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

2584-12

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

A Genome as a Toolbox: Species-centered laws and models

Marco Cosentino Lagomarsino
Université Pierre et Marie Curie
Paris

A Genome as a Toolbox: Species-centered laws and models

June 3rd 2014

Spring School, Trieste

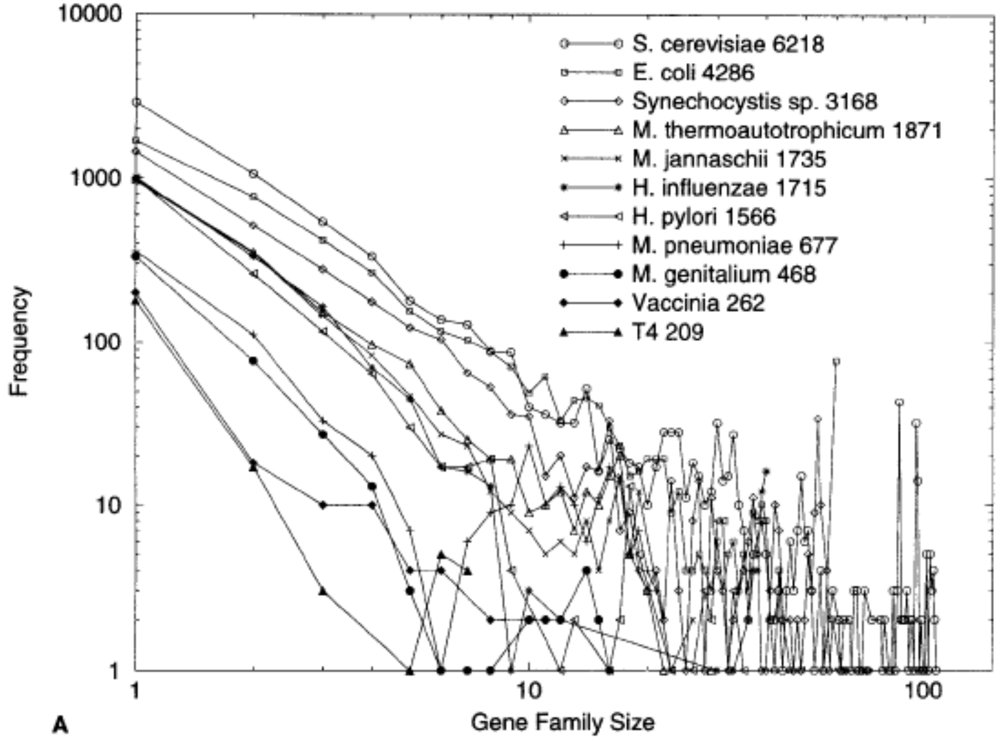
Marco Cosentino Lagomarsino

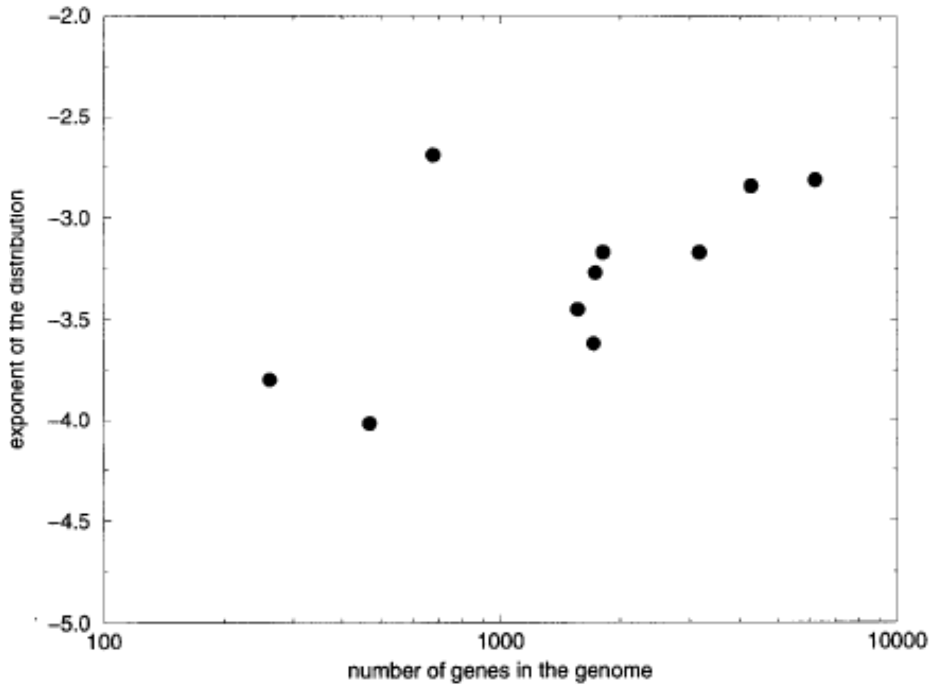
Génophysique / Genomic Physics Group

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

0) Where we left yesterday...

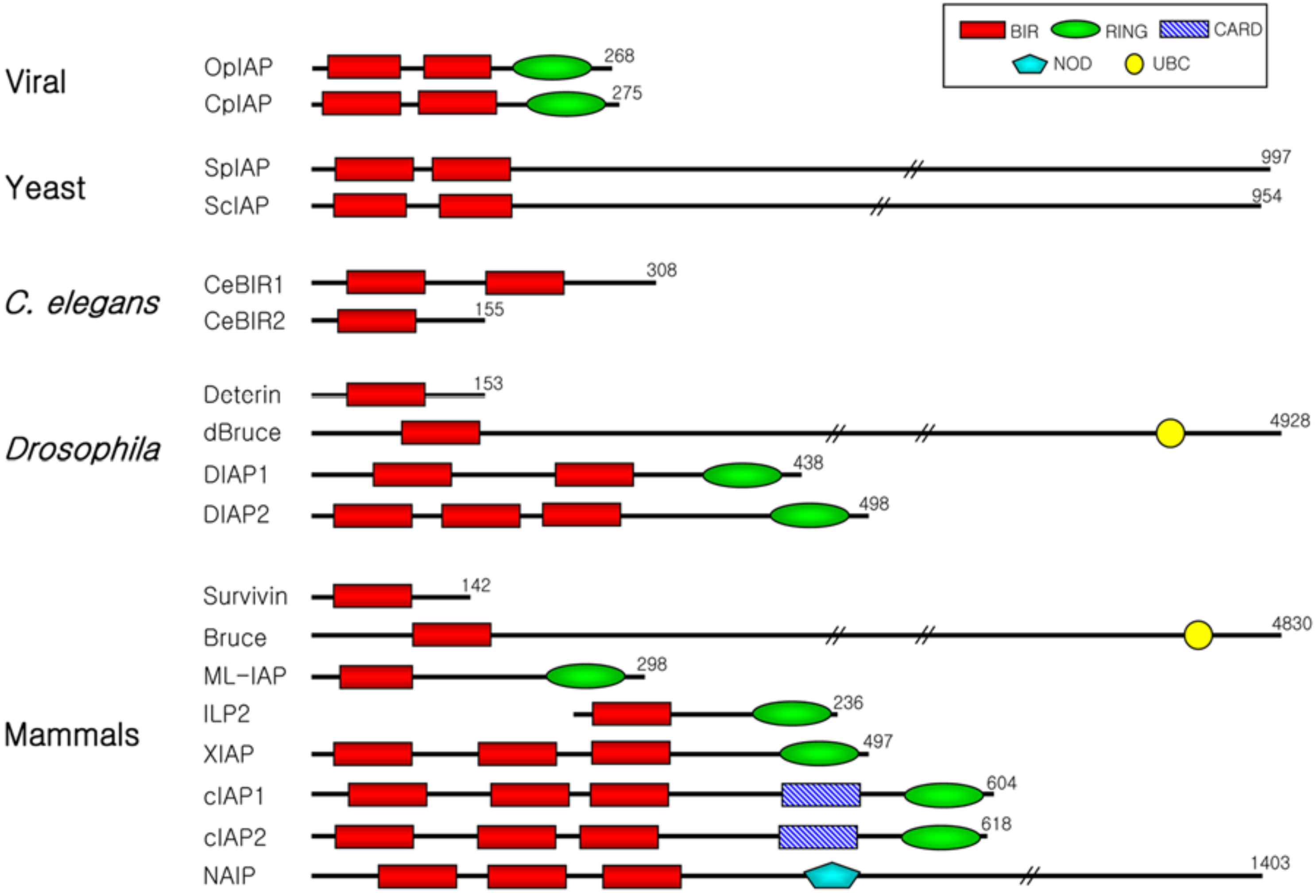
1st "law", gene-family size distributions





