The Abdus Salam
International Centre
for Theoretical Physics
50th Anniversary 1964-2014

2584-10

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

A Genome as a Toolbox: intro

Marco Cosentino Lagomarsino
Université Pierre et Marie Curie
Paris

A Genome as a Toolbox: intro

June 2nd 2014

Spring School, Trieste

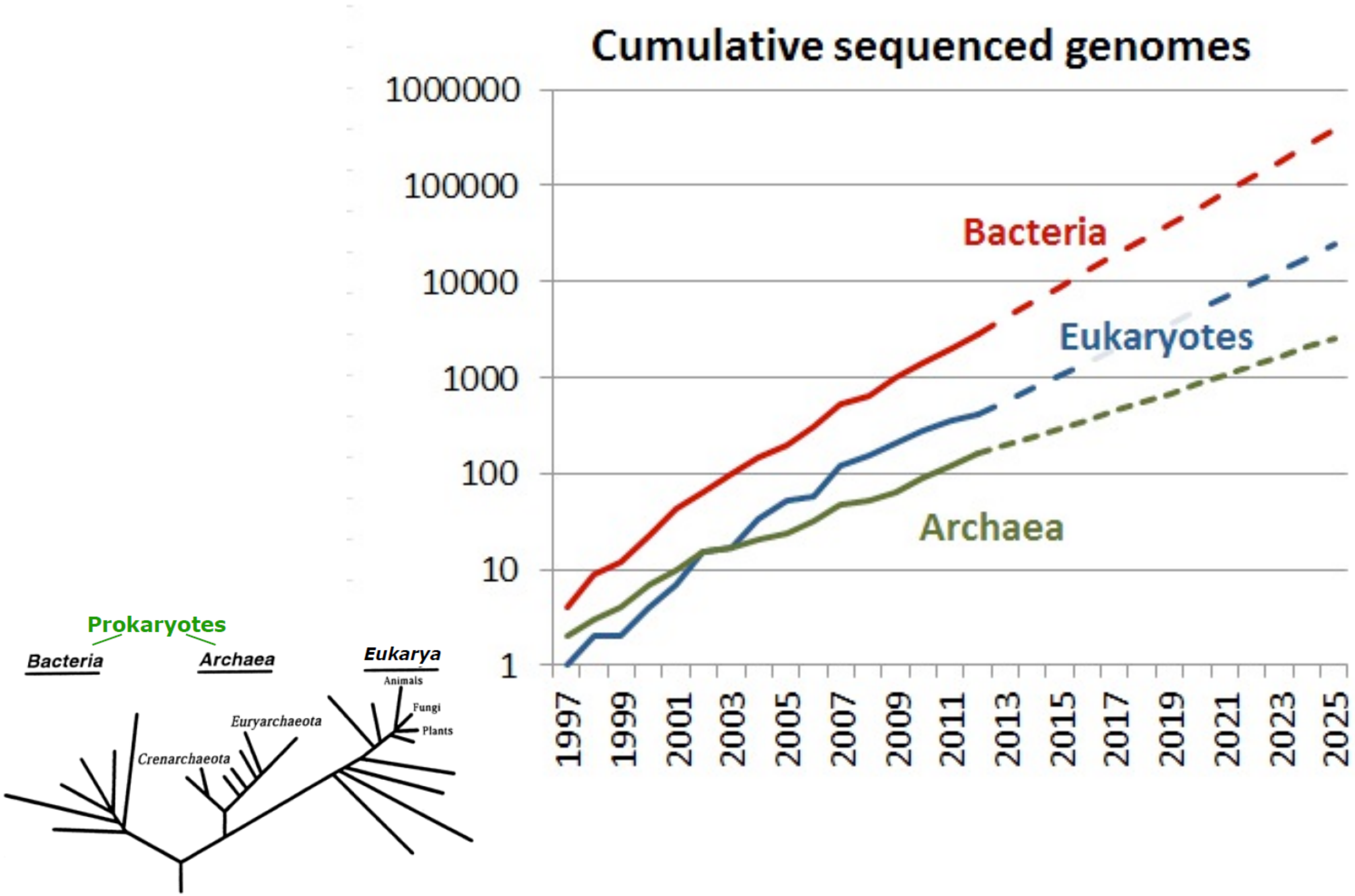
Marco Cosentino Lagomarsino

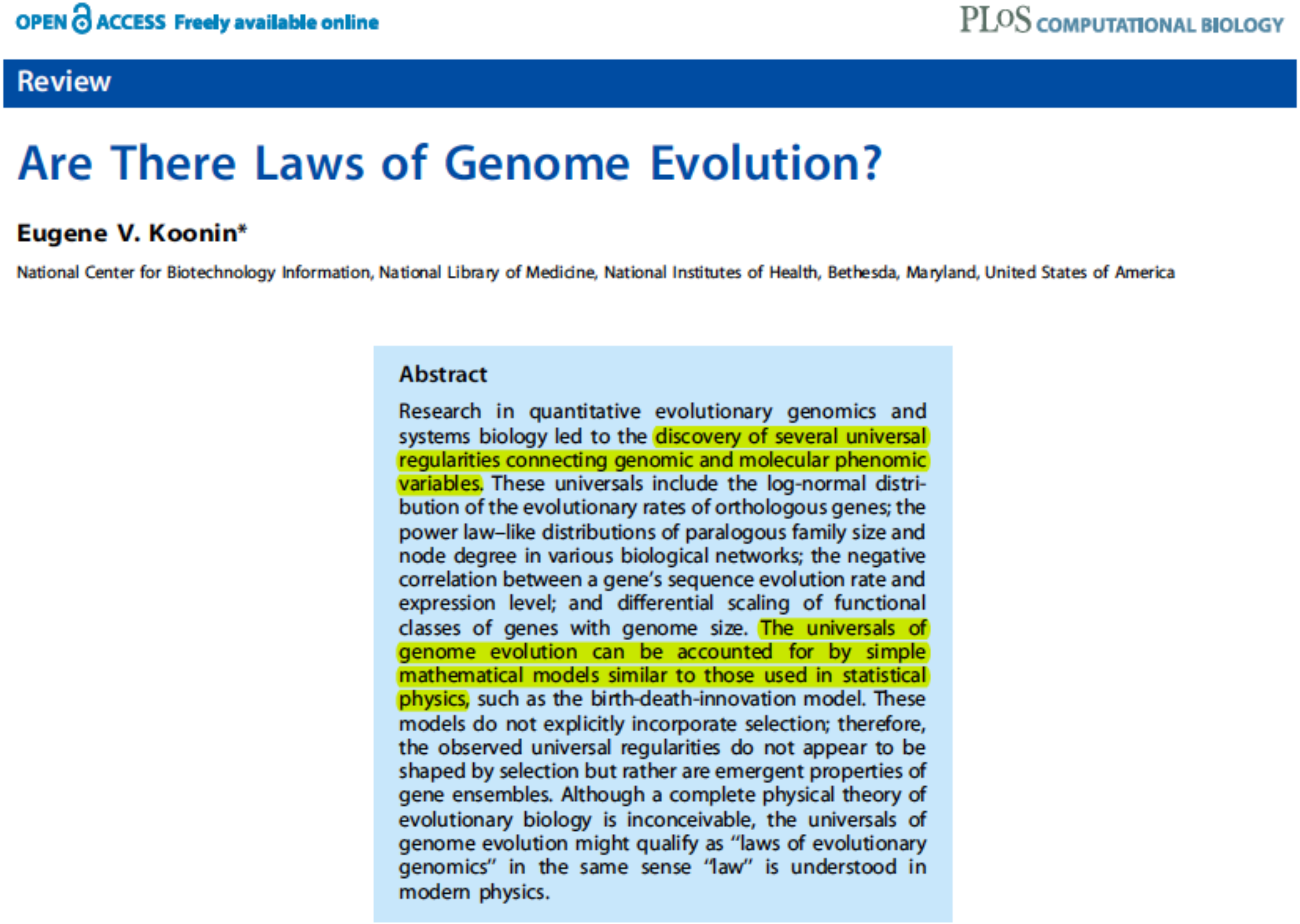
Génophysique / Genomic Physics Group

CNRS “Microorganism Genomics” UMR7238 Laboratory
Université Pierre et Marie Curie, Paris

0) Are there “laws” in genome evolution?

Genomes give abundant data





OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Review

Are There Laws of Genome Evolution?

Eugene V. Koonin*

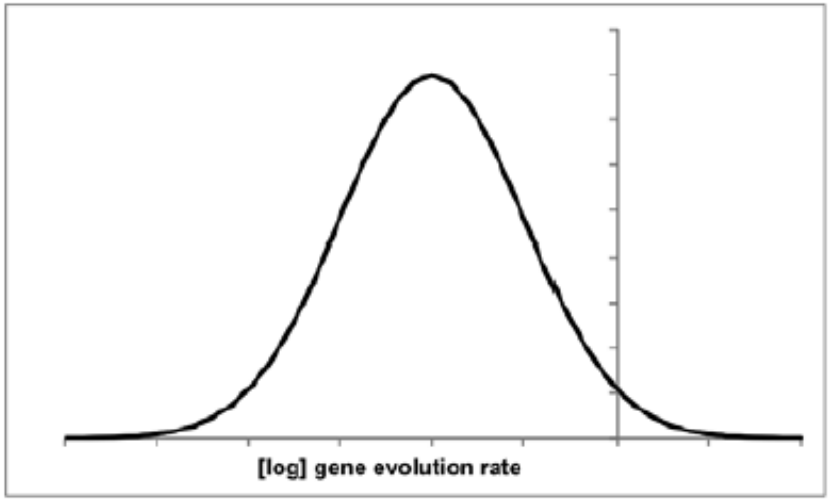
National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America

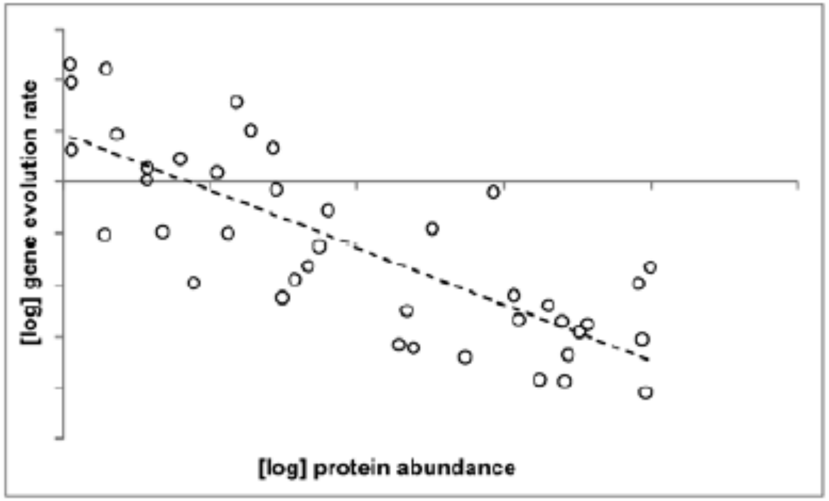
Abstract

Research in quantitative evolutionary genomics and systems biology led to the discovery of several universal regularities connecting genomic and molecular phenomic variables. These universals include the log-normal distribution of the evolutionary rates of orthologous genes; the power law–like distributions of paralogous family size and node degree in various biological networks; the negative correlation between a gene's sequence evolution rate and expression level; and differential scaling of functional classes of genes with genome size. The universals of genome evolution can be accounted for by simple mathematical models similar to those used in statistical physics, such as the birth-death-innovation model. These models do not explicitly incorporate selection; therefore, the observed universal regularities do not appear to be shaped by selection but rather are emergent properties of gene ensembles. Although a complete physical theory of evolutionary biology is inconceivable, the universals of genome evolution might qualify as "laws of evolutionary genomics" in the same sense "law" is understood in modern physics.

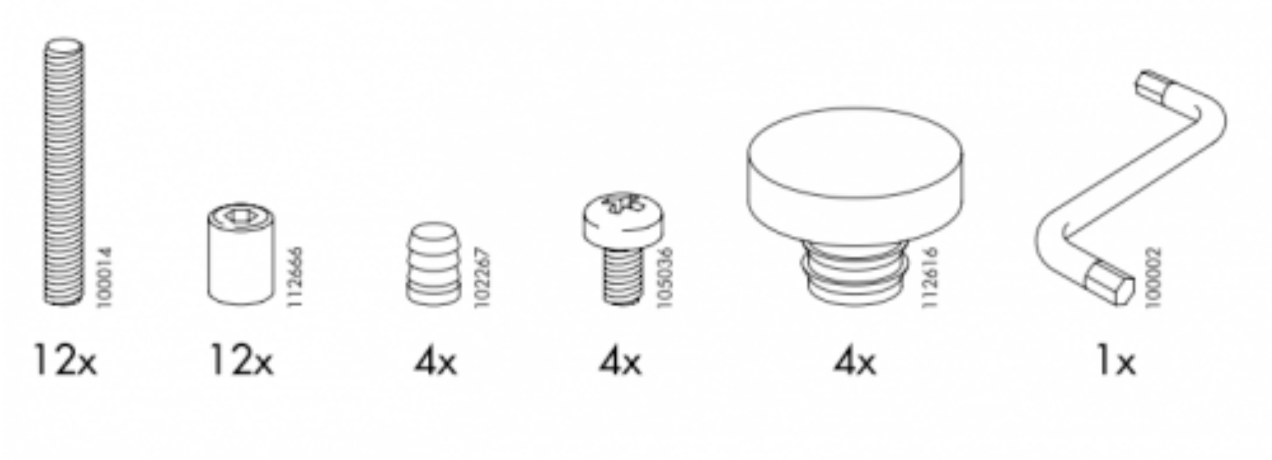
Some interesting "laws"

(Koonin, Hurst, Drummond & Wilke)

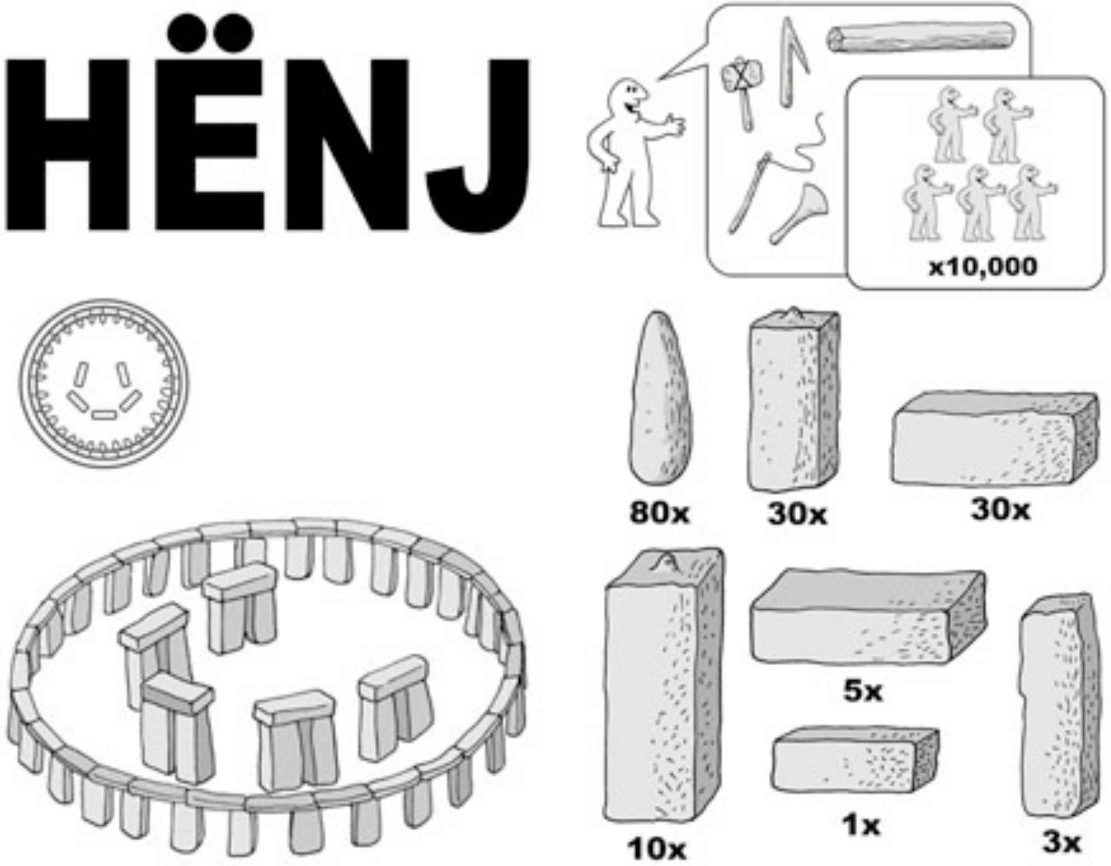




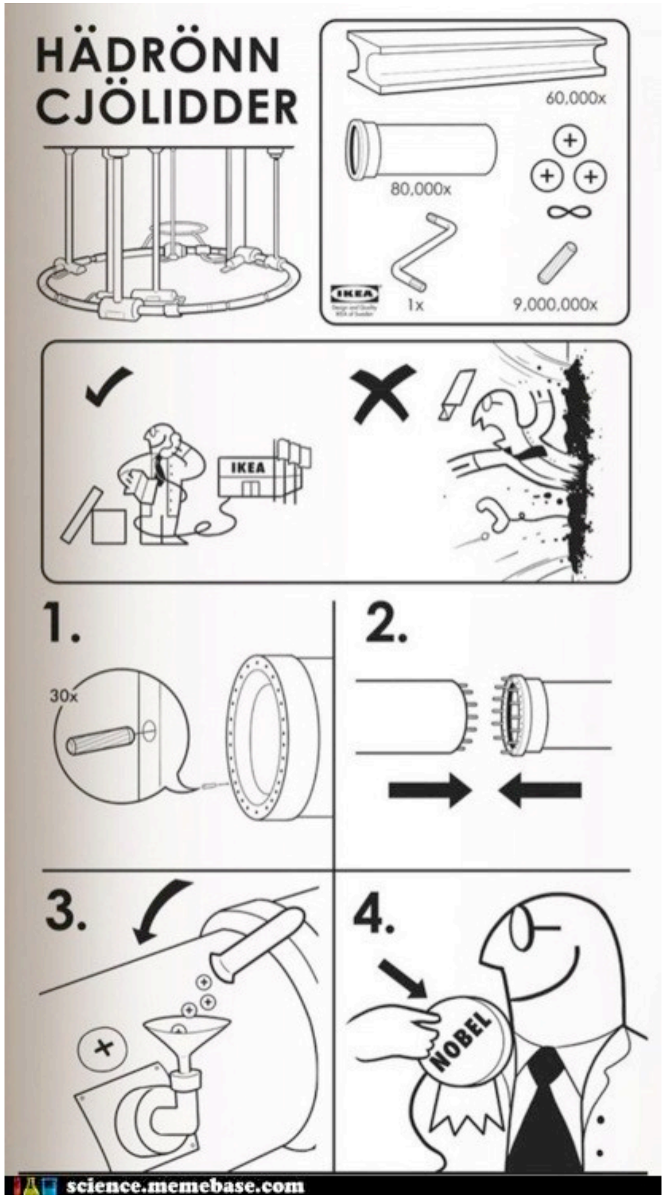
Laws in a genome “parts list”?



