Orthologs, paralogs and evolutionary genomics

Eugene Koonin

Annu. Rev. Genet. 2005. 39:309–38

Presented by Xiaoyun "Sean" Liao

COMP 571

November 20, 2008

Eugene Koonin

- Senior Investigator
  National Center for Biotechnology Information
  National Institutes of Health
  Email: koonin@ncbi.nlm.nih.gov
  www.ncbi.nlm.nih.gov/CBBresearch/Koonin/

- Research interests:
  - Empirical comparative and evolutionary genomics
 - Exploitation of genome comparisons

Outline

- Definitions and classification of orthologs and paralogs

- Identification of orthology and paralogy

- Evolutionary patterns of orthology and paralogy

Events of gene evolution

- Vertical descent (speciation)

- Gene duplication

- Gene loss

- Horizontal gene transfer (HGT)

- Rearrangements of gene structure.

Homology, orthology and paralogy

- **Homologs**:
  Genes sharing a common origin.

- **Orthologs**:
  Genes derived from a single ancestral gene in the last common ancestor (LCA) of the compared species.

- **Paralogs**:
  Genes related via duplication.
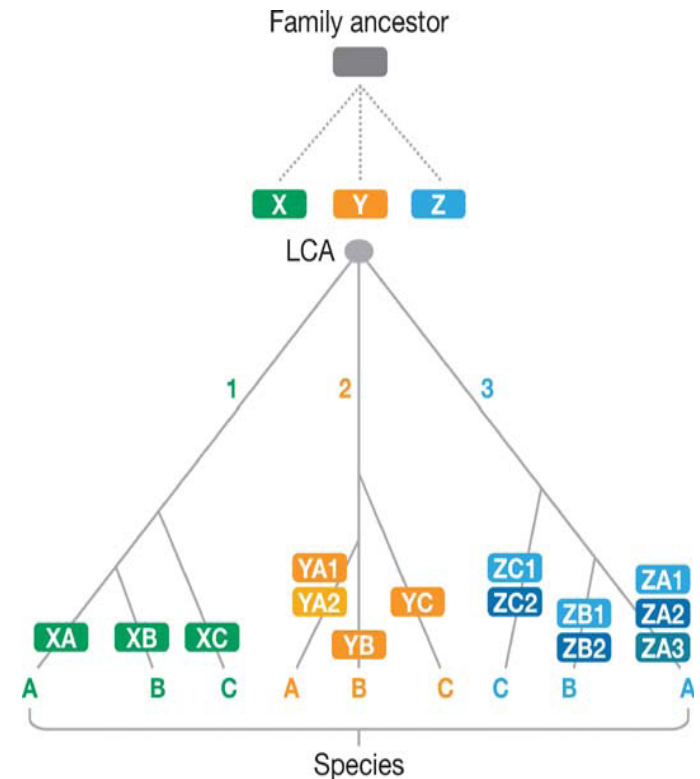
Effect of lineage-specific gene duplication

