Bioinformatique M2: Lecture 2

P. Derreumaux

II. From genome to protein function

COGS - Clusters of orthologous groups (Koonin et al., NAR)

* All-against-all sequence comparison of the proteins encoded in completed genomes (paralogs/orthologs)

* For a given protein “a” in genome A, if there are several similar proteins in genome B, the most similar one “b” is selected

* If when using the protein “b” as a query, protein “a” in genome A is selected as the best hit “a” and “b” can be included in a COG

* Proteins in a COG are more similar to other proteins in the COG than to any other protein in the compared genomes

* A COG is defined when it includes at least three homologous proteins from three distant genomes

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 44 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain. Proteins from two eukaryotic genomes were assigned to COGs and can be reached from each individual COG page

Science 1997 Oct 24;278(5338):631-7,
Nucleic Acids Res 2001 Jan 1; 29(1):22-28.

Help

COGnitor

Protein/Gene name:

Text search:

Code	Name	Proteins	in COGs
A	*Archaeoglobus fulgidus*	2420	1872
O	*Halobacterium sp. NRC-1*	2605	1701
M	*Methanococcus jannaschii*	1786	
	Methanobacterium thermoautotrophicum	1873	
P	*Thermoplasma acidophilum*	1482	
	Thermoplasma volcanium	1499	1243
K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
Z	*Aeropyrum pernix*	1841	1178
Y	*Saccharomyces cerevisiae*	5955	2290
	Candida albicans	9168	2720
Q	*Aquifex aeolicus*	1560	1329
V	*Thermotoga maritima*	1858	1527
D	*Deinococcus radiodurans*	3187	2226
R	*Mycobacterium tuberculosis*	3927	2585
	Mycobacterium leprae	1605	1134
L	*Lactococcus lactis*	2267	1618
	Streptococcus pyogenes	1697	1211
B	*Bacillus subtilis*	4118	2870
	Bacillus halodurans	4066	2878
C	*Synechocystis*	3167	2159
	Escherichia coli K12	4275	3414
E	*Escherichia coli O157*	5315	3662

Principal component analysis of genomes

List of COGs

Distribution

Functional Categories

Phylogenetic pattern search

Functional categories

J K L

D O M N P T

G C E F H I Q

R S

Pathways and functional systems

FTP

K = transcription

T = Signal transduction mechanisms

Information in COGS

* Annotation of proteins by members of known structure/function

* Phylogenetic patterns - presence or absence of proteins in a given organism --> Enables following metabolic pathways

* Multiple alignments

Méthodes de prédiction fonctionnelle existantes II

Inférences par corrélation

✓ La variation d'organisation des gènes entre organismes

→ Méthode de la pierre de Rosette *(Marcotte et al. (1999), Science 285, 751-753)*

✓ La variation de l'ordre des gènes entre organismes

→ Méthode des gènes voisins *(Dandekar et al. (1998) TIBS 23, 324-328; Overbeek et al. (1999) PNAS 96, 2896-2901)*

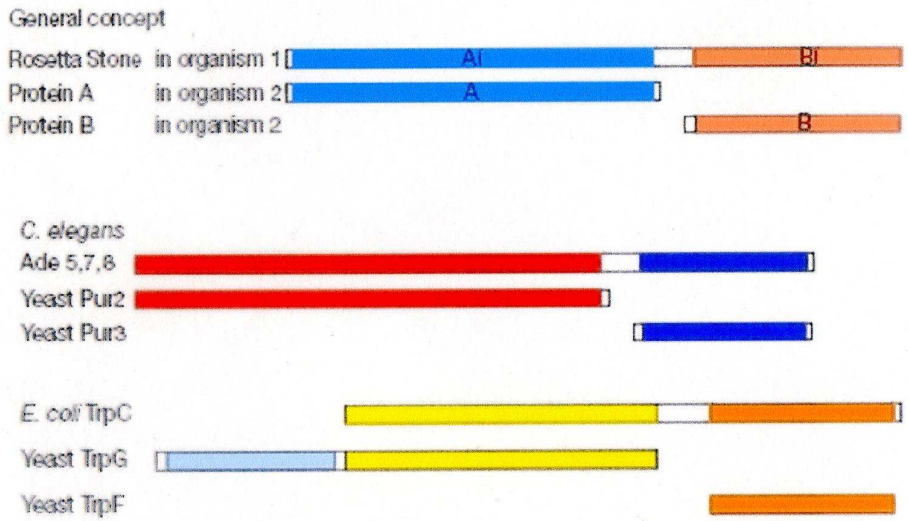
✓ La variation du contenu en gènes entre organisme

