



Predicting protein linkages in bacteria: which method is best depends on task

Anis Karimpour-Fard¹, Sonia M. Leach¹, Ryan T. Gill², and Lawrence Hunter¹

¹ University of Colorado School of Medicine

² Department of Chemical and Biological Engineering, University of Colorado, Boulder

anis.karimpour-fard@ucdenver.edu

<http://compbio.ucdenver.edu/Hunter>

Dec 11, 2009

The problem

More than 1000 Microbial genomes are fully sequenced and the key limitations are:

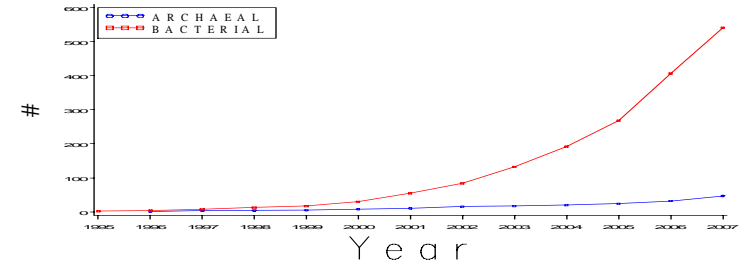
- Understanding the functions of genes

For example:

E. coli K12

KEGG	COG	TIGR
43%	38%	42%

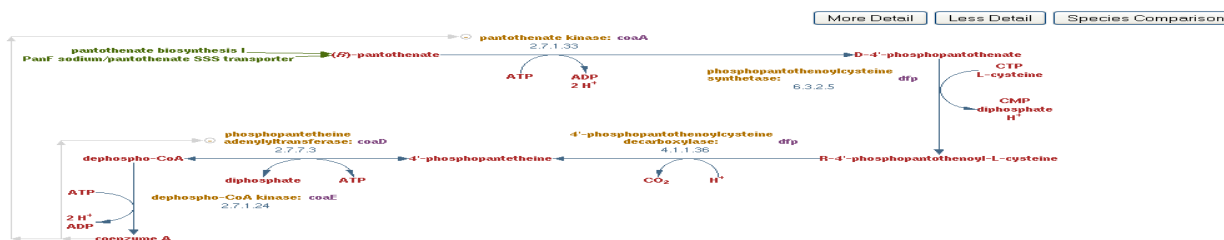
- No significant homology to another protein with known function
- No prediction based upon domains



<http://www.genomesonline.org/>

- Identifying groups of genes that contribute to common phenomena
- Other genes involved in a pathway?

Escherichia coli K-12 substr. MG1655 Pathway: coenzyme A biosynthesis

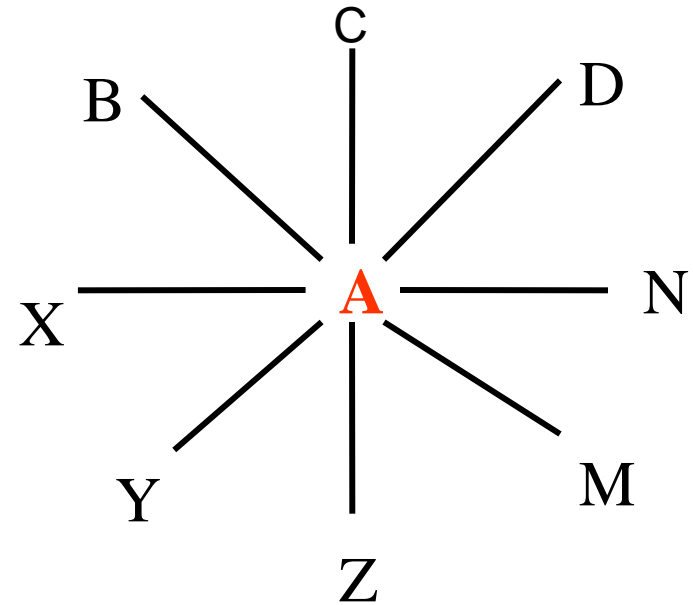


The meaning of protein function



The function of protein **A** is its action on **Substrate** to form a **Product**

Biochemical view



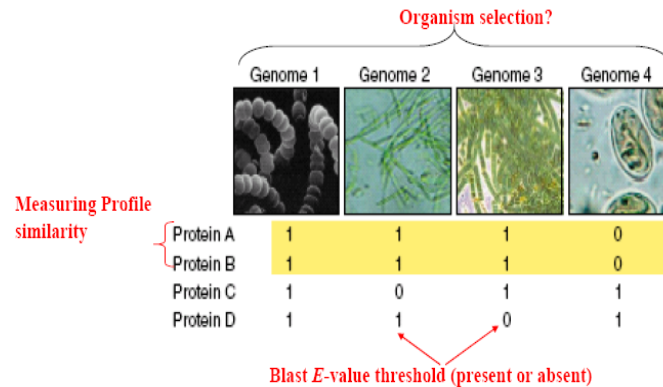
The function of **A** is the context of its interactions with other proteins in the cell

