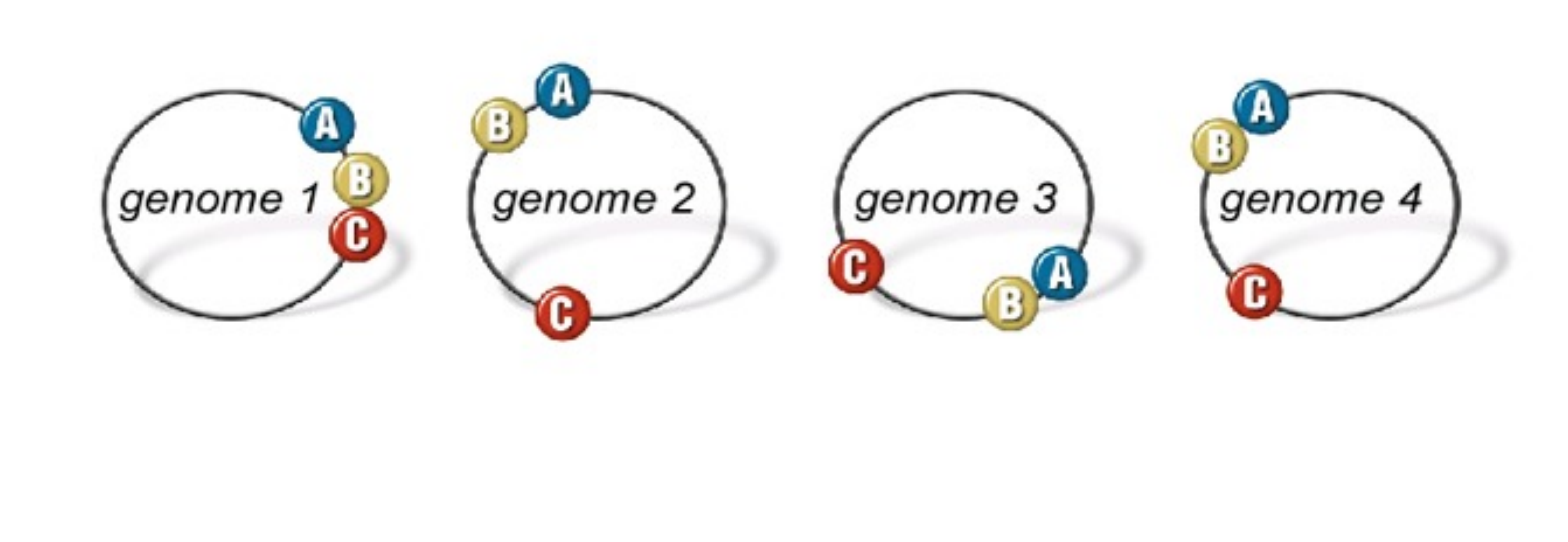
How often do we observe an overlap of k elements when we draw two lists of size m and n from a population of size N?

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

where

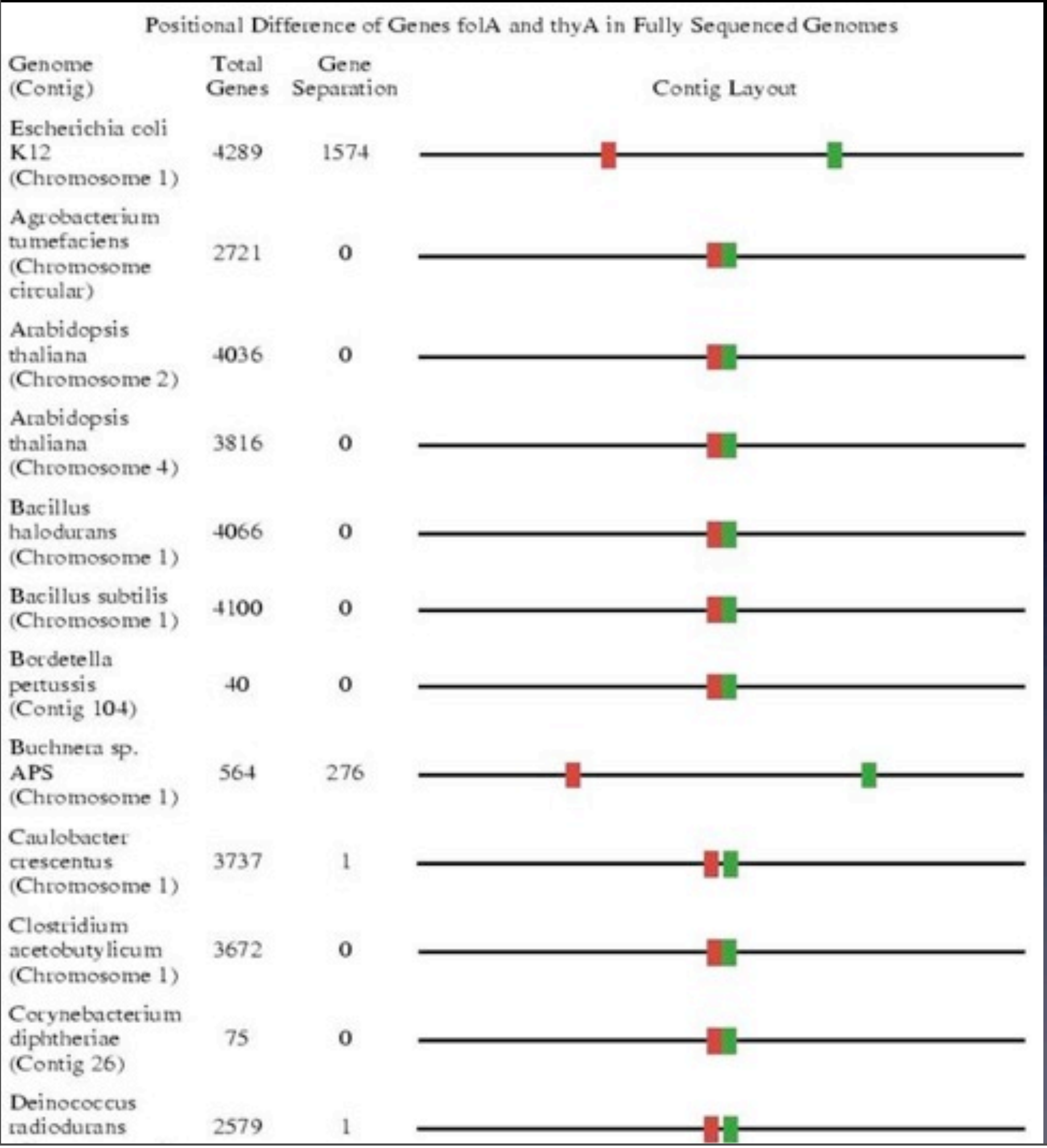
$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Gene Neighbor Method



Pellegrini M, Thompson MJ, Fierro J, Bowers P, A Computational Method to Assign Microbial Genes to Pathways.
Journal of Cellular Biochemistry Suppl 37:106-9, 2001

Linking Dihydrofolate reductase and Thymidilate synthase



Positional Difference of Genes folA and thyA in Fully Sequenced Genomes

Genome (Contig)	Total Genes	Gene Separation	Contig Layout
Escherichia coli K12 (Chromosome 1)	4289	1574	
Agrobacterium tumefaciens (Chromosome circular)	2721	0	
Arabidopsis thaliana (Chromosome 2)	4036	0	
Arabidopsis thaliana (Chromosome 4)	3816	0	
Bacillus halodurans (Chromosome 1)	4066	0	
Bacillus subtilis (Chromosome 1)	4100	0	
Bordetella pertussis (Contig 104)	40	0	
Buchnera sp. APS (Chromosome 1)	564	276	
Caulobacter crescentus (Chromosome 1)	3737	1	
Clostridium acetobutylicum (Chromosome 1)	3672	0	
Corynebacterium diphtheriae (Contig 26)	75	0	
Deinococcus radiodurans	2579	1	

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,

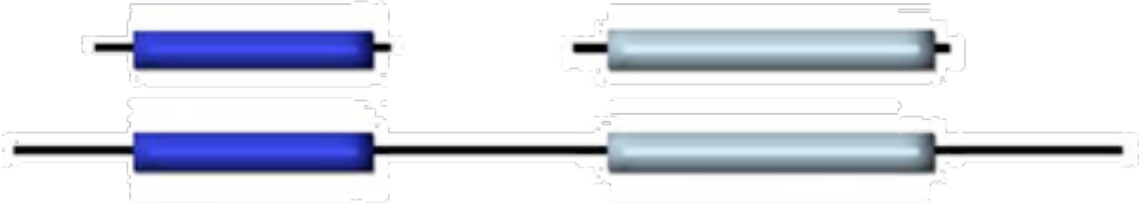
$$P(d_{ij}^* \le d_{ij}) = \frac{2d_{ij}}{n_k - 1}$$

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

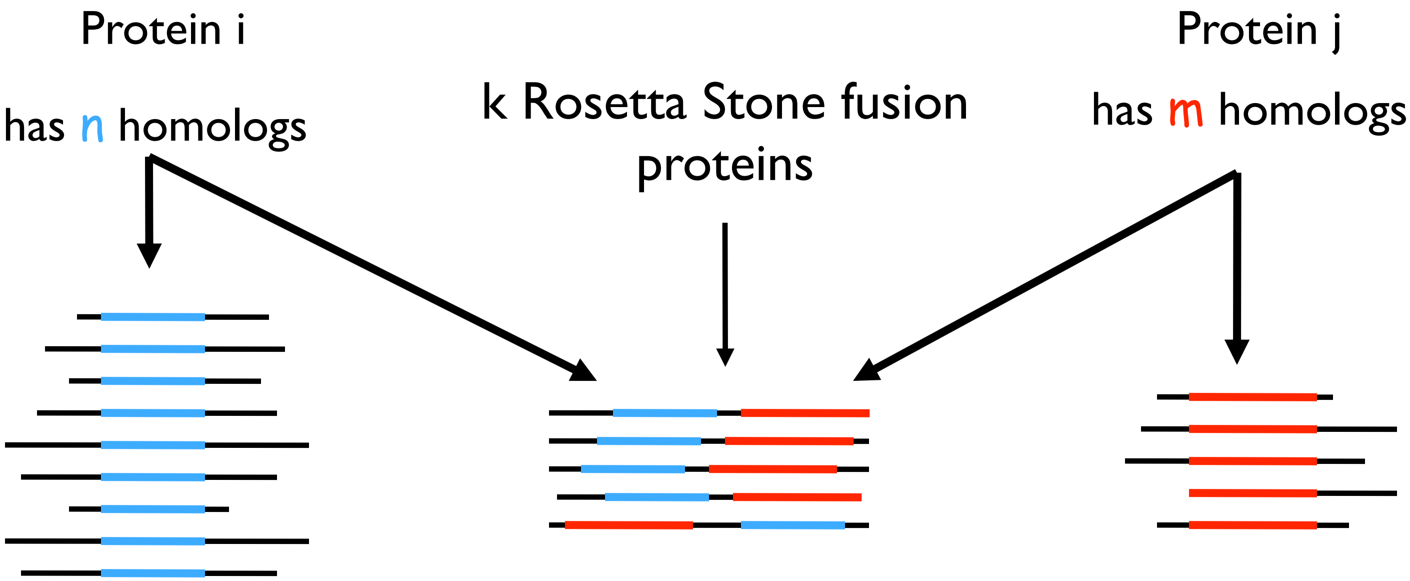
Rosetta Stone Method Identifies Protein Fusions