- **Co-orthologs**:
  Two or more genes in one lineage that are orthologous to one or more genes in another lineage due to a lineage-specific duplication(s).
- **Inparalogs** (symparalogs): Paralogous genes resulting from a lineage-specific duplication(s) subsequent to a given speciation event.
- **Outparalogs** (alloparalogs): Paralogous genes resulting from a duplication(s) preceding a given speciation event .
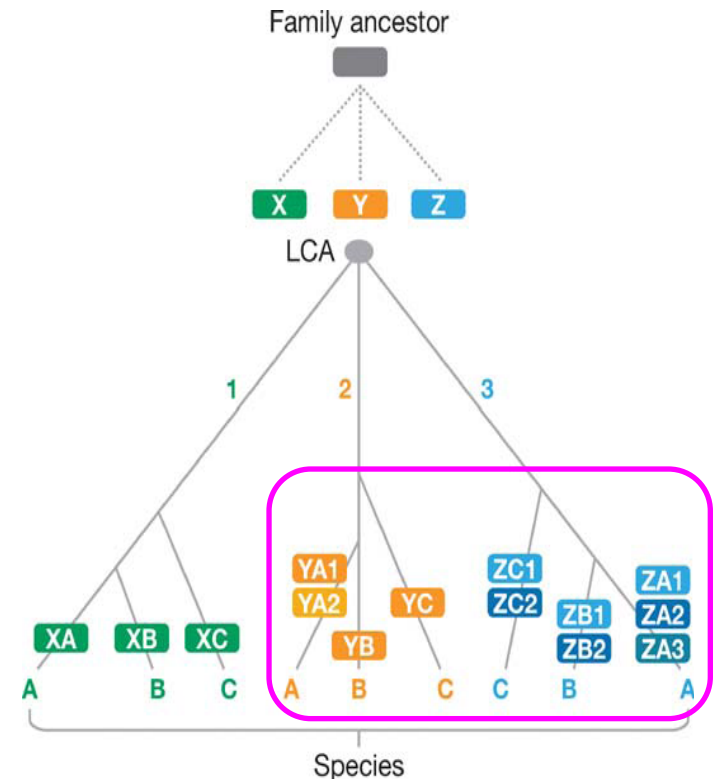
Effect of lineage-specific gene loss

- **Pseudoorthologs**:
  Genes that actually are paralogs but appear to be orthologous due to differential, lineage-specific gene loss.
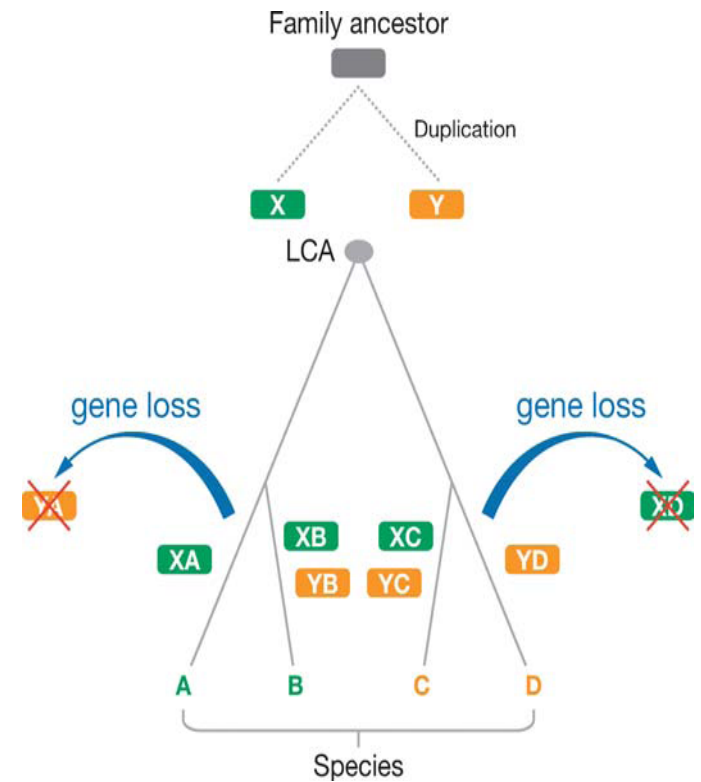
Effect of horizontal gene transfer

- Xenologs: Homologous genes acquired via XGD by one or both of the compared species but appearing to be orthologous in pairwise genome comparisons.