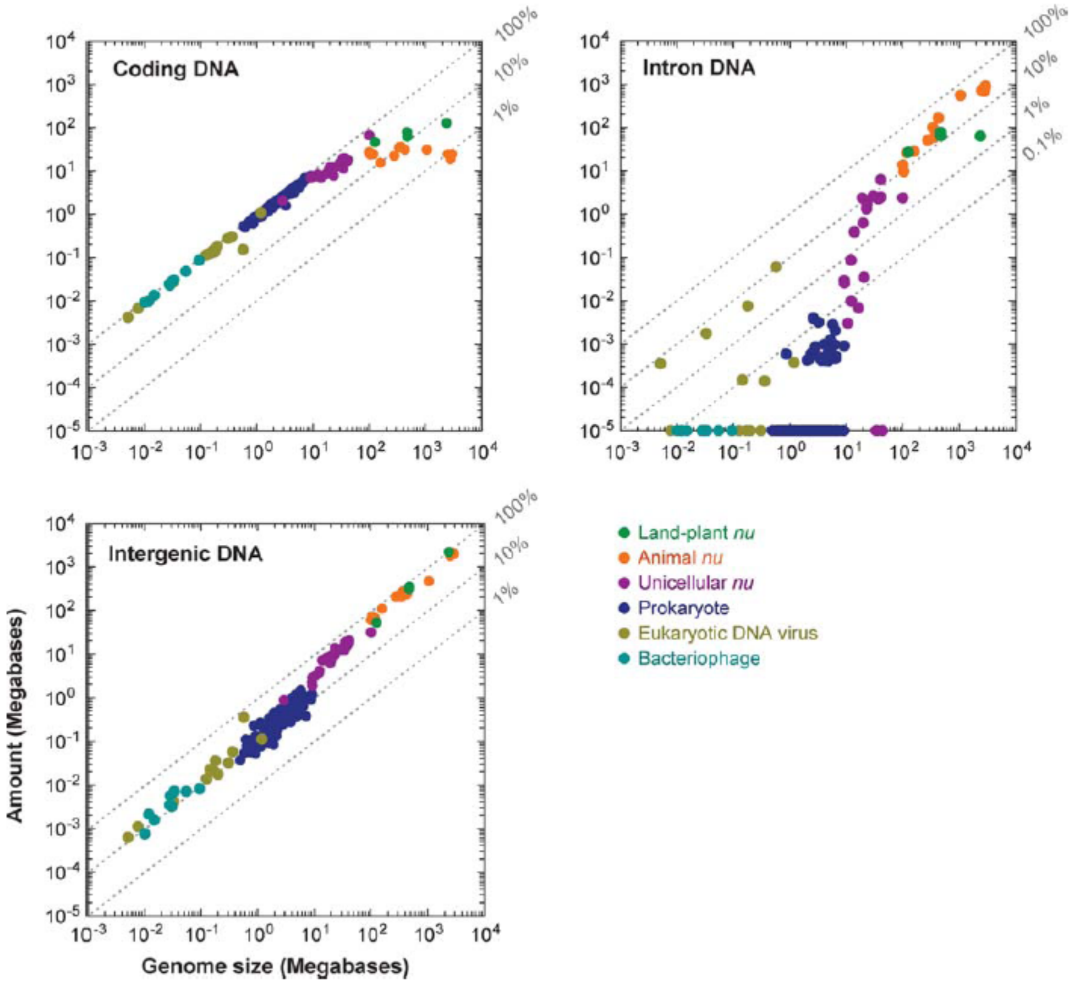
Laws in a genome "parts list"?



Laws in a genome “parts list”?

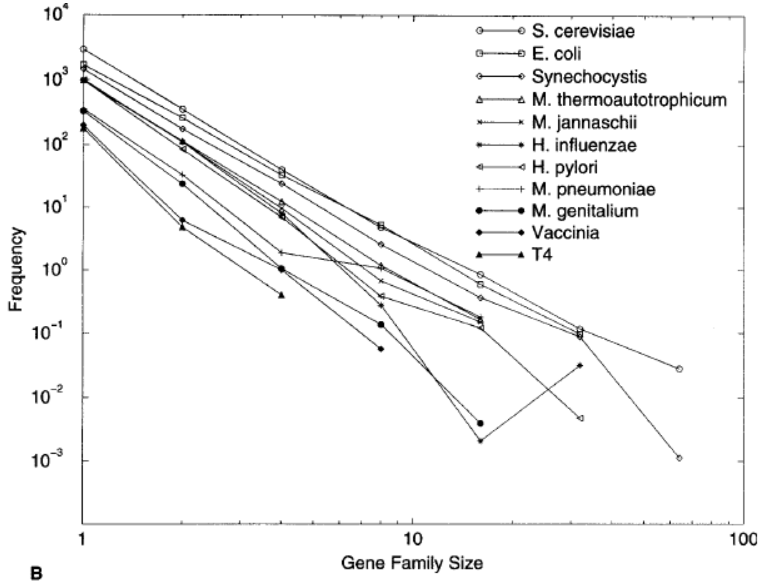


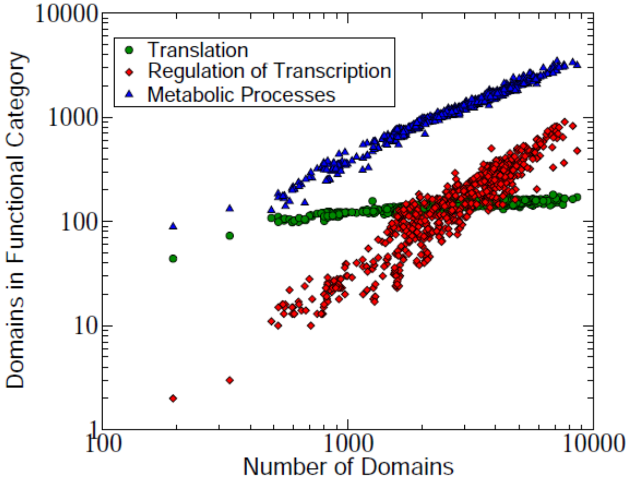
Genomes Show Common Behavior



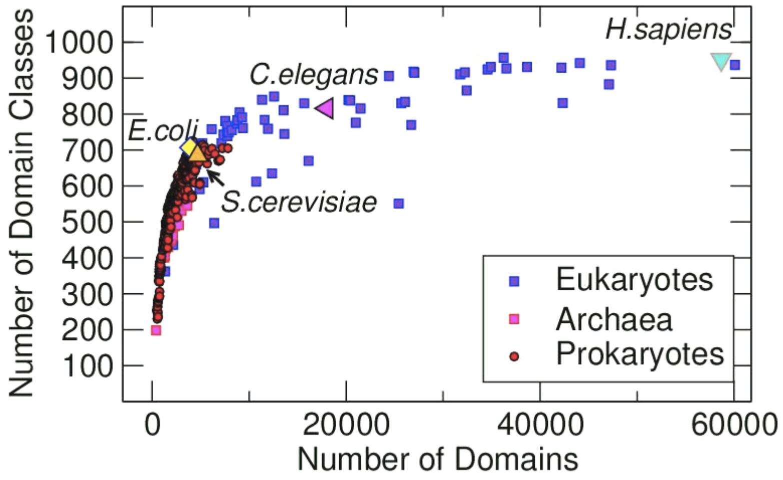
(M. Lynch)

“Laws” in gene content





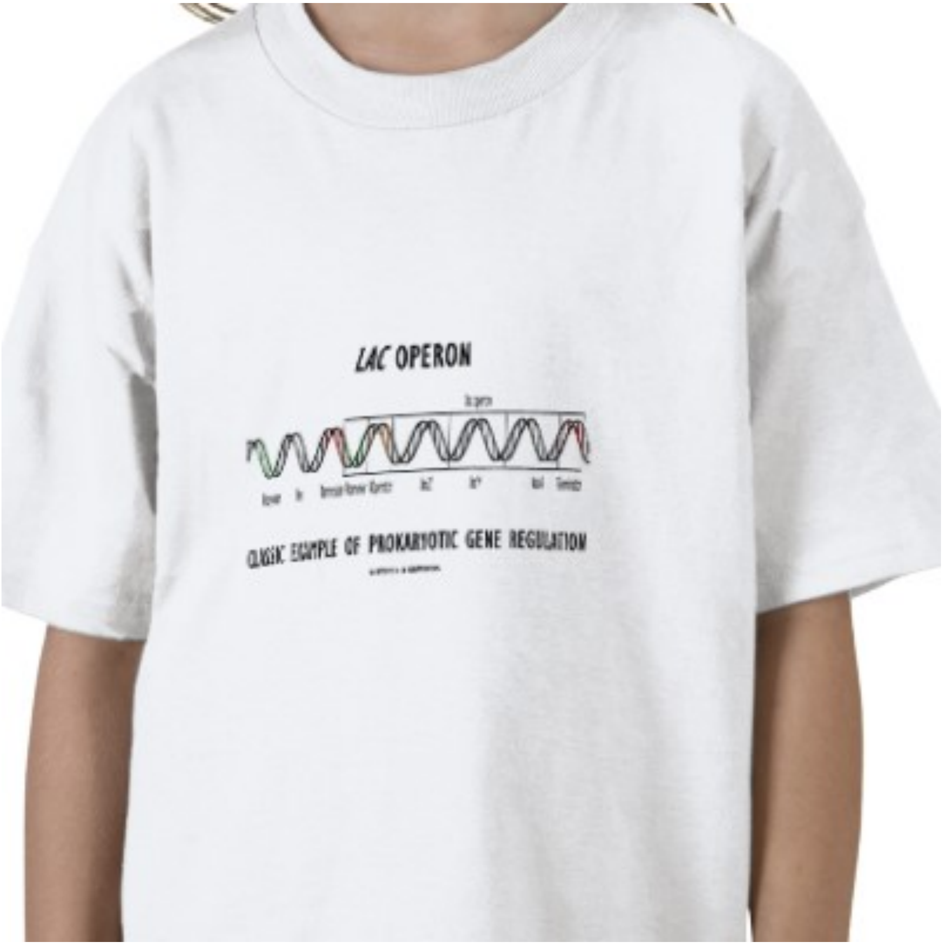
(E.van Nimwegen)



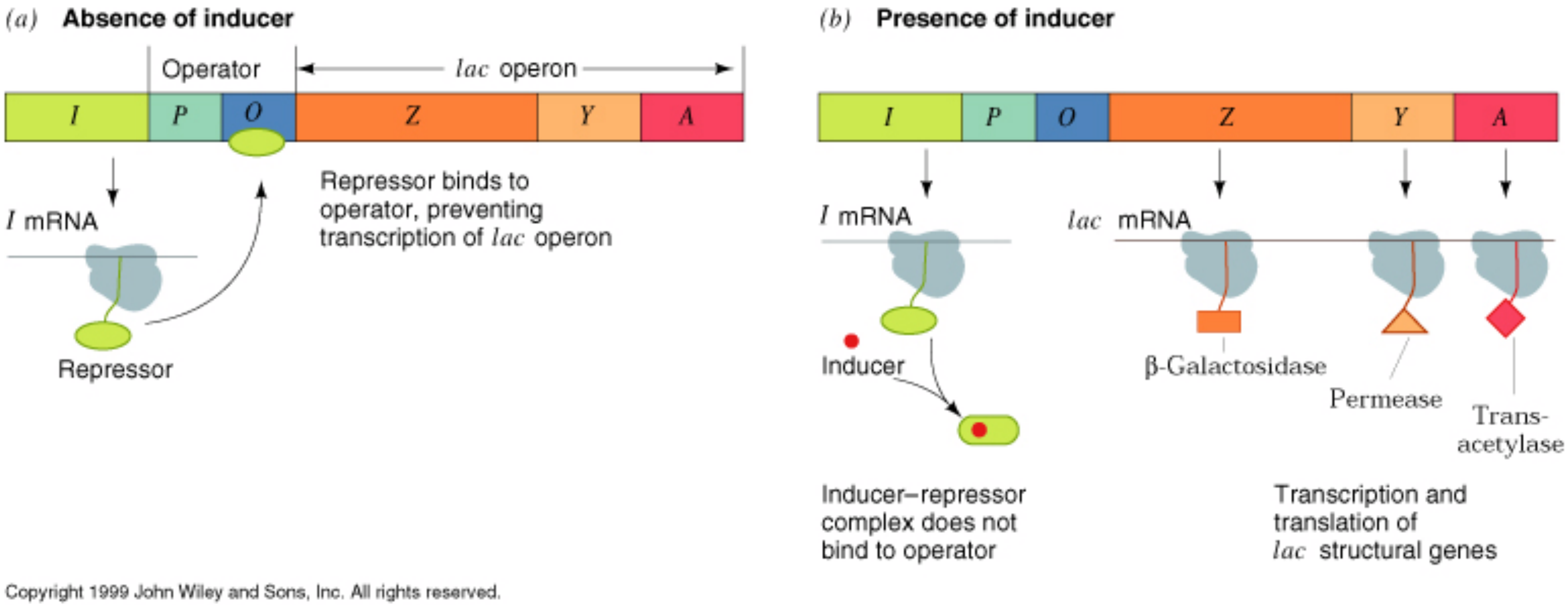
1) Partitioning of a genome into functional categories

(Monod at the genome scale)

Let us start from the Lac Operon



Let us start from the Lac Operon



(*a*) Absence of inducer
Operator
lac operon
I
P
O
Z
Y
A
I mRNA
Repressor
Repressor binds to operator, preventing transcription of *lac* operon
(*b*) Presence of inducer
I
P
O
Z
Y
A
I mRNA
lac mRNA
Inducer
β-Galactosidase
Permease
Trans-acetylase
Inducer–repressor complex does not bind to operator
Transcription and translation of *lac* structural genes

Three functional ingredients

Metabolism (Lactose)

Transcription (Repressor-Operon)

Translation (Physiology / Growth Rate)

REGULATION Information Flow

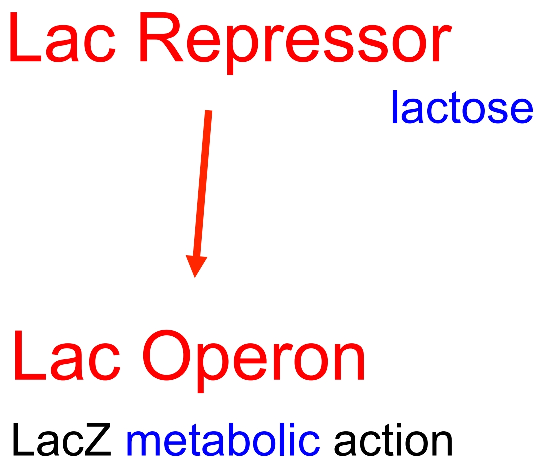
Operon Model

(Jacob & Monod, JMB 1961)

Structural genes
do stuff

Regulatory genes
decide who does what

Parenthesis: Hierarchic vs Circular



Lac Repressor

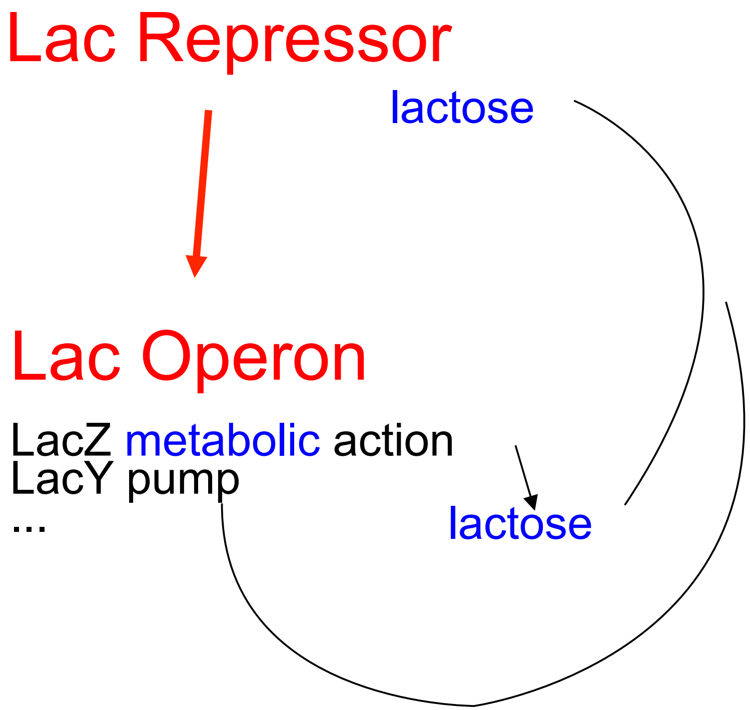
lactose

Lac Operon

LacZ metabolic action

REGULATION Information Flow

Parenthesis: Hierarchic vs Circular



Lac Repressor

lactose

Lac Operon

LacZ metabolic action
LacY pump
...

lactose

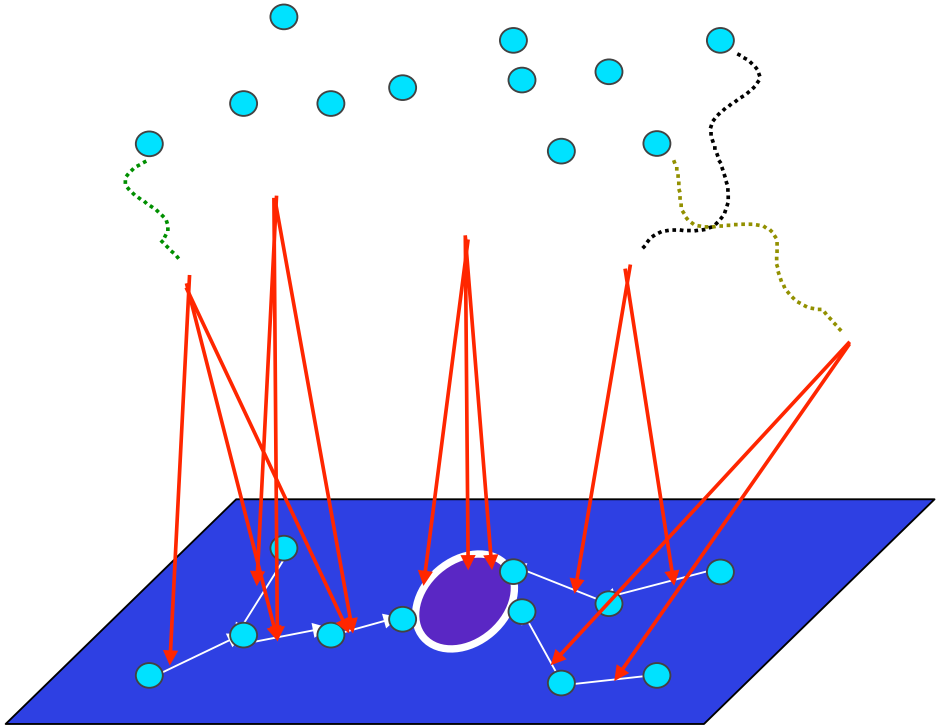
REGULATION Information Flow

Monod at genome scale,
NEEDS STATISTICS

Metabolites

Transcriptional Regulation

Metabolism



(Translation, Physiology / Growth Rate)