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

Rosetta Stone Probability



Protein i

has n homologs

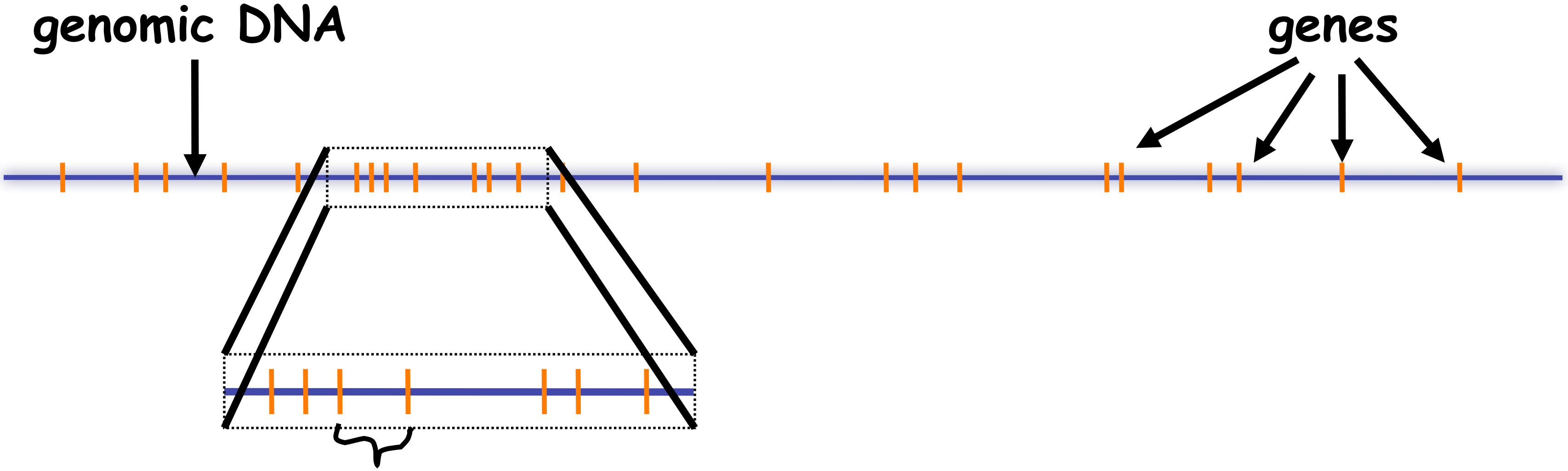
k Rosetta Stone fusion proteins

Protein j

has m homologs

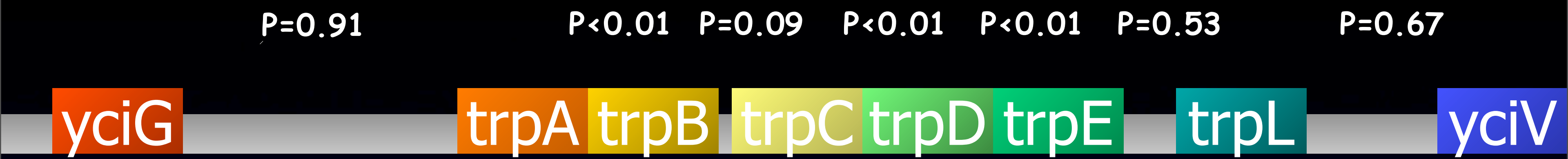
As in the case of Phylogenetic Profiles we can use the Hypergeometric distribution to estimate the statistical significance of the overlap

Gene Cluster



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

Tryptophan Operon



P=0.91 P<0.01 P=0.09 P<0.01 P<0.01 P=0.53 P=0.67

yciG trpA trpB trpC trpD trpE trpL yciV

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:

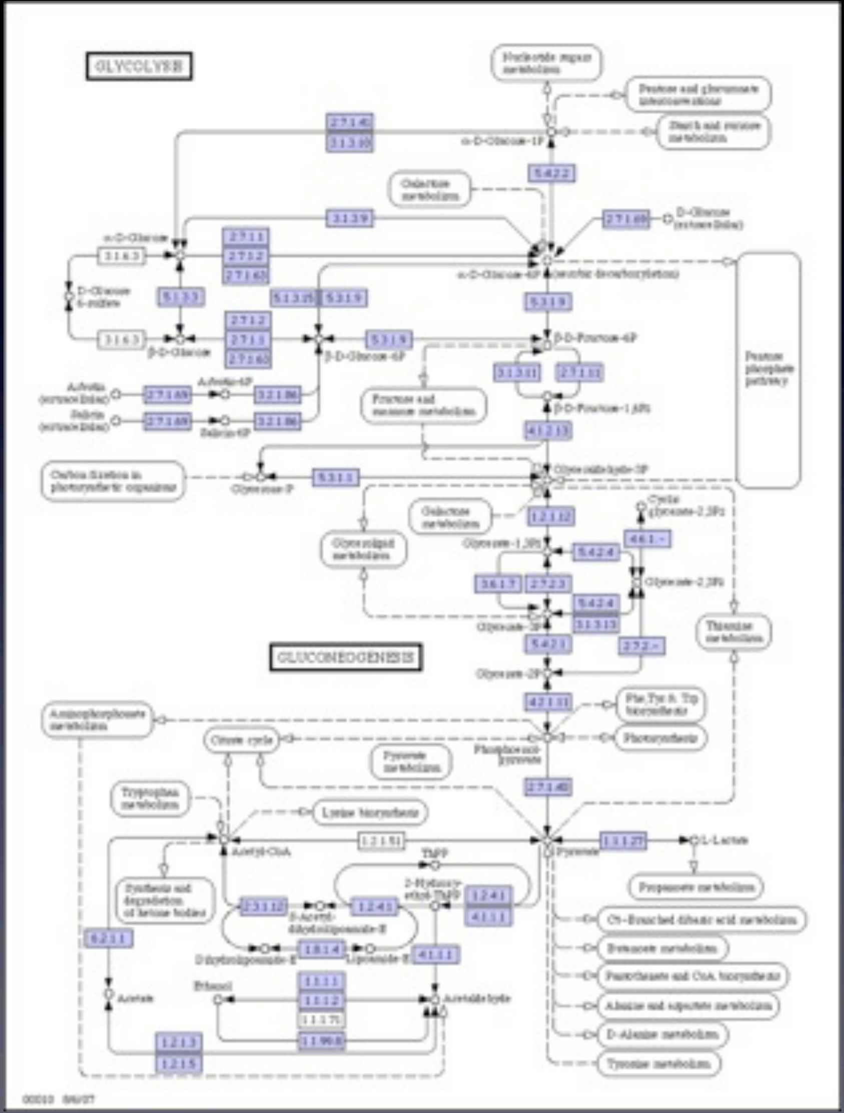
$P = \min(PP, RS, GN, OP)$

This allows us to generate networks where proteins are linked if any one method generates a statistically significant link

Testing the validity of Our Network

- 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 vales 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 = .000001
- 1000 300   P = 0.00002
- 3456 423   P = 0.00004
- ....
- 57 399 P = 1

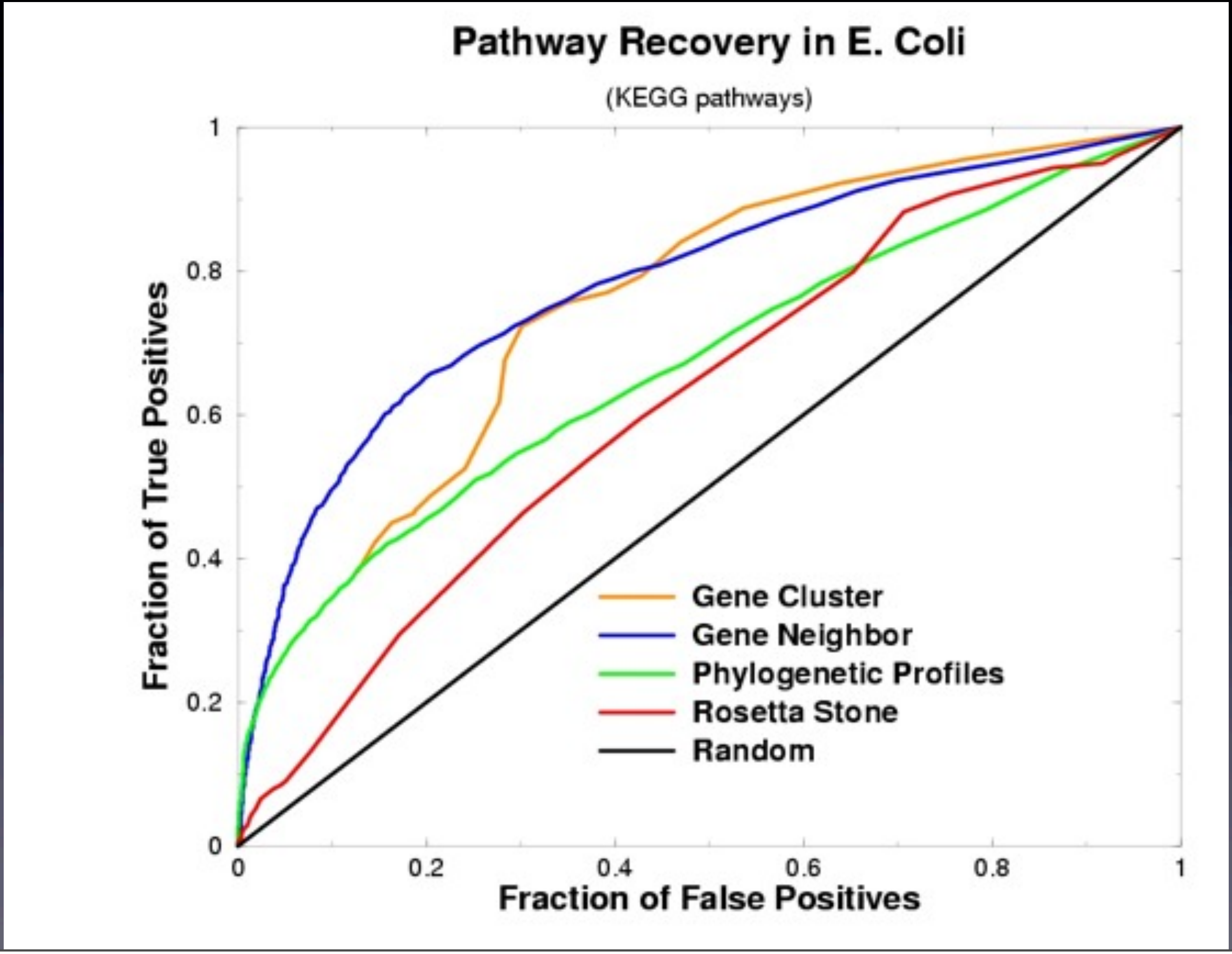
Benchmarking using Receiver Operator curves

- Determine whether each pair is a TP or FP association (based on pathways)
- 101 234   P = .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 = .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

