



The topology of the bacterial co-conserved protein network and its implications for predicting protein function

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

Dec 7, 2008

anis.karimpour-fard@uchsc.edu

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

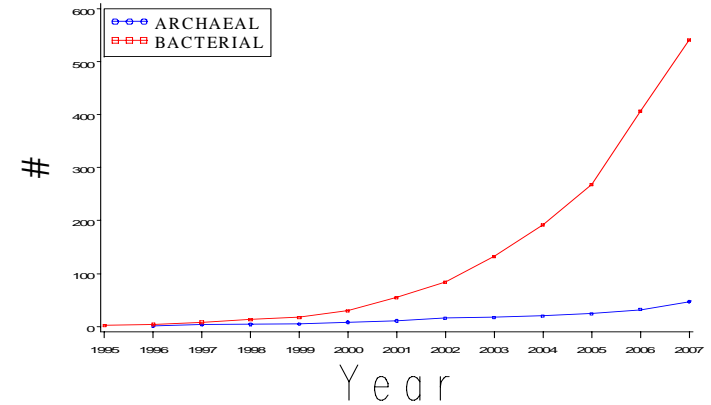
The problem

More than 600 Microbial genomes are fully sequenced and the key challenges are:

- Understanding protein function

For example: *E. coli* K12

KEGG 43%	COG 38%	TIGR 42%
--------------------	-------------------	--------------------



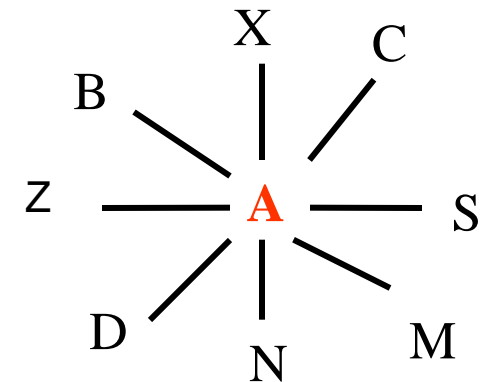
<http://www.genomesonline.org/>

- Identifying proteins that contribute to common phenomena

The meaning of protein function



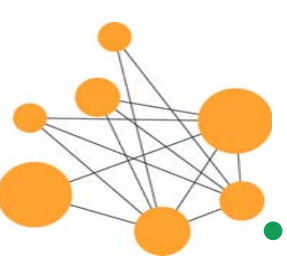
Biochemical view



Post genomic view

Eisenberg, D. et. al. Nature 2000

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



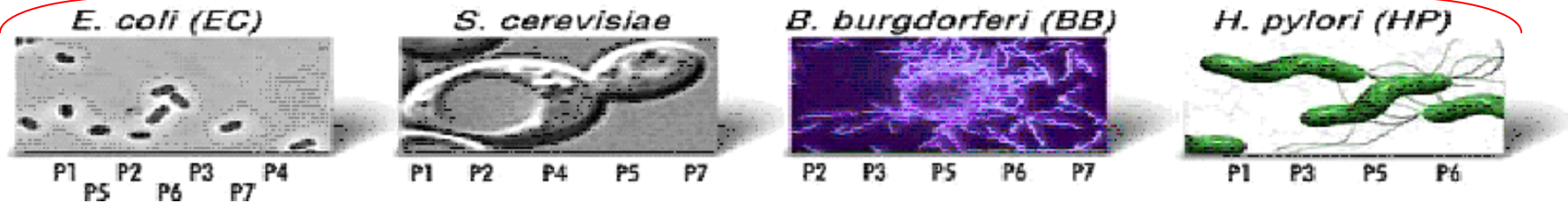
Prediction protein function

- Prediction by homology based methods (gives partial understanding about protein role)
 - *Simple sequence similarity searches (BLAST)*
 - *Profile searches (PSI-BLAST)*
 - *Databases of conserved domains (Pfam, SMART)*
- Prediction from high-throughput experimental data
 - *Microarray gene expression data*
 - *Protein-protein interaction screens*
 - ...
- Prediction from genomic context
 - **Phylogenetic profile**
 - Gene cluster
 - Gene neighbor
 - Rosetta Stone

Phylogenetic Profile

Select sets of genomes as a reference set

Reference selection?