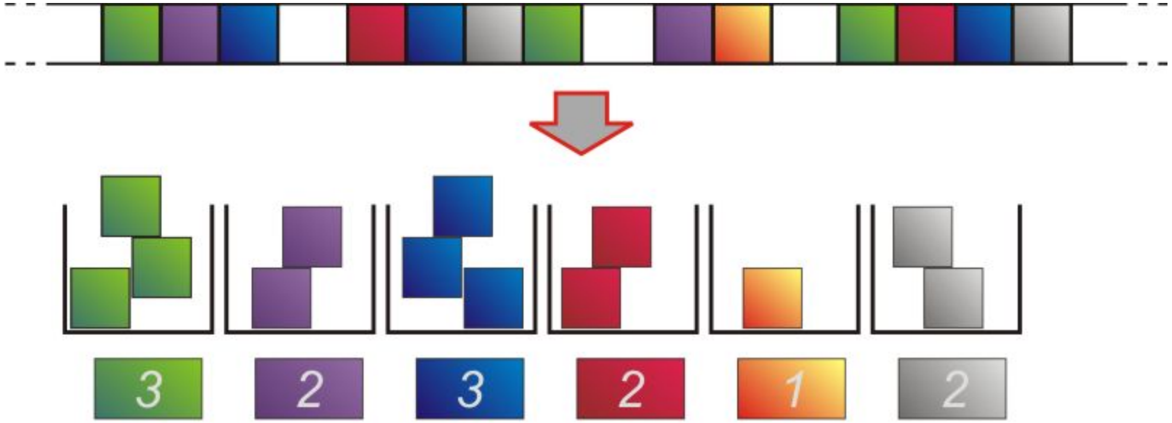
(Huynen Nimwegen MBE '98)

Protein domains
as coarse-grained view of proteins



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

Protein domains
as coarse-grained view of genomes

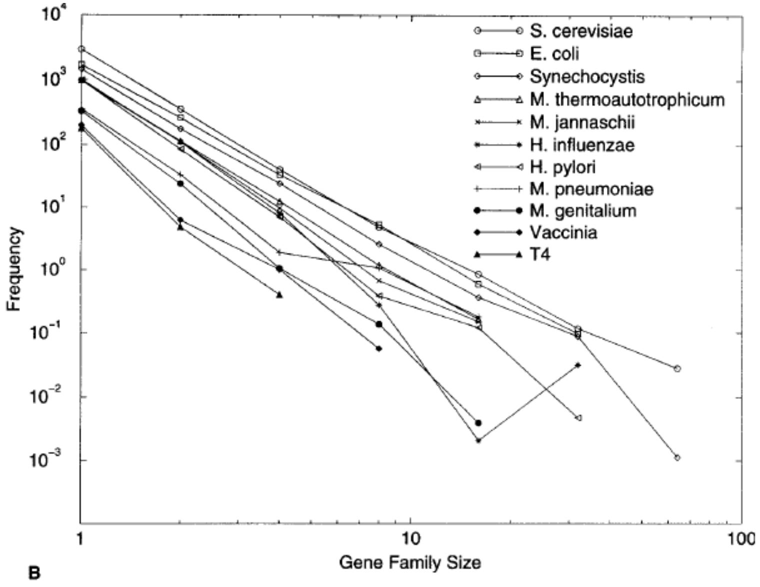


n domains

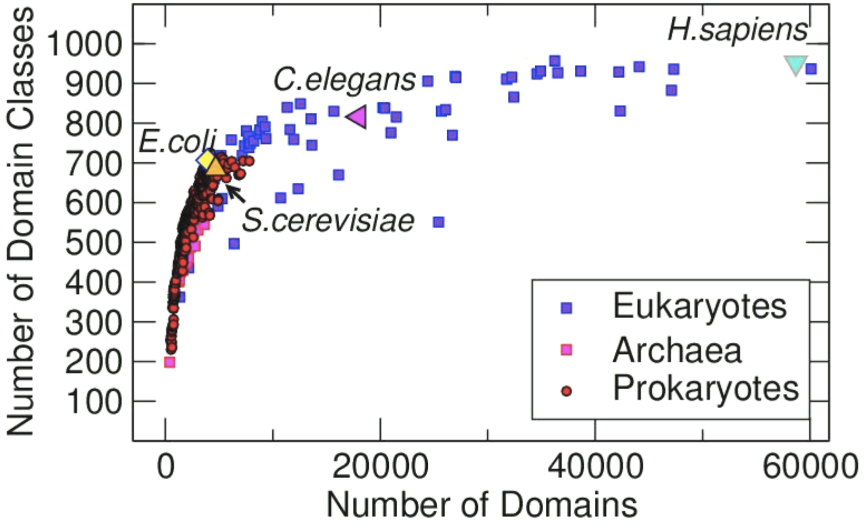
F domain families

Scaling Laws = Common Trends

gene family histogram (1998)

