Genome Evolution

演講者 鍾智明
組員 葉記達 王思寰 康証傑
陳浩志 蘇致瑋

What we can get from evolution?

- Conserved sequences
- Mutations and functions
- Insertion,replacement and deletion
- Mathematics and statistics
- bioinformatics

From today's assay

- Comparative analysis of complete genomes reveals gene loss,acquisition and acceleration of evolutionary rate in metazoa, suggests a prevalence of evolution via gene acquisition and indicates that the evolutionary rates in animal tend to be conserved
- Vladimir N. Babenko and Dmiri M. Krylov
- National center for biotechnology information

  NLM, NIH MD

  Nucleic Acid Research, 2004, Vol. 32 No. 17

  5029-5035

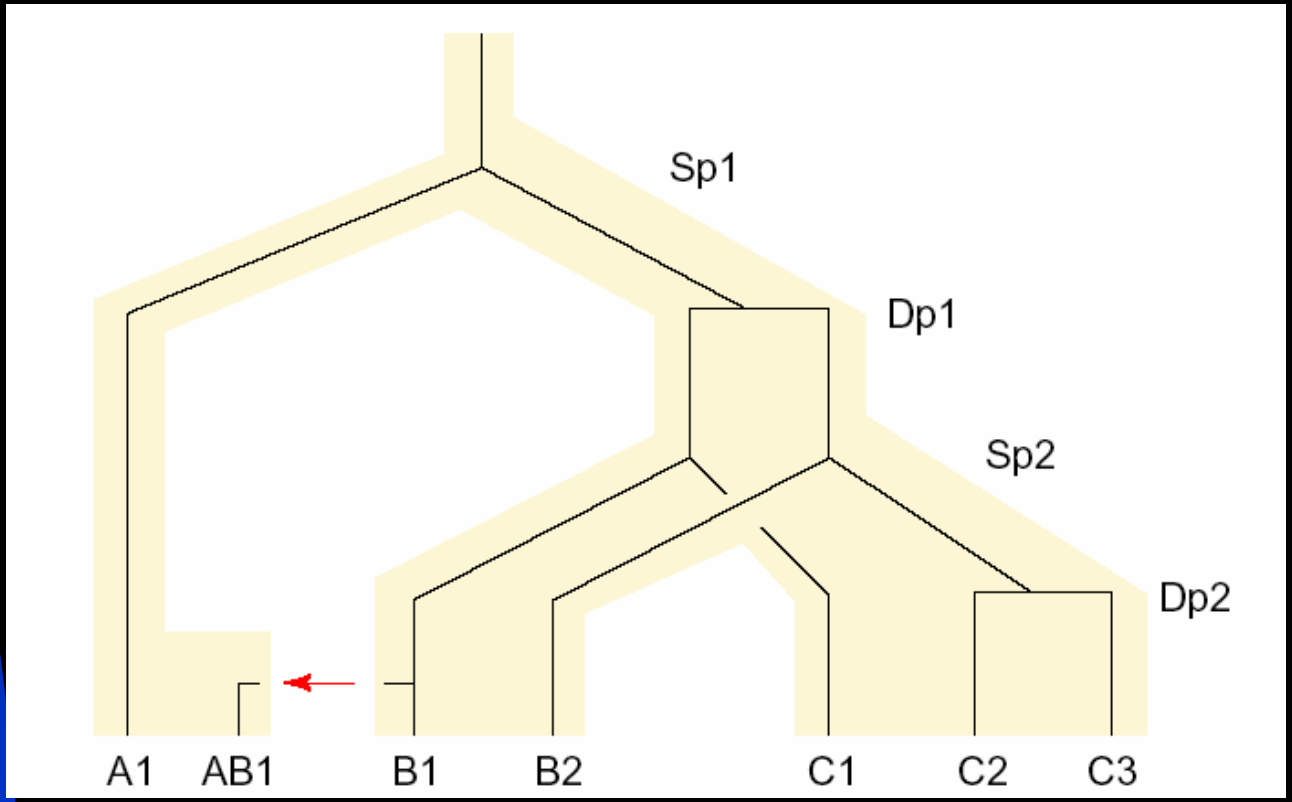
Homology: General Definition

- Homology designates a relationship of common descent between entities
- Two genes are either homologs or not
  - it doesn't make sense to say "two genes are 43% homologous"

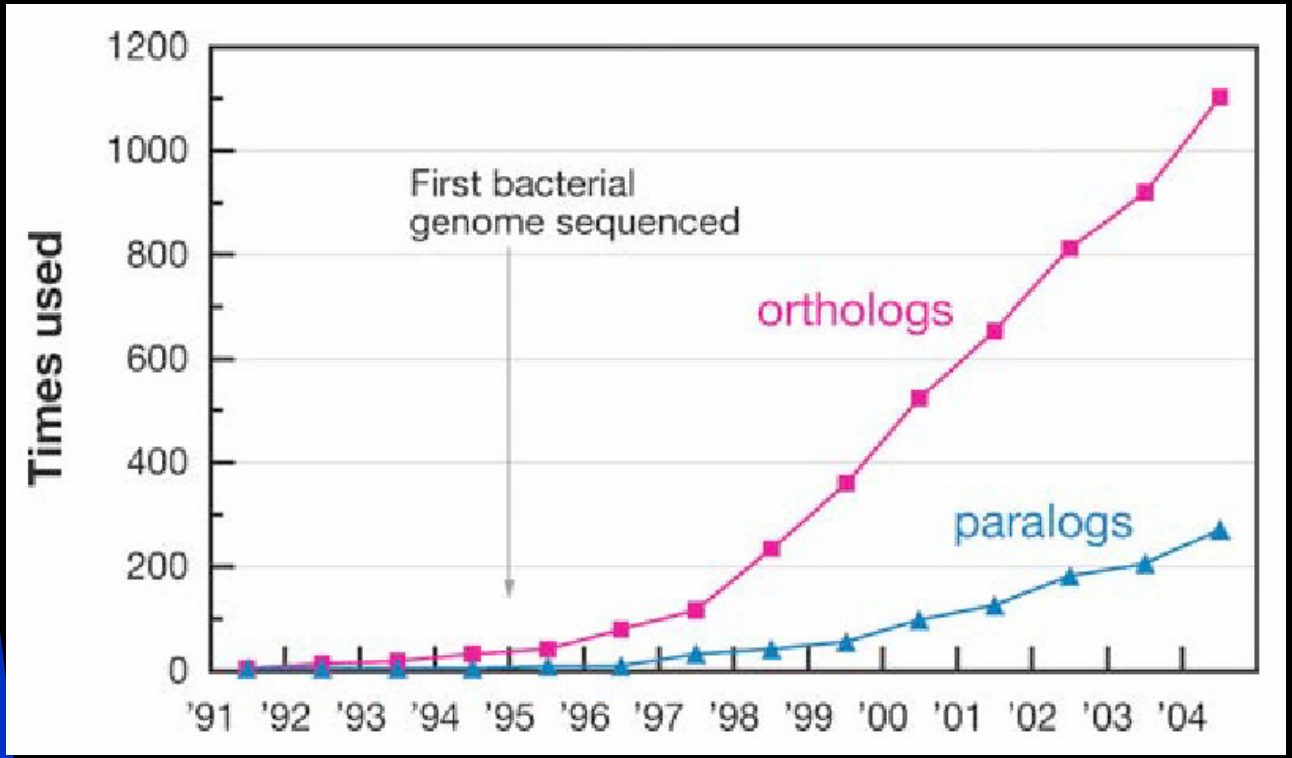
Orthology vs. Paralogy

- Two genes are orthologs if they originated from a single ancestral gene in the most recent common ancestor of their respective genomes
- Two genes are paralogs if they are related by duplication

Orthology vs. Paralogy
(Figure from Fitch, Trends in Genetics, 2000. 16(5):227-231)