Does the selection of the reference genomes influence the prediction? if so? How?

Create phylogenetic profile matrix for target organism:

- Do one-against-all BLAST search to identify all homologous target proteins in diverse reference organisms.

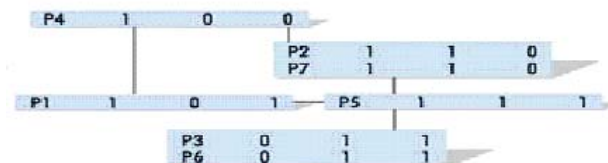
	EC	SC	BB	HP
P1	1	0	1	1
P2	1	1	0	0
P3	0	1	1	1
P4	1	0	0	0
P5	1	1	1	1
P6	0	1	1	1
P7	1	1	0	0

Measure profile similarities

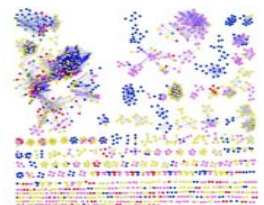
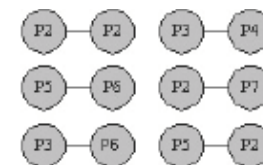
Protein X: **110001111001001110001111**

Protein Y: **111000111100000110001111**

19 matching bits out of 24



Generate Phylogenetic profile protein-protein interactions



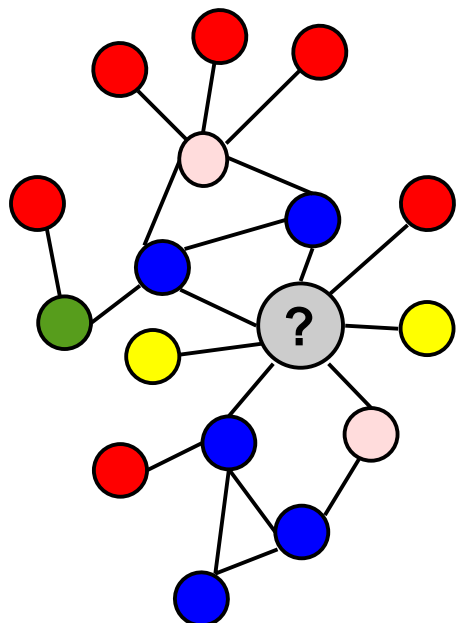


Is network topology useful for prediction?

- Can we predict function?
 - We showed that knowing neighbors helps.
 - Does knowing the local topology helps? (e.g. node degree, clustering coefficient,..) **Yes.**
 - Does knowing the global topology help? (e.g. error tolerance, diameter, average shortest path,..) **Yes.**
- Can we predict essential proteins using connectivity? **Yes.**
- Can we predict protein complexes using connectivity? **No.**

