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
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 eucaryotic genomes were assigned to COGs and can be reached from each individual COG page.

<table>
<thead>
<tr>
<th>Code</th>
<th>Name</th>
<th>Proteins in COG</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Archaeoglobus fulgidus</td>
<td>2420 1872</td>
</tr>
<tr>
<td>O</td>
<td>Halobacterium sp. NRC-1</td>
<td>2605 1701</td>
</tr>
<tr>
<td>M</td>
<td>Methanosarcina mazei</td>
<td>1786 1379</td>
</tr>
<tr>
<td>M</td>
<td>Methanobacterium thermoautotrophicum</td>
<td>1873</td>
</tr>
<tr>
<td>P</td>
<td>Thermoplasma acidophilum</td>
<td>1482</td>
</tr>
<tr>
<td>P</td>
<td>Thermoplasma volcanium</td>
<td>1499 1243</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus horikoshii</td>
<td>1800 1378</td>
</tr>
<tr>
<td>K</td>
<td>Pyrococcus abyssi</td>
<td>1768 1456</td>
</tr>
<tr>
<td>Z</td>
<td>Aeropyrum pernix</td>
<td>1841 1178</td>
</tr>
<tr>
<td>Y</td>
<td>Saccharomyces cerevisiae</td>
<td>5955 2290</td>
</tr>
<tr>
<td>Q</td>
<td>Candida albicans</td>
<td>9168 2720</td>
</tr>
<tr>
<td>V</td>
<td>Aquifex aeolicus</td>
<td>1560 1329</td>
</tr>
<tr>
<td>D</td>
<td>Thermotoga maritima</td>
<td>1858 1527</td>
</tr>
<tr>
<td>D</td>
<td>Desulfovibrio desulfuricans</td>
<td>3187 2226</td>
</tr>
<tr>
<td>R</td>
<td>Mycobacterium tuberculosis</td>
<td>3927 2585</td>
</tr>
<tr>
<td>R</td>
<td>Mycobacterium leprae</td>
<td>1605 1134</td>
</tr>
<tr>
<td>L</td>
<td>Lactococcus lactis</td>
<td>2267 1618</td>
</tr>
<tr>
<td>L</td>
<td>Streptococcus pyogenes</td>
<td>1697 1211</td>
</tr>
<tr>
<td>B</td>
<td>Bacillus subtilis</td>
<td>4118 2870</td>
</tr>
<tr>
<td>C</td>
<td>Bacillus halodurans</td>
<td>4066 2878</td>
</tr>
<tr>
<td>C</td>
<td>Synechocystis</td>
<td>3167 2159</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli K12</td>
<td>4275 3414</td>
</tr>
<tr>
<td>E</td>
<td>Escherichia coli O157</td>
<td>5315 3662</td>
</tr>
</tbody>
</table>

**Functional Categories**

- **K** = transcription
- **T** = Signal transduction mechanisms

**Principal component analysis of genomes**
- **List of COGs**
- **Distribution**
- **Phylogenetic pattern search**
- **Functional categories**
- **Pathways and functional systems**
- **FTP**
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\(^{12,16}\) 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 catalyze 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 TrpG and TrpF are functionally linked, if we did not know already that they both catalyze steps in the biosynthesis of tryptophan.

The probability that two proteins have fused can be calculated using a hypergeometric distribution.
Box 3

The method of correlated gene neighbours for inferring functional linkage

Observed gene locations

Genome 1  Genome 2  Genome 3

Inferred 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

\[ P_m(\leq X) \approx X \sum_{k=0}^{m-1} \left( -\ln X \right)^k \]

\[ \frac{1}{k!} \]
La méthode des profils phylogénétiques

Principe : utilise la variation du contenu en gènes entre organismes

Profile clusters: Identical profiles are clustered in boxes, and profiles differing by one bit are 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 < 10^-10.

The result of this calculation across N genomes yields an N-dimensional vector of 0 and 1 for the query protein. Construction: Phylogenetic Profile -> Profile 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 modeled by a Poisson distribution, and $n = \text{total number of genes divided by the number of intergeneous nucleotides}$,

$$P(\text{separation} < n) = 1 - e^{-mn}$$
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.

Search by Database Identifier

Sequence ID : 33362

Show Protein

OR

Search by Protein Characteristic

Number of Criteria to Display: 1, 3, 6, 9, 12, 15

Genome: Escherichia coli K12

Gene Name contains mure

Reset Criteria Search Proteins
Figure 6
A comparison of graphs generated by querying the String database and Proteome Navigator to identify proteins in the ATP synthase complex. COG0036, 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.