Functional Annotations

Transcriptional Regulation

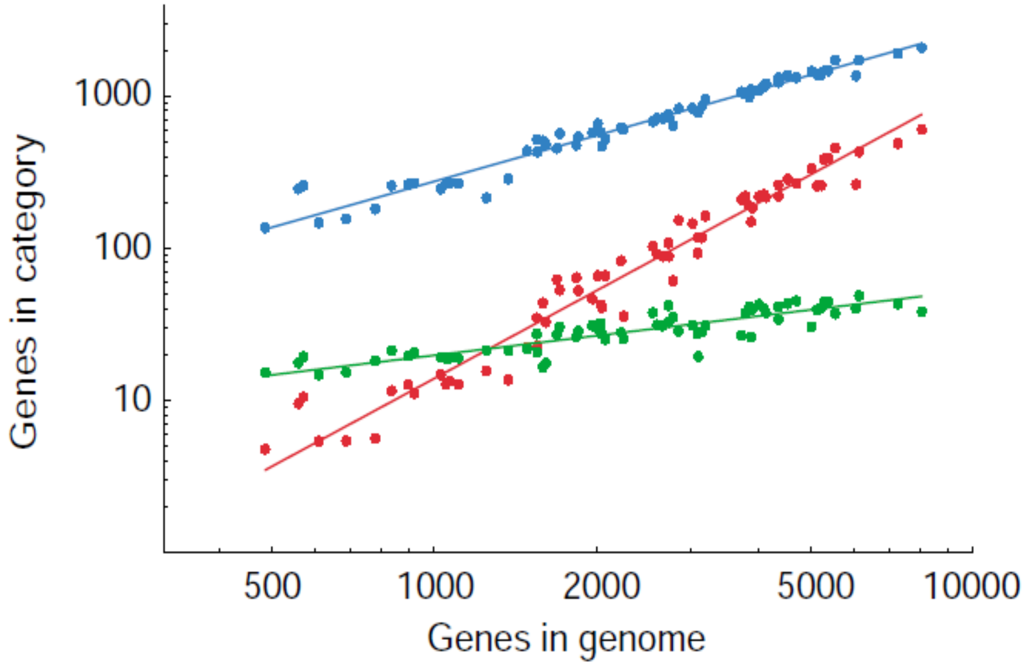
Metabolism

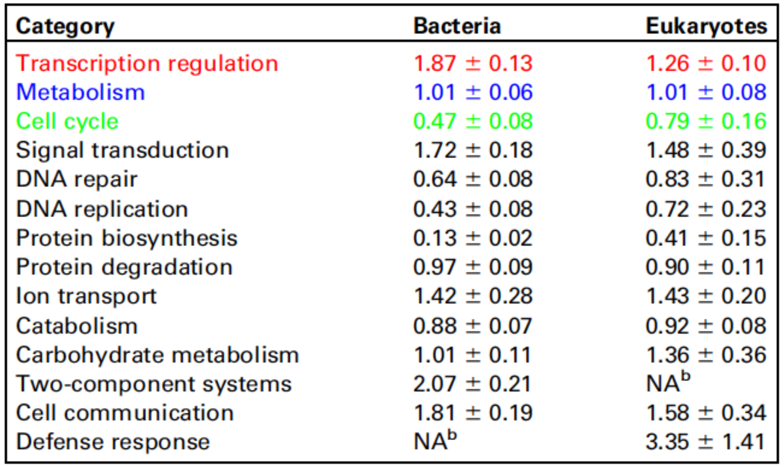
Translation

…

Category counts for many genomes

(E.van Nimwegen, 2003)



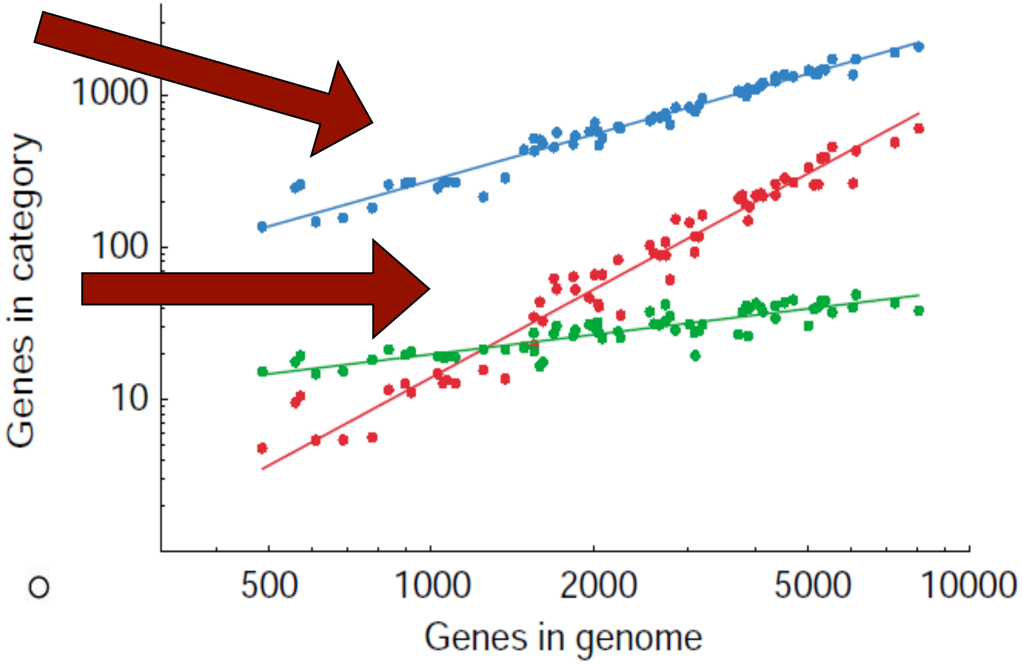


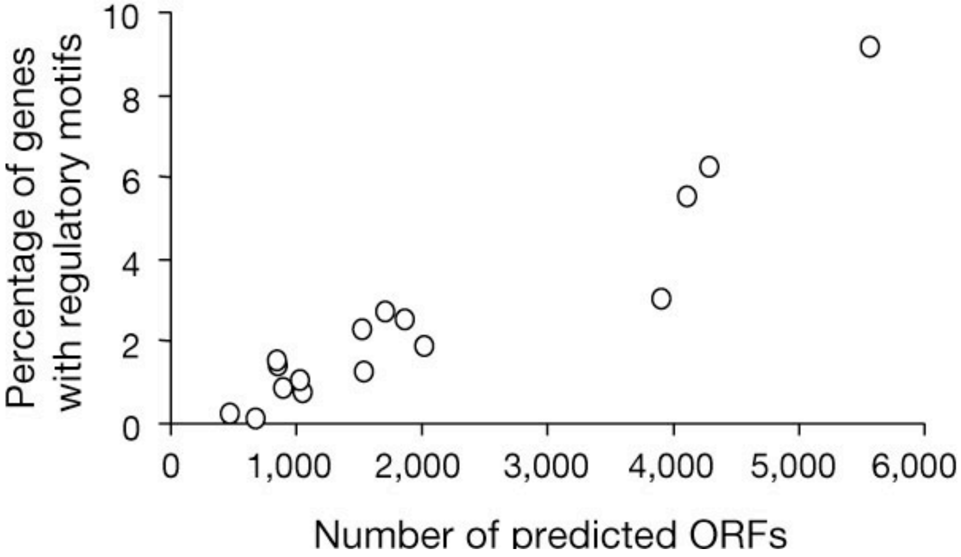
Category	Bacteria	Eukaryotes
Transcription regulation	1.87 ± 0.13	1.26 ± 0.10
Metabolism	1.01 ± 0.06	1.01 ± 0.08
Cell cycle	0.47 ± 0.08	0.79 ± 0.16
Signal transduction	1.72 ± 0.18	1.48 ± 0.39
DNA repair	0.64 ± 0.08	0.83 ± 0.31
DNA replication	0.43 ± 0.08	0.72 ± 0.23
Protein biosynthesis	0.13 ± 0.02	0.41 ± 0.15
Protein degradation	0.97 ± 0.09	0.90 ± 0.11
Ion transport	1.42 ± 0.28	1.43 ± 0.20
Catabolism	0.88 ± 0.07	0.92 ± 0.08
Carbohydrate metabolism	1.01 ± 0.11	1.36 ± 0.36
Two-component systems	2.07 ± 0.21	NA[b]
Cell communication	1.81 ± 0.19	1.58 ± 0.34
Defense response	NA[b]	3.35 ± 1.41

Back to operon model: transcription factors and metabolic enzymes

Constant fraction of Metabolic enzymes

Exponent ~two for transcription factors



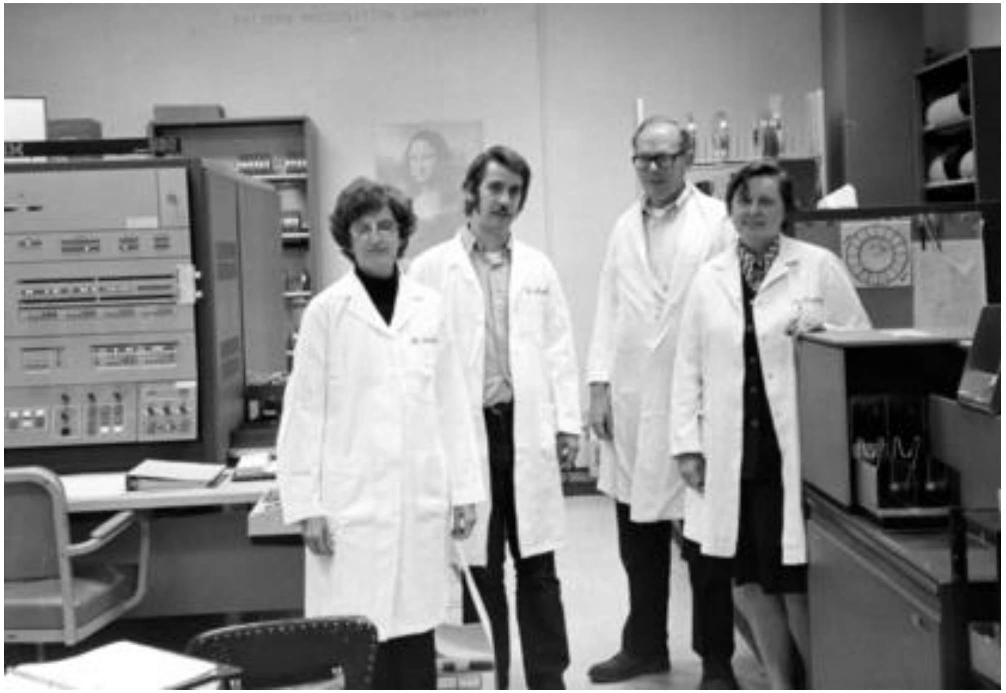


(Stover et al Nature, 2000)

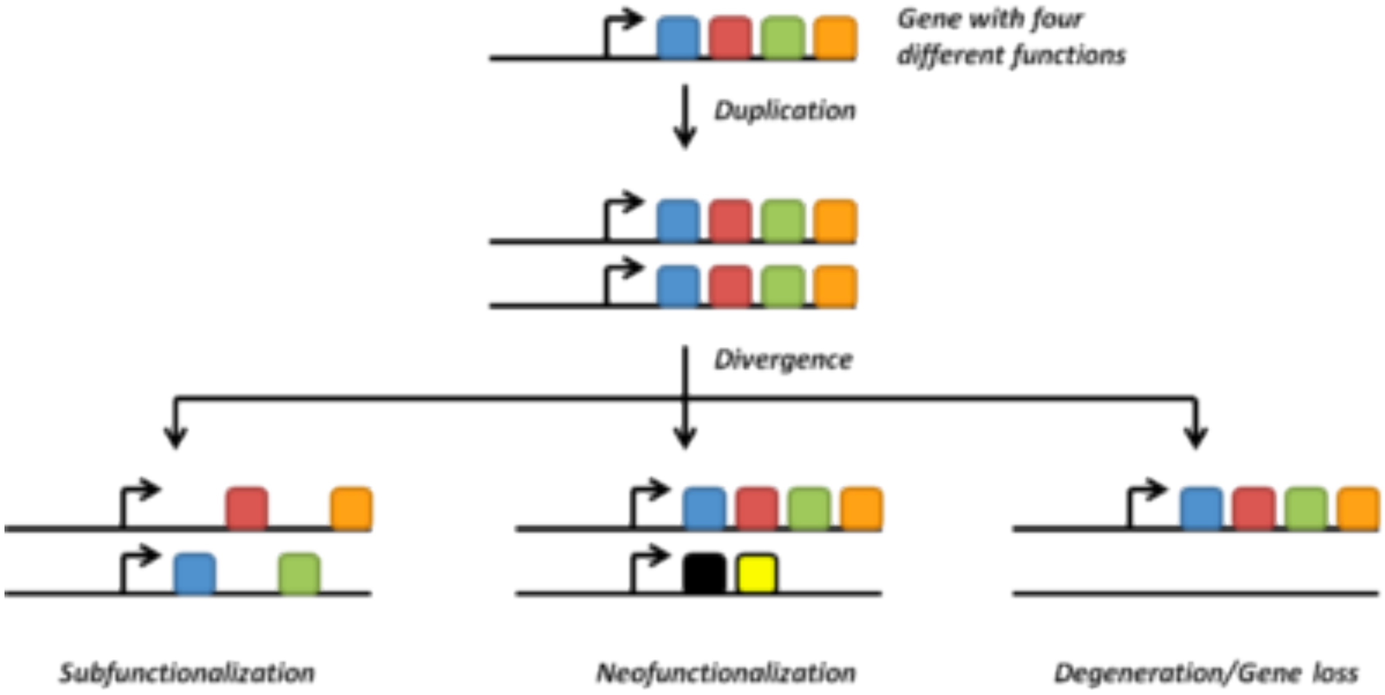
2) Partitioning of a genome into evolutionary families

(Dayhoff's Dream)