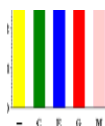


Cross validation for function prediction



1-SAMPLEUNIF = ●

6-SAMPLENEIGH = ●



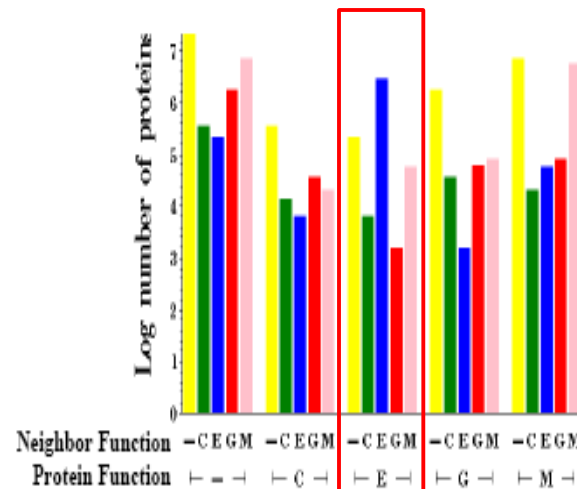
2-SAMPLEGLOBAL = ●



3-MAJORITYNEIGH = ●

4-MAJORITYCLUST = ●

5-SAMPLECONNECT = ●

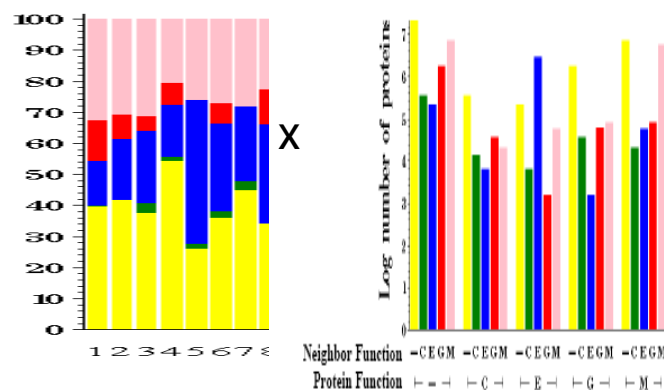
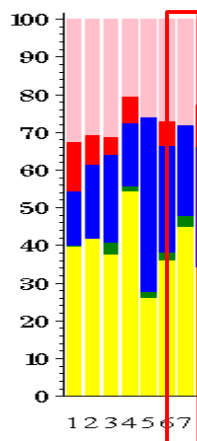


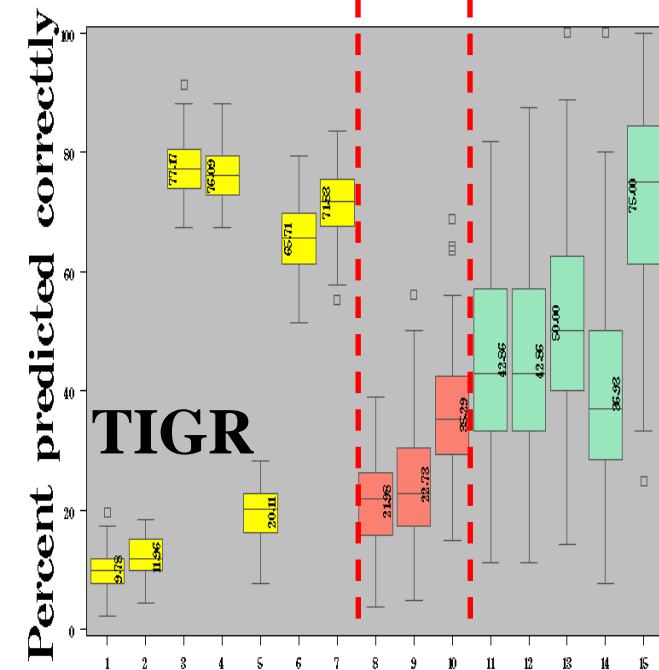
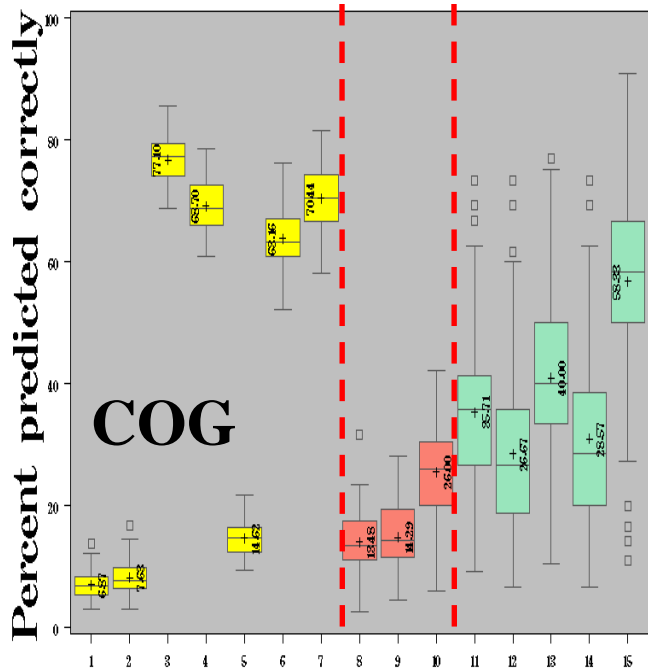
MAJORITYNEIGH