Usage of the Terms: Pre and Post Genome Era

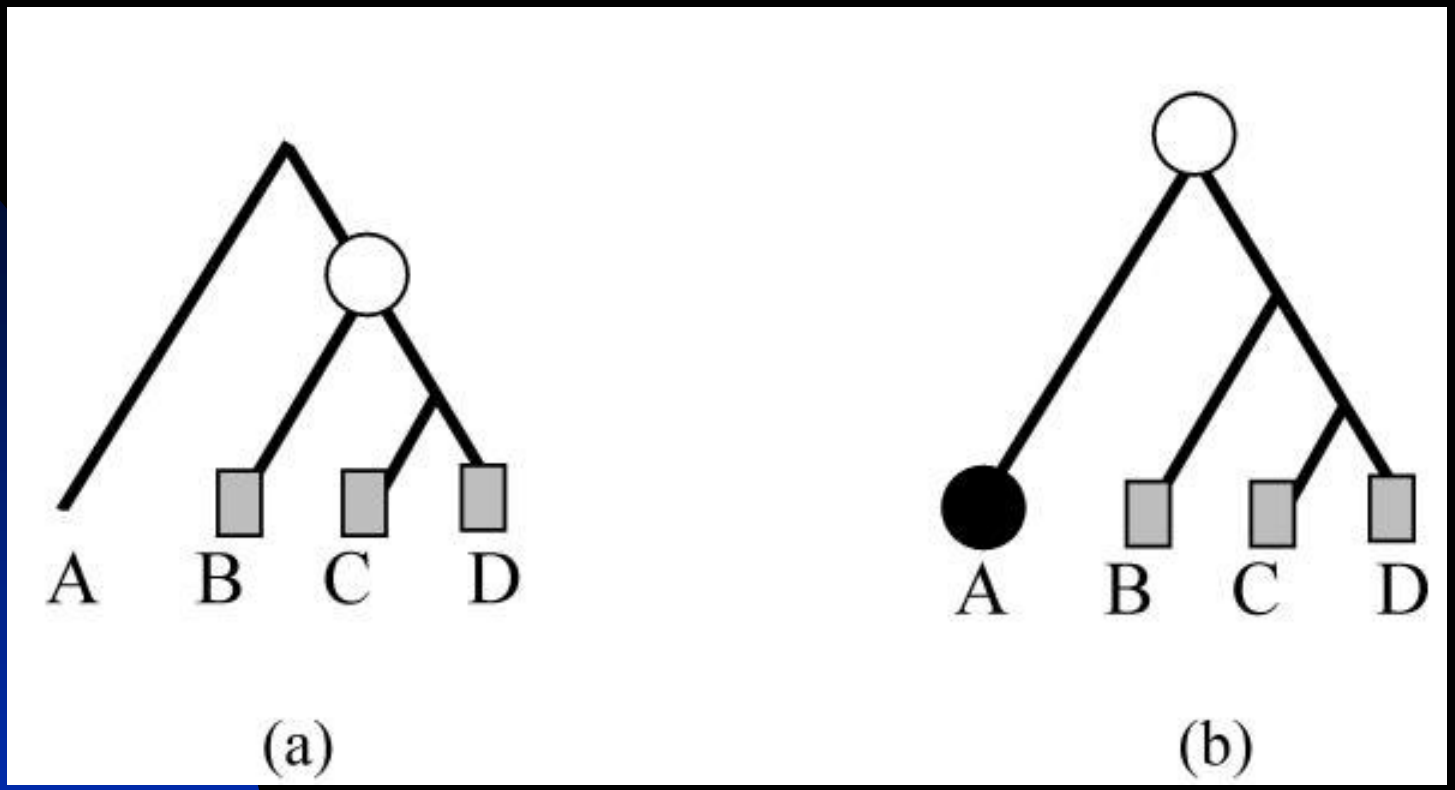


Orthology and Function

- The given definition of orthology has nothing to do with function
- However, a crucial property of orthologs, is that they typically perform equivalent functions in the respective organims

Clusters of Orthologous Groups of proteins (COGs)

- Clusters of Orthologous Groups of proteins (COGs) from the sequenced genomes of prokaryotes and unicellular eukaryotes
- COG collection currently consists of 138,458 proteins, which form 4873 COGs and comprise 75% of the 185,505 (predicted) proteins encoded in 66 genomes of unicellular organisms



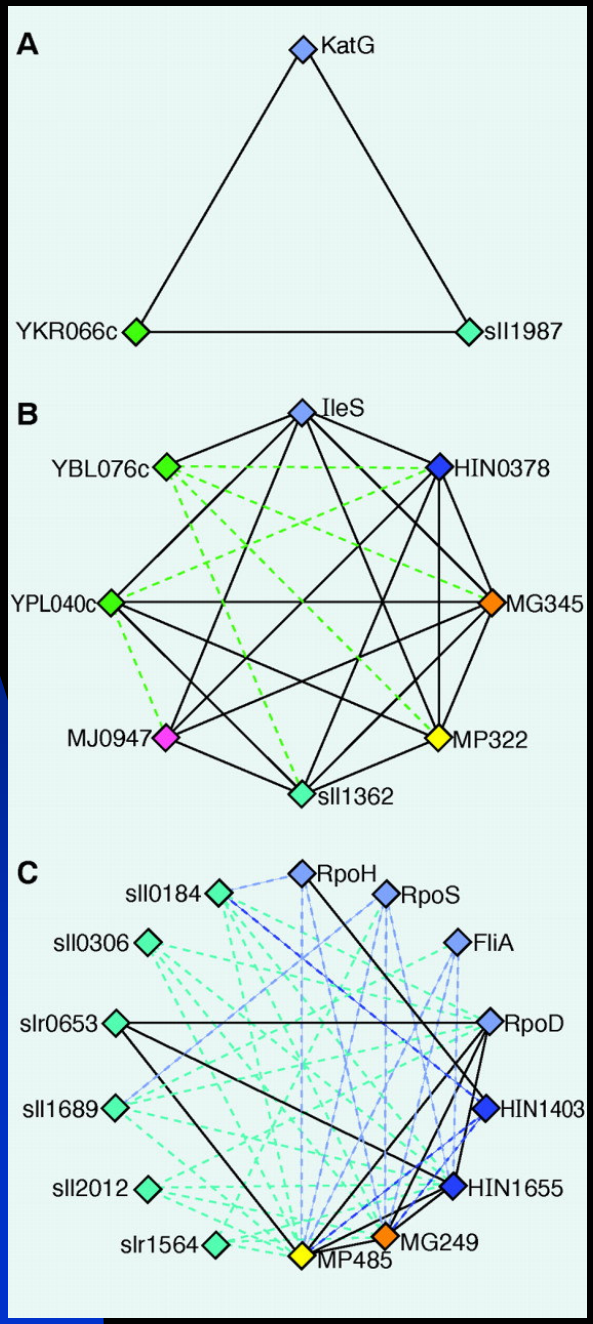
Two evolutionary scenarios leading to pattern II of Figure 1. The black circle represents gene loss and the white circle represents emergence of the COG.