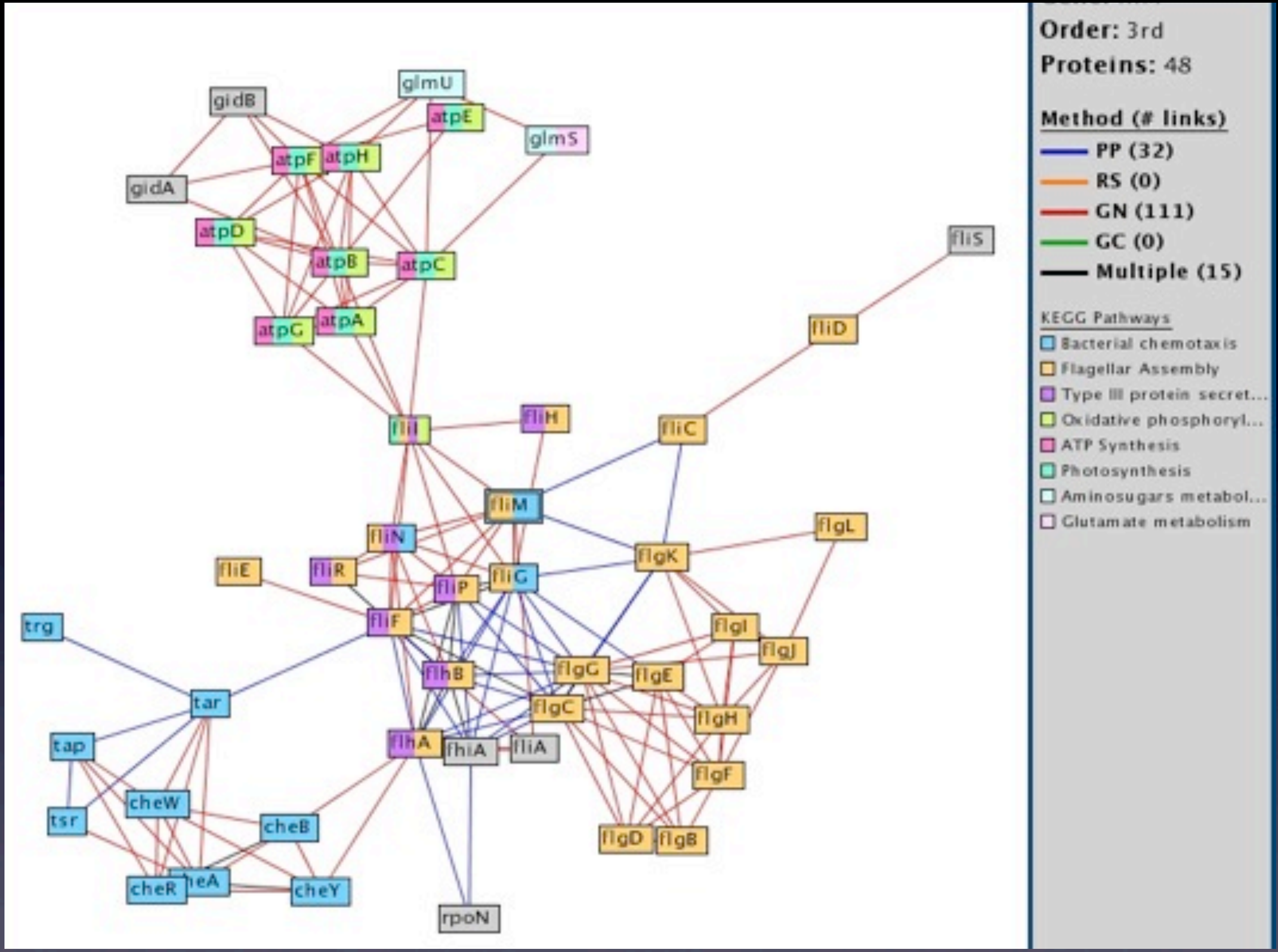


TP = same pathway
FP = different pathways

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 secret...

□ Oxidative phosphoryl...

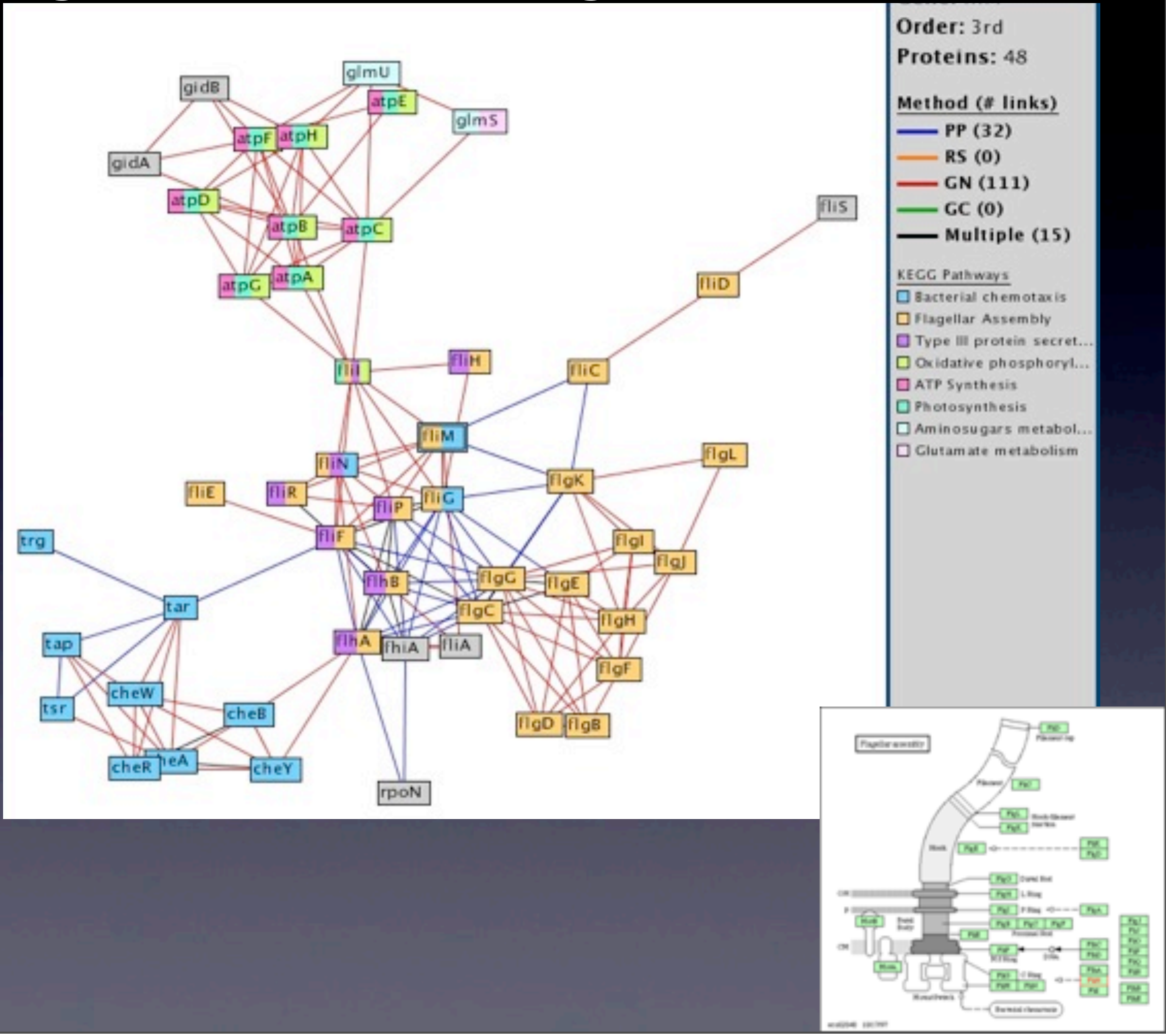
□ ATP Synthesis

□ Photosynthesis

□ Aminosugars metabol...

□ Glutamate metabolism

Bacterial Flagella Network Using Combined Methods



Order: 3rd

Proteins: 48

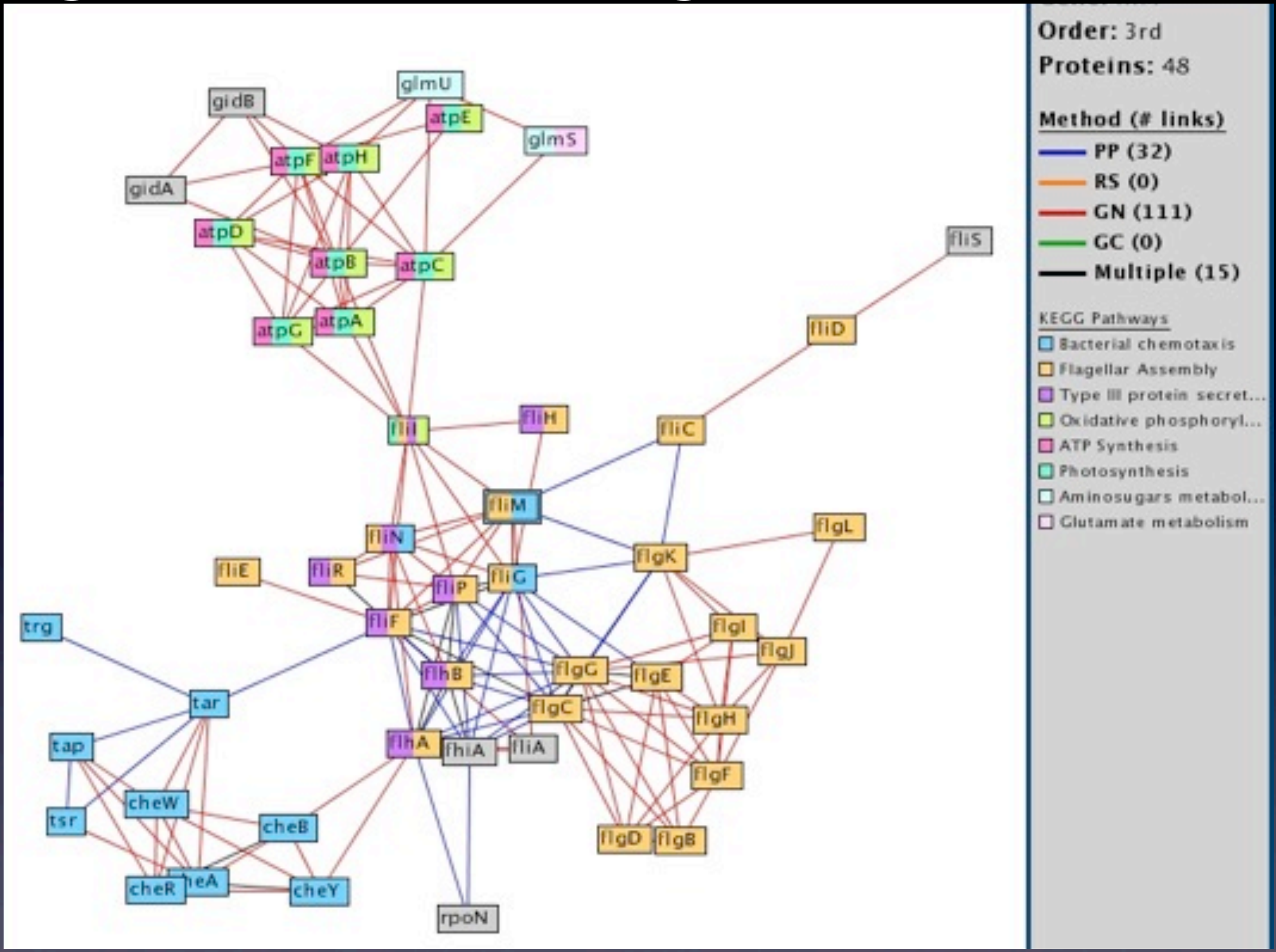
Method (# links)

- — 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

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...