7-NEIGHCONNECT = ●

KEGG Category

- Unclassified
- Cellular Processes
- Environmental Information Processing
- Genetic Information Processing
- Metabolism



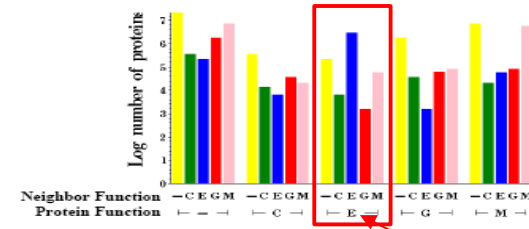
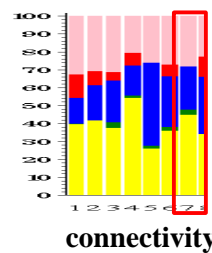


1-SAMPLEUNIF = ● 5-SAMPLECONNECT = ● 6-SAMPLENEIGH = ●

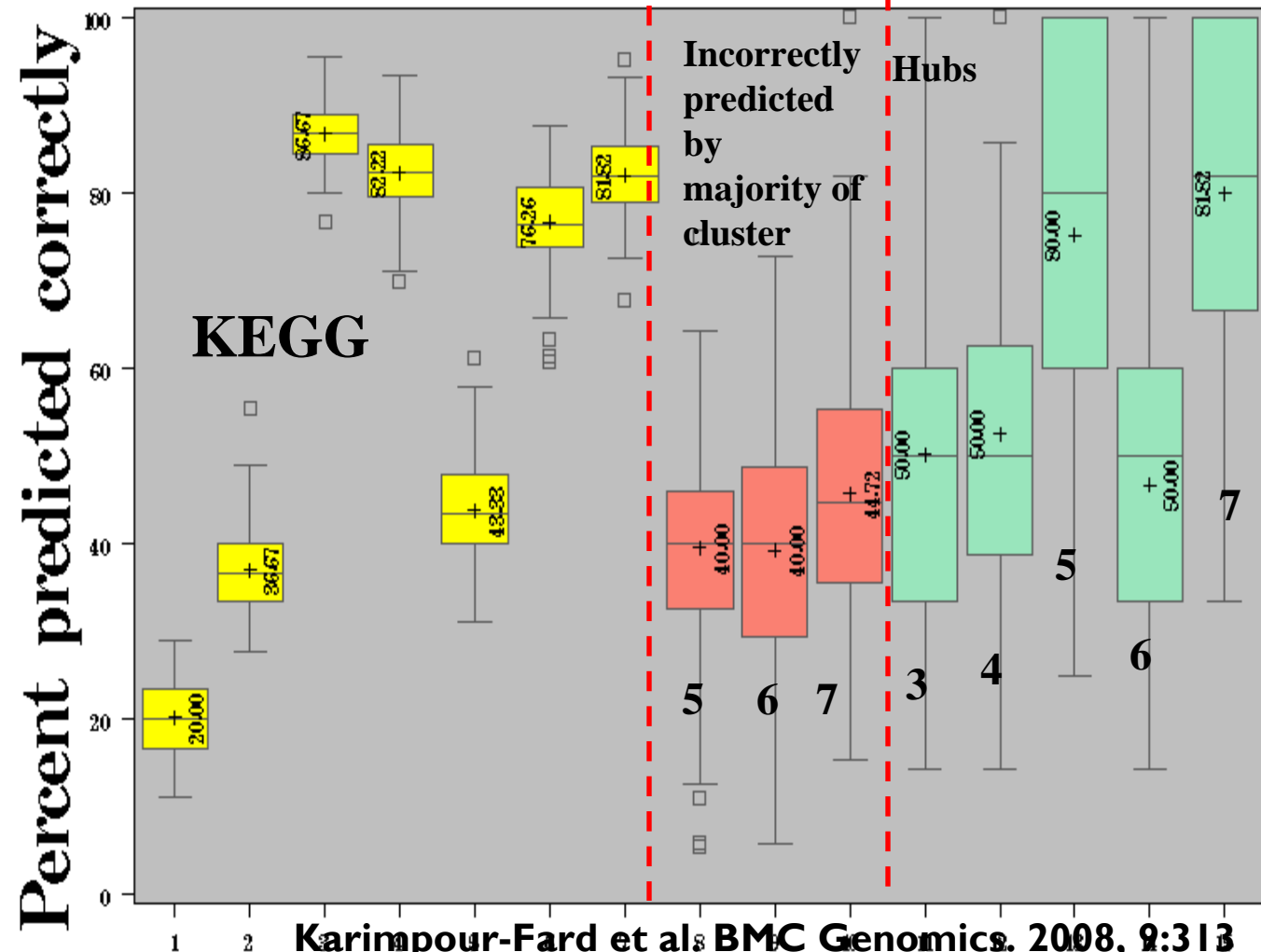
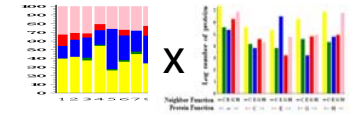
2-SAMPLEGLOBAL = ○