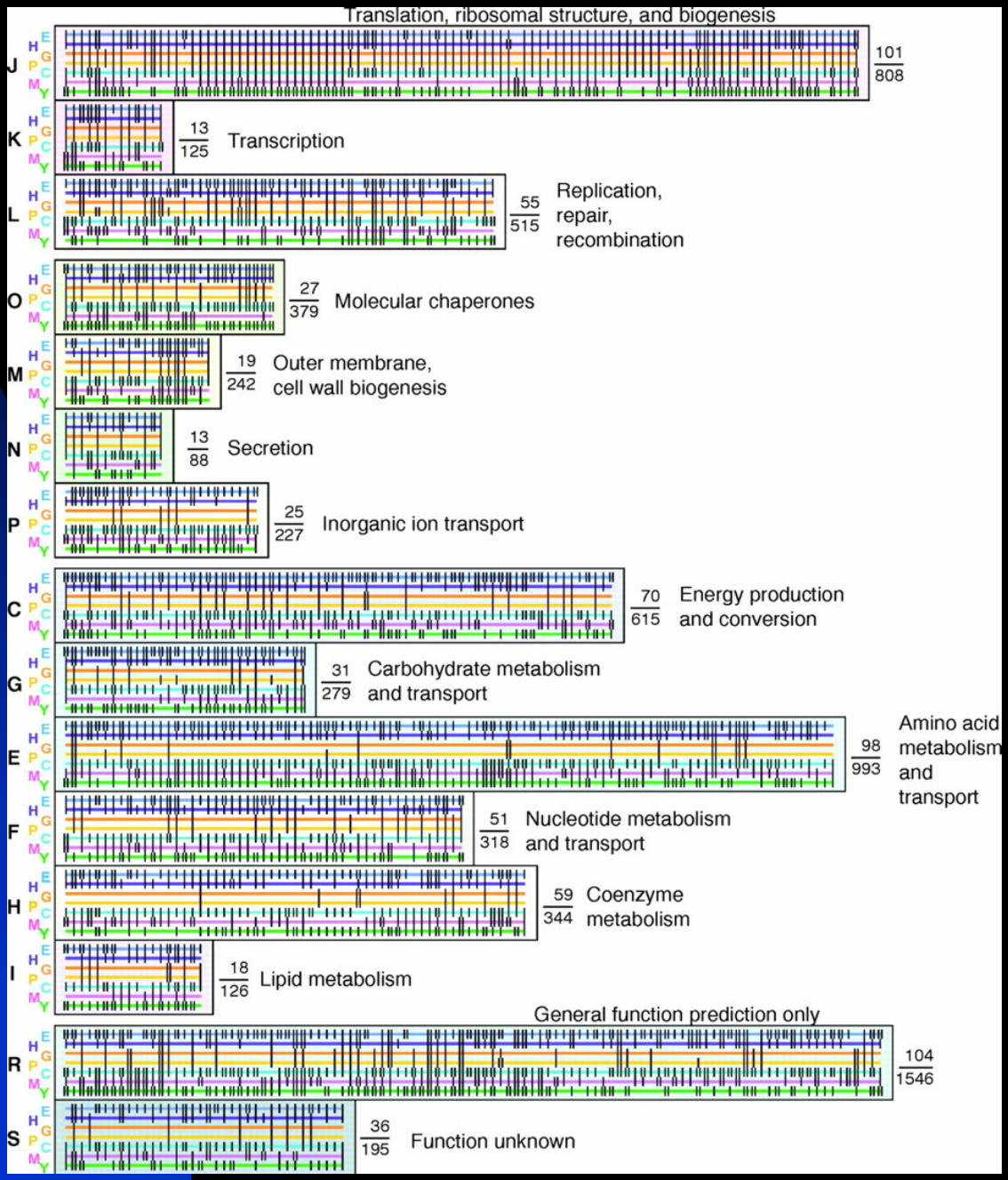
Surrogate Approaches

- Central assumption:
  - Sequences of orthologous genes are more similar to each other than they are to other genes from the compared genomes
  - In other words, sequences of orthologous genes form symmetrical best hits (SymBets)
  - Conversely, it is assumed that SymBets are most likely to be formed by orthologs

Constructing COGs

1. An all-against-all comparison of protein sequences encoded in multiple genomes (typically using BLAST)
2. Detection and clustering of obvious inparalogs (proteins from the same genome that are more similar to each other than they are any proteins from other species)
3. Identification of triangles of mutually consistent, genome-specific best hits such that clusters of inparalogs detected at step 2 are treated as single entities
4. Merging triangles with a common side to form COGs


A
KatG
YKR066c
sll1987
B
IleS
YBL076c
HIN0378
YPL040c
MG345
MJ0947
MP322
sll1362
C
RpoH
sll0184
RpoS
sll0306
FliA
slr0653
RpoD
sll1689
HIN1403
sll2012
HIN1655
slr1564
MG249
MP485

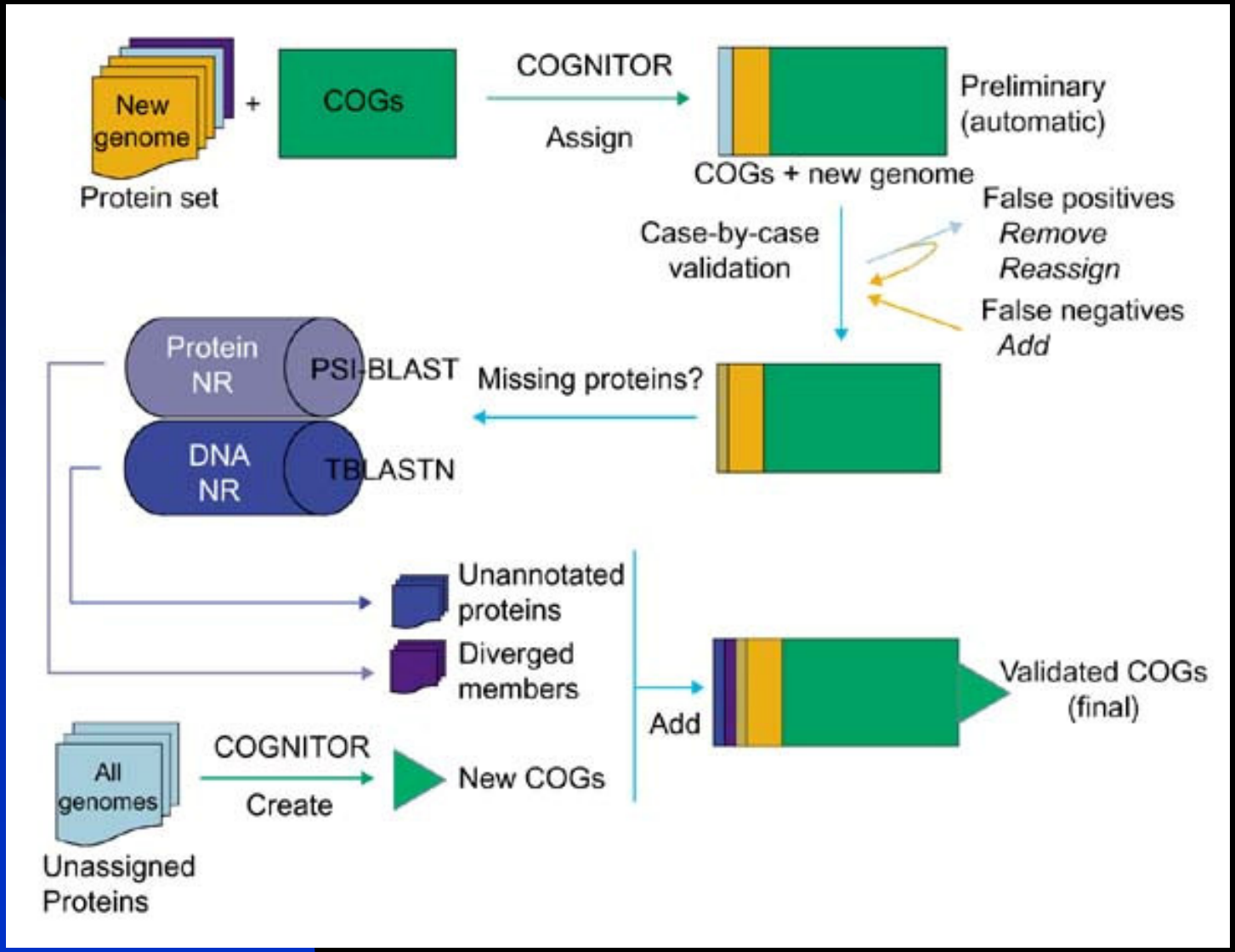

Translation, ribosomal structure, and biogenesis
J
101/808
K
13/125
Transcription
L
55/515
Replication, repair, recombination
O
27/379
Molecular chaperones
M
19/242
Outer membrane, cell wall biogenesis
N
13/88
Secretion
P
25/227
Inorganic ion transport
C
70/615
Energy production and conversion
G
31/279
Carbohydrate metabolism and transport
E
98/993
Amino acid metabolism and transport
F
51/318
Nucleotide metabolism and transport
H
59/344
Coenzyme metabolism
I
18/126
Lipid metabolism
General function prediction only
R
104/1546
S
36/195
Function unknown