Margaret Oakley-Dayhoff

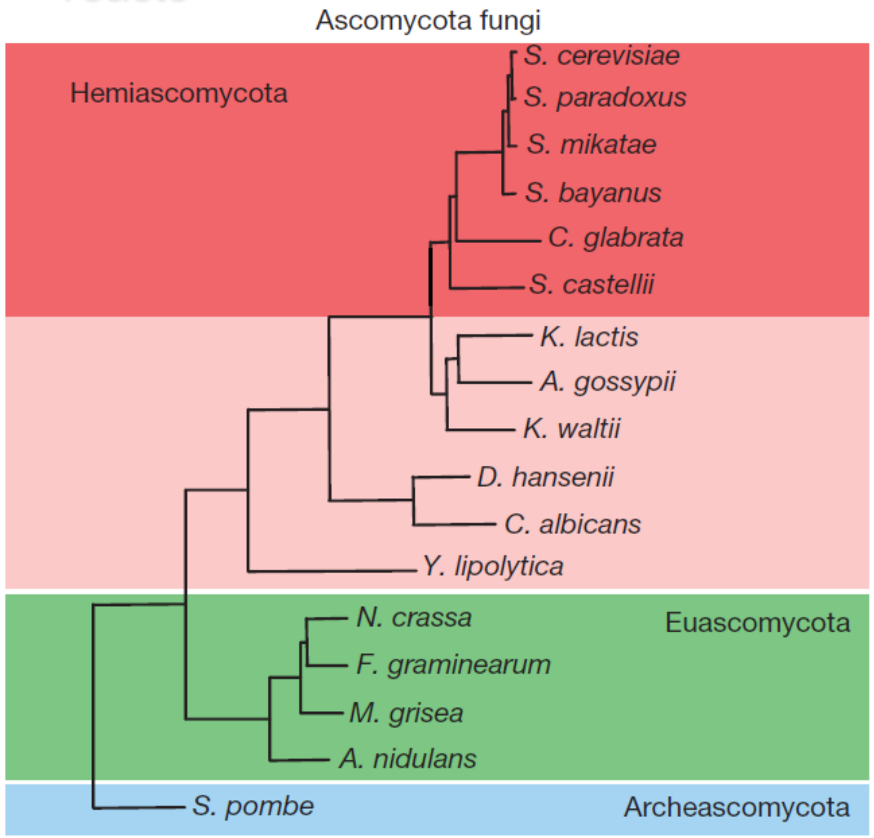


Why evolutionary families? Gene duplication

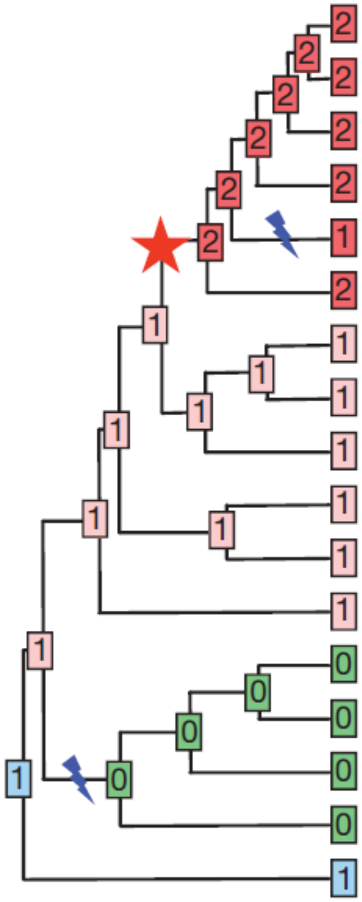


Duplication Track-record

Yeasts

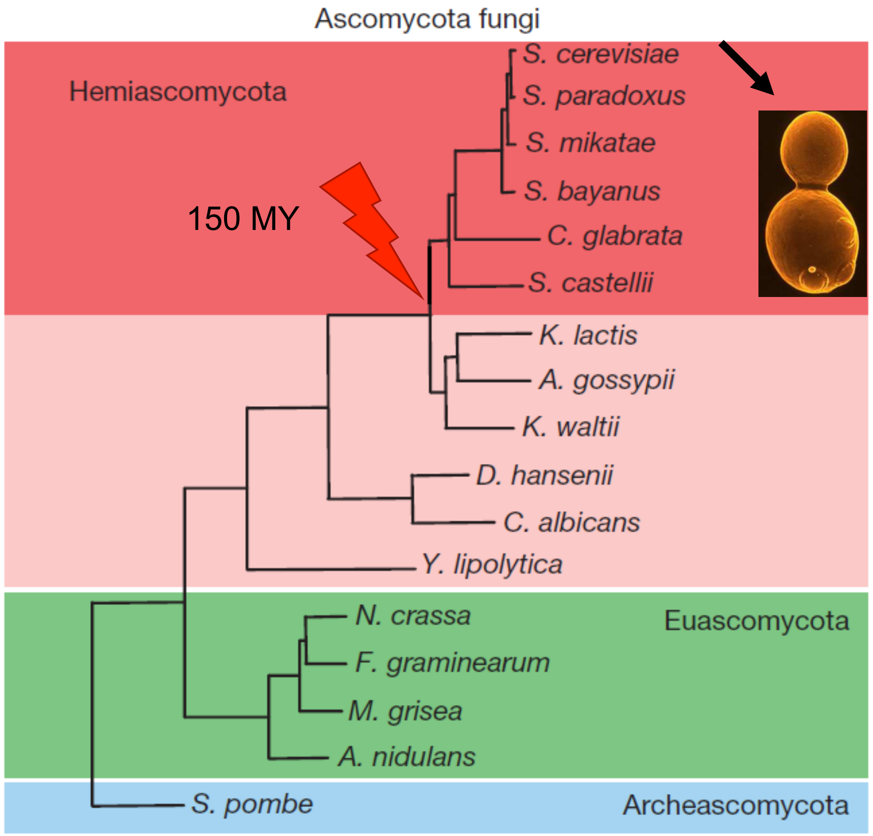


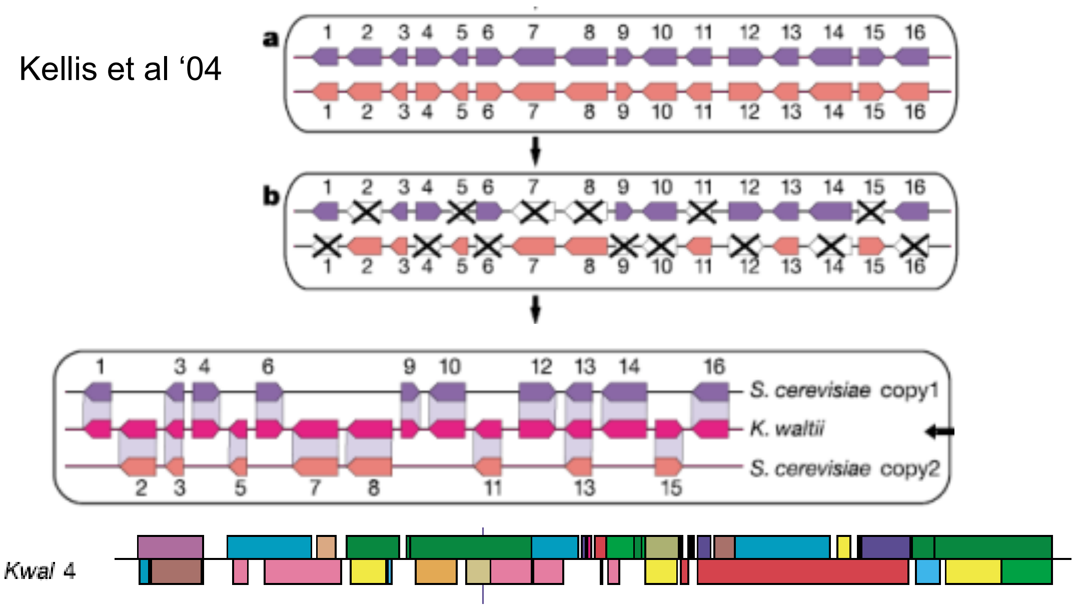
Wapinski et al ‘07



Duplications occur at all scales

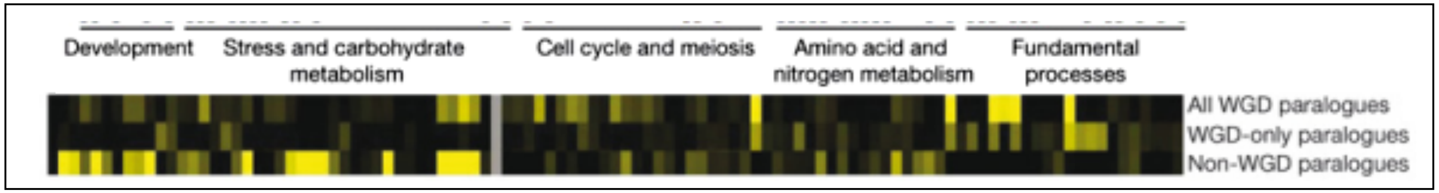
E.G. Yeast Whole-Genome Duplication



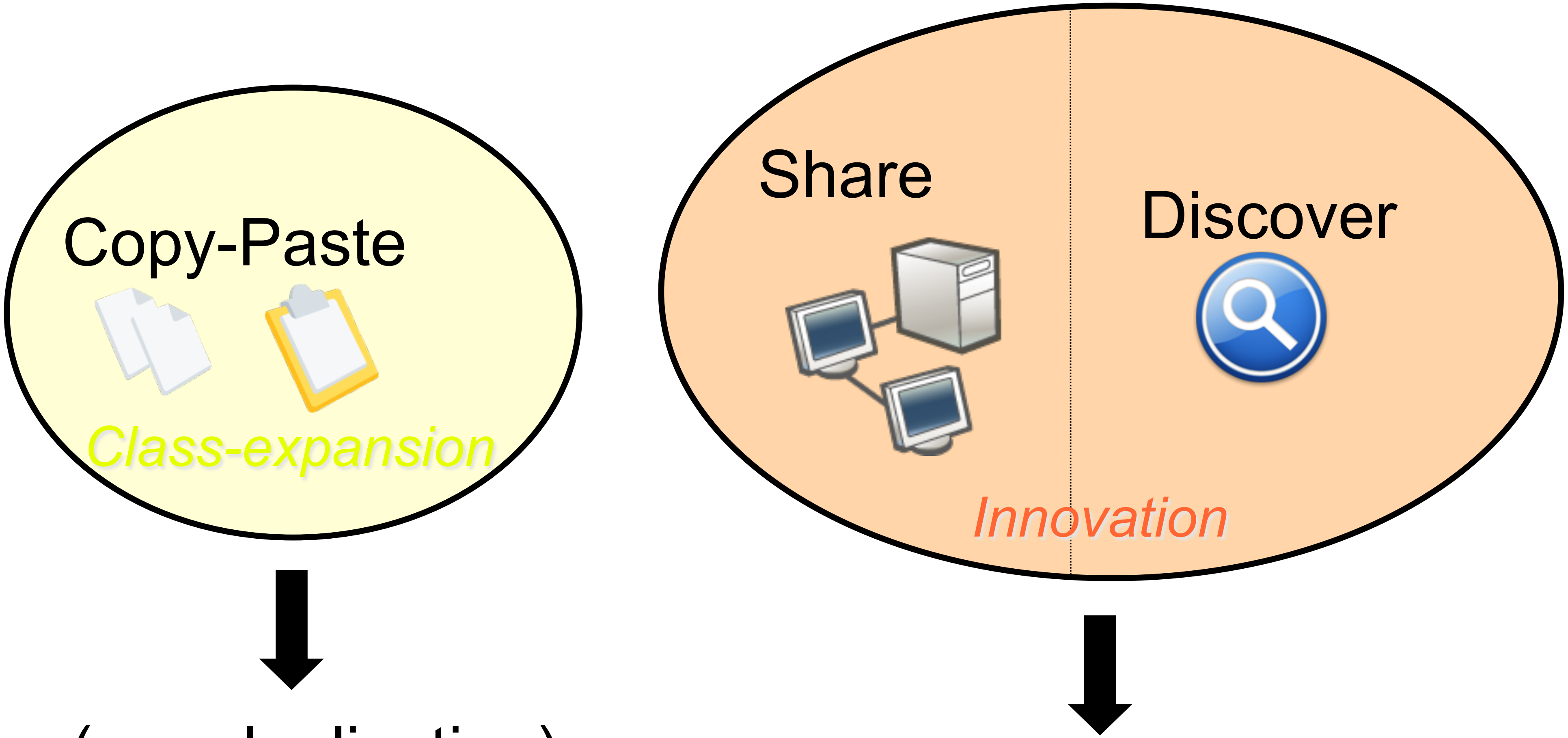

Kellis et al '04

- Important evolutionary process
- Develops new functions
- Characteristic profile of action

Wapinski et al '07



The moves of Genome Evolution



Copy-Paste

Class-expansion

Share

Discover

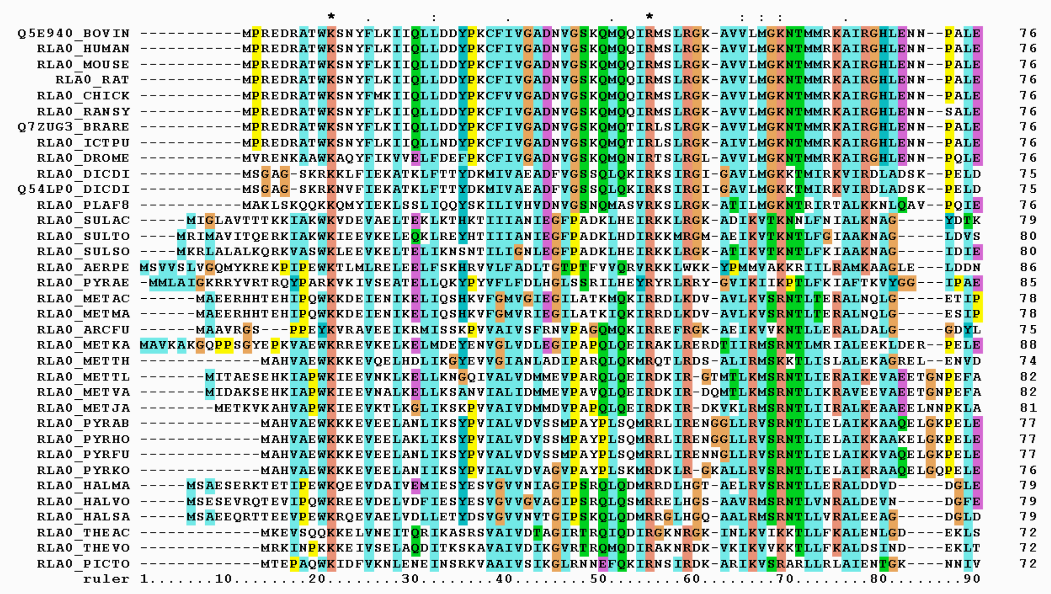
Innovation

(e.g. duplication)
Evolutionary
families of genes

New evolutionary
families of genes

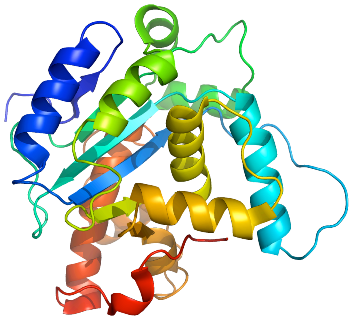
Detection of gene families

Sequence alignments
(threshold dependency)

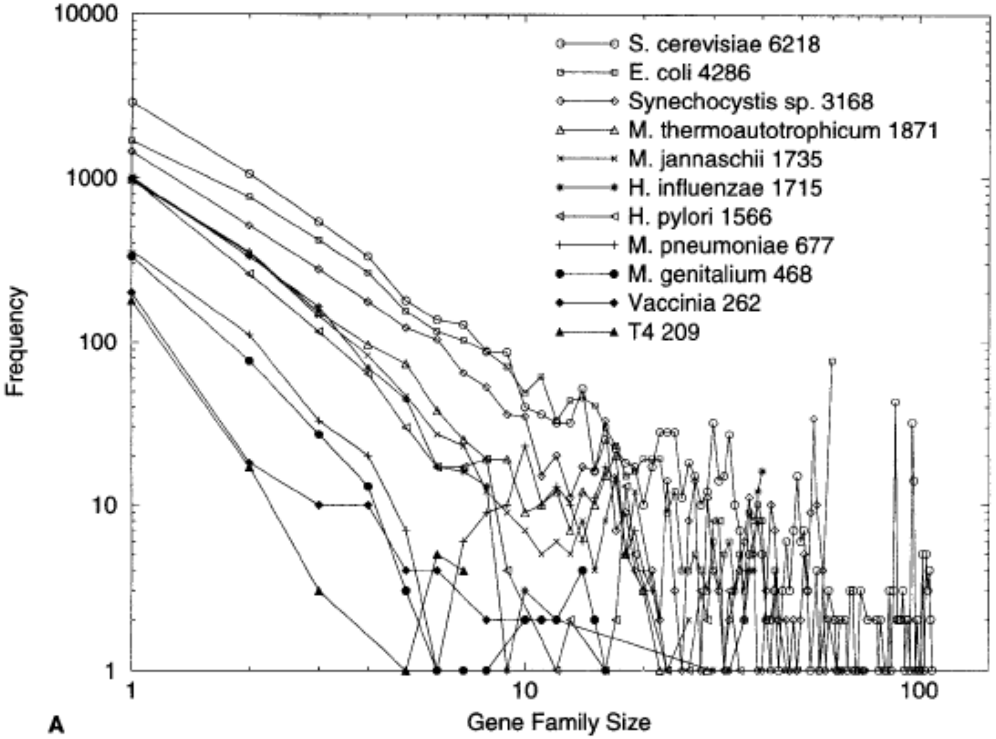


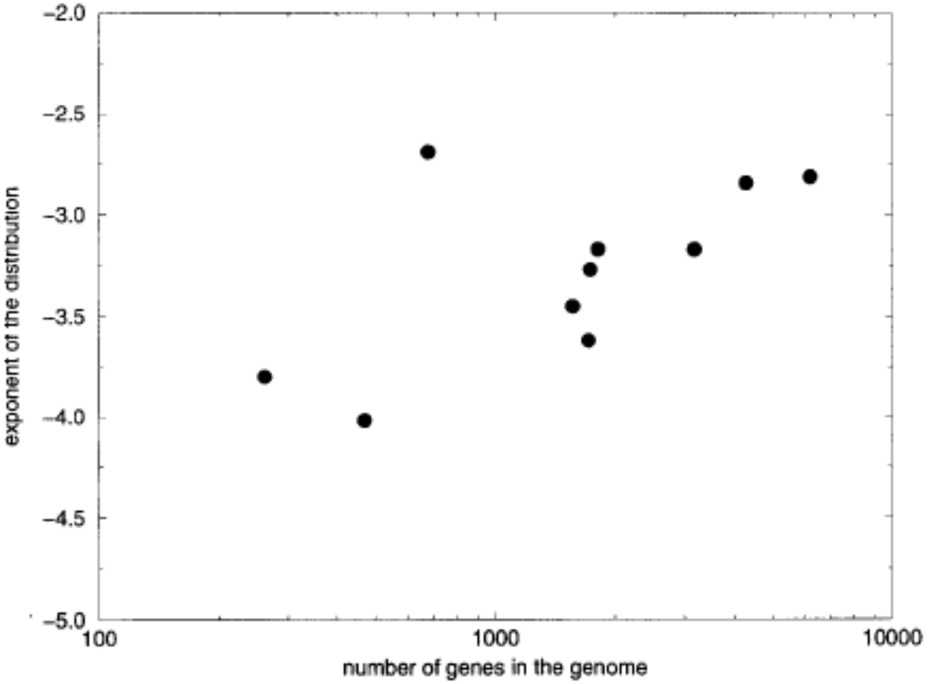
```
                                 *   .      :       .          .   *         : : :      .
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE    76
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE    76
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE    76
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD    75
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE    76
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK    79
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS    80
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE    80
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN    86
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE    85
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP    78
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP    78
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL    75
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE    88
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD    74
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA    82
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA    82
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA    81
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE    77
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE    77
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE    77
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE    76
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE    79
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE    79
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD    79
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS    72
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT    72
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV    72
       ruler 1.......10........20........30........40........50........60........70........80........90
```

But also: structural information



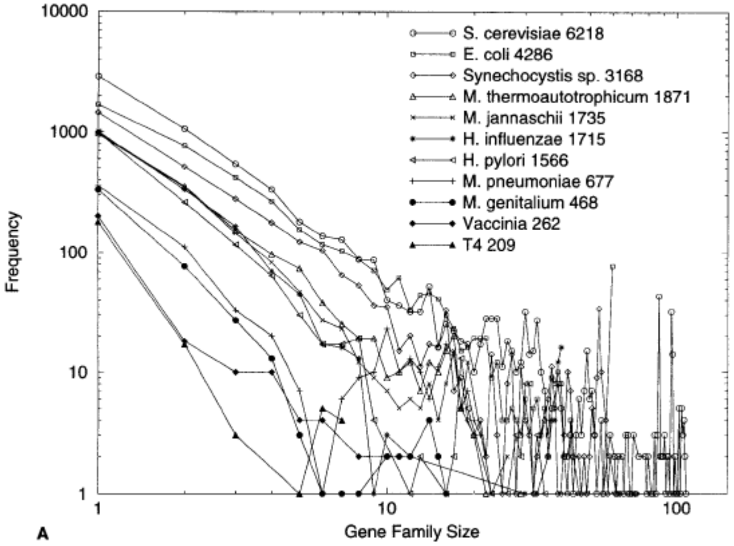
Gene-family size distributions

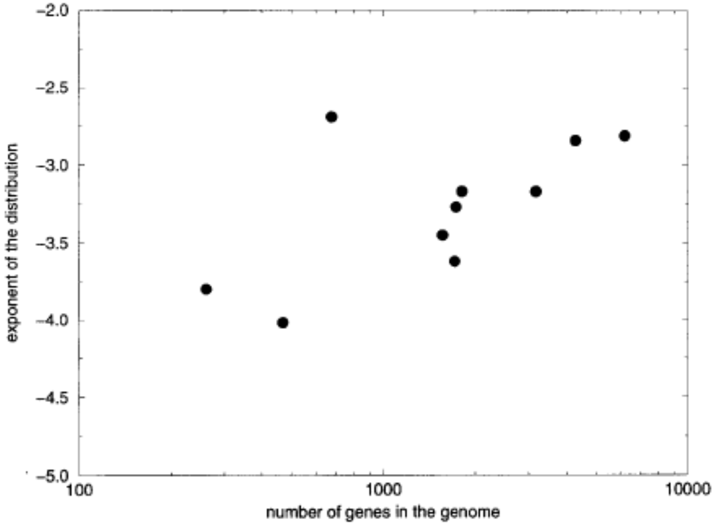




(Huynen Nimwegen MBE '98)

Gene-family size distributions





Early 2000s focus on wide tails
two main explanations

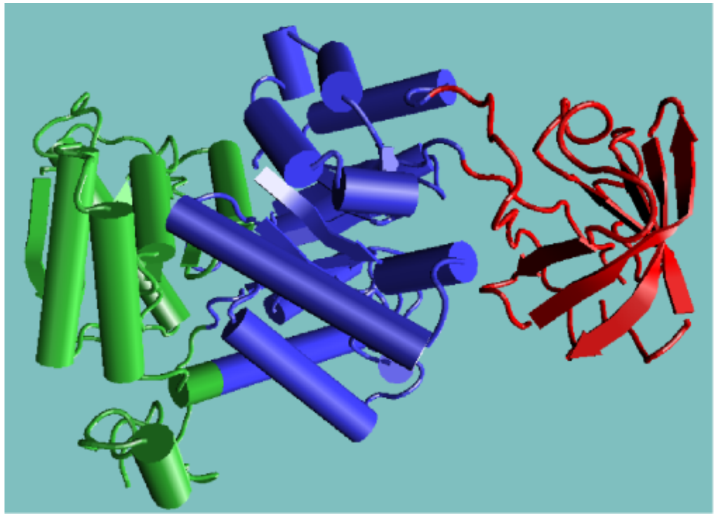
a) “designability” (e.g. Shakhnovich)

b) “genome growth” (e.g. Koonin)

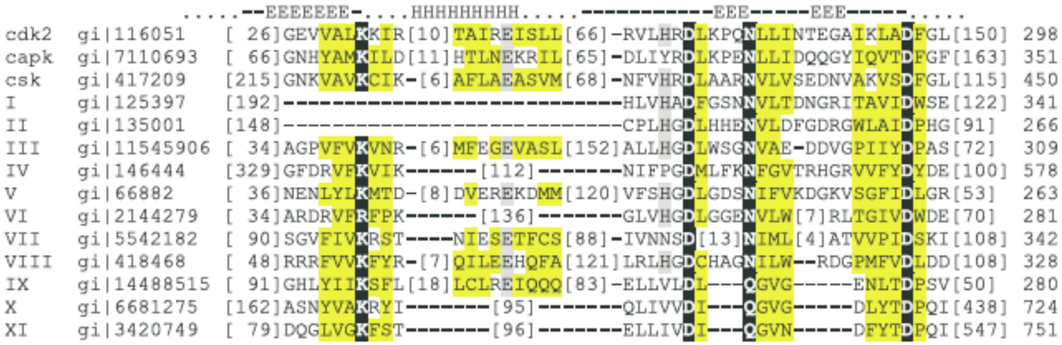
No focus on common scaling
with genome size
Until late 2000s

Homology and Protein Domains

- Basic stable sub-shapes of proteins
- Conserved in evolution
- Determine possible protein functions
- Modular