Post genomic view

Eisenberg, D. et. al. Nature 2000

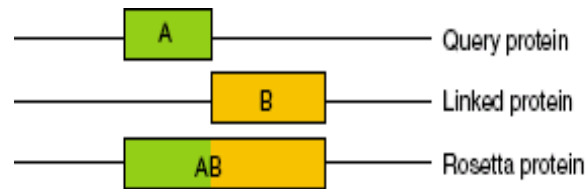
Computation method prediction from genomic context

Phylogenetic profile



Pairs of proteins that are present or absent together in genomes.

Rosetta Stone



Proteins that are separate in one organism but are fused into one protein in another organism.

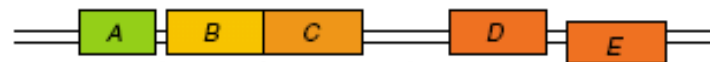
- Protein A and B are nonhomologous
- Protein AB align at least over 70% of different portion of another protein in another organisms.

Gene neighbor



Pairs of genes that are coded nearby in multiple organisms.

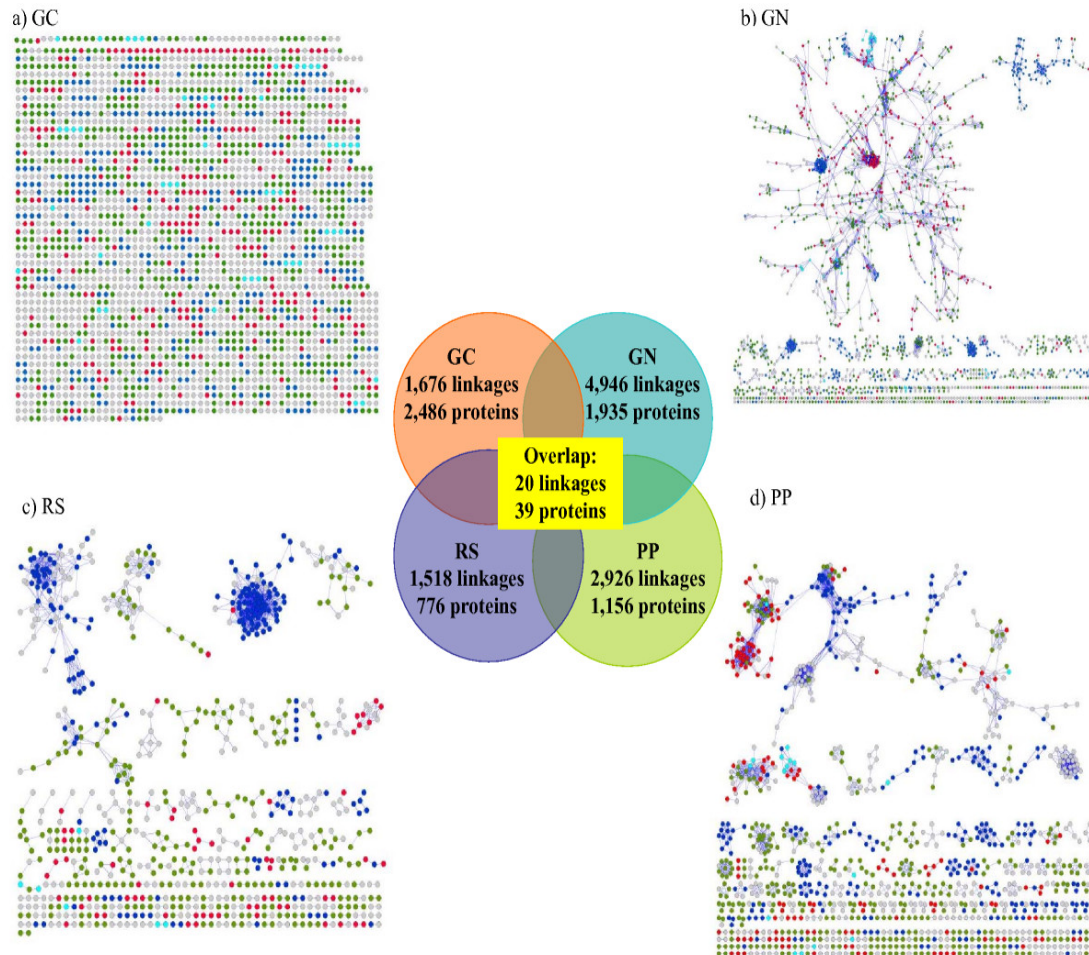
Gene cluster



Gene proximity within genome.