- Pseudoparalogs: Homologous genes that come out as paralogs in a single-genome analysis but actually ended up in the given genome as a result of a combination of vertical inheritance and HGT.
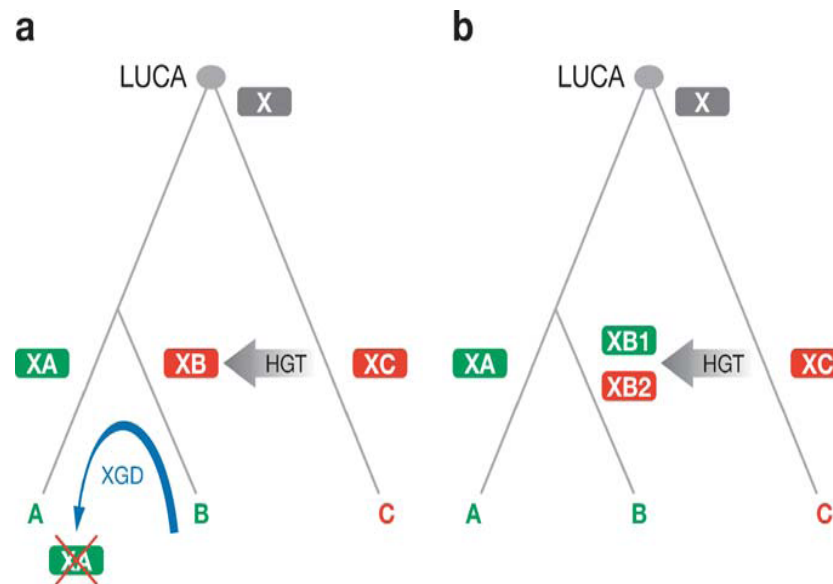
Distinction between orthologs and paralogs

- Functional annotation of genomes

- Reconstruction of genome evolution.

Identification of orthologs and paralogs

- Comparative genomics
  - Delineate sets of orthologs

- Phylogenetic analysis and tree reconcilation
  - HGT

SymBet hypothesis

- The sequences of orthologous genes (proteins) are more similar to each other than they are to any other genes (proteins) from the compared genomes, i.e., they form symmetrical best hits (SymBets).

- SymBets are most likely to be formed by orthologs, suggesting a very simple and straightforward method for identification of orthologs.
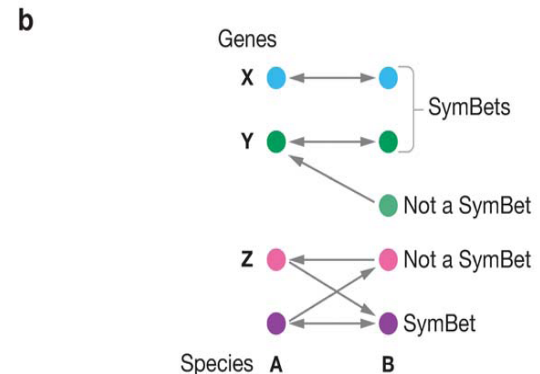
Orthology and genome-specific best hits

- A gene from one genome is most similar not to its ortholog but to a paralog from another genome.
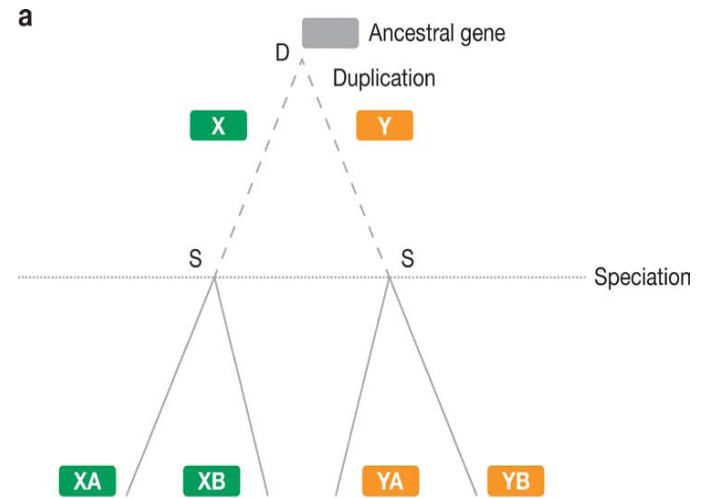
- A one-to-one orthologous relationship is a major rather than a minor pattern in genome evolution.