3-MAJORITYNEIGH = ●

4-MAJORITYCLUST = ●



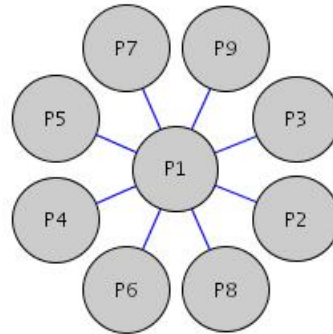
MAJORITYNEIGH





Functional classification

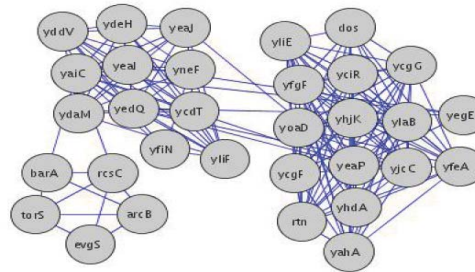
Use majority of neighbor



Unclassified

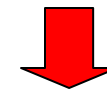
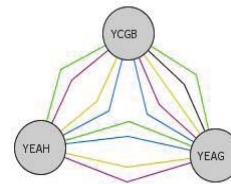
KEGG 369 proteins all neighbors unknown

Use cluster information



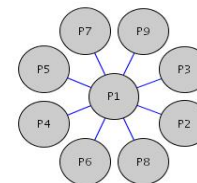
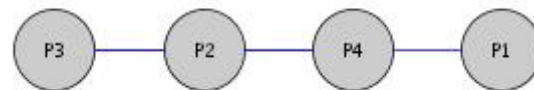
KEGG 87 clusters (271 proteins)
all proteins unknown

Cross-species



KEGG 57 clusters (159 proteins)
all proteins unknown

Use connectivity



Only 26 of those proteins have interactions in DIP database

Use combination of majority of neighbour, cluster information, cross species and connectivity to predict protein function