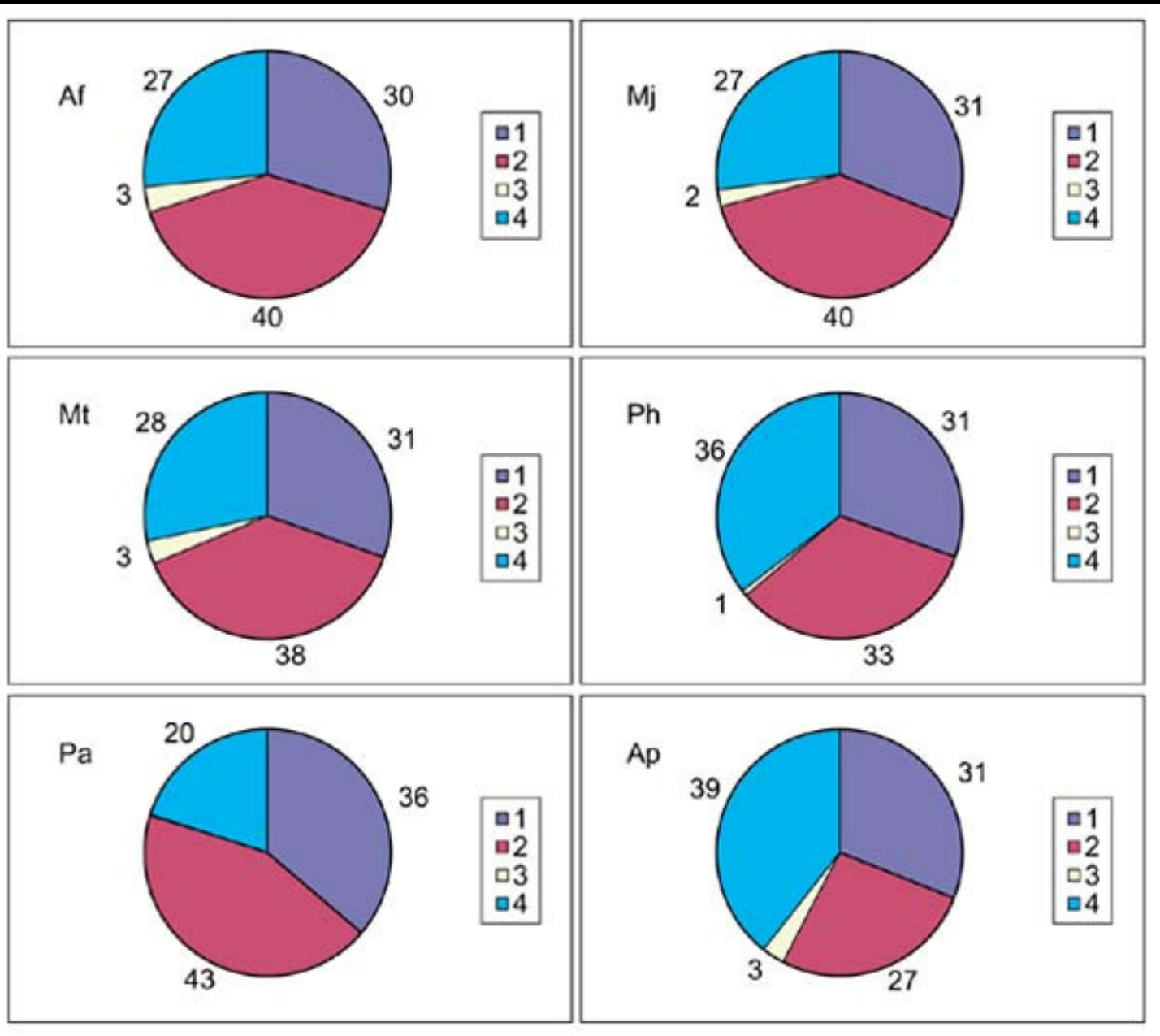
1.INFORMATION STORAGE AND PROCESSING
[J] Translation, ribosomal structure and biogenesis [A] RNA processing and modification [K] Transcription [L] Replication, recombination and repair [B] Chromatin structure and dynamics

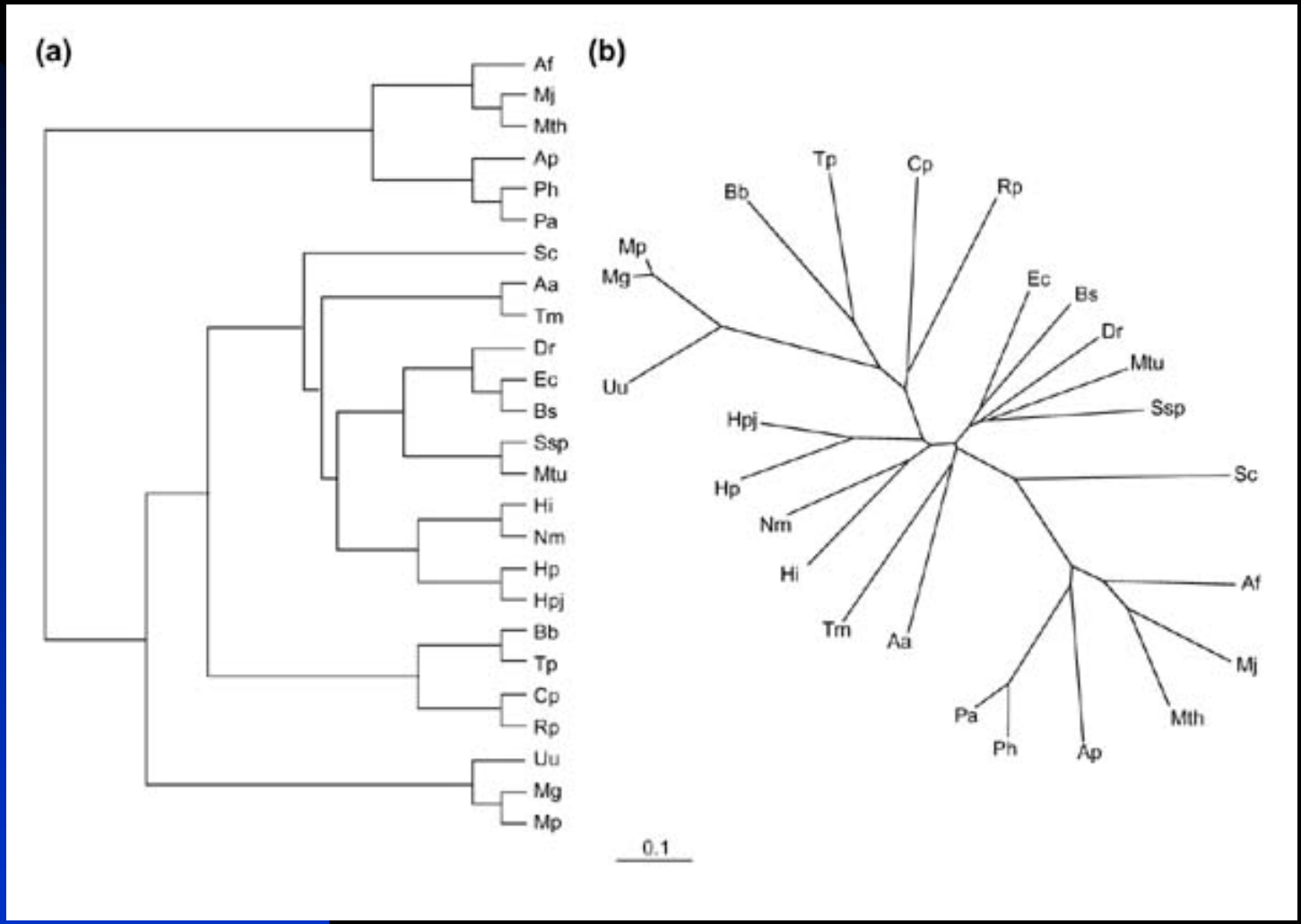
2.CELLULAR PROCESSES AND SIGNALING
[D] Cell cycle control, cell division, chromosome partitioning [Y] Nuclear structure [V] Defense mechanisms [T] Signal transduction mechanisms [M] Cell wall/membrane/envelope biogenesis [N] Cell motility [Z] Cytoskeleton [W] Extracellular structures [U] Intracellular trafficking, secretion, and vesicular transport [O] Posttranslational modification, protein turnover, chaperones