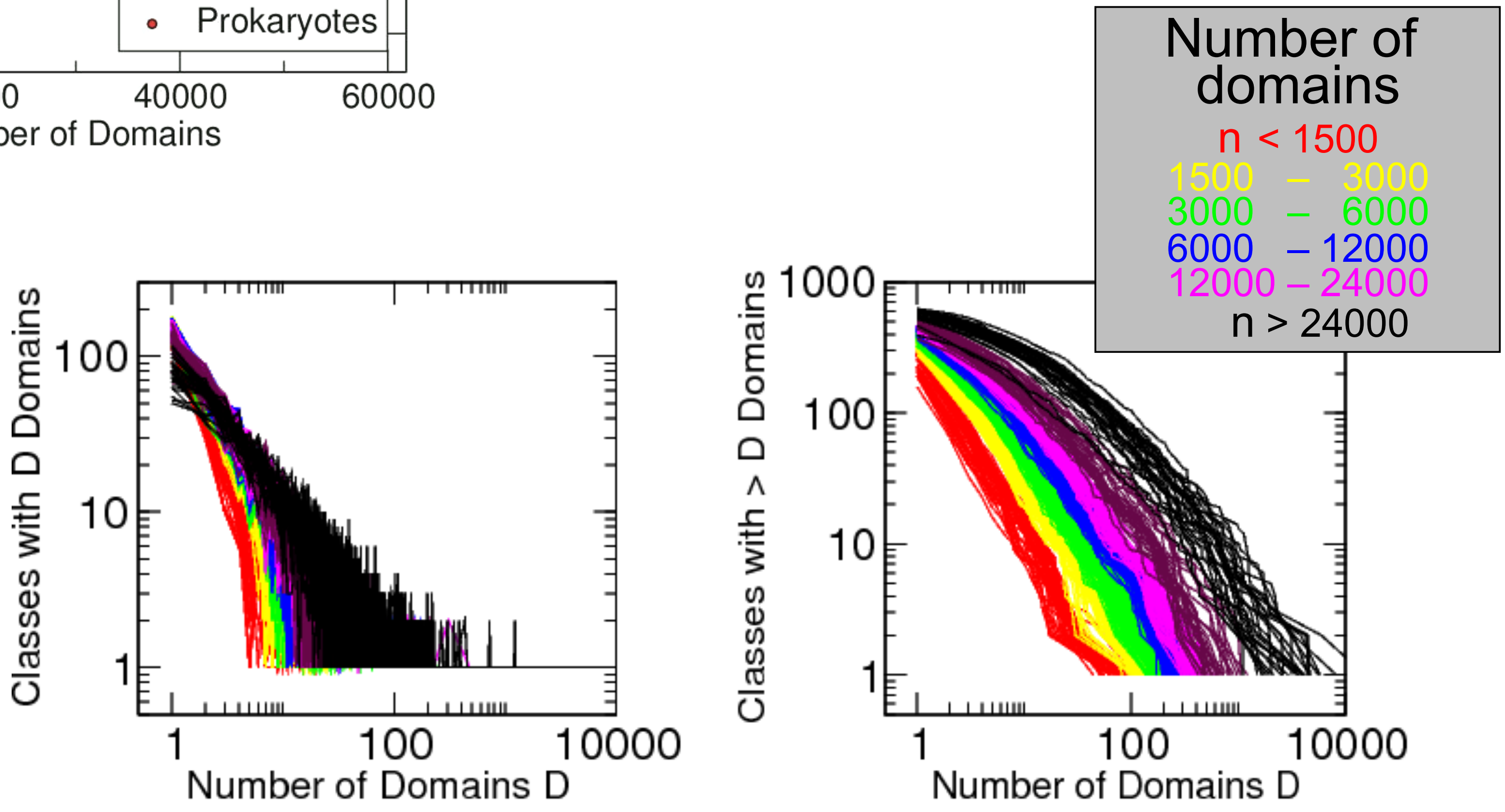


Scaling Laws = Common Trends

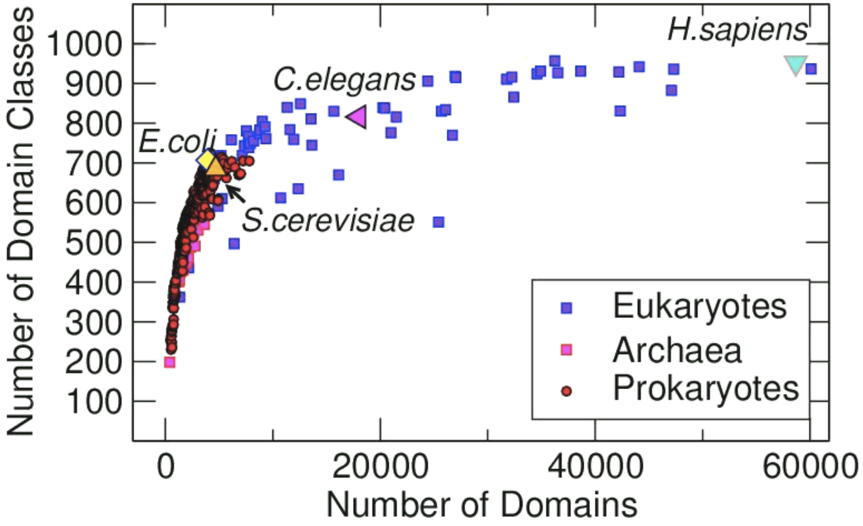


# domain families F vs domains n

domain family histogram

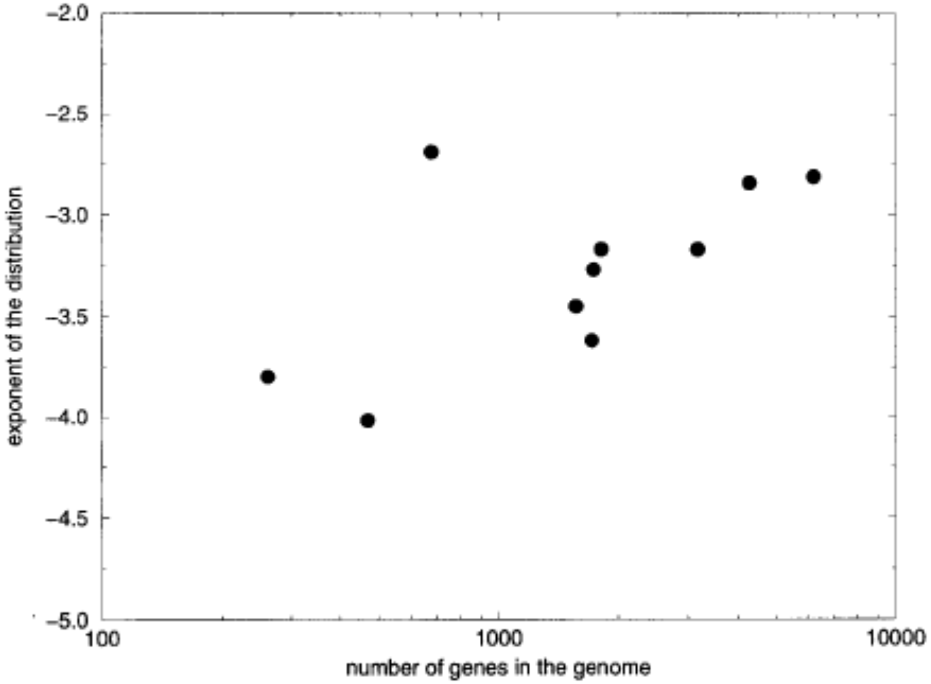


Scaling Laws = Common Trends

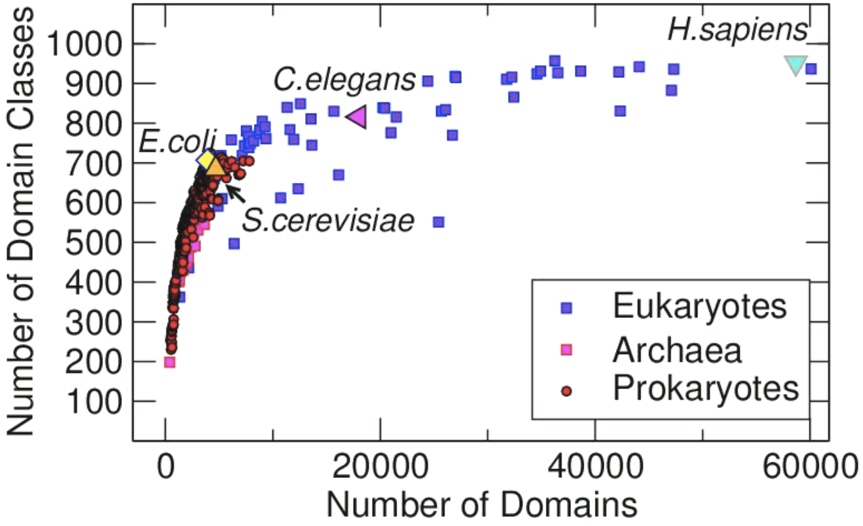


# domain families F
vs domains n

gene family histogram exponent (1998)

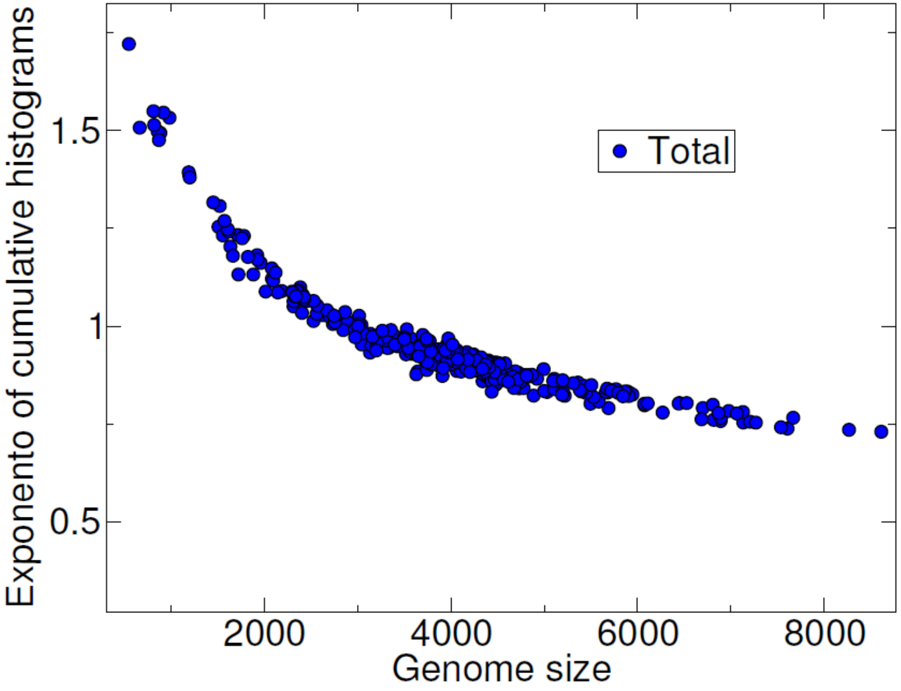


Scaling Laws = Common Trends



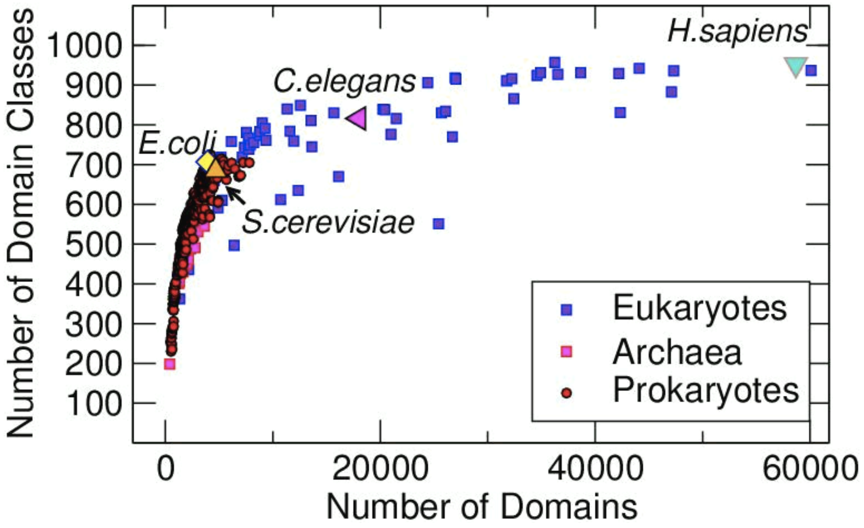
# domain families ***F***
vs domains ***n***

domain family histogram exponent

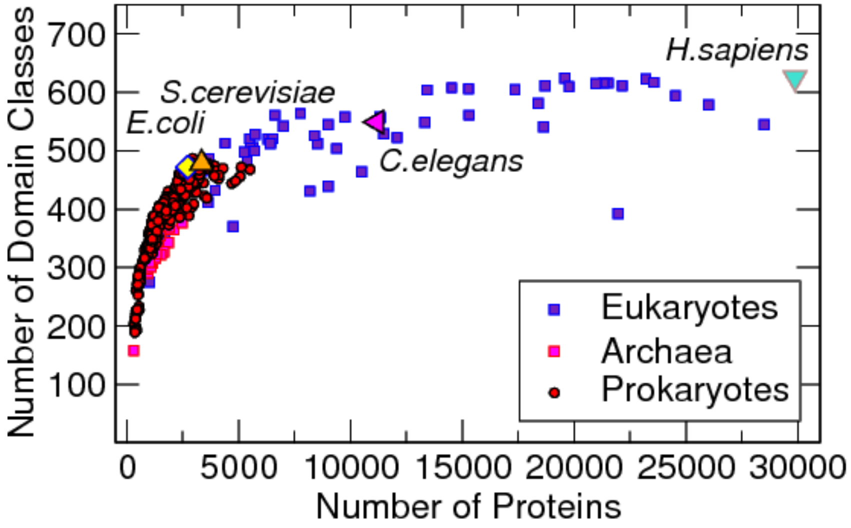


Scaling Laws – Superfamilies & Folds

Superfamilies



Folds



Trend is not dependent
on domain taxonomy level

Functional Annotations

Transcriptional Regulation

Metabolism

Translation

…

Data Structure – One Species

Data Structure – Many Species

	FUNCTION 1					FUNCTION C
	★	★	★	■		⬟
	family 1	family 2	family 3	family 4	...	family F
genome 1	5	0	2	21		5
genome 2	7	0	3	32		7
genome 3	12	2	2	23		2
...	...	...	...	...	...	...
genome G	2	4	2	24		3