Table 2 Symmetrical best hits between selected prokaryotic genomes[a]

	Ec	Yp	Hp	Bs	Mg	Aa	Tm	Mj	Ma	Ta
Ec-4289		0.584	0.456	0.305	0.527	0.519	0.428	0.24	0.151	0.308
Yp-4083	2385		0.432	0.28	0.525	0.496	0.423	0.218	0.144	0.275
Hp-1566	714	677		0.403	0.442	0.396	0.321	0.181	0.211	0.176
Bs-4100	1251	1144	631		0.648	0.495	0.465	0.239	0.153	0.306
Mg-480	253	252	212	311		0.469	0.515	0.235	0.281	0.238
Aa-1553	806	771	615	768	225		0.449	0.279	0.33	0.256
Tm–1846	808	780	503	858	247	697		0.245	0.294	0.265
Mj-1770	425	385	284	423	113	434	433		0.489	0.362
Ma-4540	649	589	330	627	135	513	543	866		0.415
Ta-1478	455	406	260	453	114	378	392	535	614	

[a]The bottom half of the table shows the number of SymBets for each pair of genomes and the upper half shows the fraction of proteins in the smaller genome that form SymBets with the putative orthologs from the larger genome. Species abbreviations are as follows. Proteobacteria: Ec, *Escherichia coli*; Yp, *Yersinia pestis*; Hp, *Helicobacter pylori*: Gram-positive bacteria: Bs, *Bacillus subtilis*; Mg, *Mycoplasma genitalium*: deep-branching, hyperthermophilic bacteria: Aa, *Aquifex aeolicus*; Tm, *Thermotoga maritima*; archaea: Mj, *Methanocaldococcus jannaschii*; Ta, *Thermoplasma acidophilum*; Ma, *Methanosarcina acetivorans*. For each species, the total number of protein-coding genes is indicated.

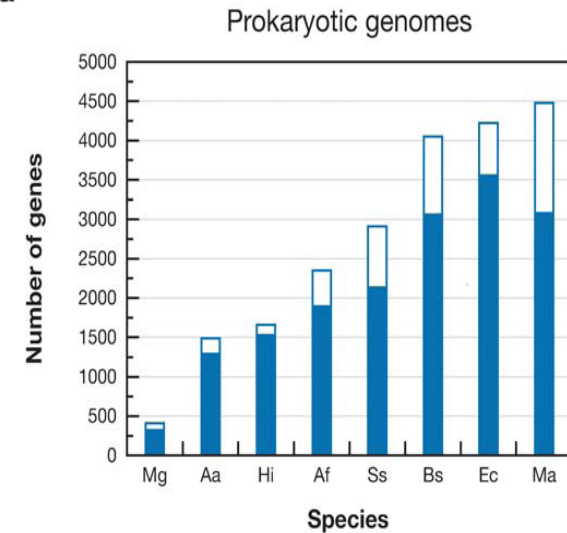# Clusters of Orthologous Groups of proteins (COGs)

- Generalize and extend the notion of a genome-specific best hit.

- The system captured not only one-to-one orthologous relationships but also coorthologous relationships between inparalogs.

- Definitions and classification of orthologs and paralogs

- Identification of orthology and paralogy

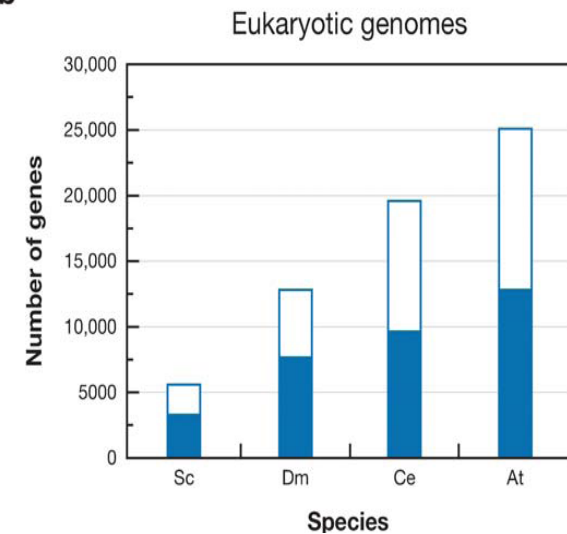- Evolutionary patterns of orthology and paralogy