→ Méthode des profils phylogénétiques *(Pellegrini et al. (1999) PNAS 96,4285-4288)*

La Méthode de la "Pierre de Rosette"
(Marcotte et al., Science, 285, 751-753, 1999)

Box 2

The Rosetta Stone method for detecting functional linkage



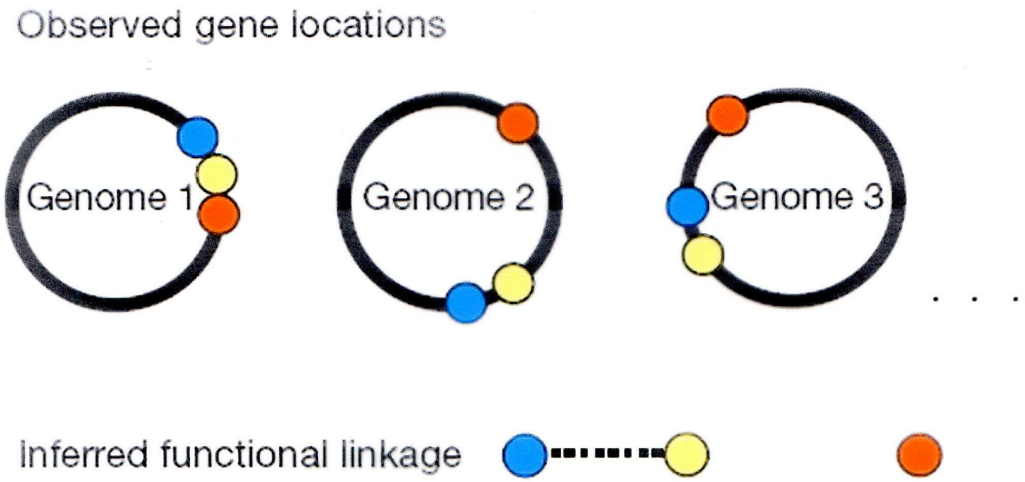
The domain fusion or Rosetta Stone method for detecting functional linkage[12,15] is illustrated here by three examples. The top sequence in all three triplets of proteins is the fused domain or Rosetta Stone sequence; it is homologous to two separate sequences in another species. In the middle example, the genes *Pur2* and *Pur3* of yeast both encode enzymes that catalyse steps in the purine biosynthetic pathway. If it were not previously known from biochemical and genetic experiments that these enzymes are functionally linked, the linkage would be apparent from the Rosetta stone sequence Ade5,7,8 from *Caenorhabditis elegans*. Similarly, in the lower example, the fused sequence of TrpC in the *Escherichia coli* genome would inform us that the yeast proteins TrG and TrpF are functionally linked, if we did not know already that they both catalyse steps in the biosynthesis of tryptophan.

The probability that two proteins have fused can be calculated using a hypergeometric distribution.

Méthode des gènes voisins

Box 3

The method of correlated gene neighbours for inferring functional linkage



If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.

probability that two genes are separated by fewer than d genes (N: total number of genes in the genome

$$P(\leq d) = \frac{2d}{N-1}$$

If the two genes have homologs in $m-1$ organisms we compute the product of P_i

$$X = \prod_{i=1}^{m} P_i(\leq d_i)$$

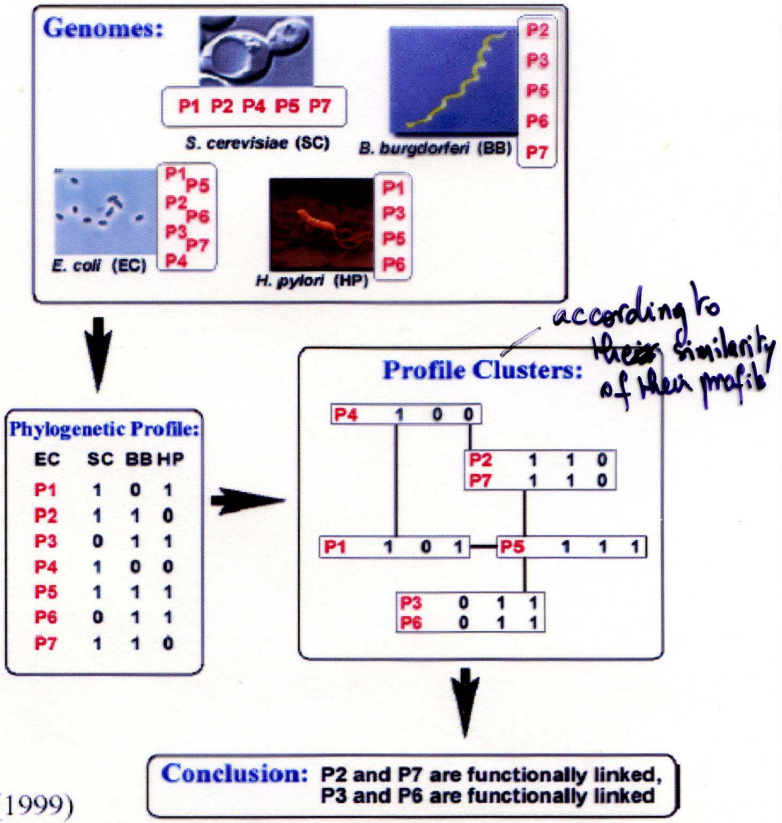
The probability of obtaining a value of X < the observed one:

$$P_m(\leq X) \simeq X \sum_{k=0}^{m-1} \frac{(-\ln X)^k}{k!}$$

e.g. 4 Genomes, 7 proteins in E.coli (P_1, P_7)

La méthode des profils phylogénétiques

Principe : utilise la variation du contenu en gènes entre organisme



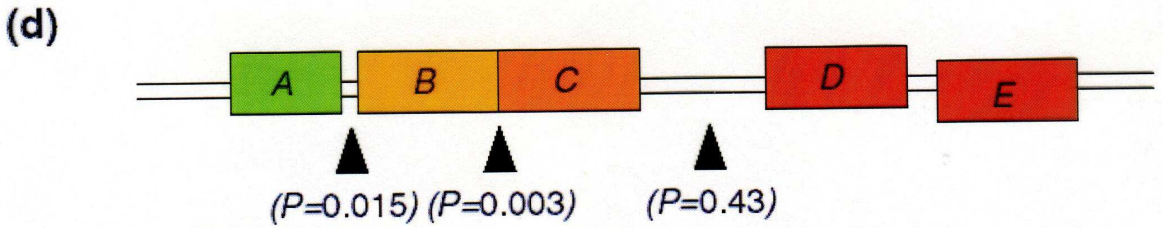
Profile clusters : Identical profiles are clustered in boxes, and profiles differing by one bit connected by lines.

Pellegrini et al. PNAS **96**, 4285-4288 (1999)

We define a homolog of a query protein to be present in a secondary genome if E-value using BLAST $\leq 10^{-10}$.
The result of this calculation across N genomes yields an N-dimensional vector of 0 and 1 for the query protein. → construction Phylogenic Profile → Profil Clusters.
The probability that two proteins coevolve can be calculated using a hypergeometric distribution

Gene cluster or operon method

This method identifies closed spaced genes, and assigns a probability P of observing a particular gap distance.

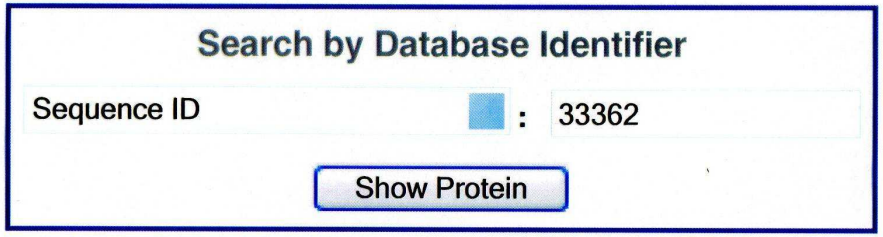


Assuming that gene start positions can be modelled by a Poisson distribution, and m = total number of genes divided by the number of intergenic nucleotides