- Adjacent
- On the same strand

Complete protein-protein networks of *E. coli* K12



How to validate the interactions?

- **Gold standard**

- EcoCyc (curated database)
- RegulonDB (curated database)
- TIGR

- 18 role category

- KEGG functional category and subcategory**

- **Metabolism (11 subcategory)**

- **Genetic information processing (2 subcategory)**

- **Environmental information processing (4 subcategory)**

- **Cellular processes (2 subcategory)**

- NCBI COG functional category**

- **17 functional category**

- **The choice of gold standard impacted the outcome**

- Incomplete**
- Biased**

- Little overlap of linkages and proteins (link confidence > 0.6)
- Predictions made for **78% of unclassified genes in KEGG** (grey nodes)

Which method is most appropriate for a given prediction task?

- Can we predict protein function? **Yes**

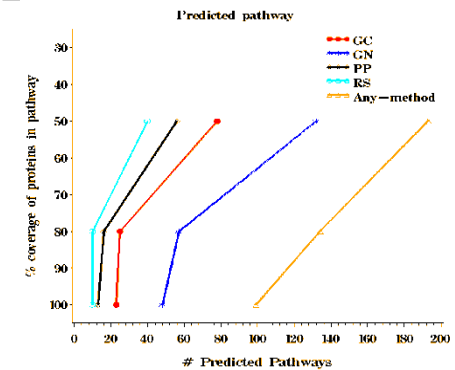
- Which method?