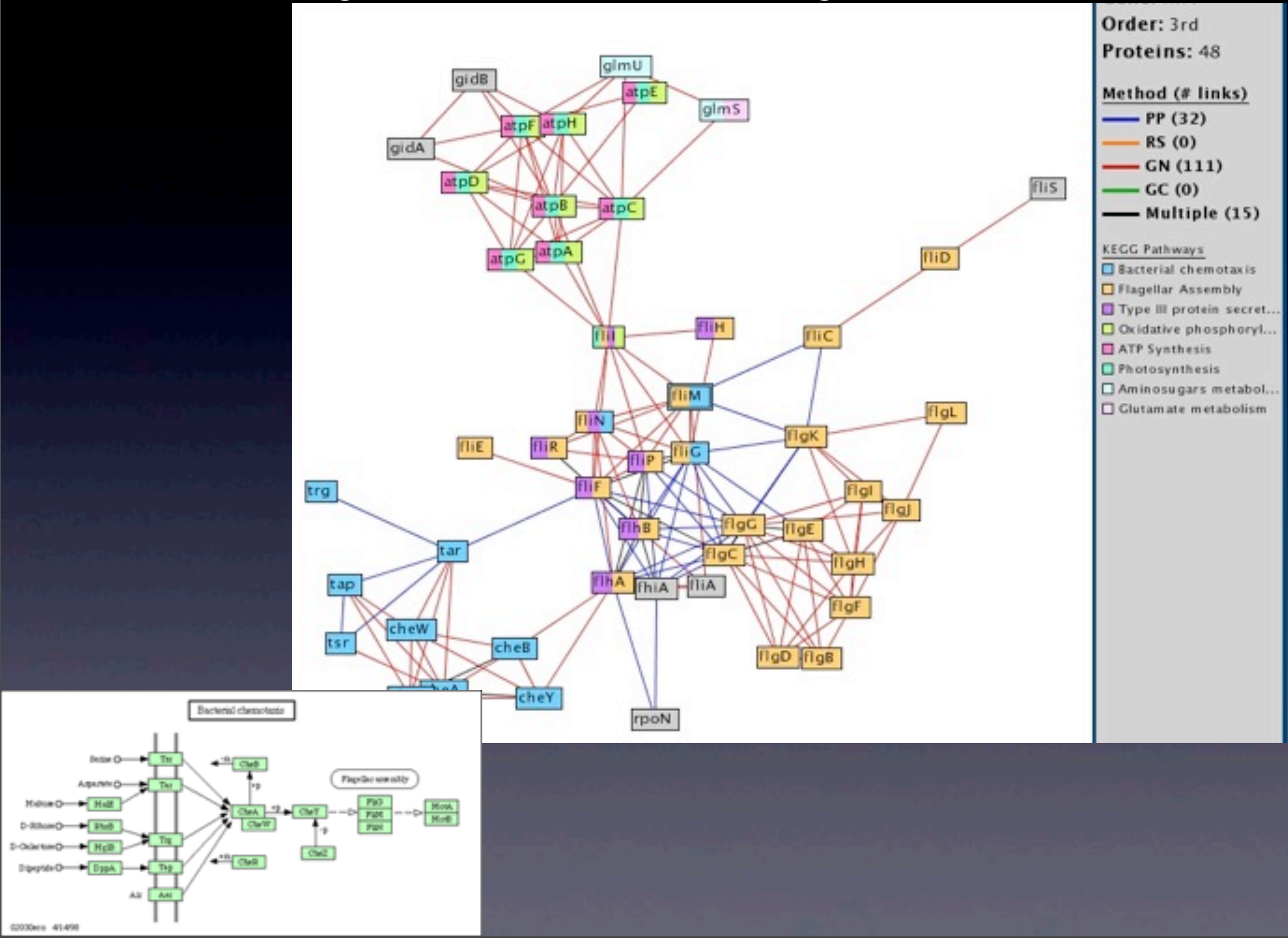
□ ATP Synthesis

□ Photosynthesis

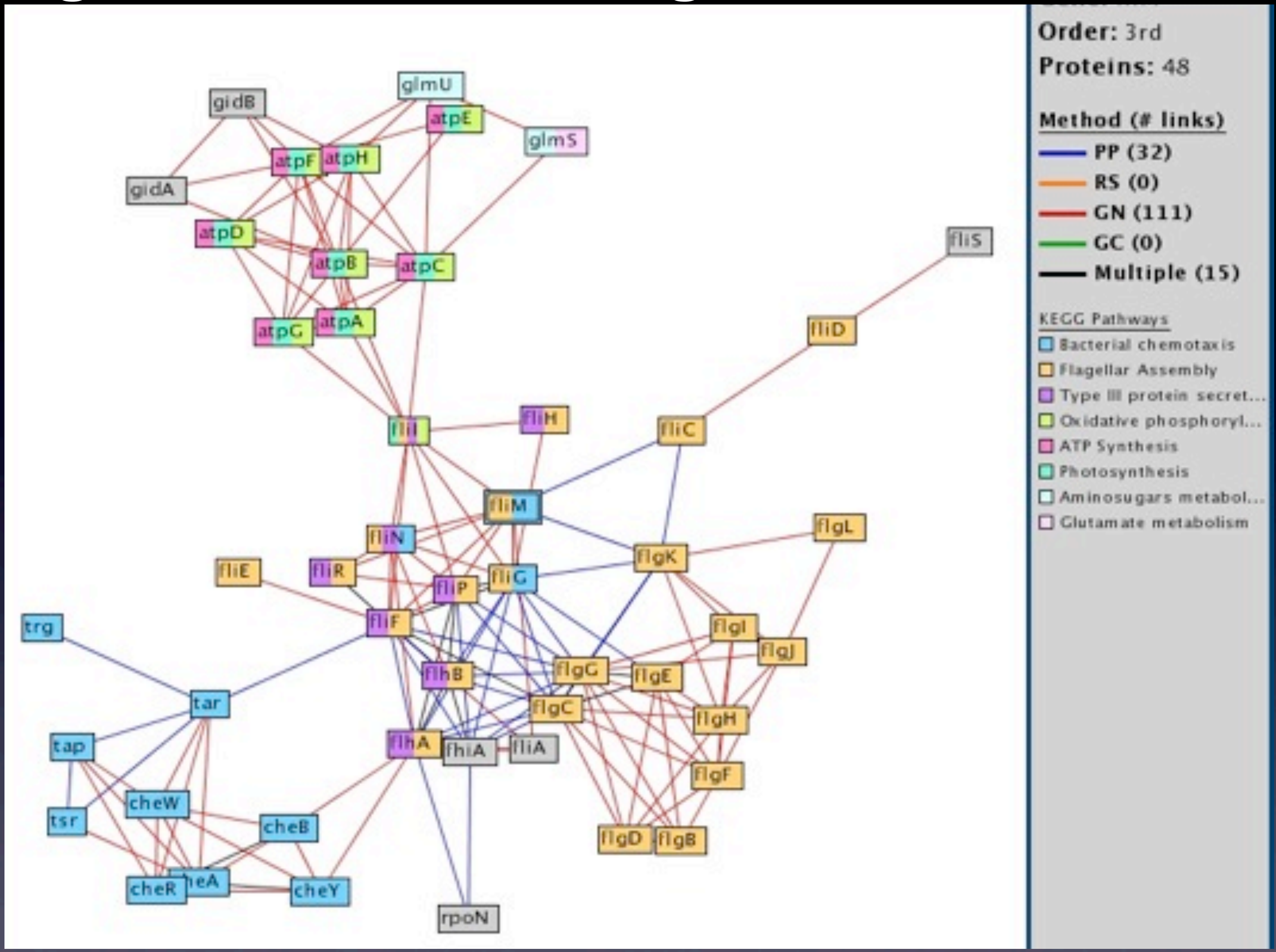
□ Aminosugars metabol...

□ Glutamate metabolism

Bacterial Flagella Network Using Combined Methods



Bacterial Flagella Network Using Combined Methods



Order: 3rd

Proteins: 48

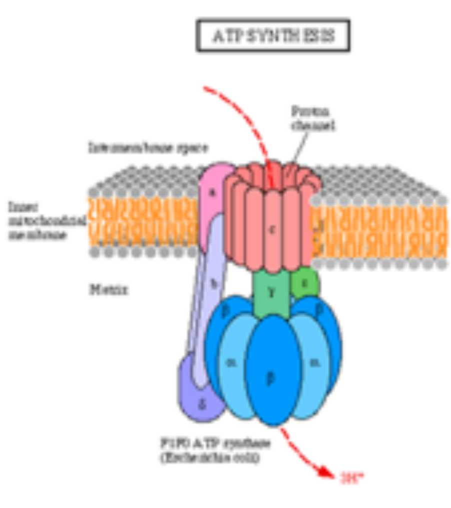
Method (# links)

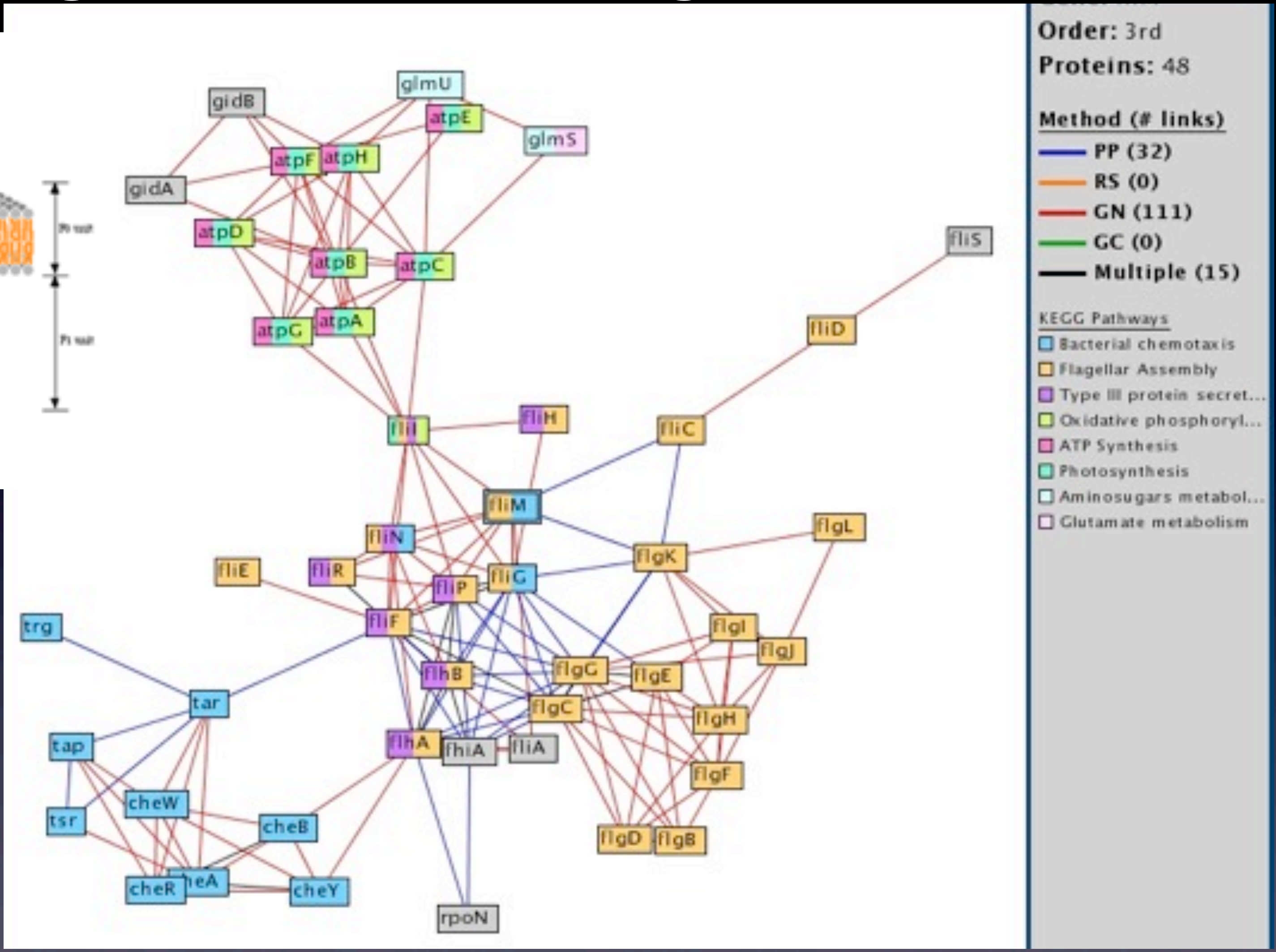
- **PP (32)**
- **RS (0)**
- **GN (111)**
- **GC (0)**
- **Multiple (15)**