row sum

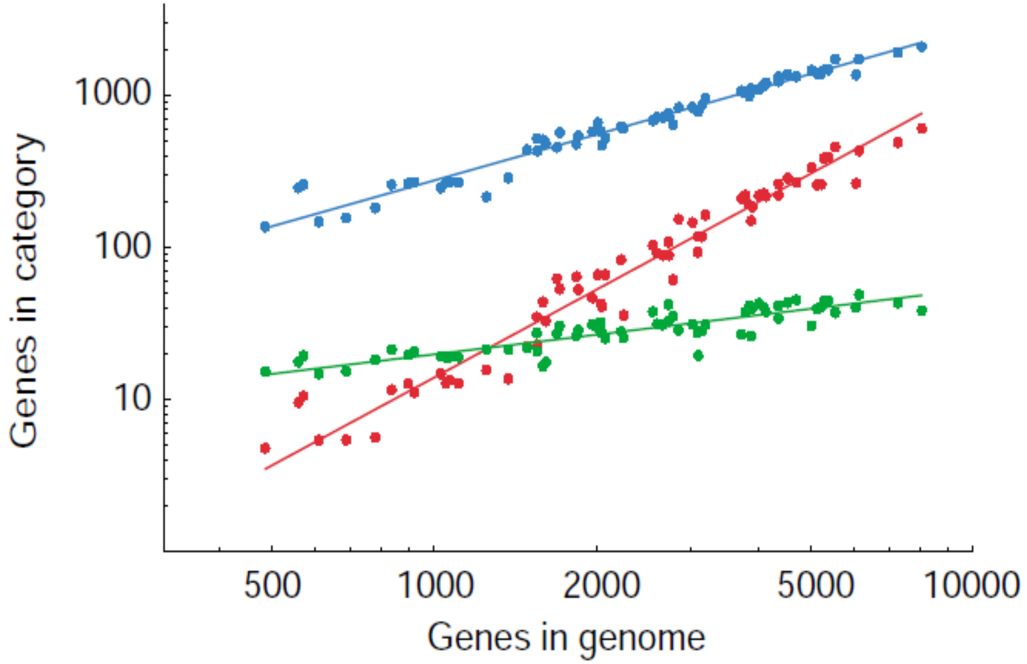
= genome “size”

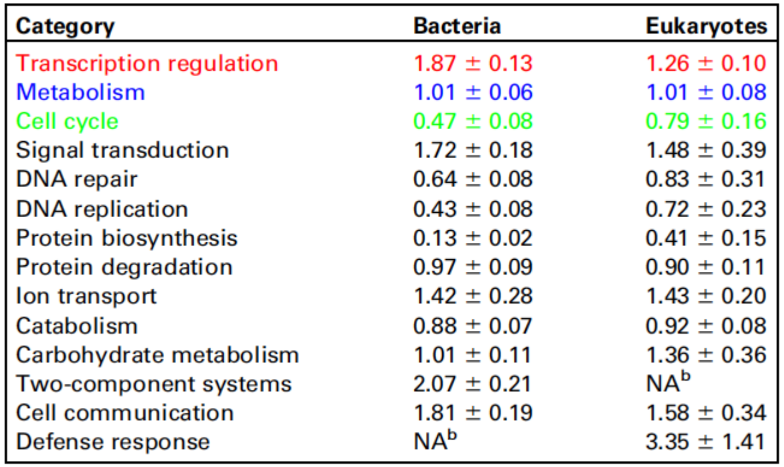
(related by phylogeny)

column sum = total family abundance

2nd “law” scaling of functional categories

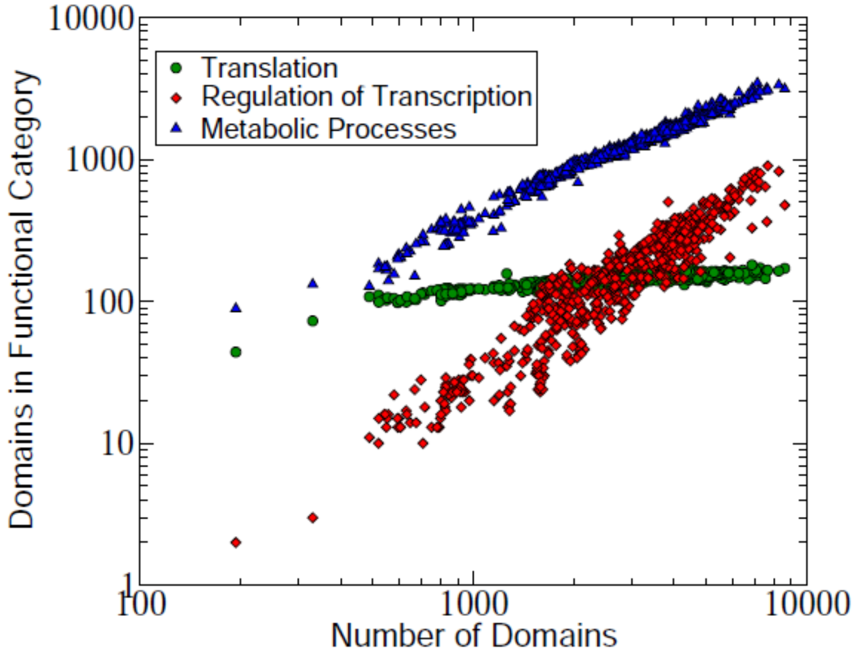
(E.van Nimwegen, 2003)

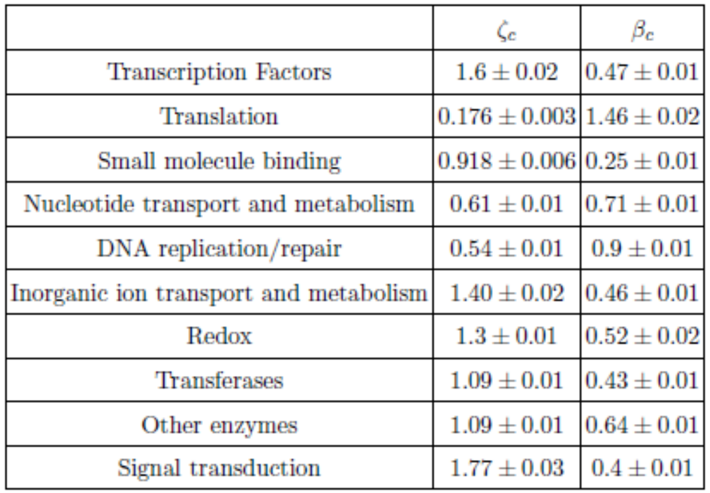




Category	Bacteria	Eukaryotes
Transcription regulation	1.87 ± 0.13	1.26 ± 0.10
Metabolism	1.01 ± 0.06	1.01 ± 0.08
Cell cycle	0.47 ± 0.08	0.79 ± 0.16
Signal transduction	1.72 ± 0.18	1.48 ± 0.39
DNA repair	0.64 ± 0.08	0.83 ± 0.31
DNA replication	0.43 ± 0.08	0.72 ± 0.23
Protein biosynthesis	0.13 ± 0.02	0.41 ± 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

2nd "law" scaling of functional categories





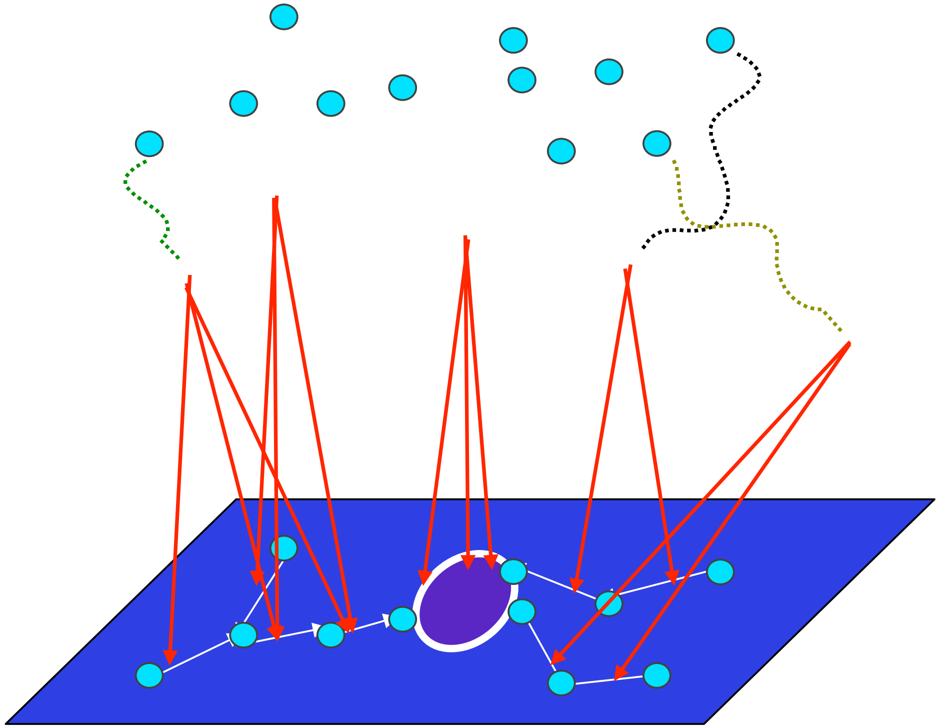
	ζ_c	β_c
Transcription Factors	1.6 ± 0.02	0.47 ± 0.01
Translation	0.176 ± 0.003	1.46 ± 0.02
Small molecule binding	0.918 ± 0.006	0.25 ± 0.01
Nucleotide transport and metabolism	0.61 ± 0.01	0.71 ± 0.01
DNA replication/repair	0.54 ± 0.01	0.9 ± 0.01
Inorganic ion transport and metabolism	1.40 ± 0.02	0.46 ± 0.01
Redox	1.3 ± 0.01	0.52 ± 0.02
Transferases	1.09 ± 0.01	0.43 ± 0.01
Other enzymes	1.09 ± 0.01	0.64 ± 0.01
Signal transduction	1.77 ± 0.03	0.4 ± 0.01