- **Phylogenetic profile for function prediction**

- **Rosetta Stone for prediction of transporter**

- Can we predict pathway? **Yes**

- Which method? **Gene neighbor**

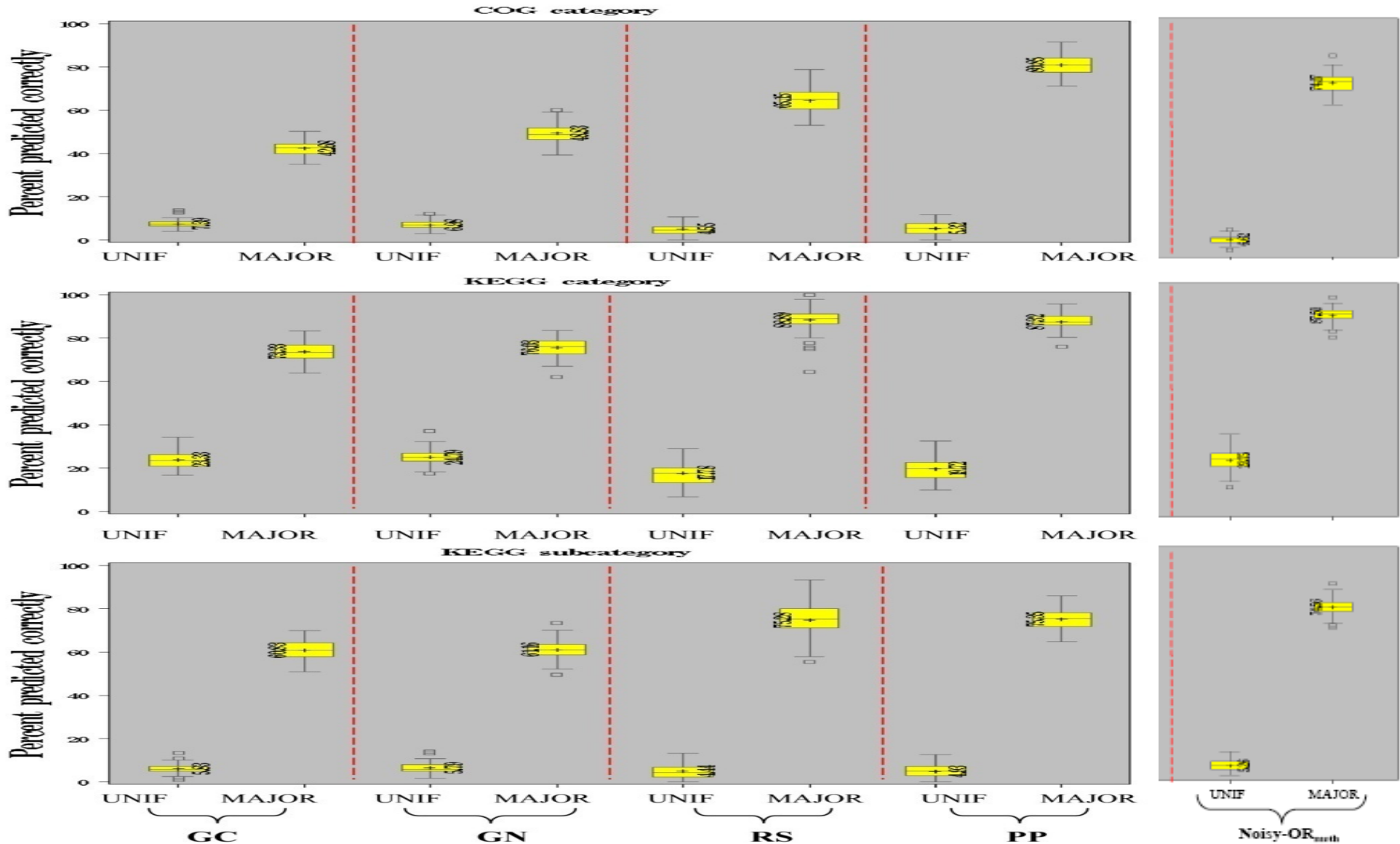


- Can we predict operons (a group of genes that are co-expressed by one promoter)? **Yes**

- Which method? **Gene cluster**

- Can we combine different methods to have a higher coverage by taking into account different features of each method? **Yes**

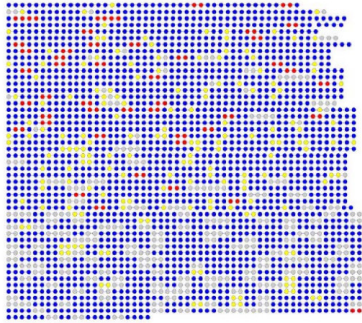
Function prediction cross-evaluation of protein linkages



- Cross validation : (10% test, 90% train, 100 runs)
- UNIF: function randomly chosen among those in cluster
- MAJOR: function is majority among immediate neighbors

Incomplete gold standard

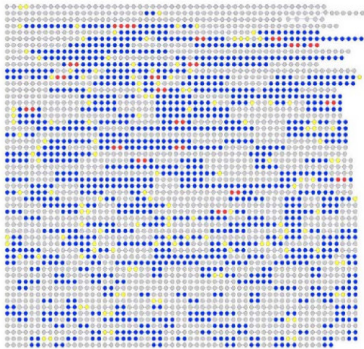
a) *E. coli* K12 (RegulonDB Jun 2007)



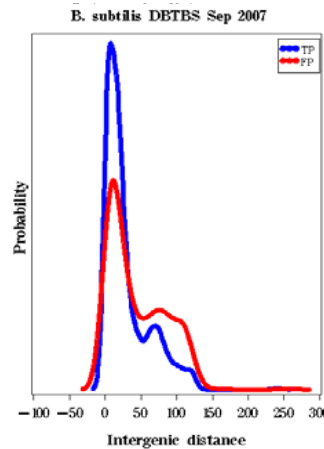
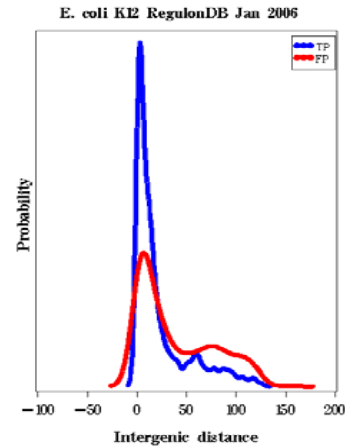
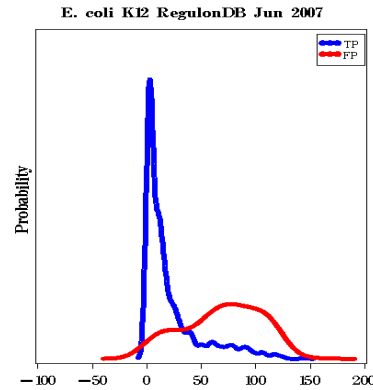
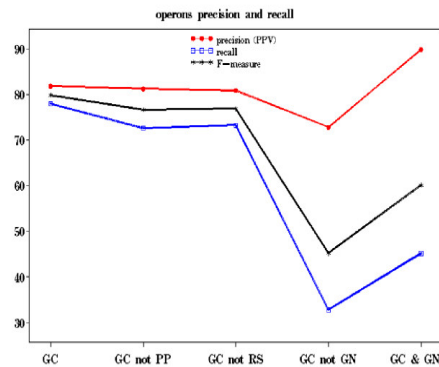
b) *E. coli* K12 (RegulonDB Jan 2006)



c) *B. subtilis* (DBTBS Sep 2007)



d) Predictive Value of Combinations in *E. coli* K12



- **Blue**: both proteins in same operon
- **Red**: both protein in operon, but not same operon
- **Yellow**: only one protein known to be in an operon
- **Gray**: neither protein classified as an operon protein

GC predictions, for most part, confirmed by database improvements

There is no well-curated database available to do a complete evaluation