KEGG Pathways

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

Bacterial Flagella Network Using Combined Methods





Order: 3rd

Proteins: 48

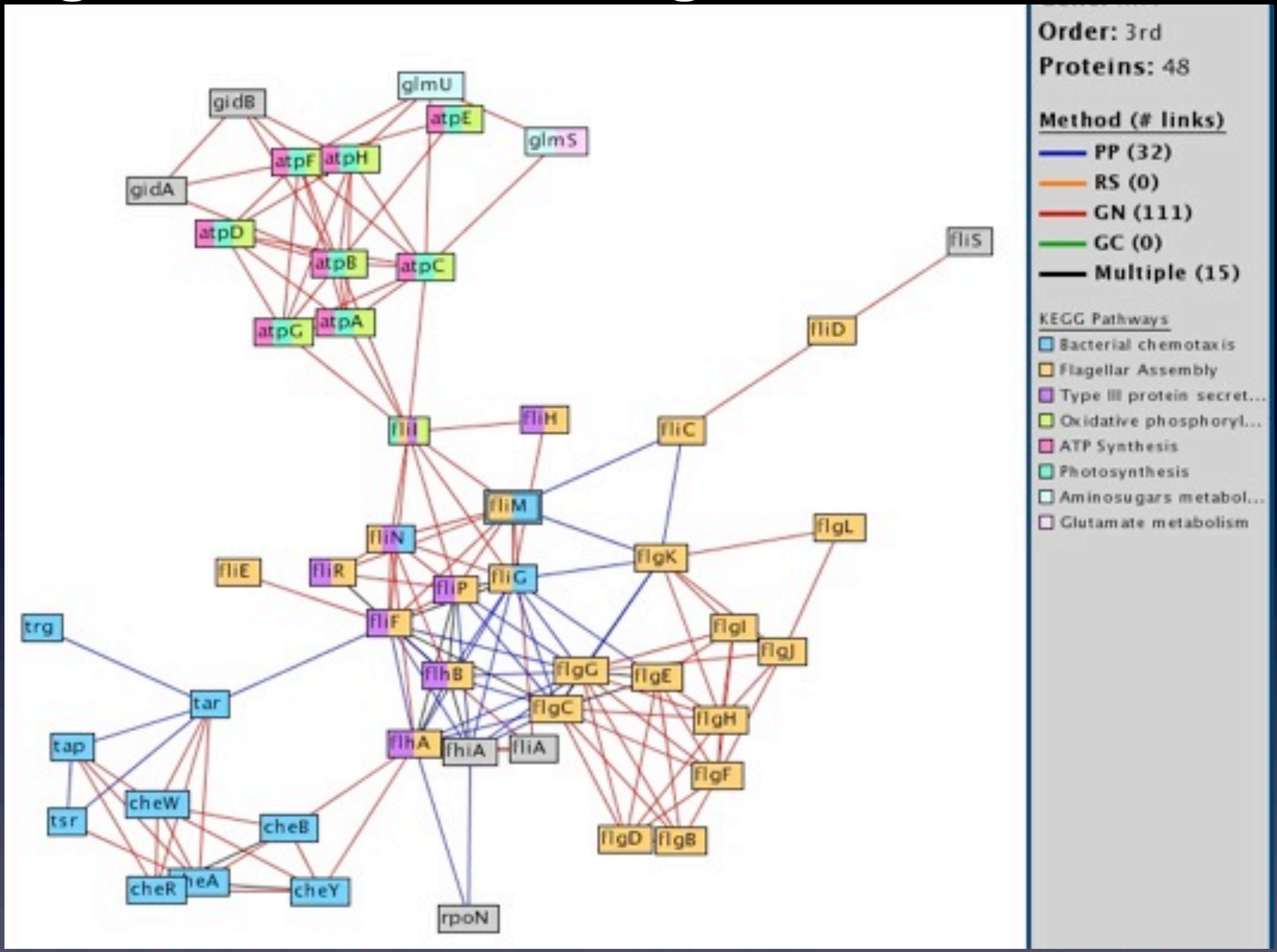
Method (# links)

- 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

Bacterial Flagella Network Using Combined Methods



Order: 3rd

Proteins: 48

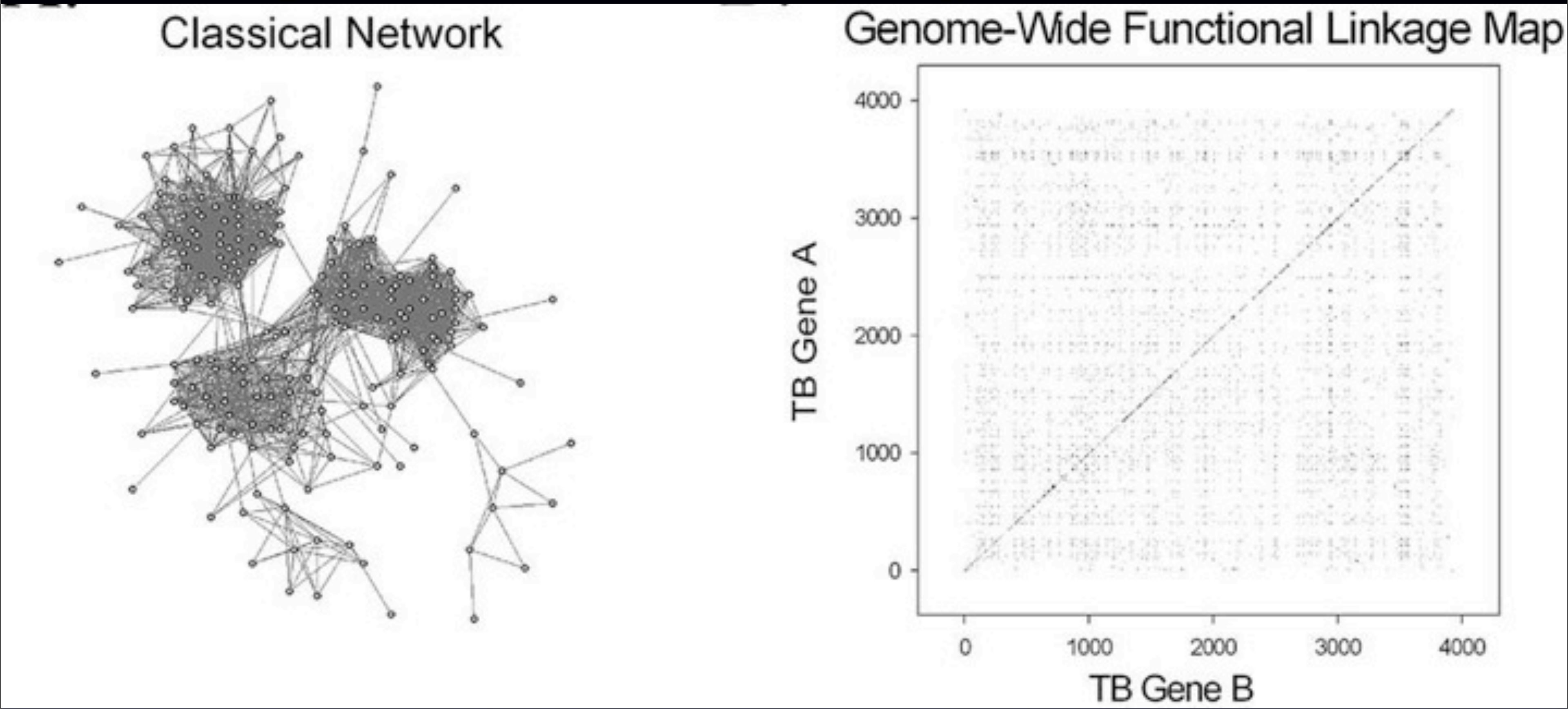
Method (# links)

- — PP (32)
- — RS (0)
- — GN (111)
- — GC (0)
- — Multiple (15)

KEGG Pathways

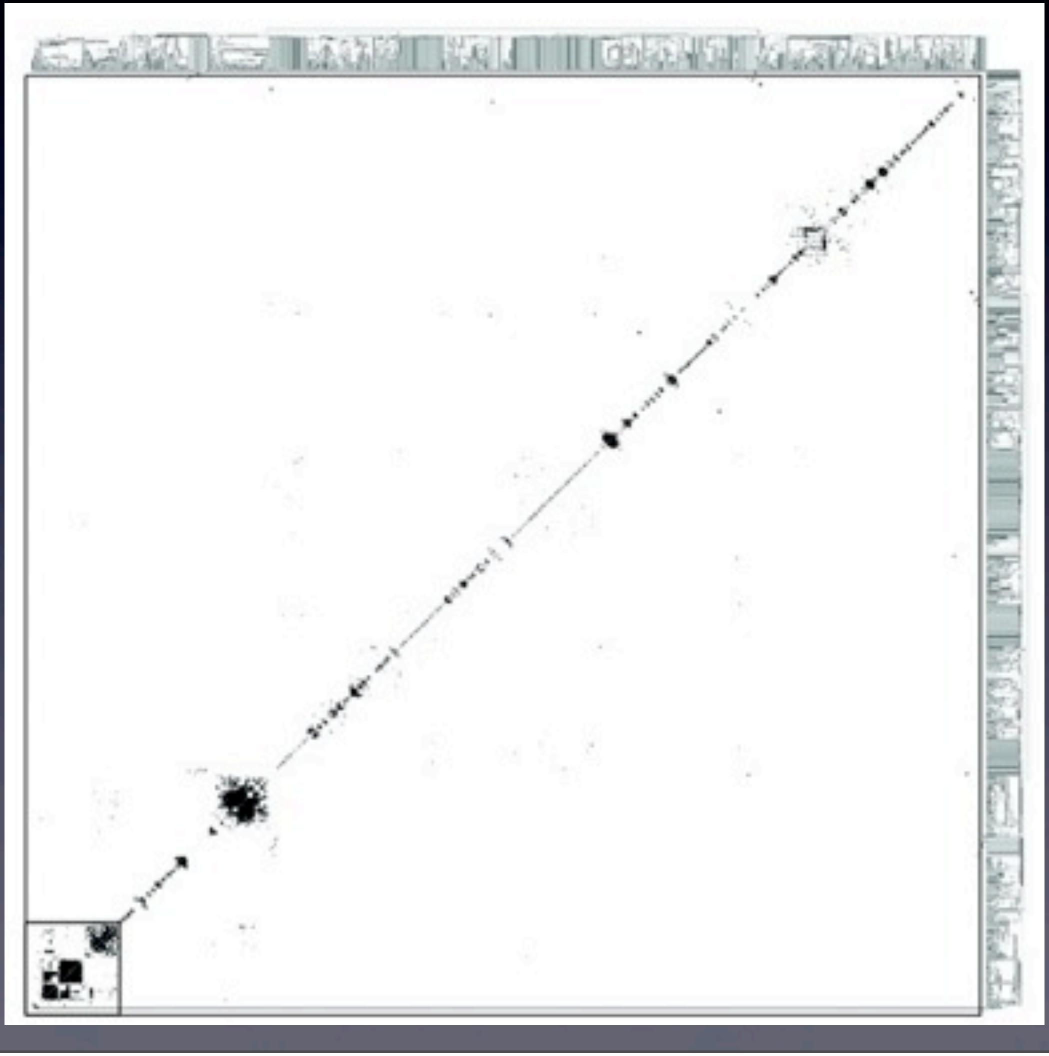
- □ Bacterial chemotaxis
- □ Flagellar Assembly
- □ Type III protein secret...
- □ Oxidative phosphoryl...
- □ ATP Synthesis
- □ Photosynthesis
- □ Aminosugars metabol...
- □ Glutamate metabolism