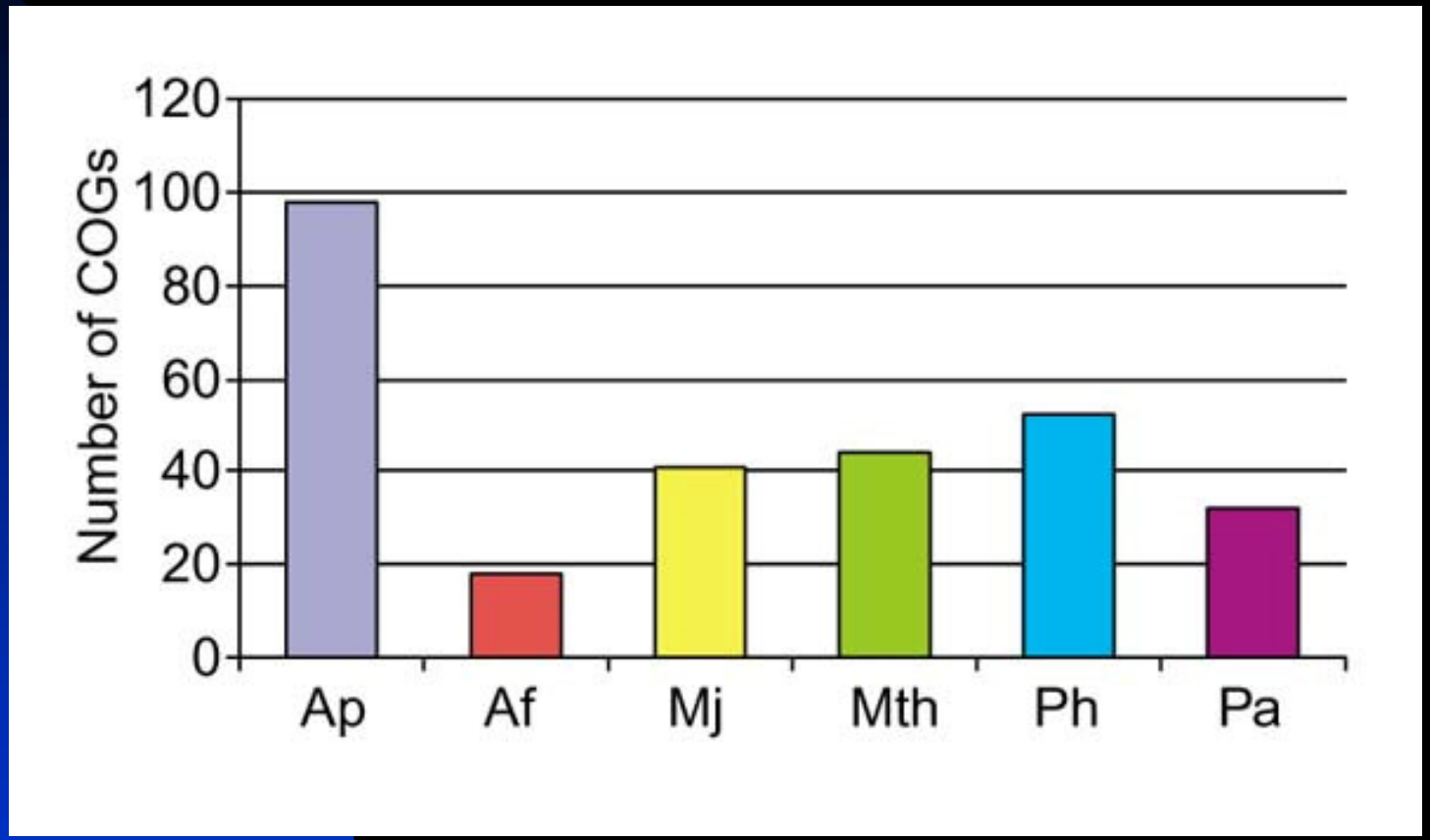
3.METABOLISM
[C] Energy production and conversion [G] Carbohydrate transport and metabolism [E] Amino acid transport and metabolism [F] Nucleotide transport and metabolism [H] Coenzyme transport and metabolism [I] Lipid transport and metabolism [P] Inorganic ion transport and metabolism [Q] Secondary metabolites biosynthesis, transport and catabolism

4.POORLY CHARACTERIZED [R] General function prediction only [S] Function unknown


New genome
Protein set
+
COGs
COGNITOR
Assign
Preliminary (automatic)
COGs + new genome
Case-by-case validation
False positives
Remove
Reassign
False negatives
Add
Protein NR
PSI-BLAST
DNA NR
TBLASTN
Missing proteins?
Unannotated proteins
Diverged members
Add
Validated COGs (final)
All genomes
Unassigned Proteins
COGNITOR
Create
New COGs


Af
27
30
3
40
1
2
3
4
Mj
27
31
2
40
1
2
3
4
Mt
28
31
3
38
1
2
3
4
Ph
36
31
1
33
1
2
3
4
Pa
20
36
43
1
2
3
4
Ap
39
31
3
27
1
2
3
4


(a)
Af
Mj
Mth
Ap
Ph
Pa
Sc
Aa
Tm
Dr
Ec
Bs
Ssp
Mtu
Hi
Nm
Hp
Hpj
Bb
Tp
Cp
Rp
Uu
Mg
Mp
(b)
Tp
Cp
Bb
Rp
Mp
Mg
Ec
Bs
Dr
Mtu
Uu
Ssp
Hpj
Sc
Hp
Nm
Af
Hi
Mj
Tm
Aa
Pa
Mth
Ph
Ap
0.1