“Spherical cow” view on metabolic and transcription networks

Metabolites

Transcriptional Regulation

Metabolism



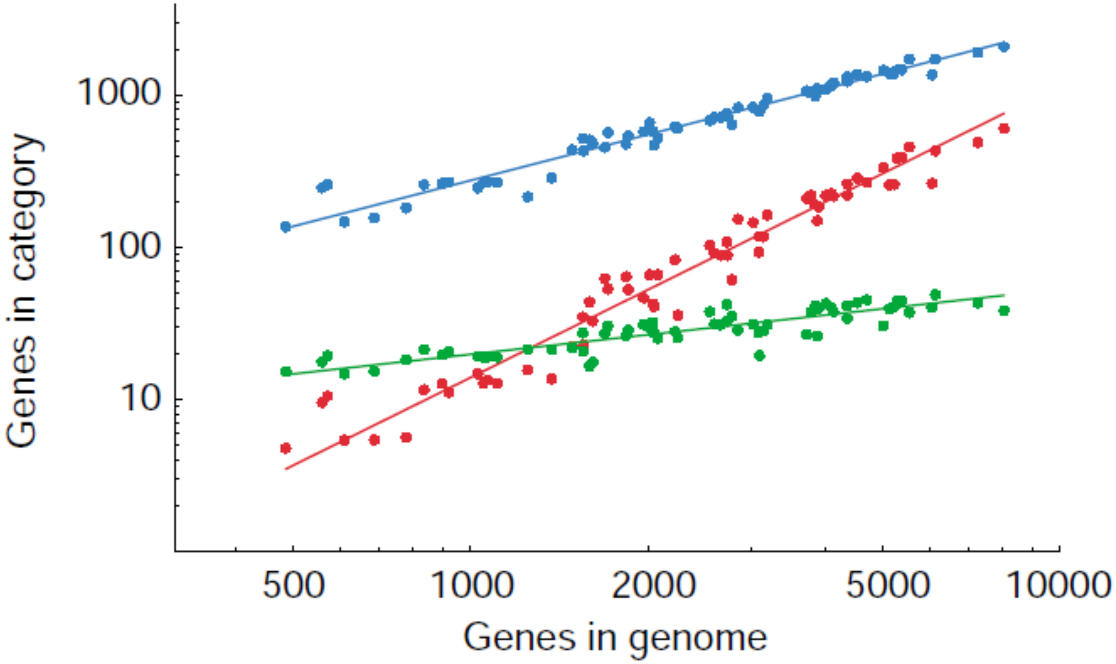
Growth by HGT:
Add pathways
Add Transcription Factors

1) Partitioning of a genome into functional categories

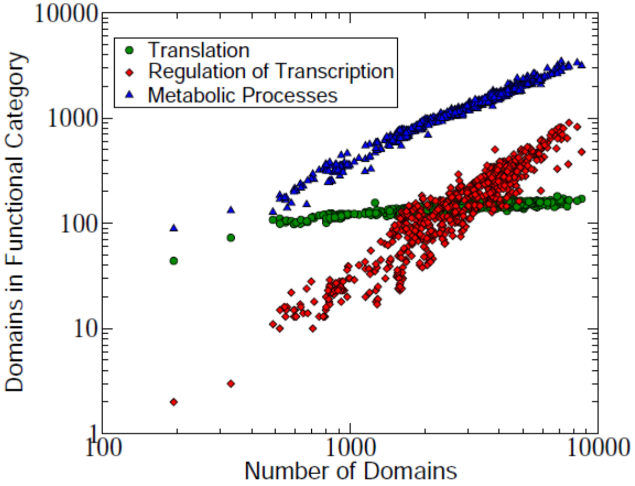
(Monod at the genome scale)

Category counts for many genomes

(E.van Nimwegen, 2003)



More recent Data:



Near-quadratic scaling for TFs

Tells us about regulatory complexity vs genome size

TF<Kout> = NG<Kin> = # edges, hence

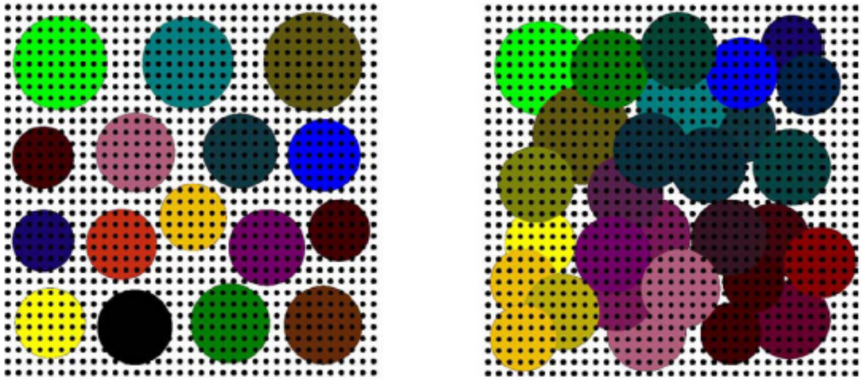
TF/NG = <Kin>/<Kout> increases with NG

<Kout> decreases: functions become more specialized
<Kin> increases: regulation becomes more interconnected

(likely both phenomena occur)

Hypotheses for the scaling of TFs = RECIPES

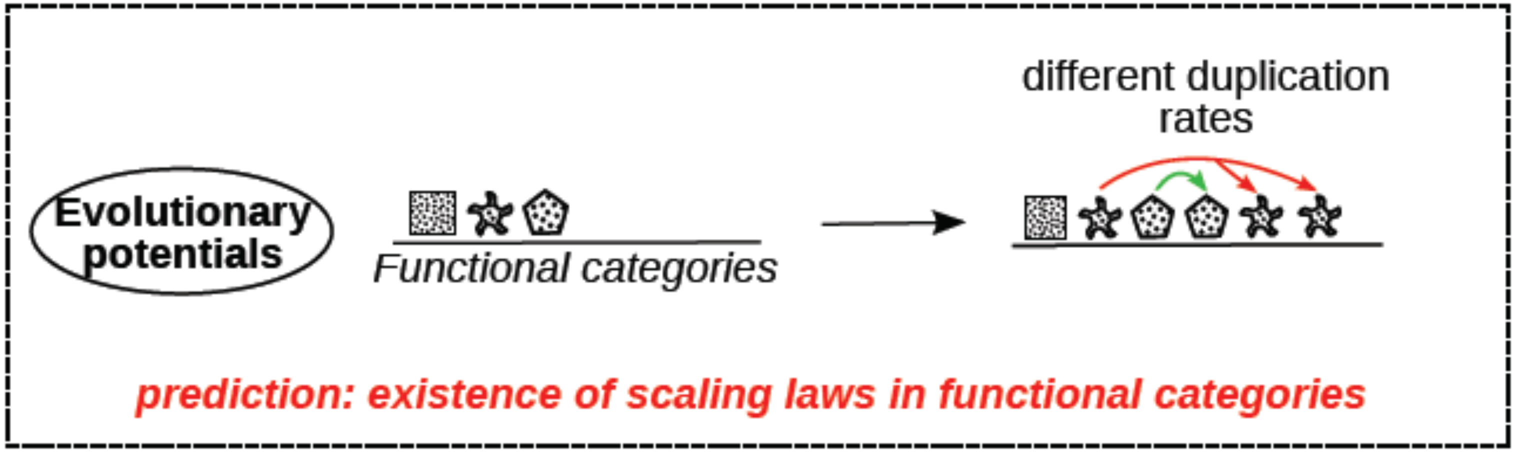
Coding limits?



Optimization of the number of expression patterns?

Constraints in genome growth?