Alternative Representations of Network

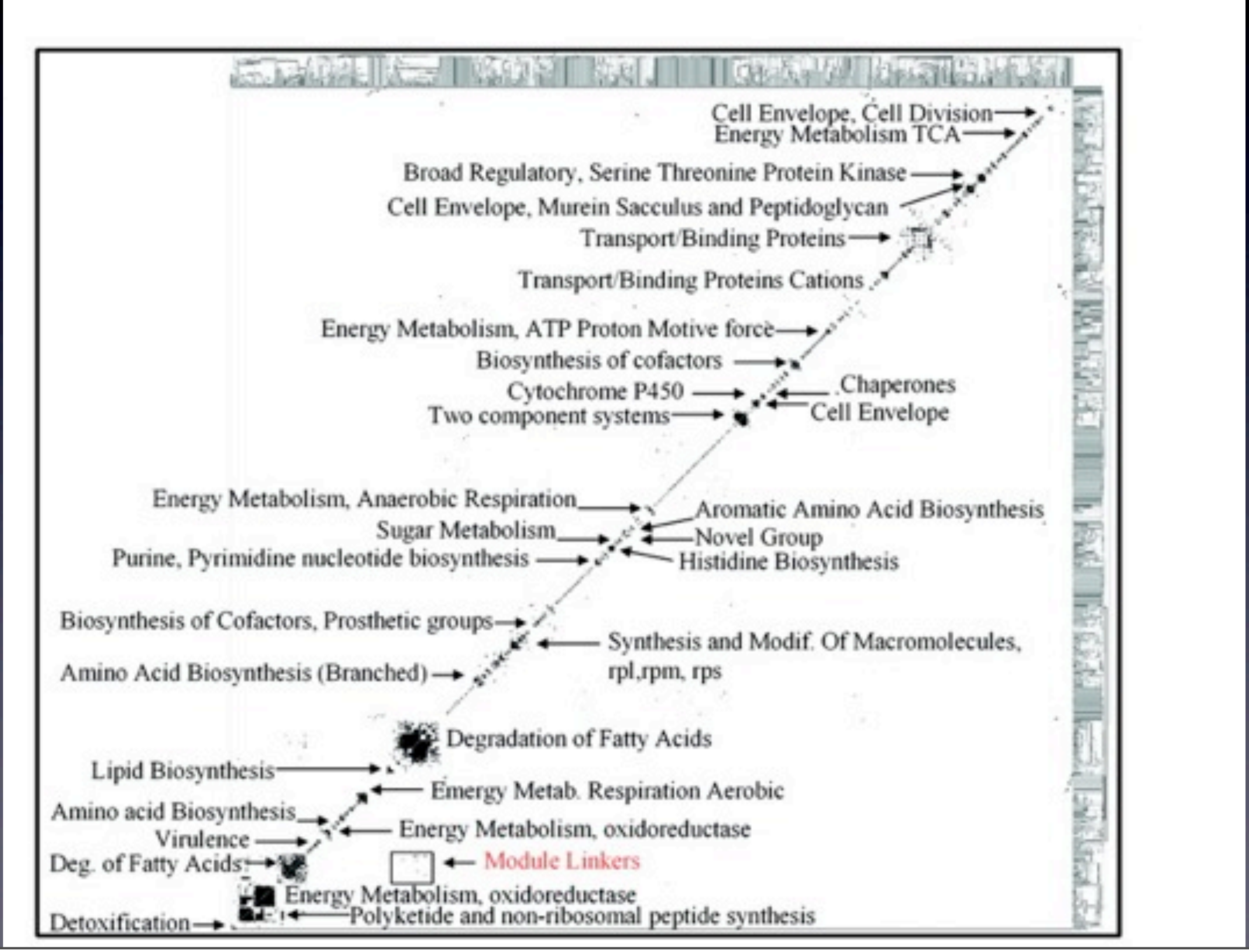


Strong M, Graeber TG, Beeby M, Pelligrini M,Thompson MJ,Yeates TO, Eisenberg D. Inference and Visualization of Protein Networks in *Mycobacterium tuberculosis* Based on Hierarchical Clustering of Whole Genome Functional Linkage Maps. Submitted to Nucleic Acids Research

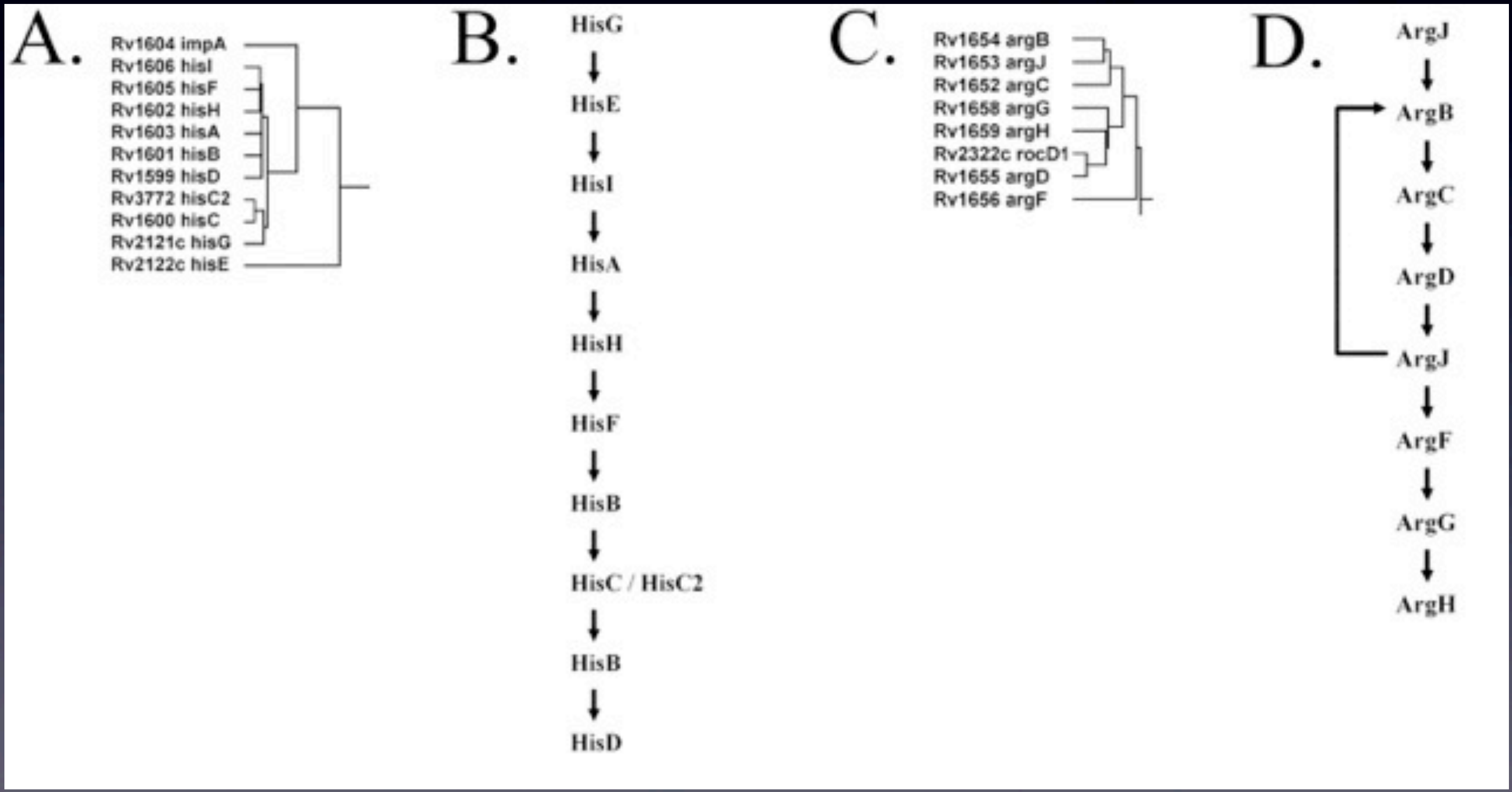
Hierarchical Clustering Reveals Modular Evolution