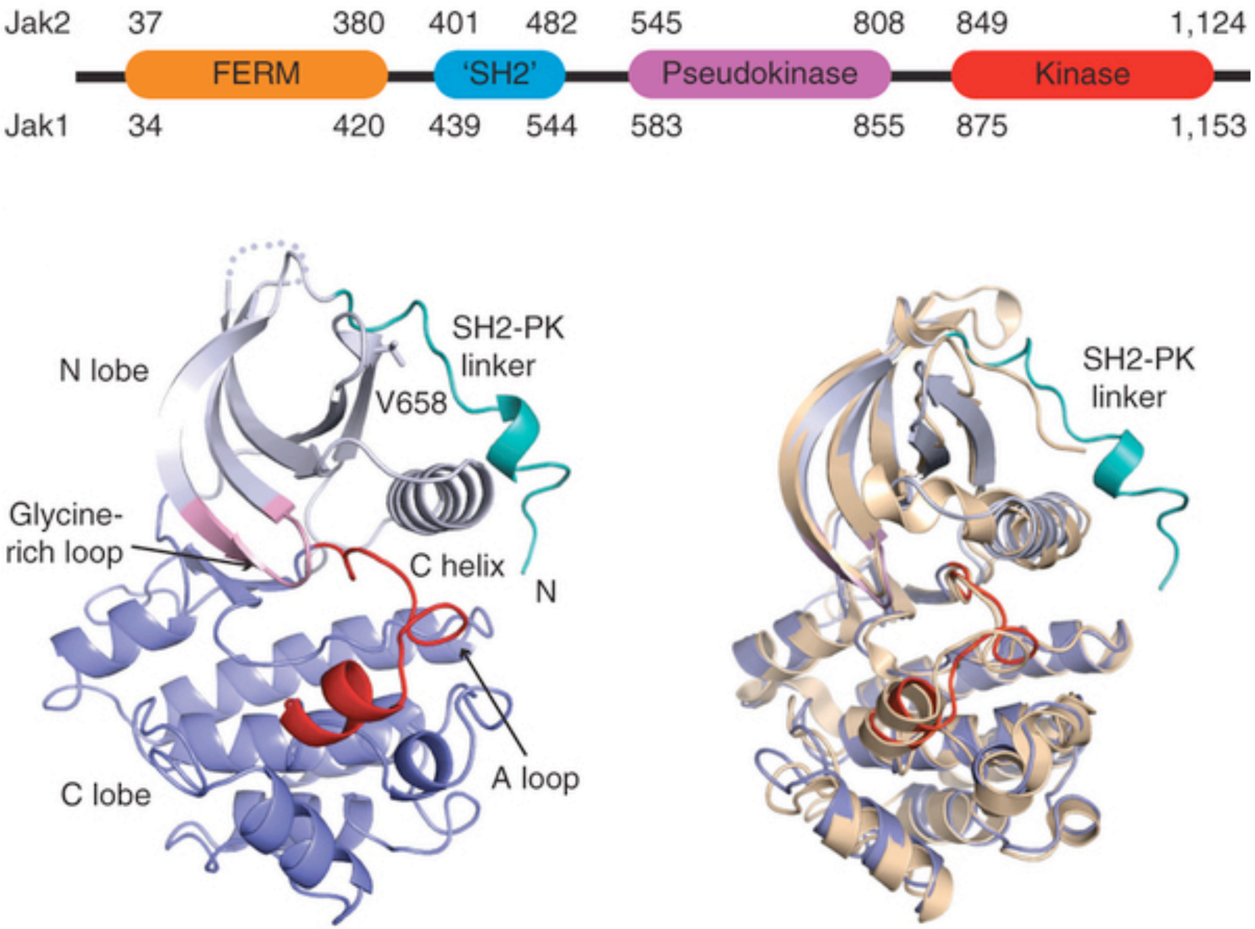


(Pyruvate kinase)

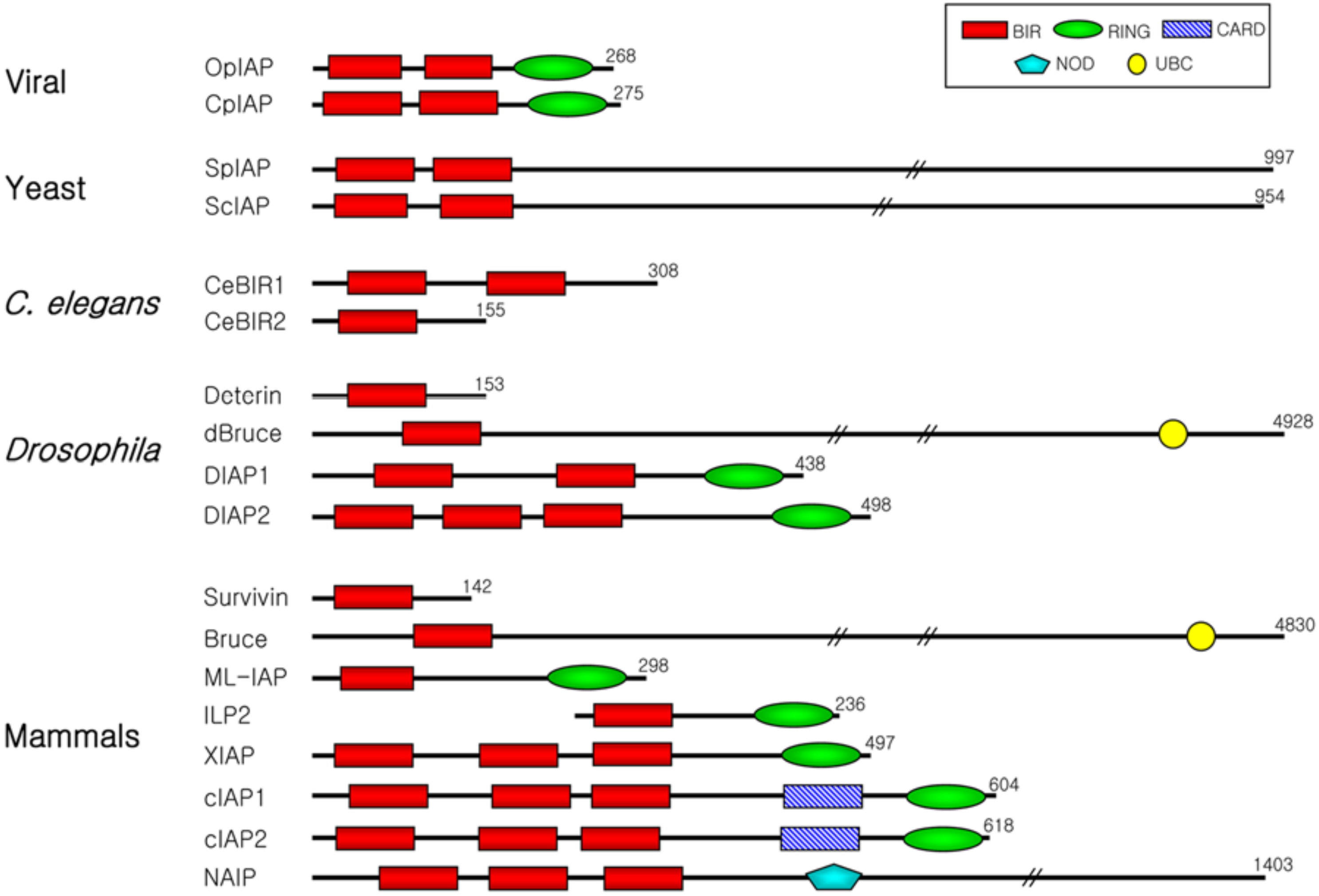


```
               .....--EEEEEEE-....HHHHHHHHH.....------------EEE-----EEE-----.....
cdk2  gi|116051    [ 26]GEVVALKKIR[10]TAIREISLL[66]-RVLHRDLKPQNLLINTEGAIKLADFGL[150] 298
capk  gi|7110693   [ 66]GNHYAMKILD[11]HTLNEKRIL[65]-DLIYRDLKPENLLIDQQGYIQVTDFGF[163] 351
csk   gi|417209    [215]GNKVAVKCIK-[6]AFLAEASVM[68]-NFVHRDLAARNVLVSEDNVAKVSDFGL[115] 450
I     gi|125397    [192]----------------------------HLVHADFGSNNVLTDNGRITAVIDWSE[122] 341
II    gi|135001    [148]----------------------------CPLHGDLHHENVLDFGDRGWLAIDPHG[91]  266
III   gi|11545906  [ 34]AGPVFVKVNR-[6]MFEGEVASL[152]ALLHGDLWSGNVAE-DDVGPIIYDPAS[72]  309
IV    gi|146444    [329]GFDRVFKVIK------[112]-------NIFPGDMLFKNFGVTRHGRVVFYDYDE[100] 578
V     gi|66882     [ 36]NENLYLKMTD-[8]DVEREKDMM[120]VFSHGDLGDSNIFVKDGKVSGFIDLGR[53]  263
VI    gi|2144279   [ 34]ARDRVFRFPK------[136]-------GLVHGDLGGENVLW[7]RLTGIVDWDE[70]  281
VII   gi|5542182   [ 90]SGVFIVKRST----NIESETFCS[88]-IVNNSD[13]NIML[4]ATVVPIDSKI[108] 342
VIII  gi|418468    [ 48]RRRFVVKFYR-[7]QILEEHQFA[121]LRLHGDCHAGNILW--RDGPMFVDLDD[108] 328
IX    gi|14488515  [ 91]GHLYIIKSFL[18]LCLREIQQQ[83]-ELLVLDL---QGVG-----ENLTDPSV[50]  280
X     gi|6681275   [162]ASNYVAKRYI-------[95]-------QLIVVDI---QGVG-----DLYTDPQI[438] 724
XI    gi|3420749   [ 79]DQGLVGKFST-------[96]-------ELLIVDI---QGVN-----DFYTDPQI[547] 751
```

Protein Domains



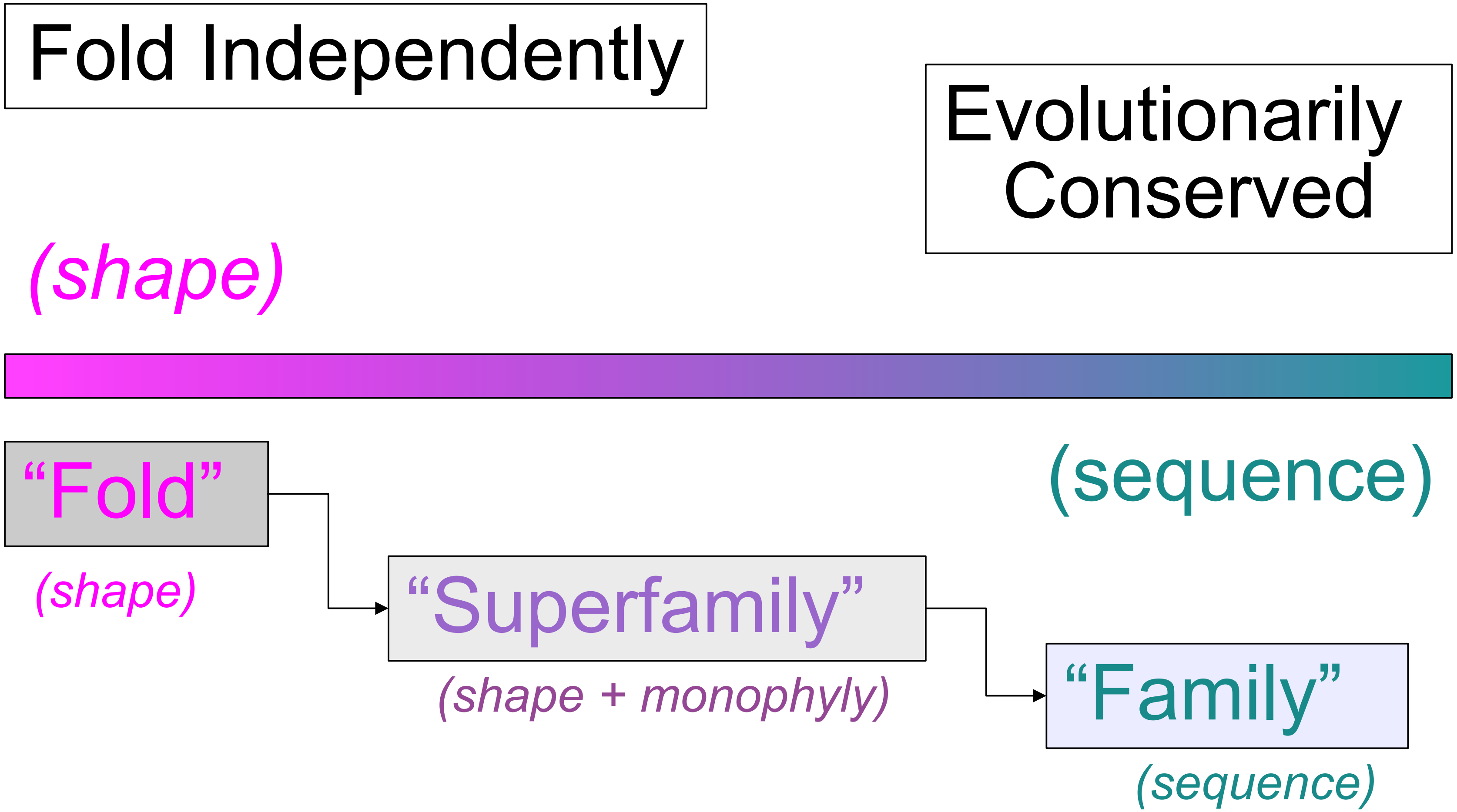
Biologist's first slide



“Coarse-grained” view of a protein
Structure / Evolution / Function

Taxonomy

2 Domain *definitions*



Fold Independently

Evolutionarily Conserved

(shape)

(sequence)

“Fold”

(shape)

“Superfamily”

(shape + monophyly)

“Family”

(sequence)

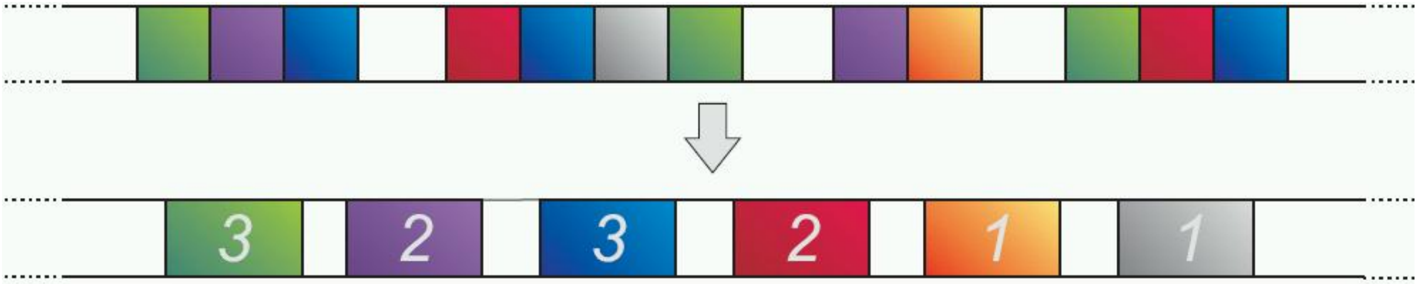
3 Hierarchic classes

Genome-scale data

Databases of structural domain families
(SCOP / SUPERFAMILY, CATH / gene3D for structure*)*

- Cover hundreds of genomes
- Typically 30-60% sequence coverage
- 50-70% proteins with at least one hit

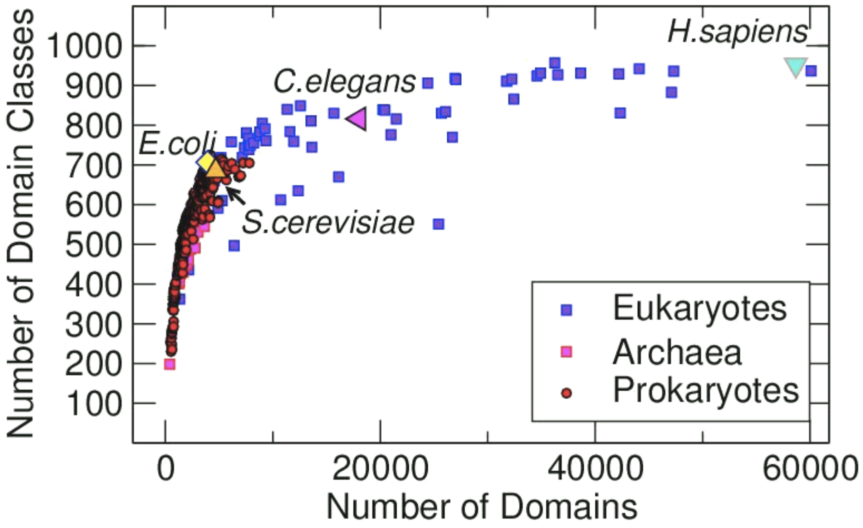
"Coarse-grained" view of a *genome*



n domains

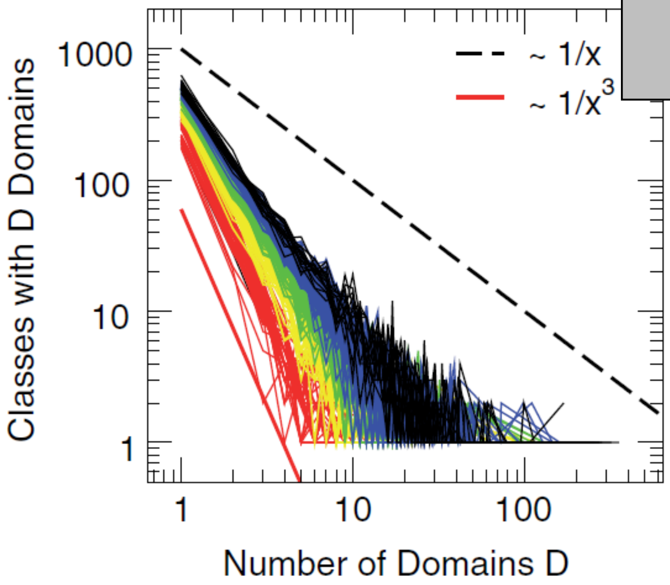
F domain classes

Scaling Laws for Evolutionary classes



Number of evolutionary families
# families F
vs genome size n

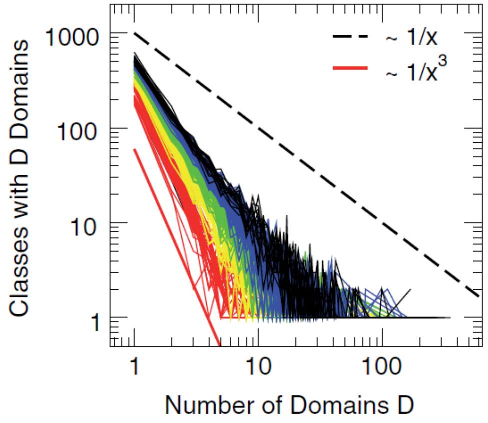
Population distribution of evolutionary families
family population histogram



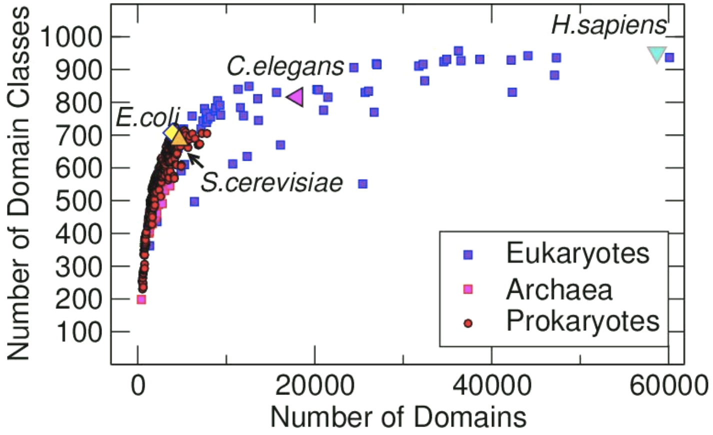
Number of domains
n < 1500
1500 – 3000
3000 – 6000
6000 – 12000
12000 – 24000
n > 24000