Number of COGs
120
100
80
60
40
20
0
Ap
Af
Mj
Mth
Ph
Pa

Construction of KOGs for 7 sequenced eukaryotic genomes

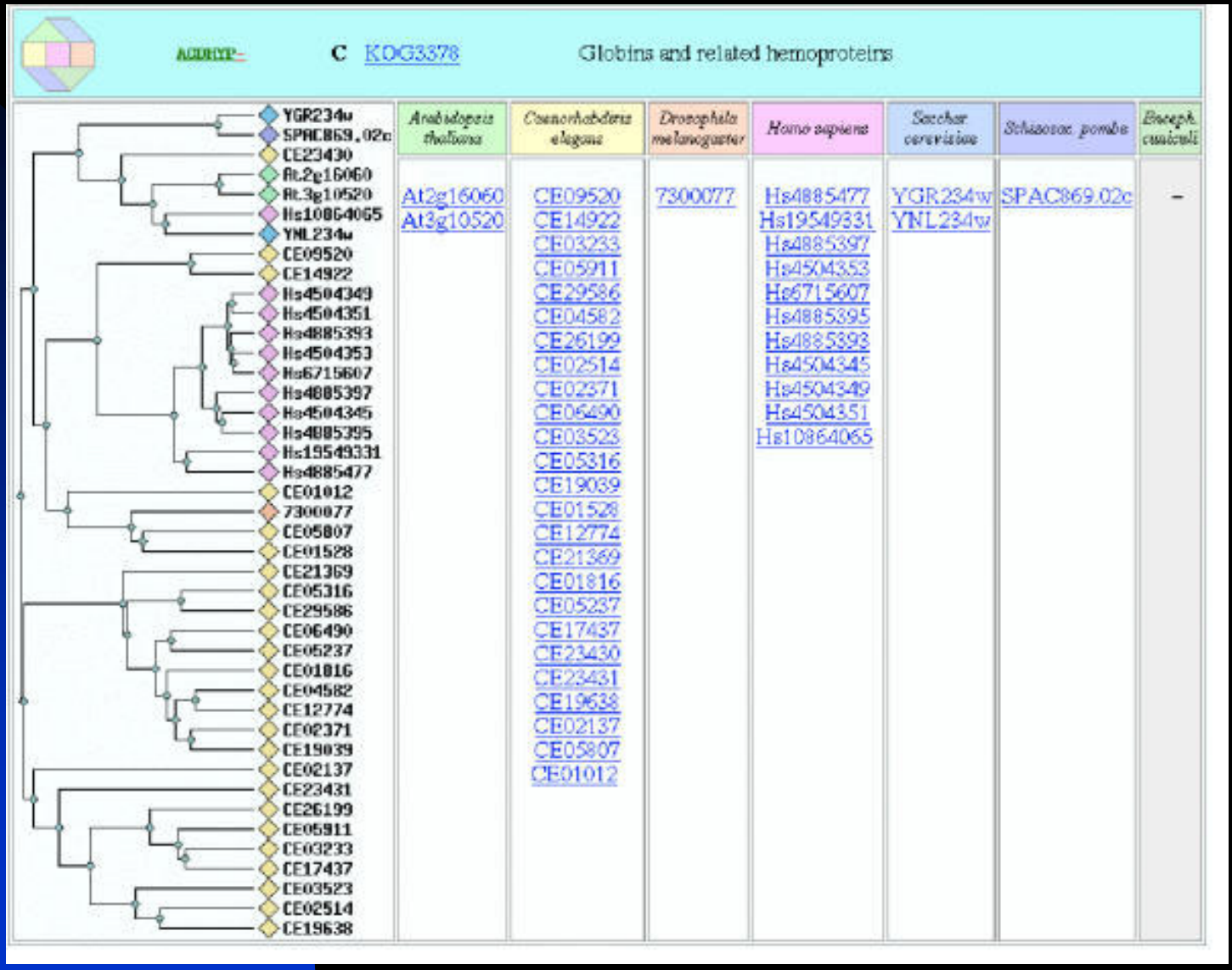
- KOGs were constructed from annotated proteins encoded in the genomes

The eukaryotic orthologous groups (KOGs)

- include proteins from 7 eukaryotic genomes: three animals (the nematode *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster* and *Homo sapiens*), one plant, *Arabidopsis thaliana*, two fungi (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*), and the intracellular microsporidian parasite *Encephalitozoon cuniculi*

KOG

- The current KOG set consists of 4852 clusters of orthologs, which include 59,838 proteins, or ~54% of the analyzed eukaryotic 110,655 gene products


C KOG3378 Globins and related hemoproteins
YGR234w
SPAC869.02c
CE23430
At2g16060
At3g10520
Hs10864065
YNL234w
CE09520
CE14922
Hs4504349
Hs4504351
Hs4885393
Hs4504353
Hs6715607
Hs4885397
Hs4504345
Hs4885395
Hs19549331
Hs4885477
CE01012
7300077
CE05807
CE01528
CE21369
CE05316
CE29586
CE06490
CE05237
CE01816
CE04582
CE12774
CE02371
CE19039
CE02137
CE23431
CE26199
CE05911
CE03233
CE17437
CE03523
CE02514
CE19638
Arabidopsis thaliana
Caenorhabditis elegans
Drosophila melanogaster
Homo sapiens
Saccharomyces cerevisiae
Schizosaccharomyces pombe
Encephalitozoon cuniculi
At2g16060
At3g10520
CE09520
CE14922
CE03233
CE05911
CE29586
CE04582
CE26199
CE02514
CE02371
CE06490
CE03523
CE05316
CE19039
CE01528
CE12774
CE21369
CE01816
CE05237
CE17437
CE23430
CE23431
CE19638
CE02137
CE05807
CE01012
7300077
Hs4885477
Hs19549331
Hs4885397
Hs4504353
Hs6715607
Hs4885395
Hs4885393
Hs4504345
Hs4504349
Hs4504351
Hs10864065
YGR234w
YNL234w
SPAC869.02c
-

General definitions and concepts of evolutionary scenarios

- i) gene loss
- ii) emergence of a new gene (COG)
- iii) acquisition of a gene (COG)

Functional categories of genes lost in Metazoa.