Growth Model for Functional Categories


Evolutionary potentials
Functional categories
different duplication rates
prediction: existence of scaling laws in functional categories

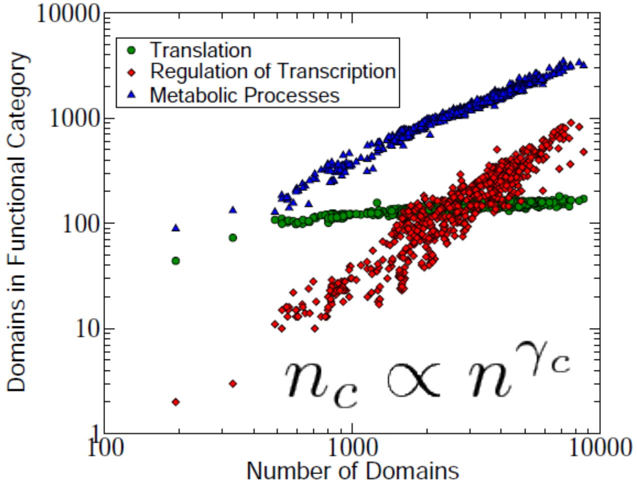
“Evolutionary Potentials”

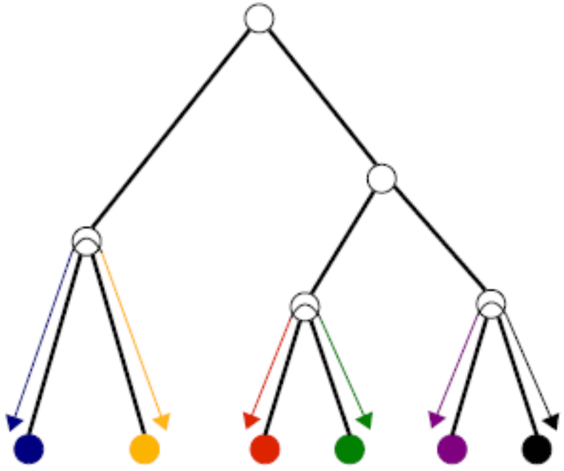
(Molina and van Nimwegen, *Trends Genet.* 2009)

“Preferential Attachment” + Specificity $\frac{dn_c}{dn} \propto \rho_c \frac{n_c}{n}$

Observed scaling law $n_c \propto n^{\gamma_c} \rightarrow \frac{dn_c}{dn} \propto \gamma_c \frac{n_c}{n}$

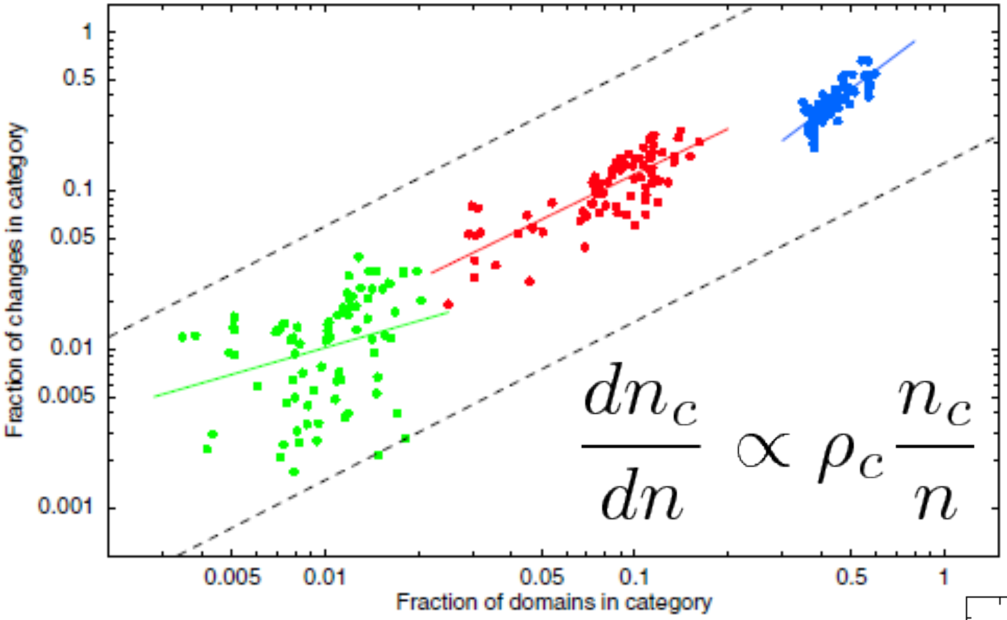
Expected equality exponent - potential $\rho_c = \gamma_c \quad \forall c$



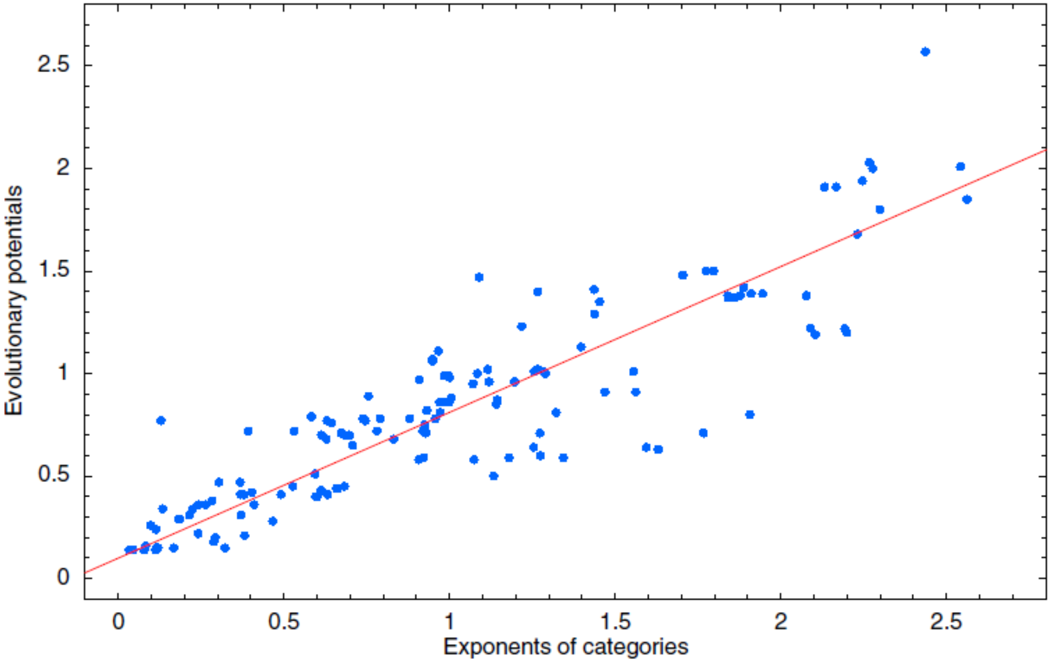


$\frac{dn_c}{dn} \propto \rho_c \frac{n_c}{n}$

Estimate of evolutionary potentials



$$\frac{dn_c}{dn} \propto \rho_c \frac{n_c}{n}$$



Note: normalization couples the growth of different functions!

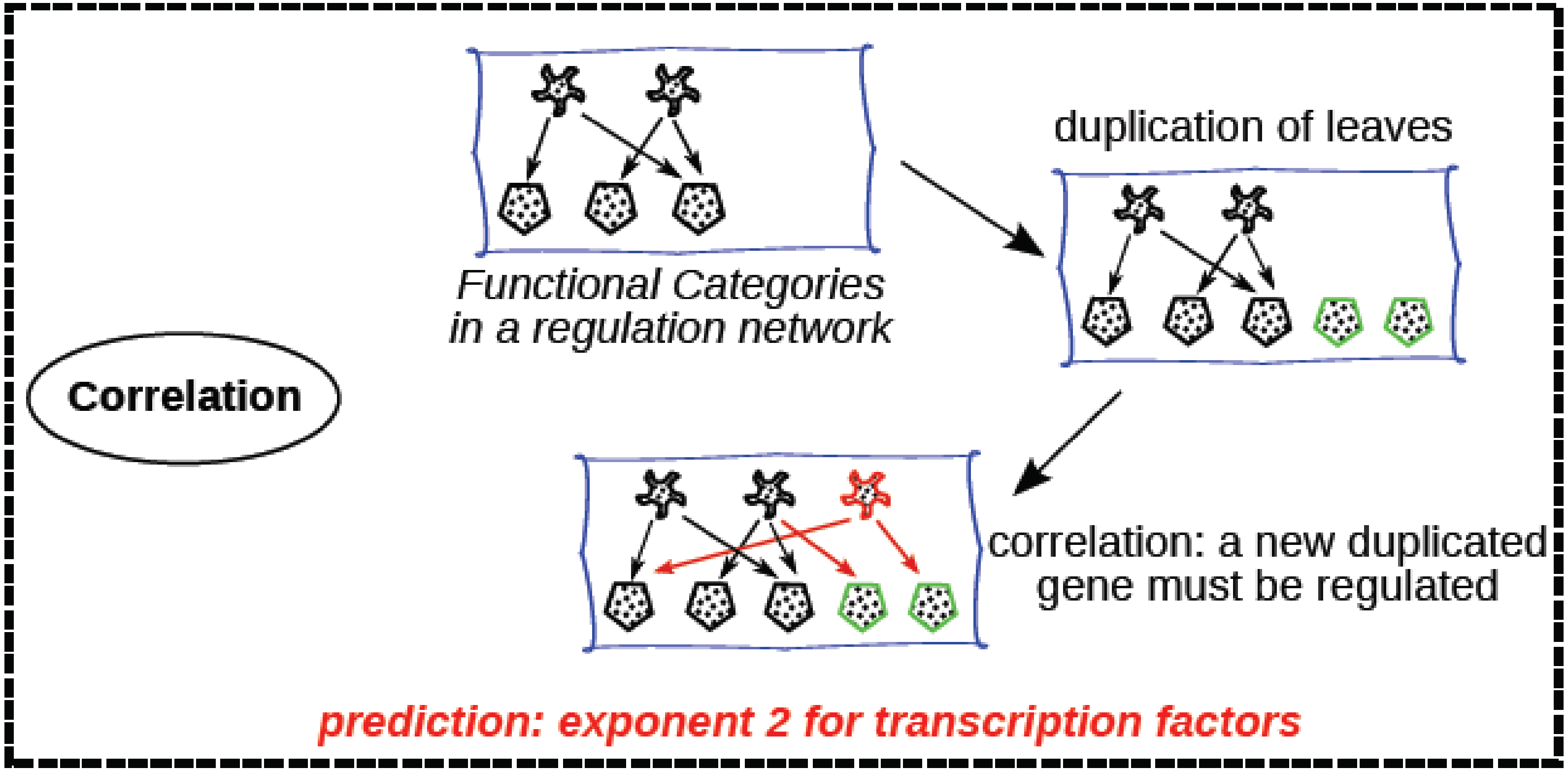
$$\frac{dn_c}{dn} = \rho_c \frac{n_c}{C(n)}$$ is consistent if $$C(n) = \sum_c \rho_c n_c$$

because $$dn = \sum_c dn_c$$

Also one needs $$C(n) \sim n$$

(more on this tomorrow…)