Exercise

- Go to www.supfam.org
- Follow “domain assignments” and click one prokaryote
- Download the “domain assignments” txt file
- Figure out the file and make this plot, for 10 bacteria with different sizes



- Find 5 partners and share data to make ~50 points of this plot



The existence of these scaling laws is
surprising

It indicates that domain class partitioning
depends on size
and not on the specific
evolutionary history of a genome

3) Horizontal Gene Transfer (HGT)

“Moves” of gene-family dynamics

Copy-Paste

Intra species HGT + Duplication

Share

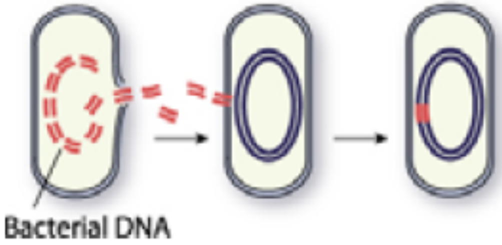
Inter-species HGT

Trash

Loss

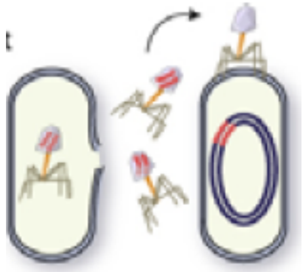
Main mechanisms of Horizontal Gene Transfer

Transformation



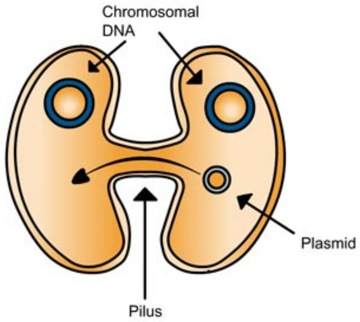
Direct DNA uptake

Transduction



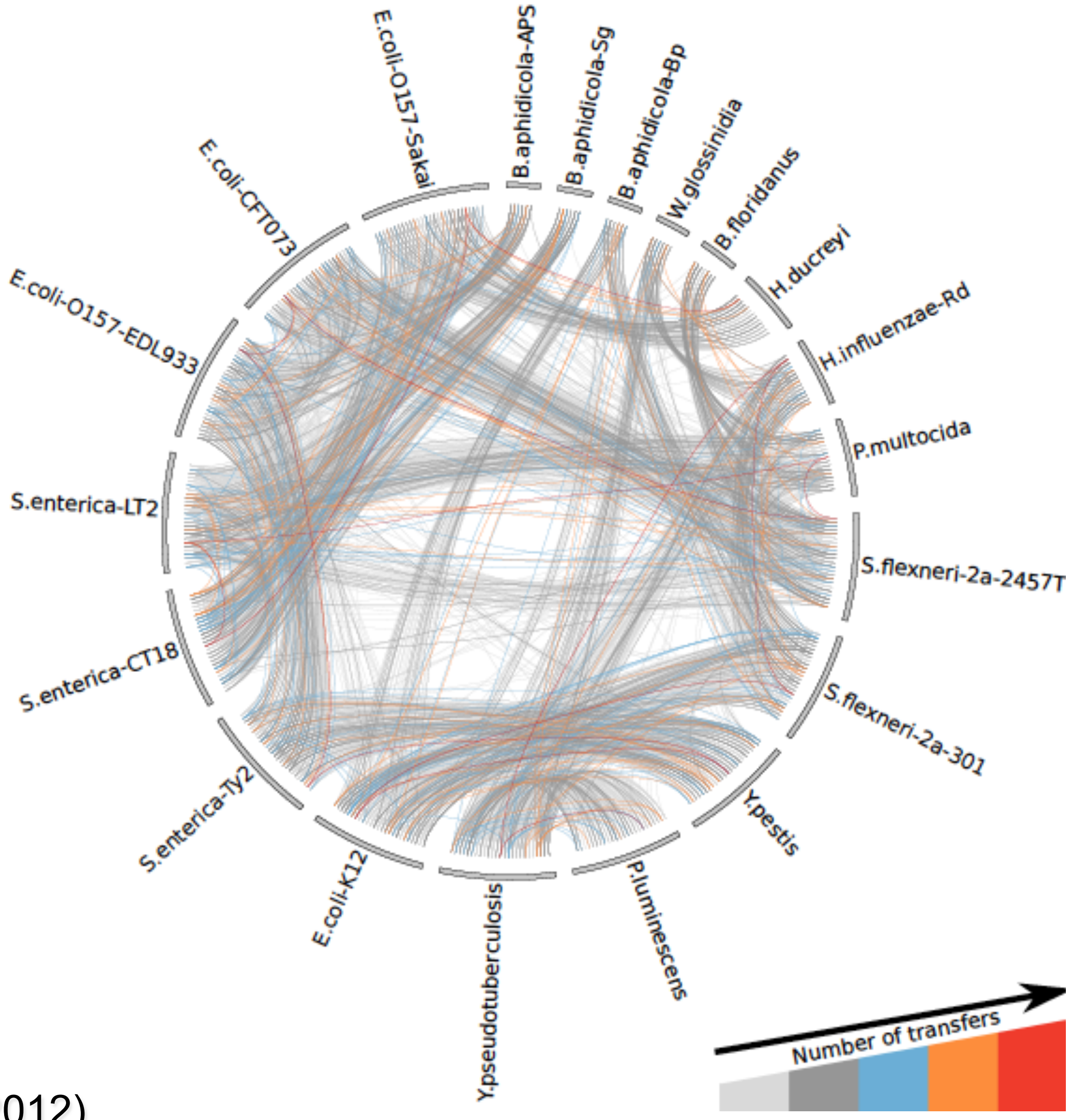
Through phages

Conjugation



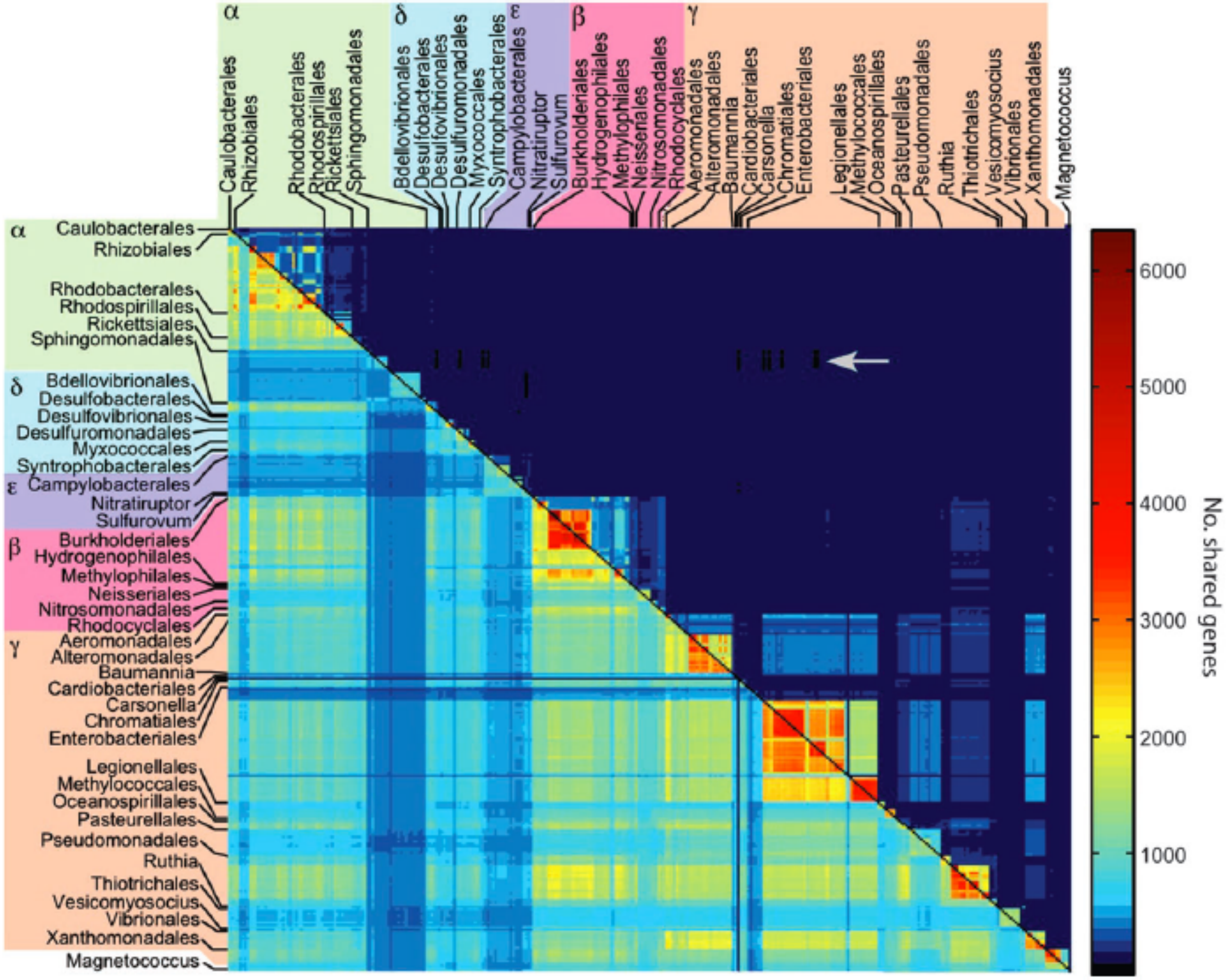
Sharing of plasmids (through contact)

Horizontal transfer of genes is a dominant force of bacterial gene-family evolution



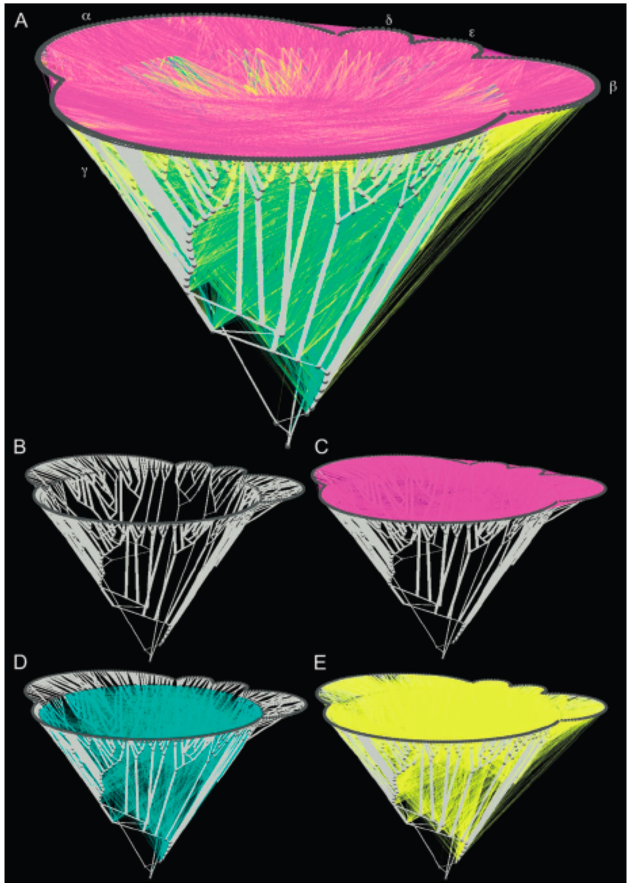
(Grassi et al MGE 2012)

Large-scale studies reveal biases/mechanisms



(Kloesges et al MBE 2010)

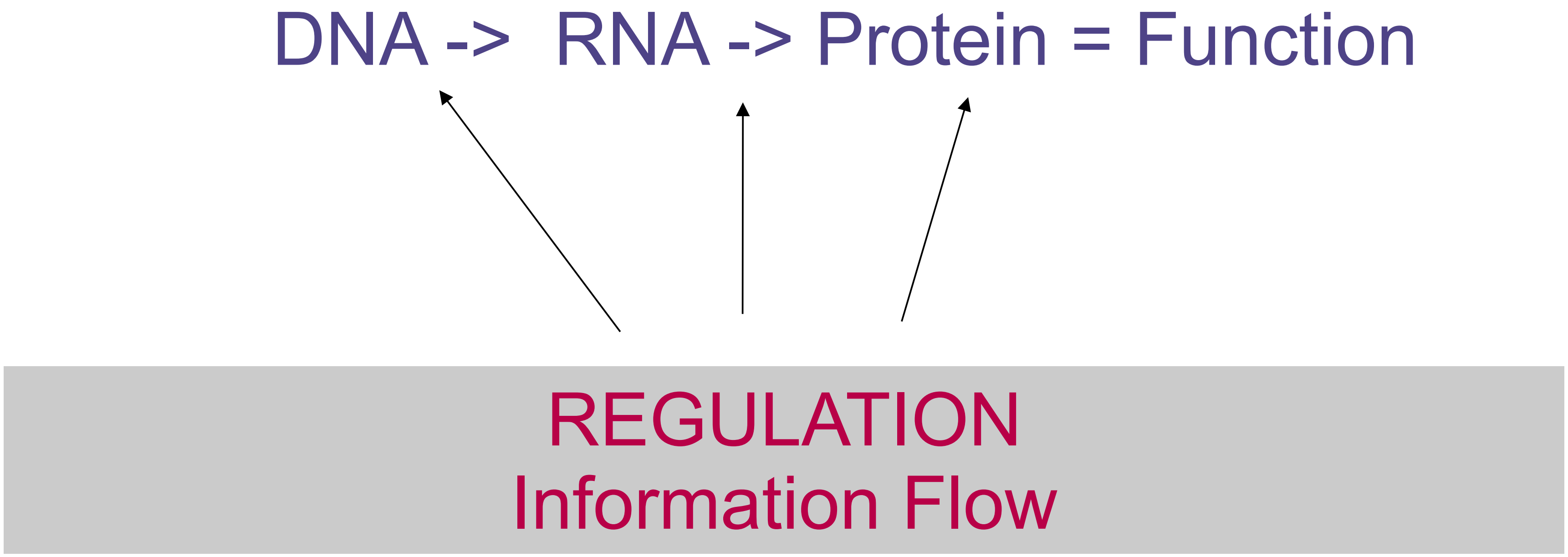
A tree or a network, or both?



(Kloesges et al MBE 2010)

4) Main biological interaction networks

“Central Dogma” of Molecular Biology

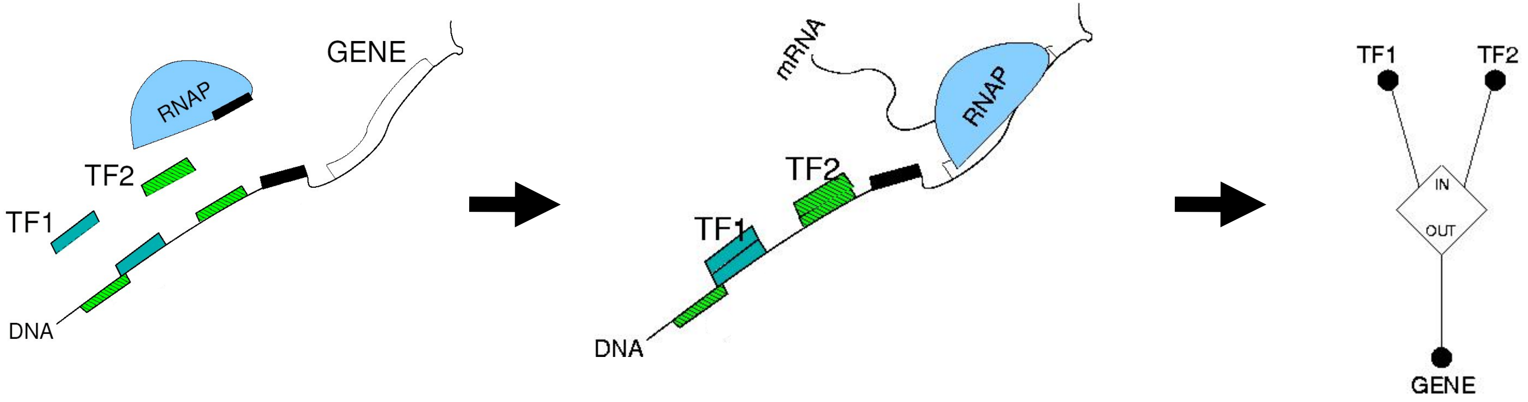


DNA -> RNA -> Protein = Function

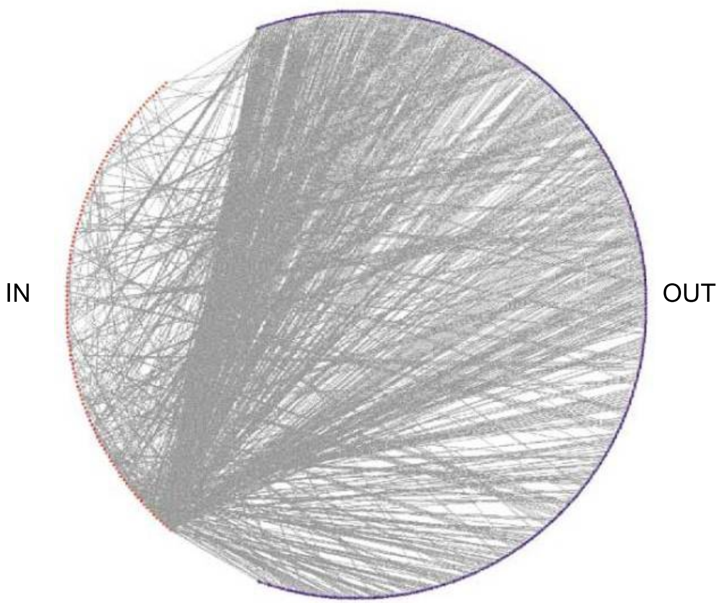
REGULATION
Information Flow

Network Approach (1) global (2) simple

Transcription Network



E.coli network



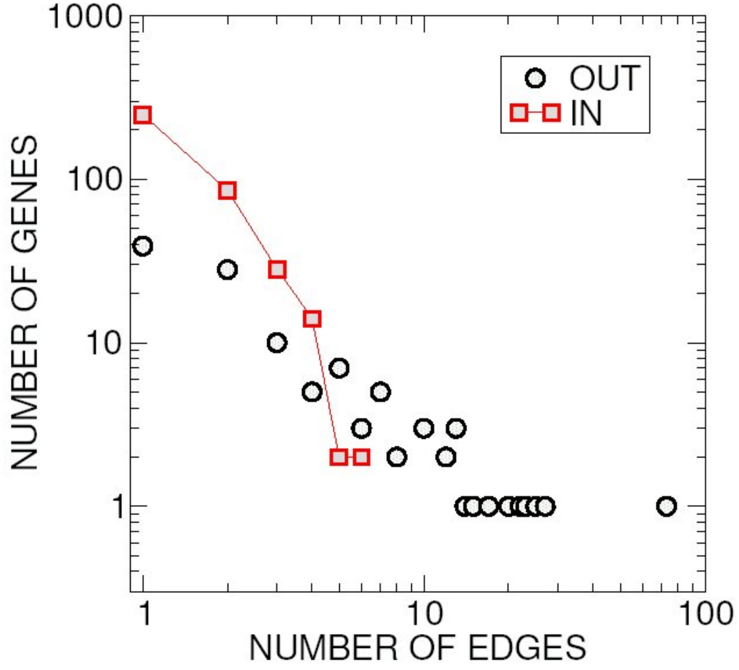
Approach :
1) global
2) simple

Transcription Network

Directed graph / Factor graph. Two kinds of nodes

Regulatory (TFs)
Targets, or “structural” genes (TGs)