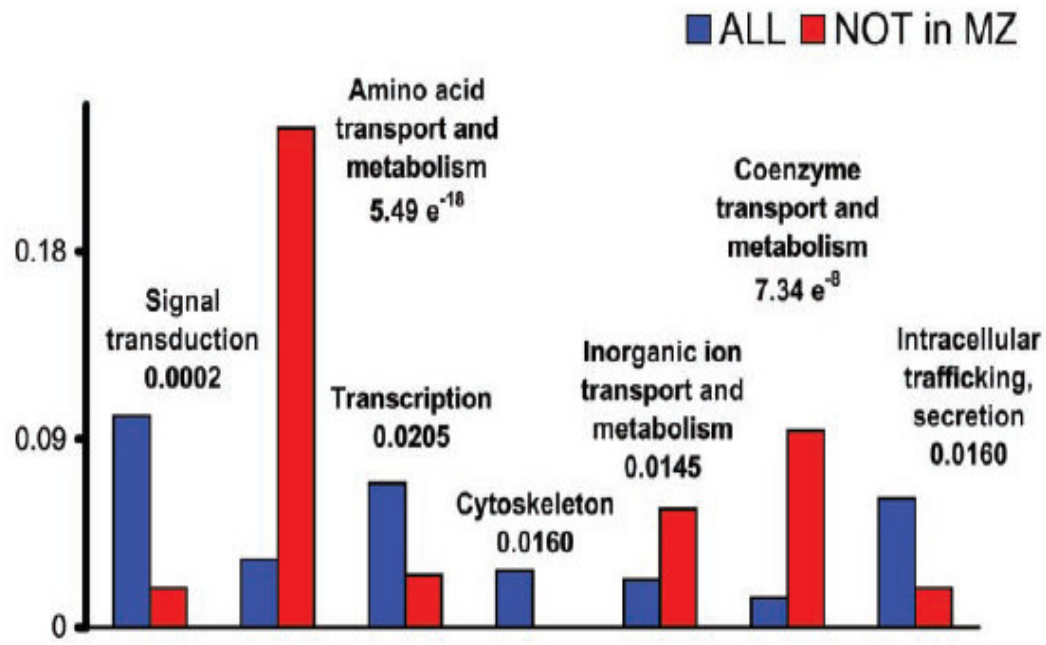
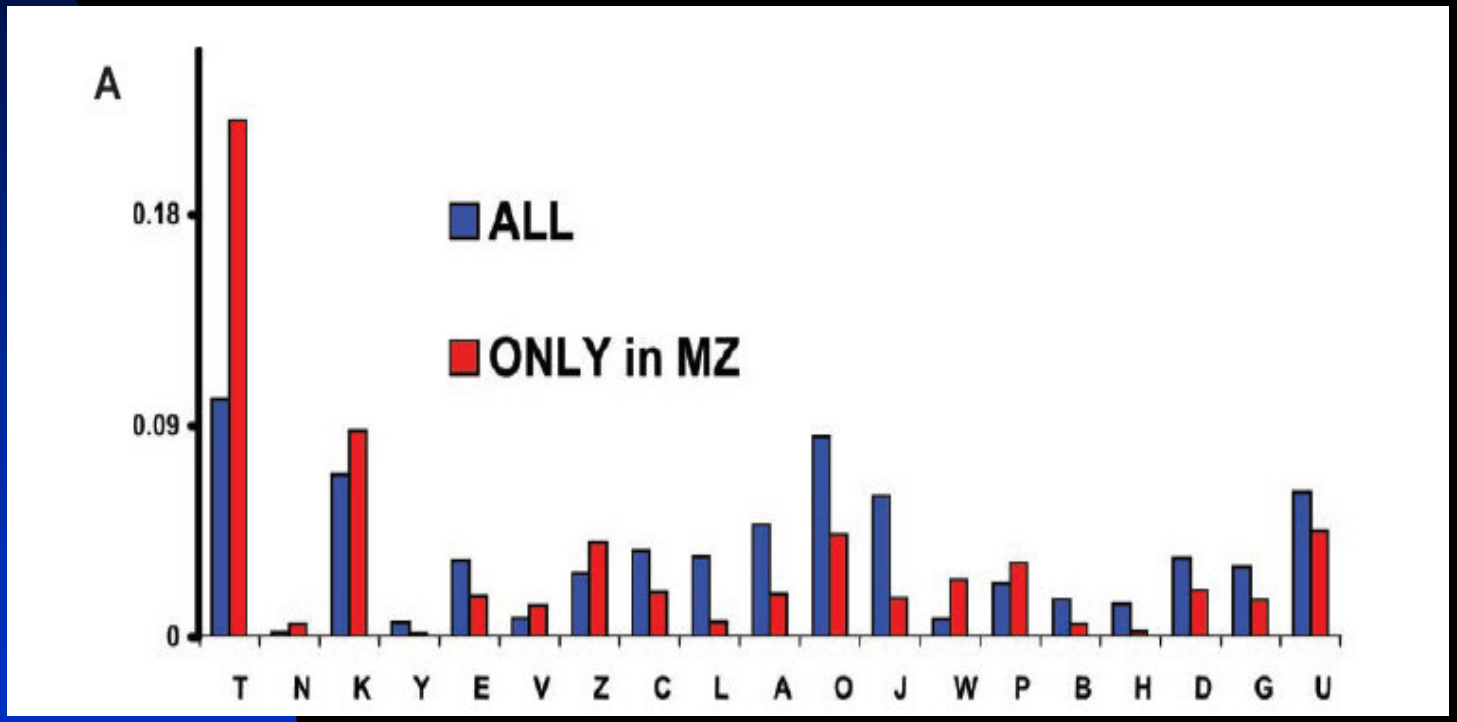
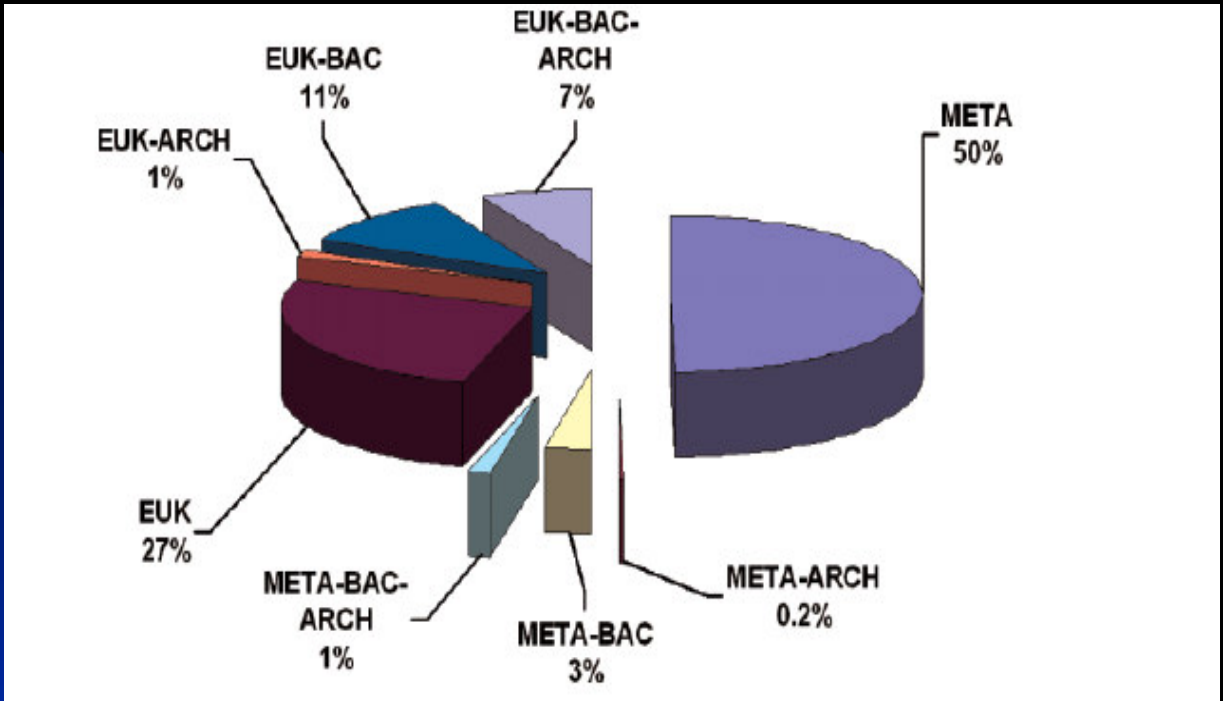


Figure 1. Functional categories of genes lost in Metazoa. Gene loss specific to Metazoa is analyzed for distribution among functional categories. The fraction of lost genes belonging to each category (red bars) is compared with the fraction of genes among all studied gene families belonging to the same functional category (blue bars). Categories for which there is a statistically significant ($P < 0.05$) difference within a functional category are shown. The significance value (P-value) is shown above the bars.