Alternative Picture: Correlated Expansion of Functional Categories



Correlation

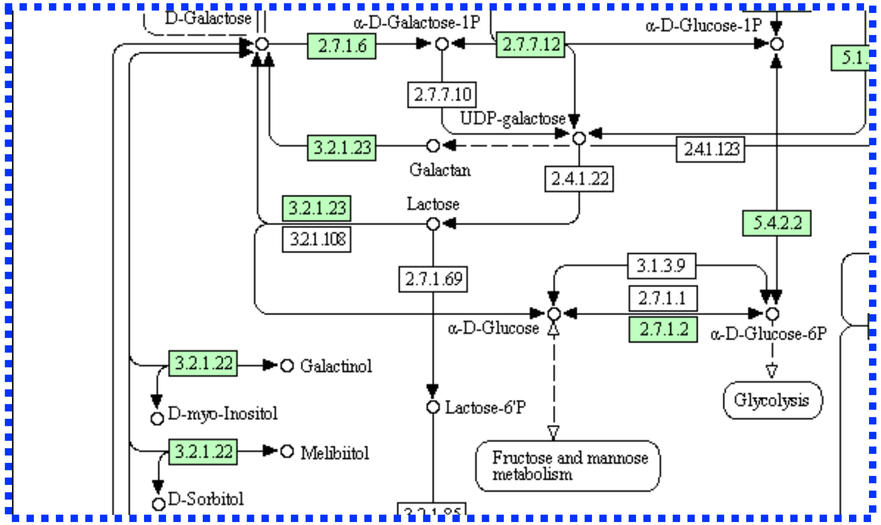
Functional Categories in a regulation network

duplication of leaves

correlation: a new duplicated gene must be regulated

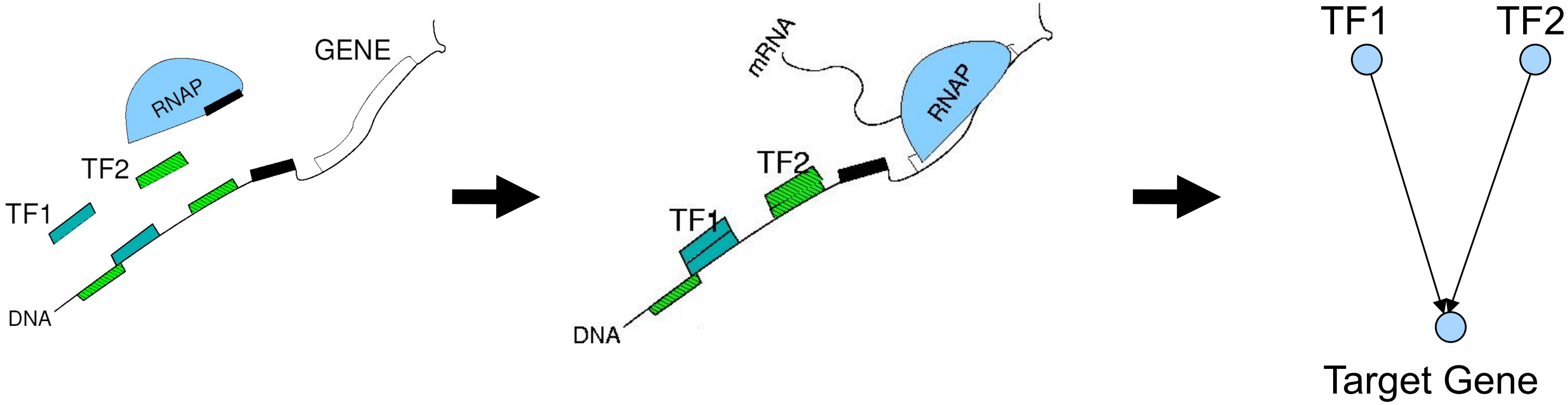
prediction: exponent 2 for transcription factors