Degree sequences



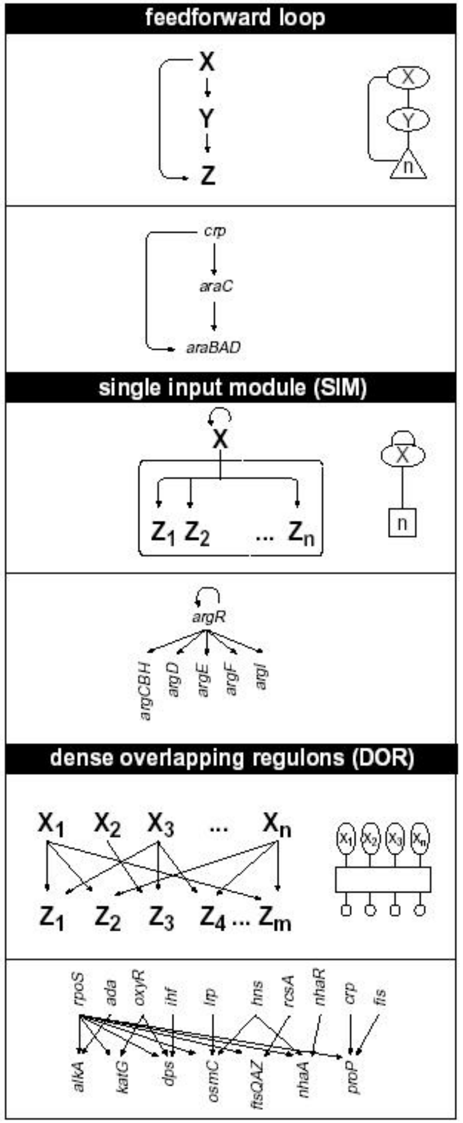
Topology approach example: network motifs

(> 500 genes, e.g. E.coli, S.cerevisiae)

Structural analysis

Example: network motifs = subgraphs that are more recurrent than in random networks

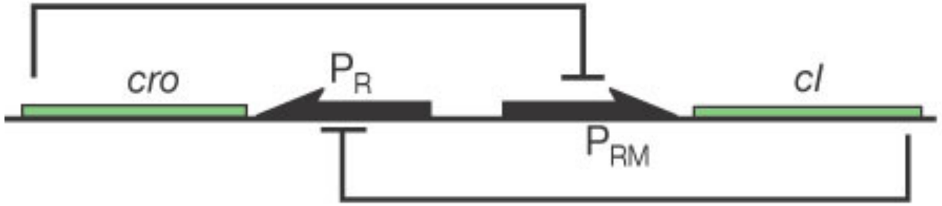
Randomizations = Ensemble of random graphs with the same degree sequences as E. coli, but shuffled links



Network motifs in E. coli
Uri Alon's group
(Shen-orr et al Nature Gen 02)

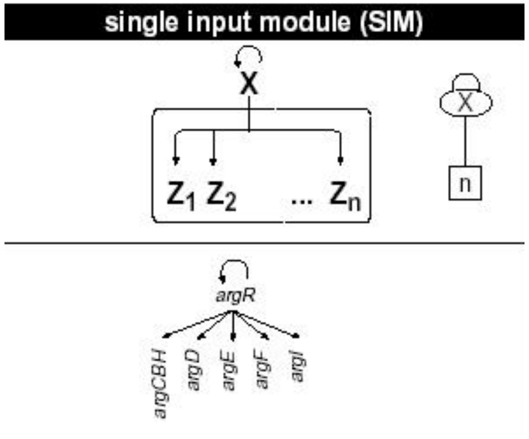
Feedback vs Hierarchy

Feedback: Multistability, periodicity,… (Thomas, Kauffman, Savageau…)
Example: Phage λ (Arkin et al Genetics 98)



Switch involves mutual
Negative feedback

Hierarchy: Organization of the transcription program
Example: SIM motif (Shen-Orr et al Nat Genet 02)

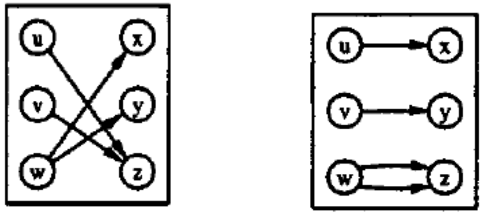


Input -> Output
Hierarchical structure

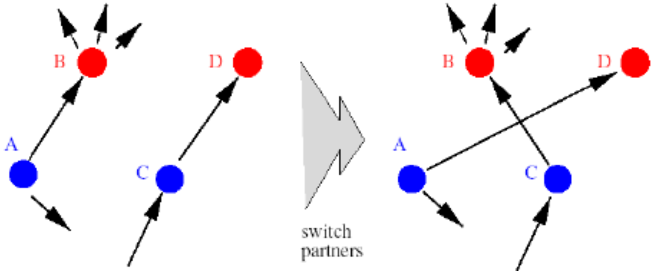
Randomization Algorithms

Randomizations =
Ensemble of random graphs with the
same degree sequences as E. coli,
but shuffled links

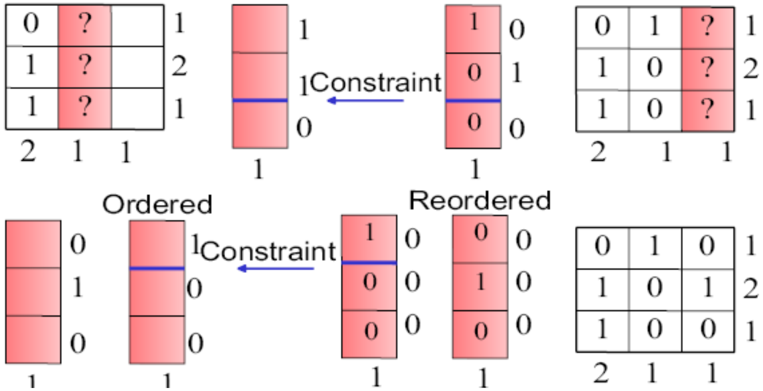
Stub Pairing (Molloy-Reed)



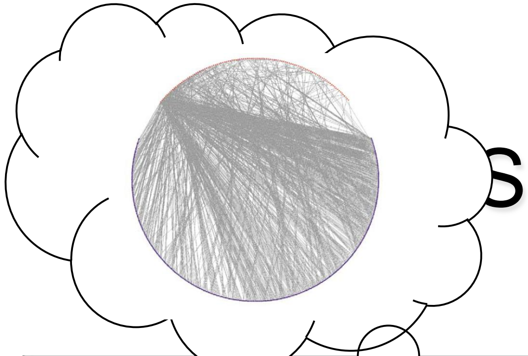
“Switches” (Maslov-Sneppen)

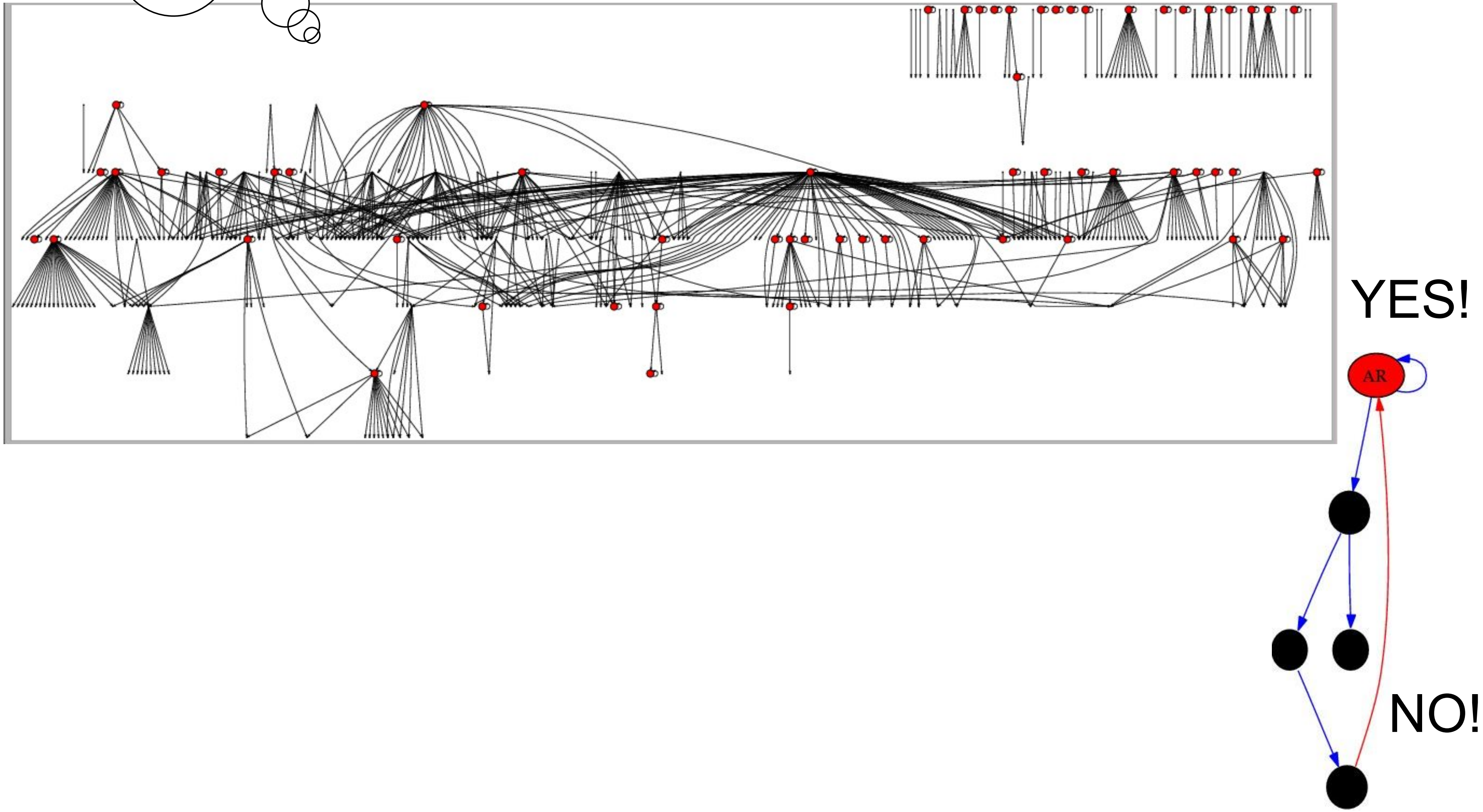


Importance Sampling Montecarlo



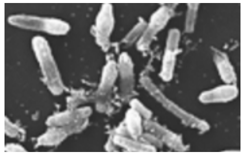
E. coli network:
Shallow hierarchy with mostly self-feedback



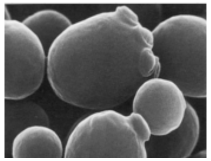


Comparing Topologies

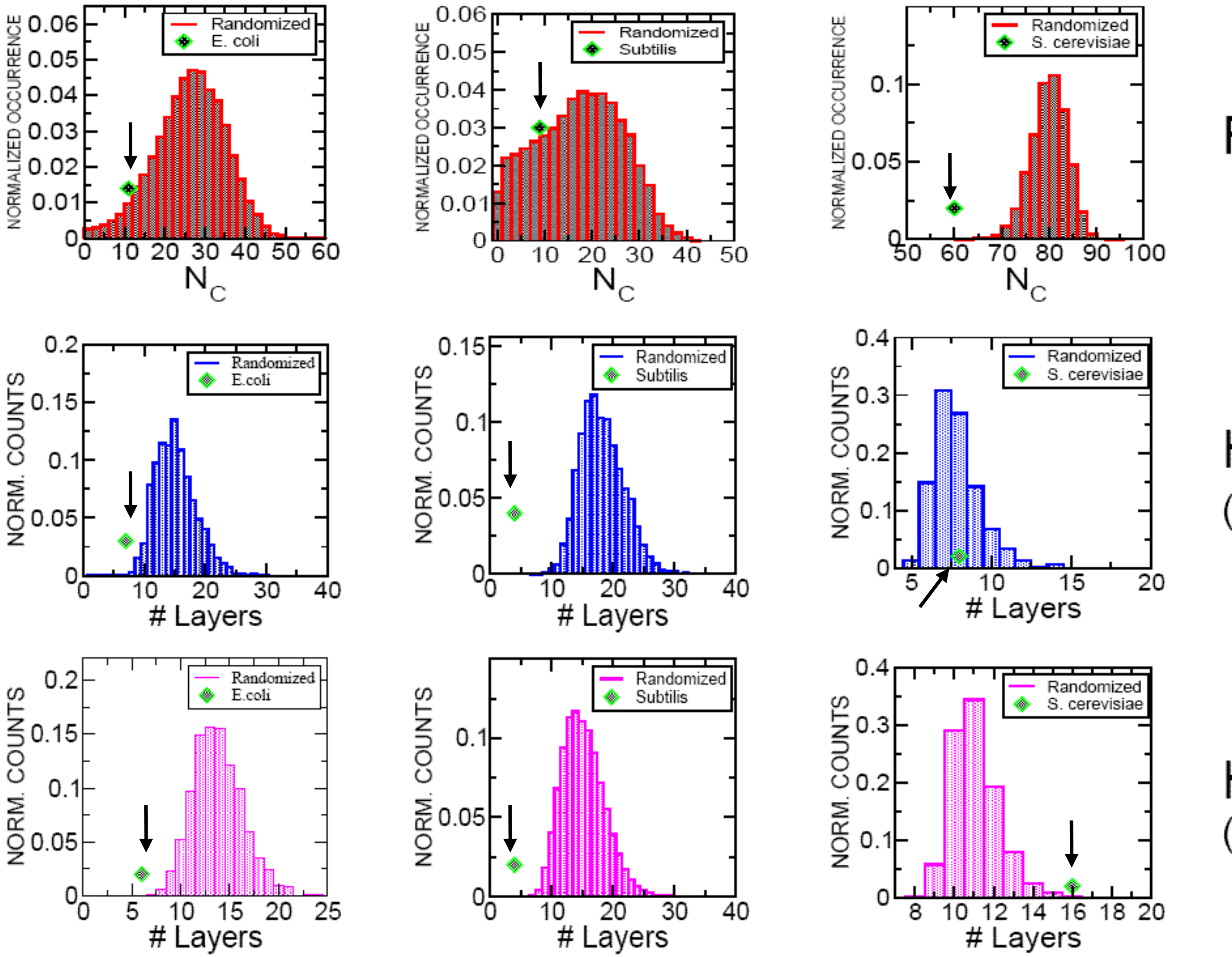
E.coli



B.subtilis



S.cerevisiae



Feedback

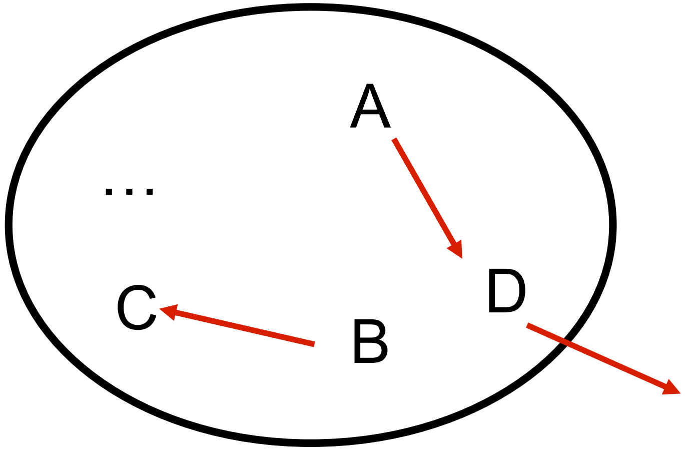
Hierarchy
(Longest Path)

Hierarchy
(Shortest Path)

Evolutionary analysis

Comparison of homology classes
with network interactions

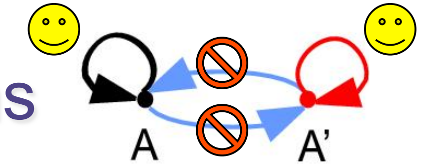
Network
interactions



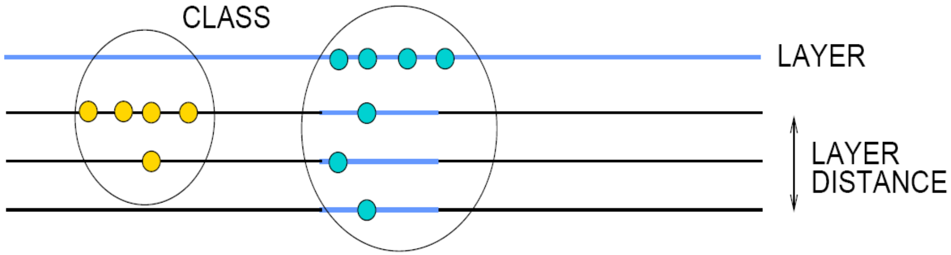
Homology class
(common ancestor)

Main results:

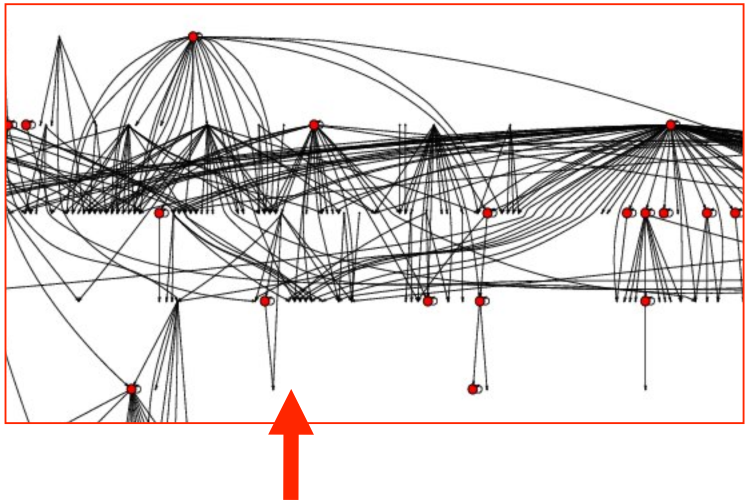
Family expansion and autoregulations



Family expansion and layers

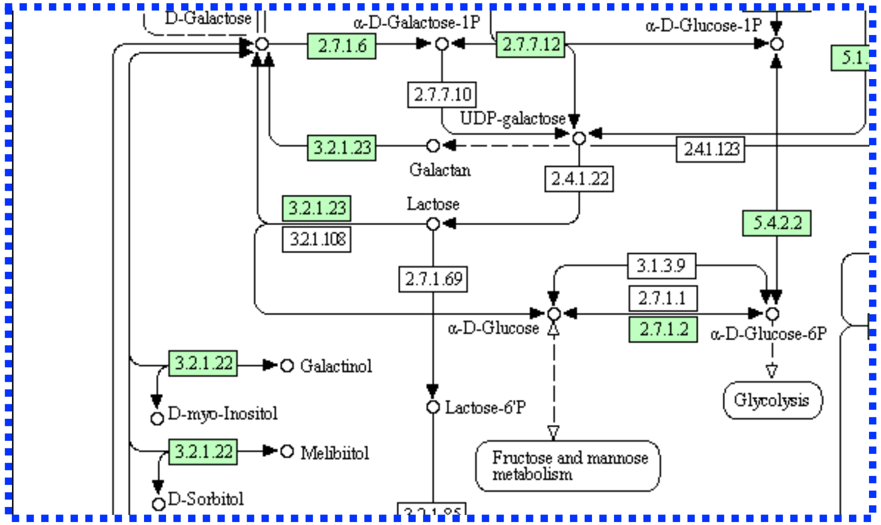


Horizontal Transfer



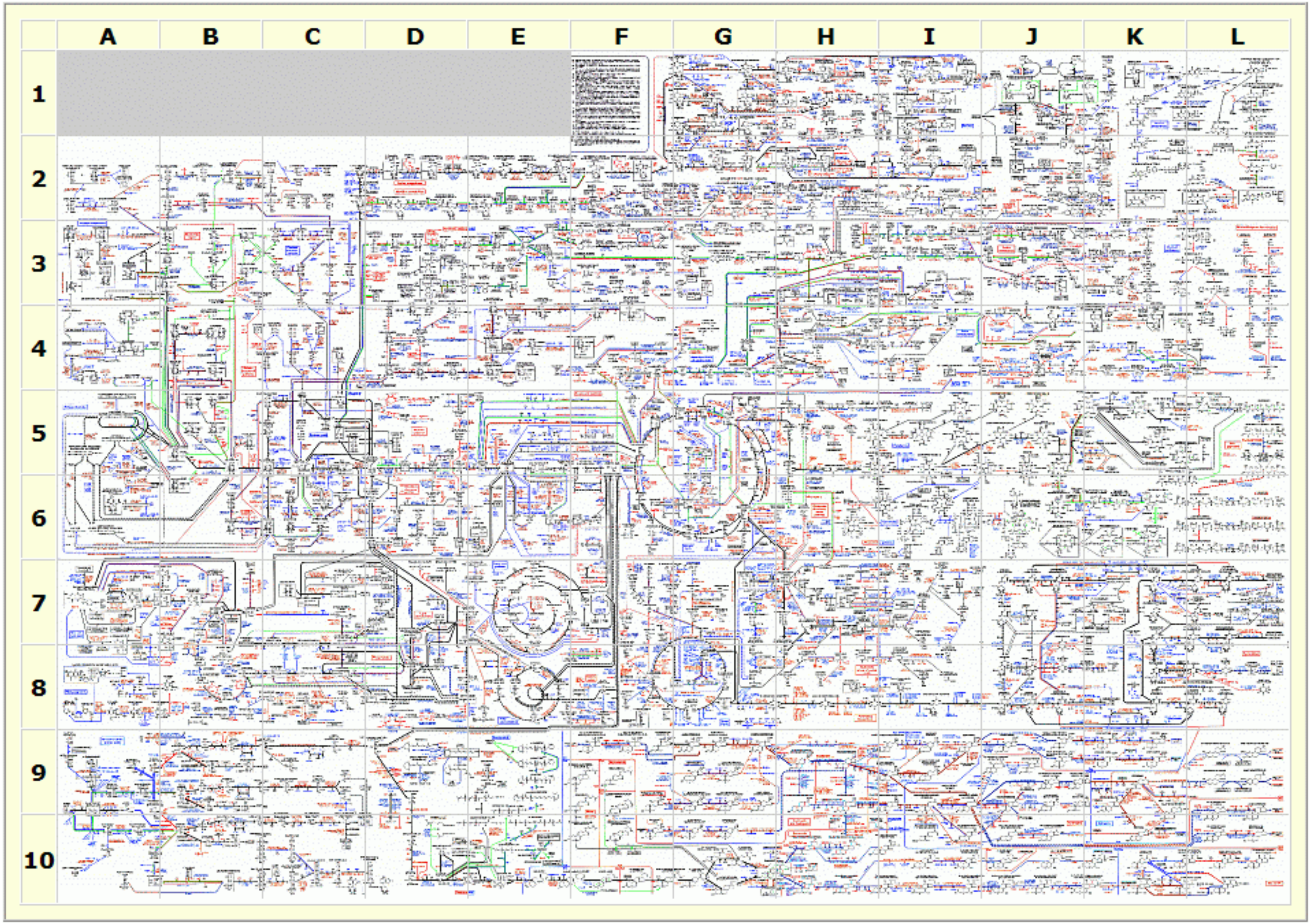
(Cosentino Lagomarsino *et al.* PNAS 2007,
Sellerio et al, Mol Biosys 2009)

Metabolism at Large Scale



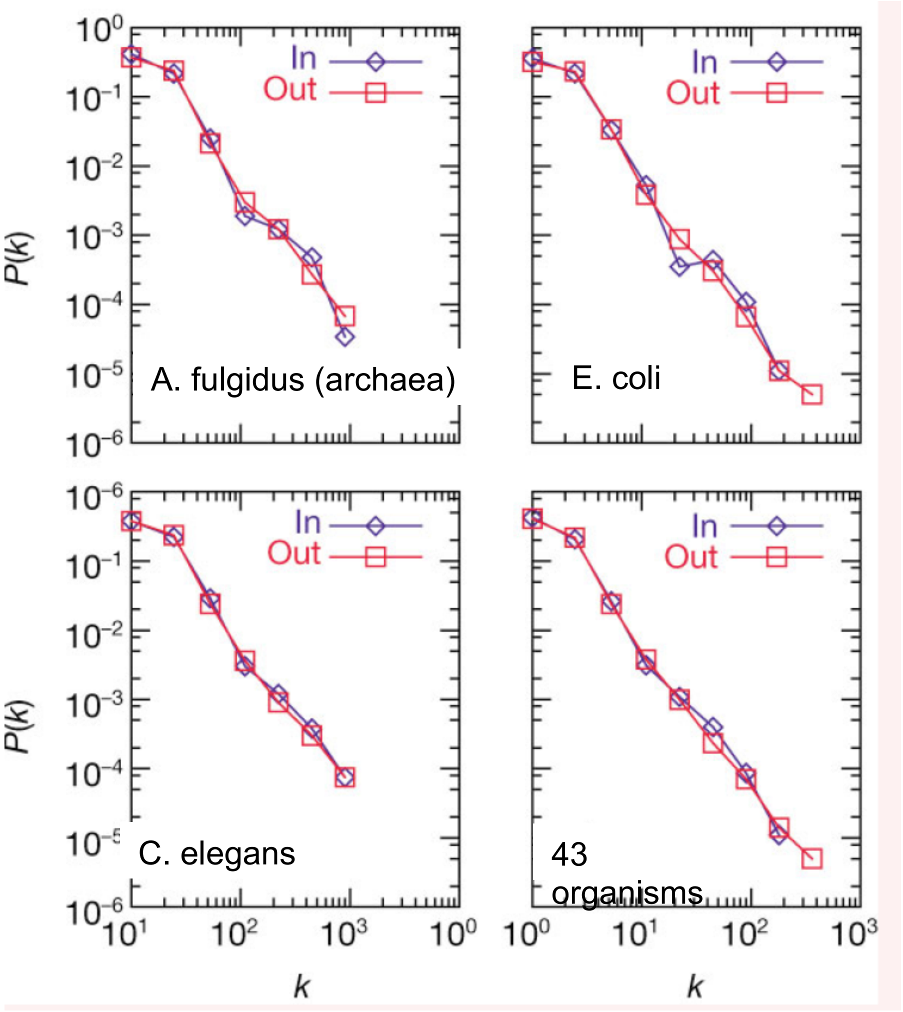
Metabolic network:

Edges = Metabolic enzymes (genes)
Nodes = Chemicla reactions

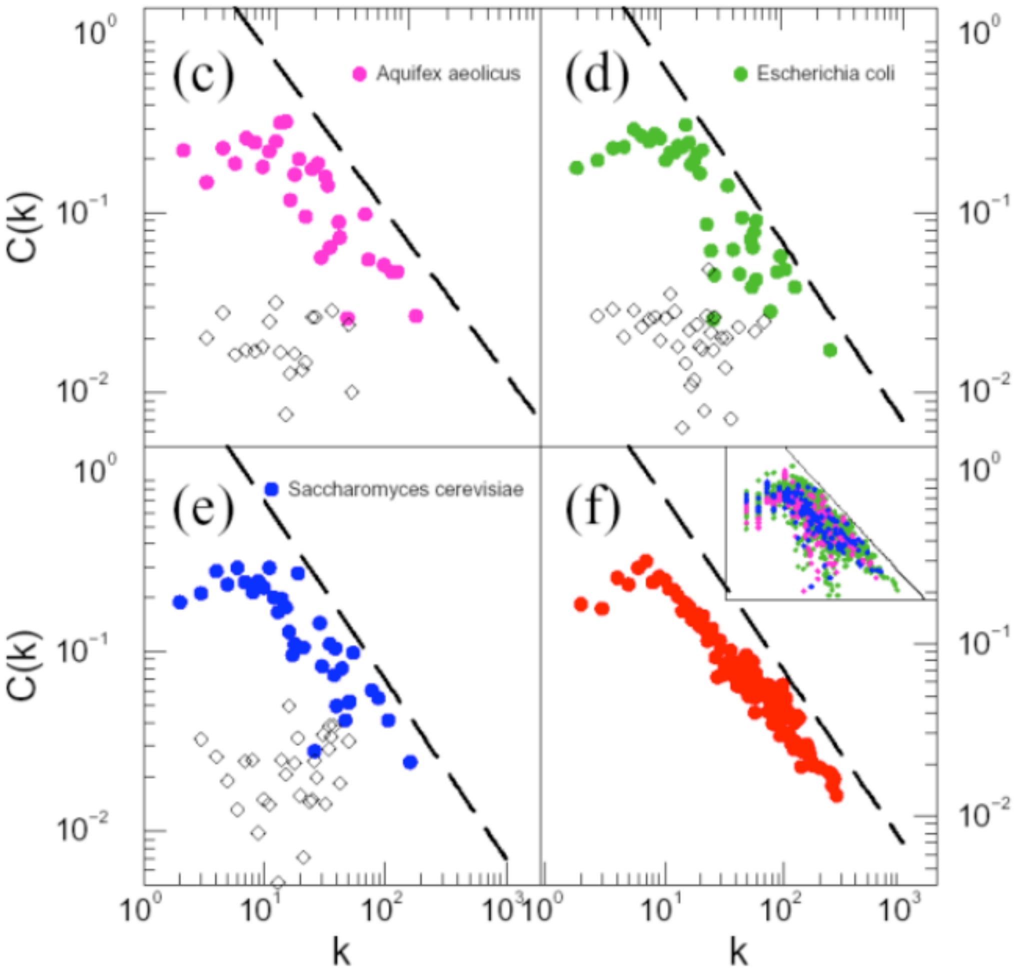


Metabolic network

Metabolic network topology



Degree distributions



Hierarchical modular structure

(Jeong et al Nature 2002)

(Ravasz et al Science 2002)

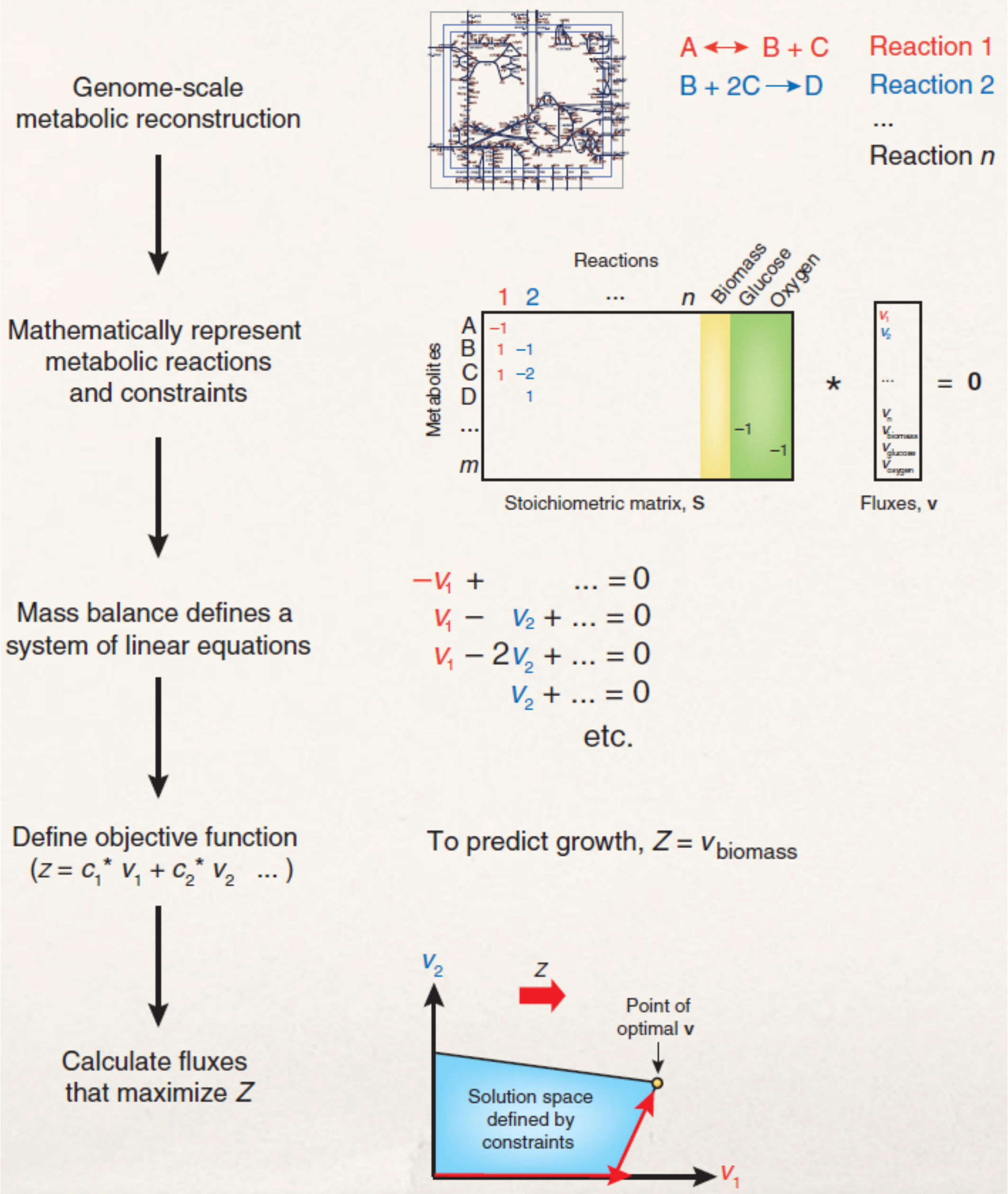
Flux-balance approach

Describe stoichiometry
as flux network

Steady-state = linear programming

Objective function is typically
Biomass

E.g. Predicts many phenotypes
In fast-growing bacteria



Genome-scale metabolic reconstruction

Mathematically represent metabolic reactions and constraints

Mass balance defines a system of linear equations

Define objective function ($z = c_1{}^* \, v_1 + c_2{}^* \, v_2 \; \ldots$)

Calculate fluxes that maximize Z

$A \leftrightarrow B + C$ Reaction 1

$B + 2C \rightarrow D$ Reaction 2

...

Reaction n

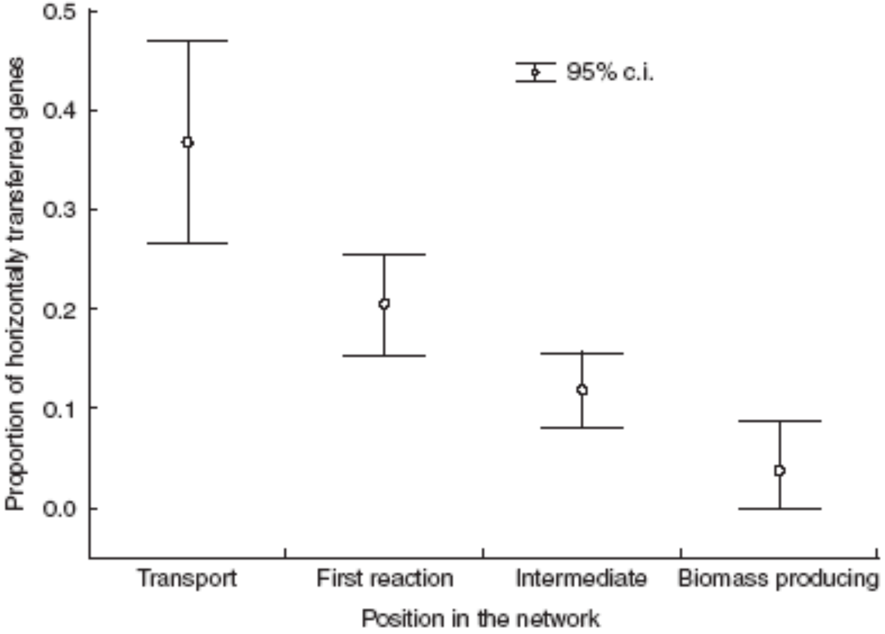
$$
\begin{aligned}
-v_1 + \quad & \ldots = 0 \\
v_1 - \quad v_2 + & \ldots = 0 \\
v_1 - 2v_2 + & \ldots = 0 \\
v_2 + & \ldots = 0
\end{aligned}
$$

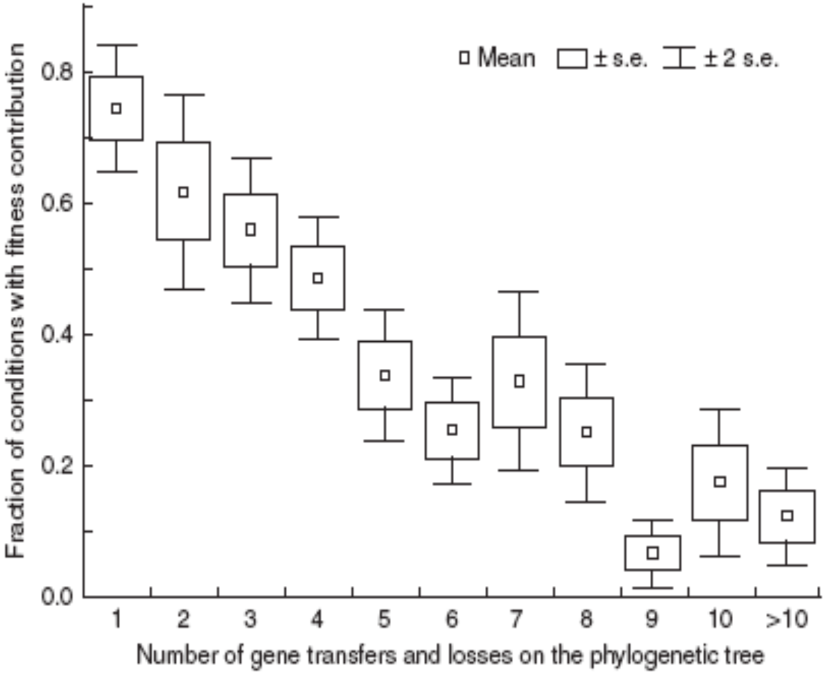
etc.

To predict growth, $Z = v_{\text{biomass}}$

Reactions
1 2 ... n Biomass Glucose Oxygen
Metabolites
A B C D ... m
−1
1 −1
1 −2
1
−1
−1
*
= 0
Stoichiometric matrix, S
v_1 v_2 ... v_n $v_{biomass}$ $v_{glucose}$ v_{oxygen}
Fluxes, v
v_2
Z
Point of optimal v
Solution space defined by constraints
v_1

Integration of HGTs in metabolism





1. peripheral reactions (nutrient uptake and first metabolic step) were more likely to be transferred (topology)

2. HGTs contributing to the evolution of metabolic networks in proteobacteria were generally environment-specific (single KO FBA with 136 conditions)

3. coupled enzymes were gained or lost together In a statistically significant manner (topology)

(Pal et al MBE 2005)

Conclusions

- Abundant data on genome composition, with striking statistical regularities
- “Laws” in the partitioning into functional and evolutionary elements
- Horizontal transfers are a dominant for adding new genes in bacteria
- New methabolic pathways can be “imported”, and controlled by a shallow hierarchy of transcription factors.