Coverage of genomes

- The coverage of genomes with COGs slowly increases with the growing number of included species.
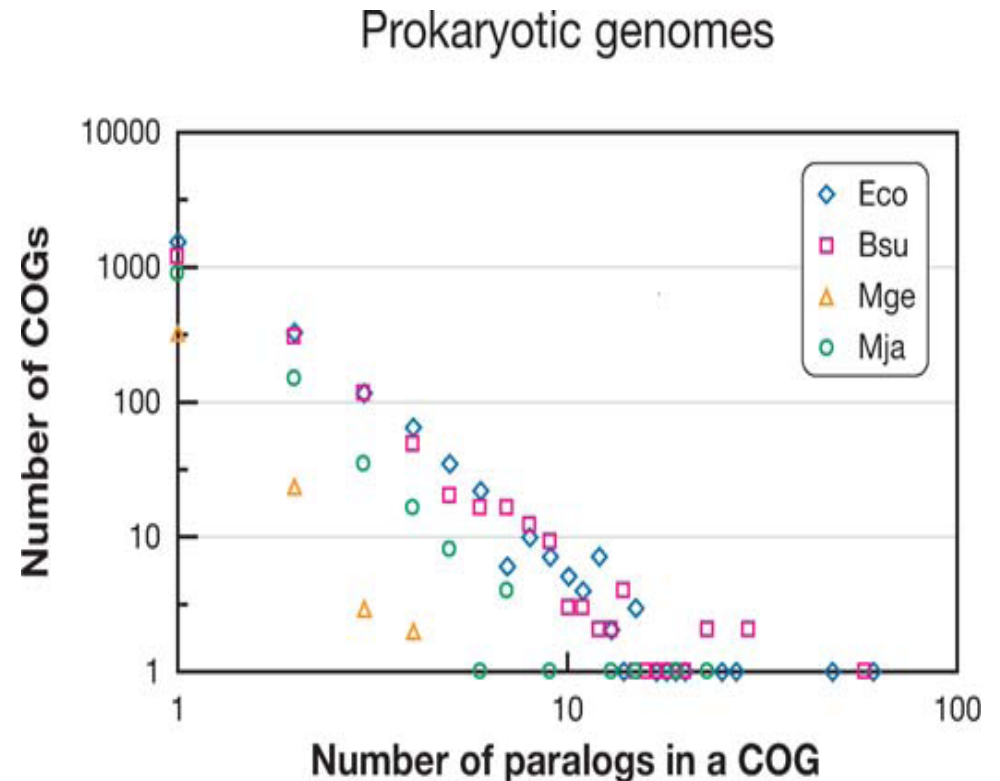

a

Prokaryotic genomes

Number of genes

Mg Aa Hi Af Ss Bs Ec Ma

Species

b

Eukaryotic genomes

Number of genes

Sc Dm Ce At

Species

16

One-to-one orthologs and inparalogs

- Compatible with the earlier quantitative analysis of lineage-specific expansions in prokaryotes.

Orthologous clusters and molecular clock

- Kimura's neutral theory
  - The rate of evolution of a gene remains the same as long as the biological function does not change.

- The majority of orthologous genes evolve in the clock-like mode as long as there was no duplication.