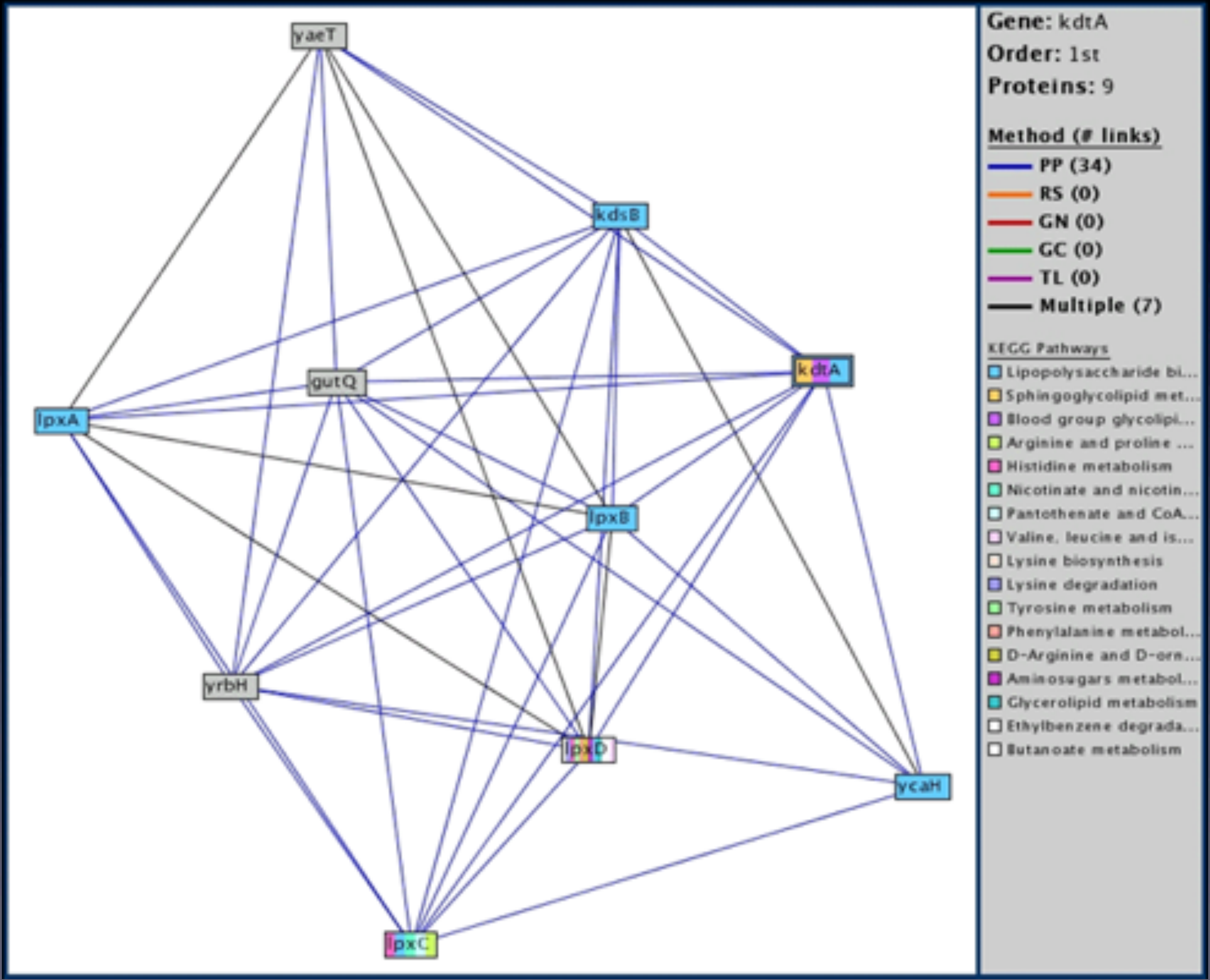
Clusters are Enriched for Pathways and Complexes



Examples of Clusters that Contain Components of Biochemical Pathways



Cluster Reveals Additional ORFs Involved in Lipopolysaccharide Biosynthesis



Gene: kdtA

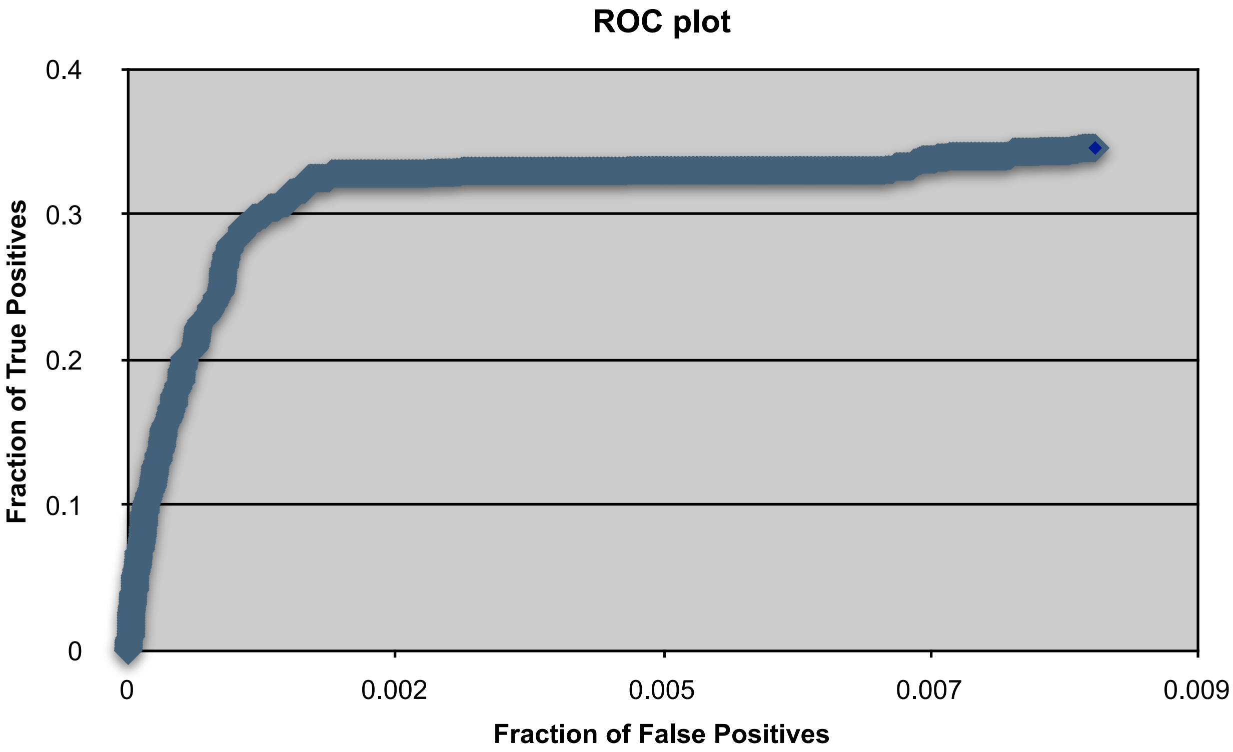
Order: 1st

Proteins: 9

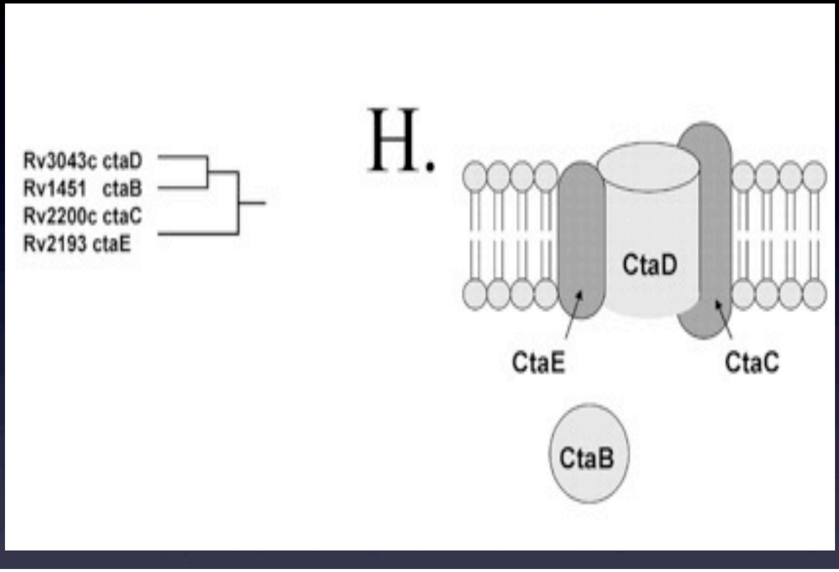
Clusters are also Enriched for Subunits of Protein Complexes

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

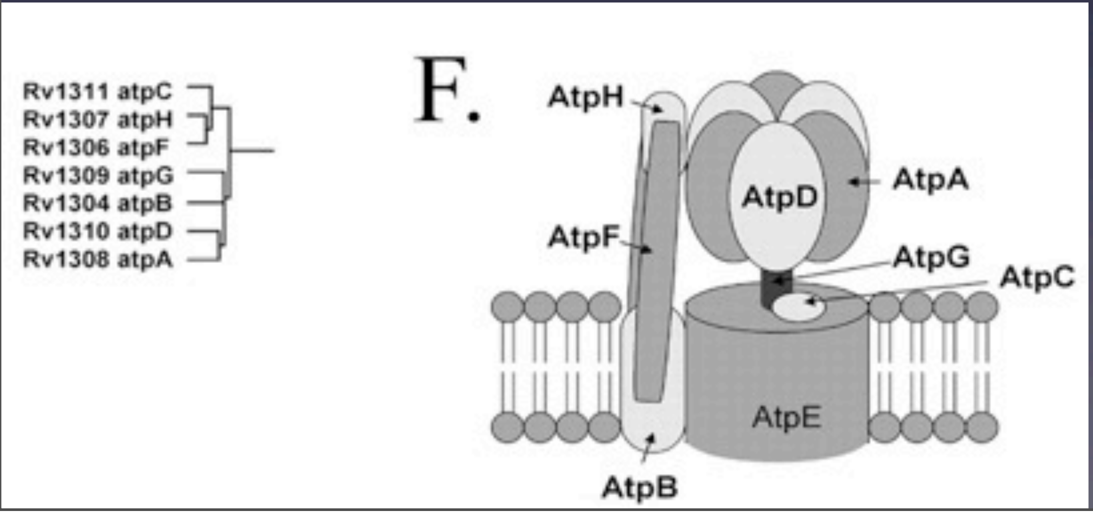
For high confidence links, we recover one third of true interactions and only one thousandth of the false positive ones