Fuctional categories of metazoa specific genes

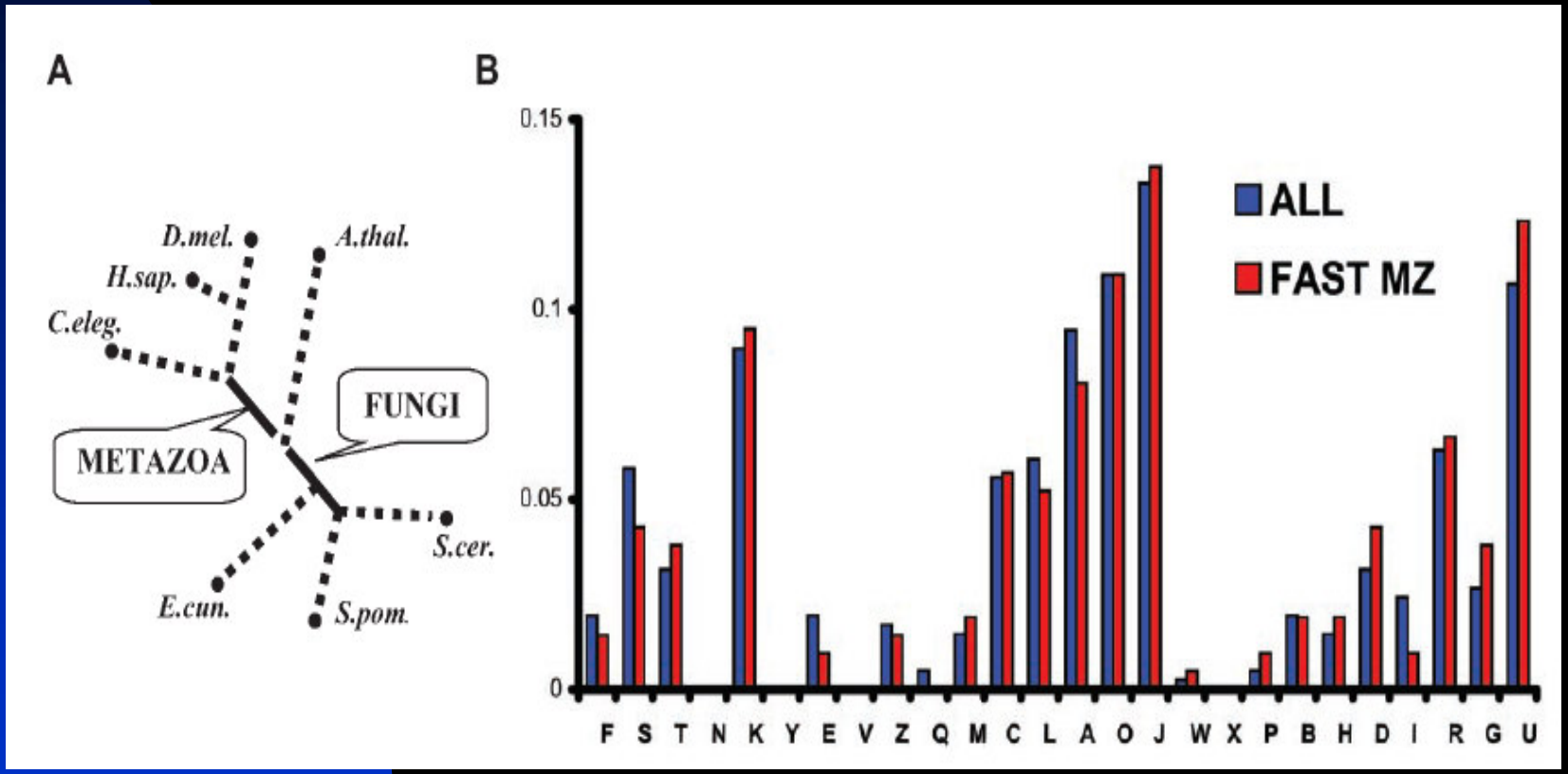


The fraction of Metazoa-specific genes belonging to each category (red bars) is compared with the fraction of genes among all studied gene families belonging to the same functional category (blue bars).

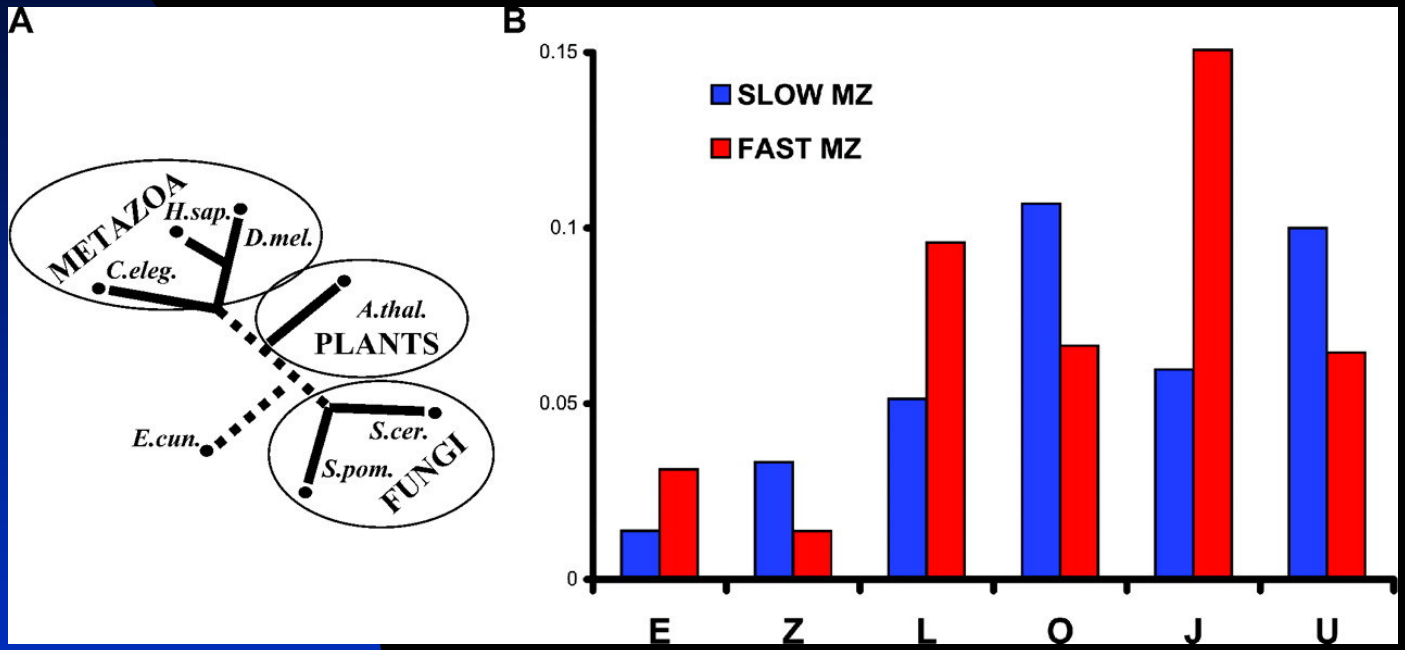


difference within a functional category are shown. The functional categories are abbreviated as follows, the significance value (*P*-value) is in parenthesis: T, Signal transduction mechanisms (1.64×10^{-21}); N, Cell motility (0.0293); K,Transcription (0.0160); Y, Nuclear structure (0.0266); E, Amino acid transport and metabolism (0.0026); V, Defense mechanisms (0.0454); Z, Cytoskeleton (0.0097); C, Energy production and conversion (0.0009); L, Replication, recombination and repair (2.47×10^{-9}); A, RNA processing and modification (8.27×10^{-7}); O, Posttranslational modification, protein turnover, chaperones (8.02×10^{-7}); J, Translation, ribosomal structure and biogenesis (2.08×10^{-11}); W, Extracellular structures (0.000002); P, Inorganic ion transport and metabolism (0.0420); B, Chromatin structure and dynamics (0.0020); H, Coenzyme transport and metabolism (0.00009); D, Cell cycle control, cell division, chromosome partitioning (0.0075); G, Carbohydrate transport and metabolism (0.0037); U, Intracellular trafficking, secretion and vesicular transport (0.0217). (B) The origin of Metazoa-specific proteins. Metazoa-specific proteins without direct orthologs in other taxa are analyzed for detectable homologues in the following groups: META, only in Metazoa; MET-BAC, in Metazoa and bacteria; META-BAC-ARCH, in Metazoa, bacteria and Archaea; EUK-ARCH, in different eukaryotes and Archaea; EUK, in different eukaryotes only; EUK-BAC, in different eukaryotes and bacteria; EUK-BAC-ARCH, in different eukaryotes, bacteria and Archaea; EUK-ARCH, in different eukaryotes and Archaea.

Accelerated evolutionary rates in Metazoan proteins



Evolutionary rates in the separate lineages of Metazoa



Nucleic Acids Research

Table 1. Correlation coefficients between protein and DNA substitutions

Protein–DNA	r with K_n	r with K_s	d.f.
Homo sapiens/ *Mus musculus*	0.284 ($<10^{-4}$)	0.068 (0.0859)	637
Homo sapiens/ *Rattus norvegicus.*	0.261 ($<10^{-4}$)	0.095 (0.0202)	597
Drosophila melanogaster/ *Anopheles gambiae*	0.339 ($<10^{-4}$)	0.094 (0.0134)	688
Ceanorhabditis elegans/ *Caenorhabditis briggsae*	0.285 ($<10^{-4}$)	0.099 (0.0109)	659

The Pearson correlation coefficients (r) between the rate of amino acid substitution averaged over three major Metazoan phyla (protein) and the rate of DNA mutations (DNA), non-synonymous (K_n) and synonymous (K_s), measured in closely related Metazoan lineages. The significance values of each (r) are provided in parentheses, while the degrees of freedom (d.f.) for the t-test are listed in the far right column.

Table 2. K_n in slow and rapidly evolving Metazoan genes

	Slow	Rapid	Probability
Homo sapiens/ *Mus musculus*	0.049	0.063	0.005
Homo sapiens/ *Rattus norvegicus*	0.057	0.068	0.046
Drosophila melanogaster/ *Anopheles gambiae*	1.333	1.381	0.006
Ceanorhabditis elegans/ *Caenorhabditis briggsae*	0.097	0.116	0.026

The average K_n value was calculated for three pairs of species, separately for 'slow' and 'rapid' genes, designated by protein sequence substitution rates. Genes with a high rate of protein substitution had a higher K_n in all three cases. This difference in K_n is shown to be statistically significant by Student's *t*-test (column 4 'Probability'). The null hypothesis was that both 'slow' and 'rapid' groups have the same mean K_n.

The End