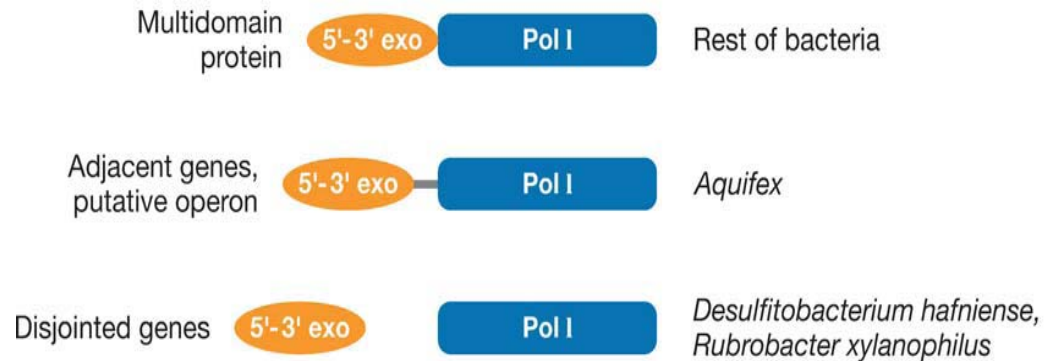# Horizontal gene transfer leading to pseudoparalogy
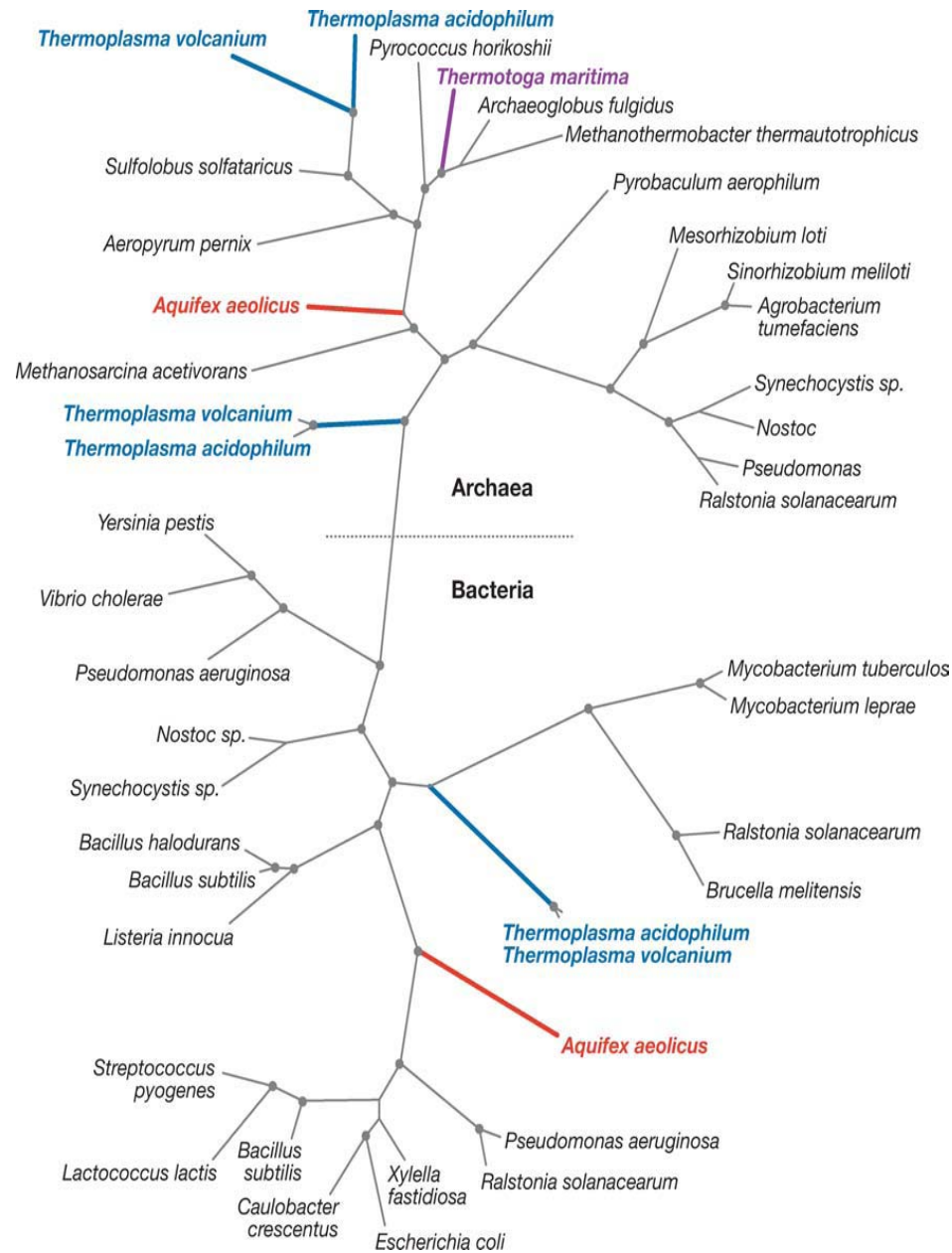
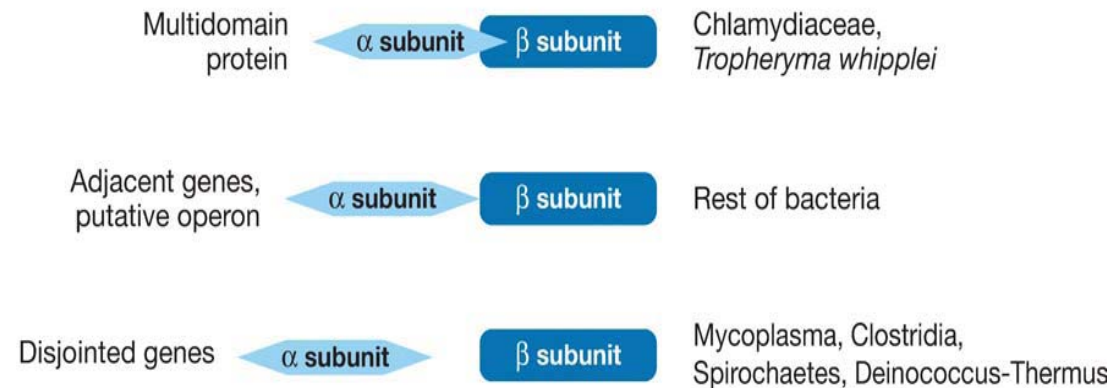# Rearrangements of gene structure and orthology
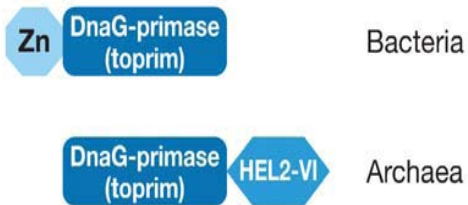
Functional correlates of orthology & paralogy

- In general, the notion that one-to-one orthologs are functionally equivalent seems to hold well.

- At greater evolutionary distances, particularly across the primary kingdom divides, there are prominent cases of apparent major differences in the functions of orthologs.

- The functional connotations of paralogy are distinct from and, in a sense, opposite to those of orthology.

Summary

1. Orthologs and paralogs are two types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication.

2. Distinguishing between orthologs and paralogs is crucial for successful functional annotation of genomes and for reconstruction of genome evolution.

3. A finer classification of orthologs and paralogs has been developed to reflect the interplay between duplication and speciation events, and effects of gene loss and horizontal gene transfer on the observed homologous relationship.

4. Methods for identification of sets of orthologous and paralogous genes involve phylogenetic analysis and various procedures for sequence similarity–based clustering.

5. Analysis of clusters of orthologous and paralogous genes is instrumental in genome annotation and in delineation of trends in genome evolution.

6. Rearrangements of gene structure confound orthologous and paralogous relationships.

Thank you for your attention.