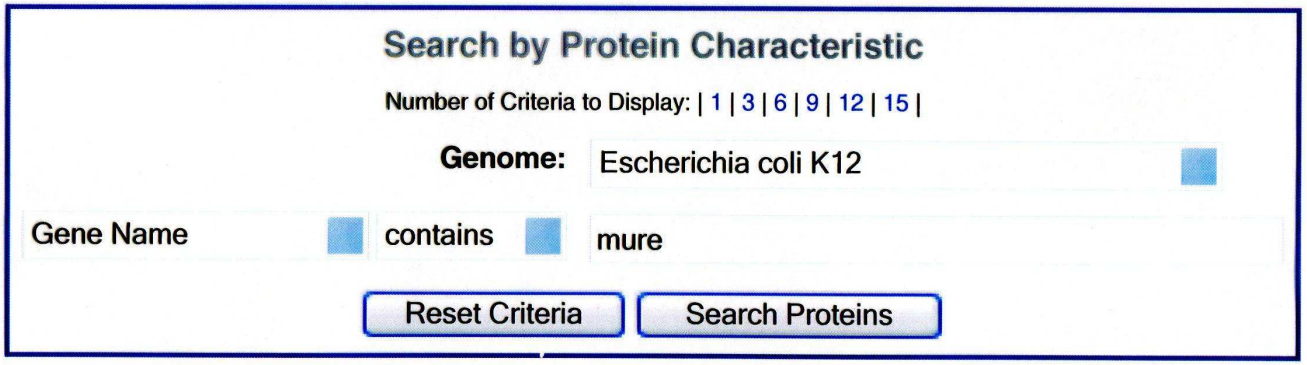
$$P(\text{separation} < N) = 1 - e^{-mN}$$

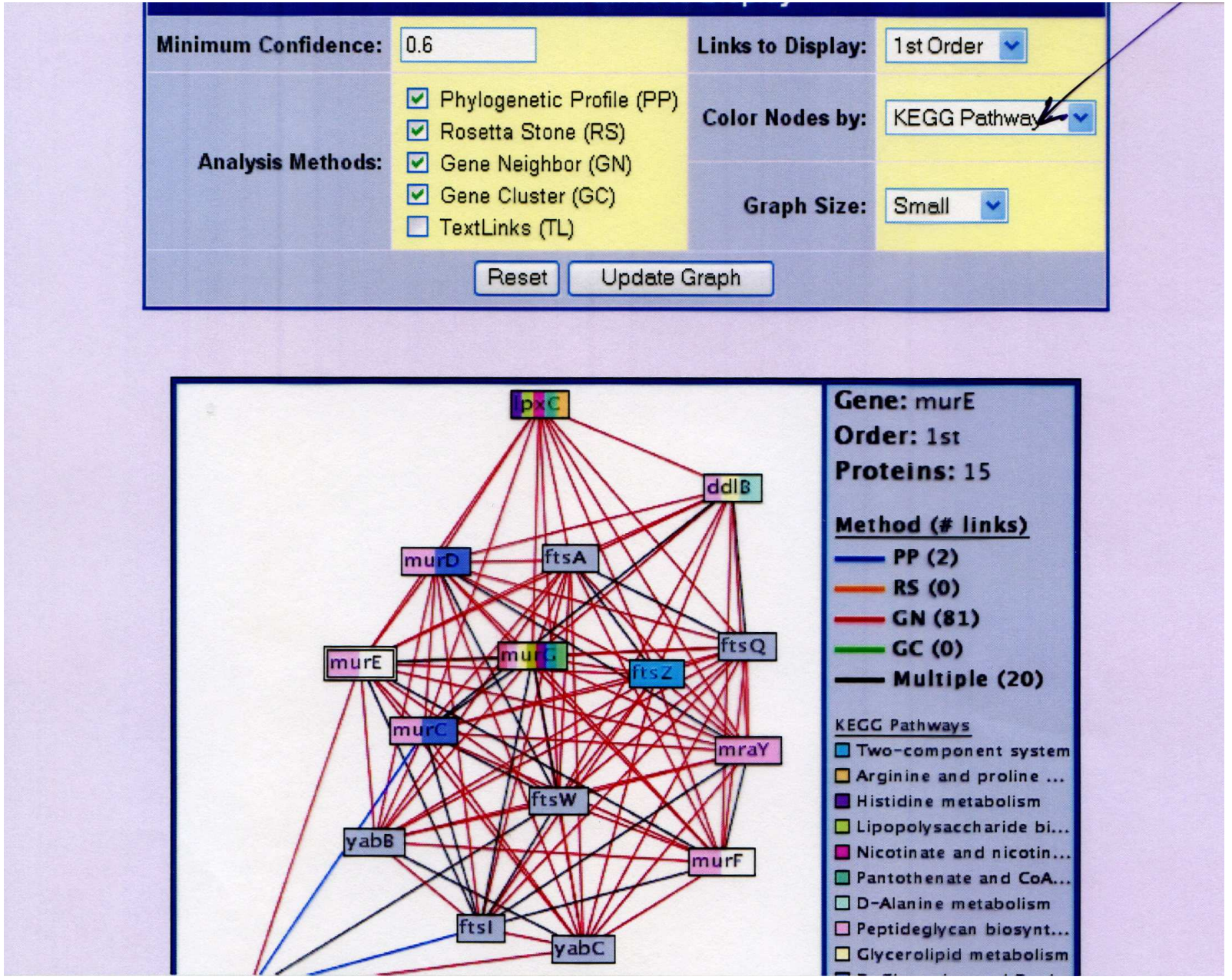
The Prolinks Database

Select a protein by entering a sequence identification number
from a public database, or the name or family of the gene.
An example is given for the *E. coli* protein murE, with seqid=33362.