Clusters Containing Subunits of Protein Complexes

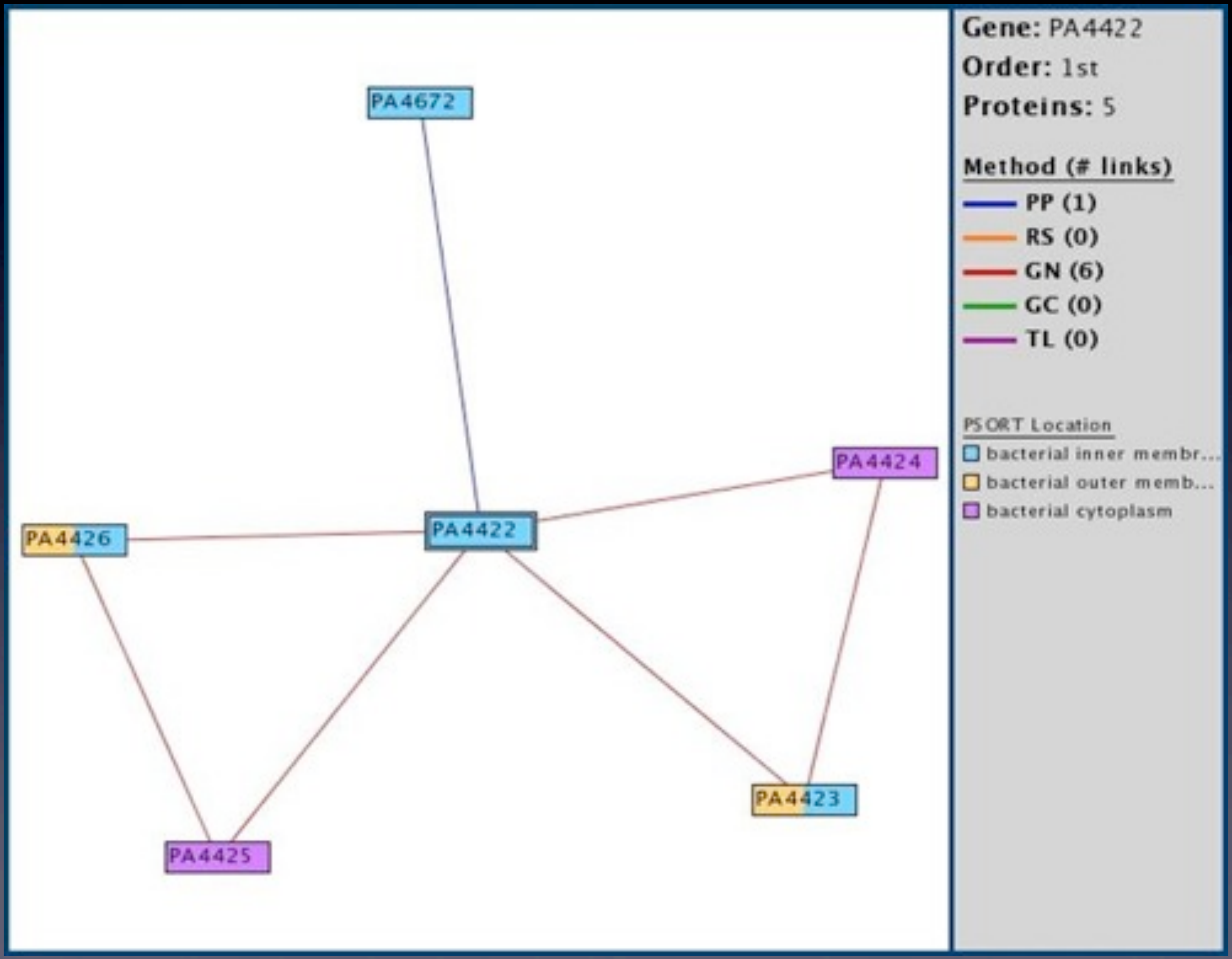


Cytochrome c oxidase controls the last step of food oxidation



ATP Synthase

Identification of an Uncharacterized Protein Complex



Gene: PA4422

Order: 1st

Proteins: 5

Method (# links)

- PP (1)
- RS (0)
- GN (6)
- GC (0)
- TL (0)

PSORT Location

- □ bacterial inner membr...
- □ bacterial outer memb...
- □ bacterial cytoplasm

Conclusions

- Protein modules appear to co-evolve across bacterial species

- Modules are enriched for proteins that participate in the same pathway or complex

PROLINKS Database

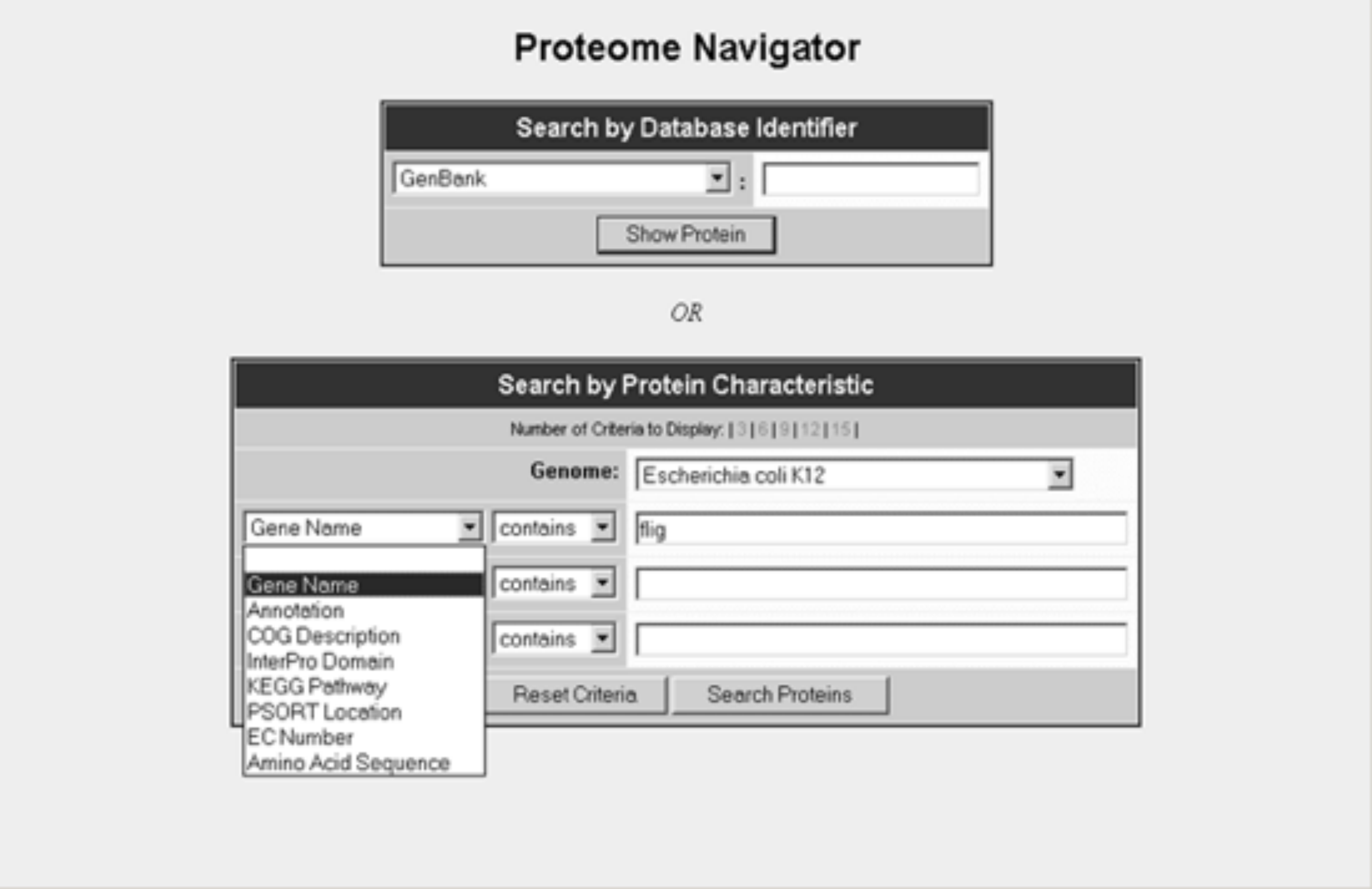
We have constructed a database that contains co-evolution links between the genes of 150 fully sequenced genomes

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

prolinks.mbi.ucla.edu/

Peter M Bowers, Matteo Pellegrini, Mike J. Thompson, Joe Fierro, Todd O. Yeates, David Eisenberg. PROLINKS: A Database of Protein Functional Linkages Derived from Co-evolution, Genome Biology, in press

Proteome Navigator Access Page



Proteome Navigator

Search by Database Identifier

GenBank :

OR

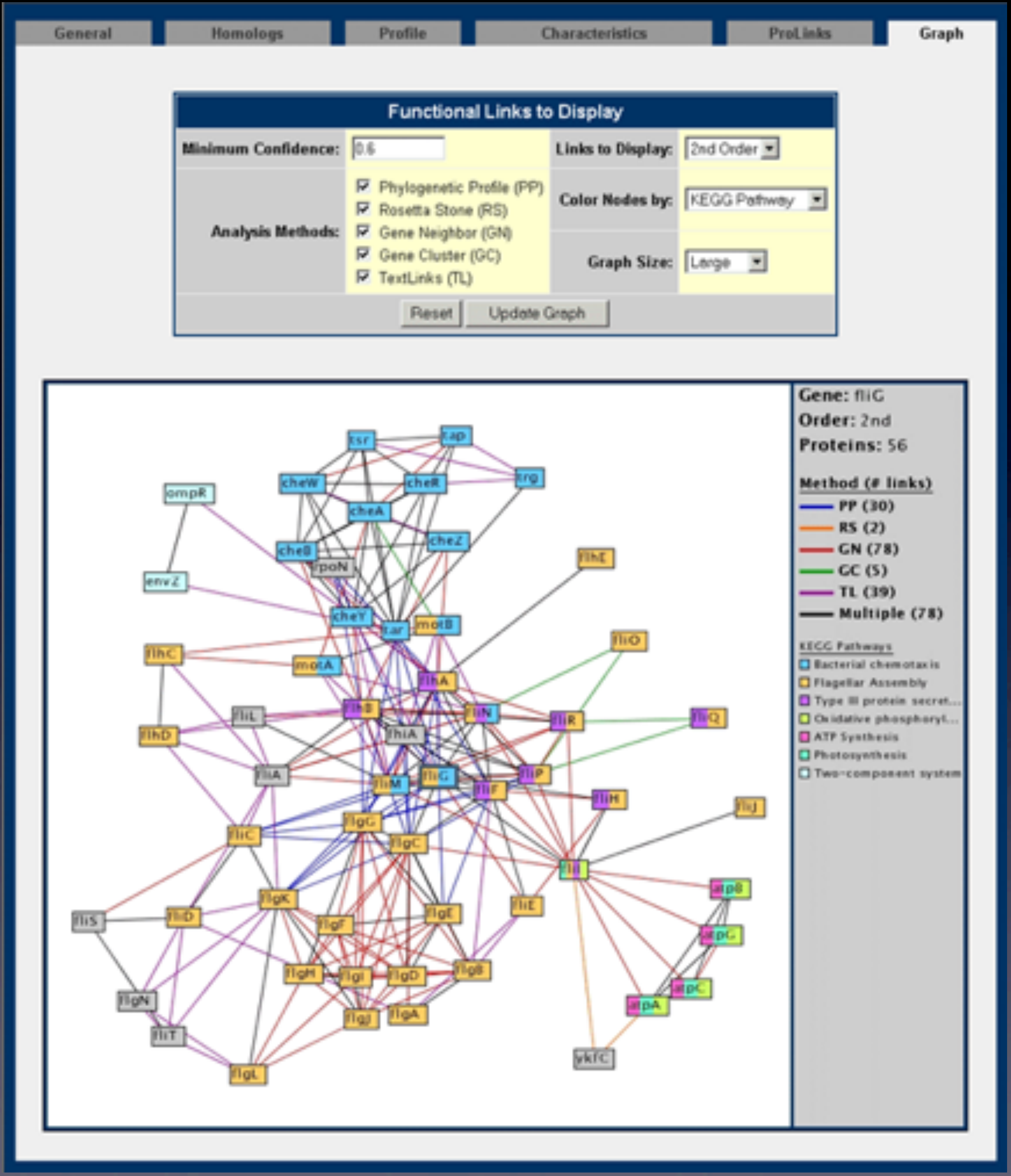
Search by Protein Characteristic

Number of Criteria to Display: | 3 | 6 | 9 | 12 | 15 |

Genome: Escherichia coli K12

Gene Name contains flig

Proteome Navigator



Functional Links to Display

Minimum Confidence: 0.6

Links to Display: 2nd Order

Analysis Methods:
- ☑ Phylogenetic Profile (PP)
- ☑ Rosetta Stone (RS)
- ☑ Gene Neighbor (GN)
- ☑ Gene Cluster (GC)
- ☑ TextLinks (TL)

Color Nodes by: KEGG Pathway

Graph Size: Large

Gene: fliG
Order: 2nd
Proteins: 56

Method (# links)
- — PP (30)
- — RS (2)
- — GN (78)
- — GC (5)
- — TL (39)
- — Multiple (78)

KEGG Pathways
- □ Bacterial chemotaxis
- □ Flagellar Assembly
- □ Type III protein secret...
- □ Oxidative phosphoryl...
- □ ATP Synthesis
- □ Photosynthesis
- □ Two-component system