OR




Minimum Confidence:
0.6
Links to Display:
1st Order
Analysis Methods:
Phylogenetic Profile (PP)
Rosetta Stone (RS)
Gene Neighbor (GN)
Gene Cluster (GC)
TextLinks (TL)
Color Nodes by:
KEGG Pathway
Graph Size:
Small
Reset
Update Graph
lpxC
ddlB
murD
ftsA
murE
murG
ftsQ
ftsZ
murC
mraY
ftsW
yabB
murF
ftsI
yabC
Gene: murE
Order: 1st
Proteins: 15
Method (# links)
PP (2)
RS (0)
GN (81)
GC (0)
Multiple (20)
KEGG Pathways
Two-component system
Arginine and proline ...
Histidine metabolism
Lipopolysaccharide bi...
Nicotinate and nicotin...
Pantothenate and CoA...
D-Alanine metabolism
Peptideglycan biosynt...
Glycerolipid metabolism

Comparison STRING and Prolinks

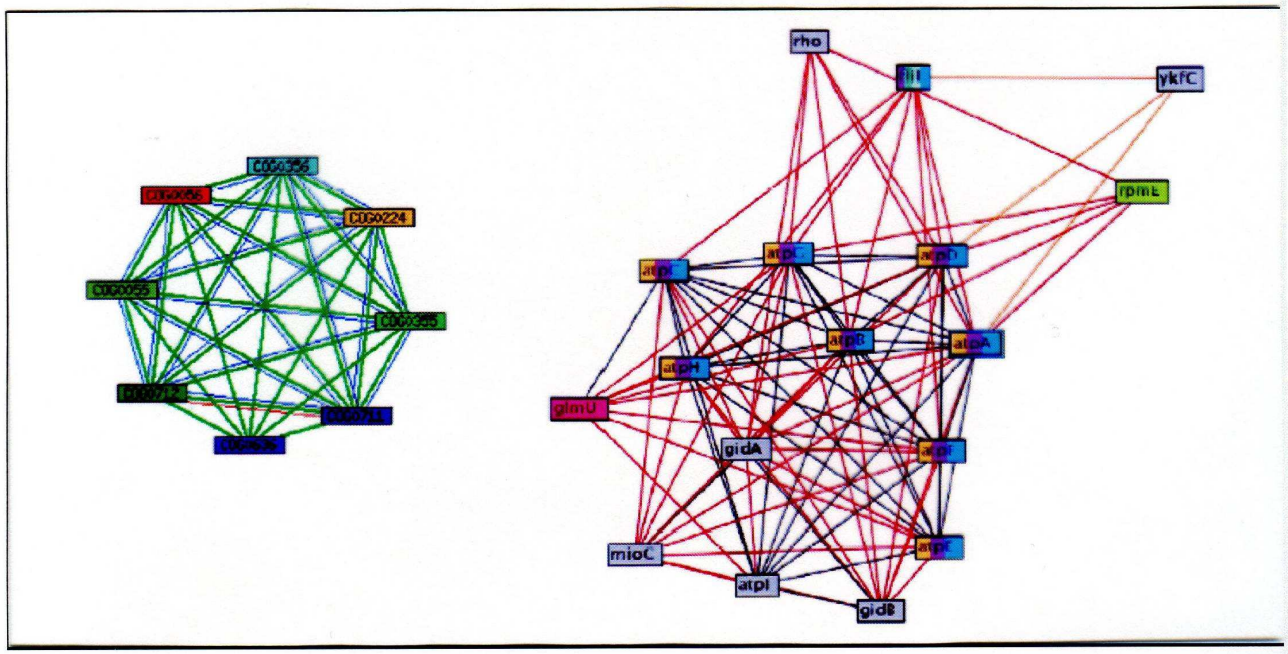


Figure 6
A comparison of graphs generated by querying the String database and Proteome Navigator to identify proteins in the ATP synthase complex. COG0056, shown in red in the String network (left), contains the *E. coli* protein AtpA, used to search each database and shown highlighted as a double-lined box in the Proteome Navigator graph (right). The Proteome Navigator network and Prolinks database identify twice the number of functionally linked proteins at the given confidence level.