Metabolism at Large Scale

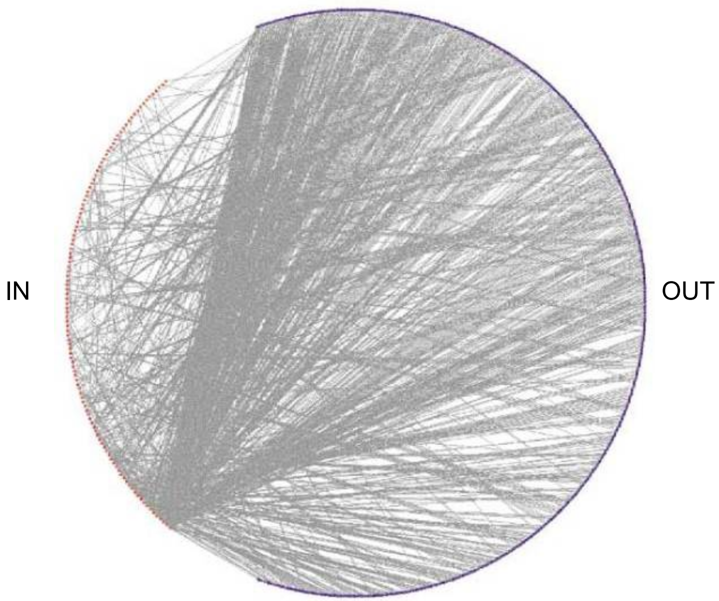


Metabolic
network

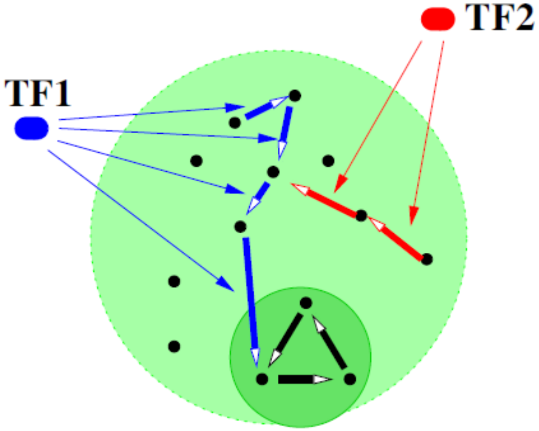
Transcription at Large Scale /1



E.coli network

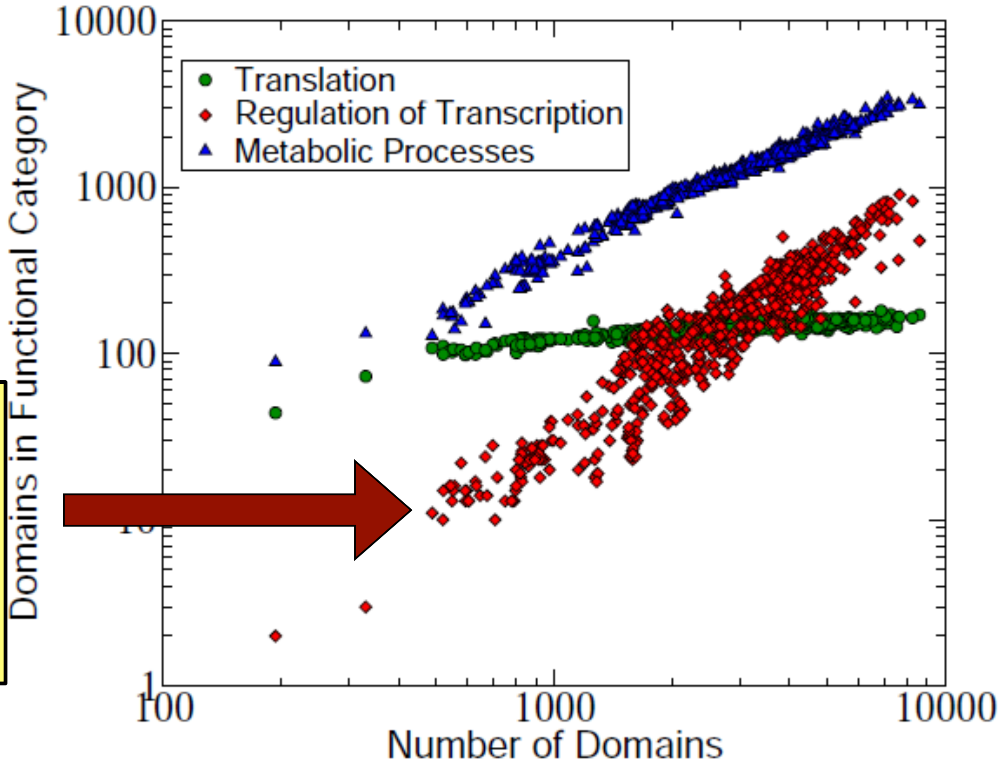


Back to operon model: transcription factors and metabolic enzymes



Related to regulatory network size needed to control *~n targets*

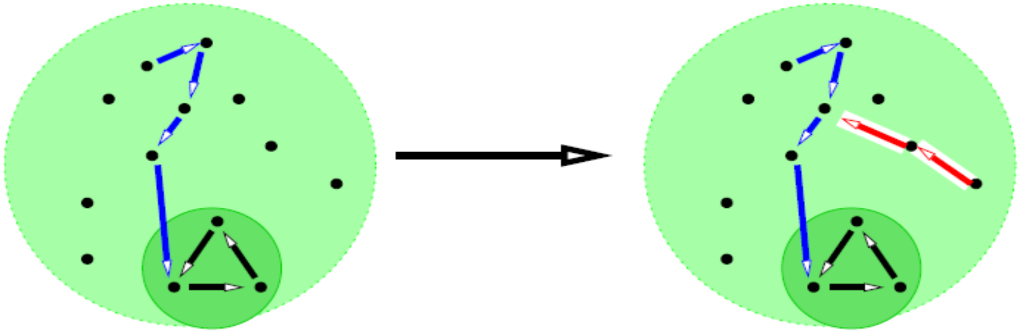
Exponent ~two for transcription factors



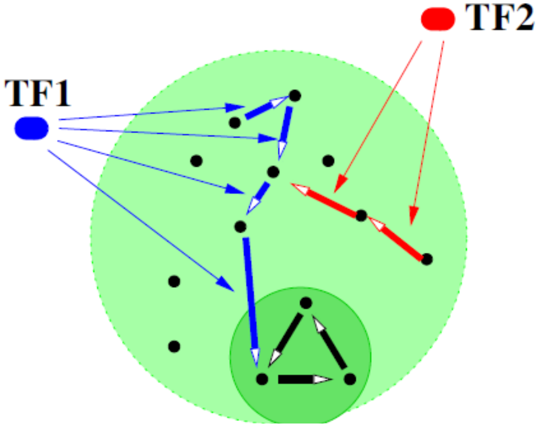
“Toolbox model” for large-scale transcription and metabolism

(Maslov et al PNAS 2009)

A universal and finite metabolic network exists
New branch = random walk



Each new branch must be regulated by a transcription factor



$$\begin{cases} \Delta n_{met} = \dfrac{U}{n_{met}} \\ \Delta n_{TF} = 1 \end{cases}$$

$$\Delta n_{TF} / \Delta n_{met} = n_{met} / U$$

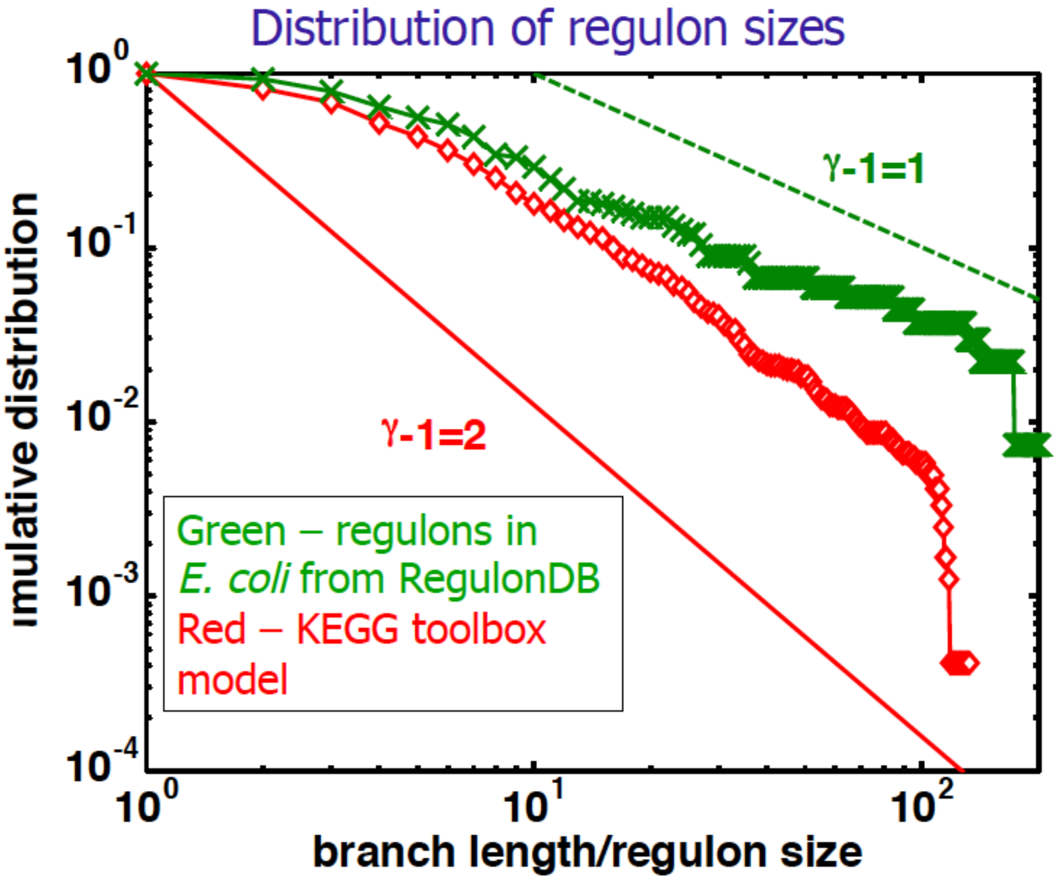
→ quadratic scaling

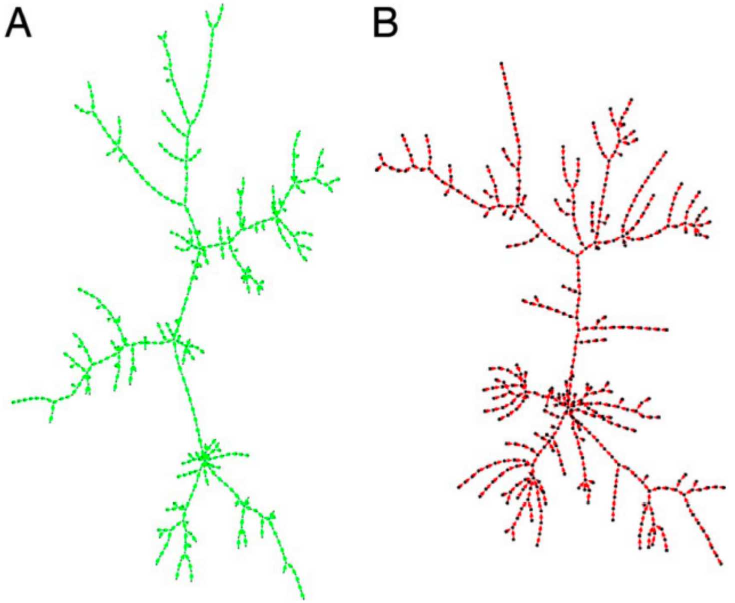
Predictions of the Toolbox model

Should work with real-world metabolism (KEGG) *works*

Power-law distribution of pathway size P(s) ~ 1/s^3

Same distribution for regulon size





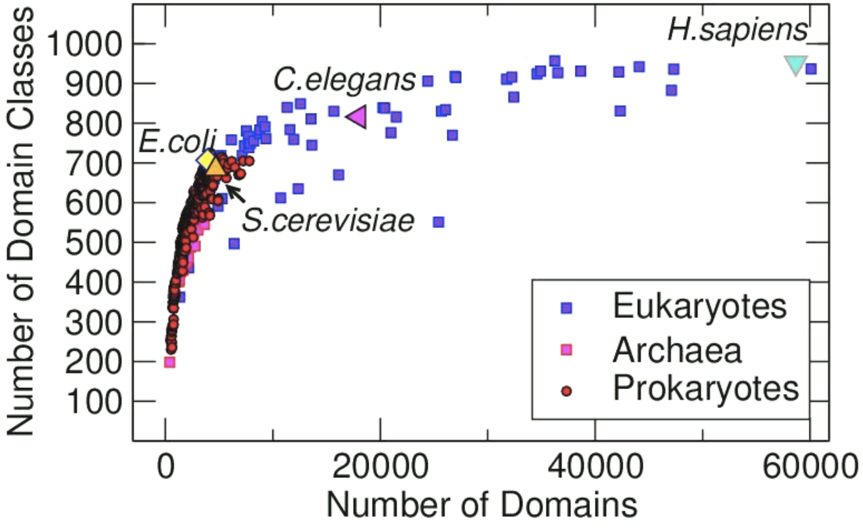
Toolbox model

E. coli metabolic network
(spanning tree)

2) Partitioning of a genome into evolutionary families

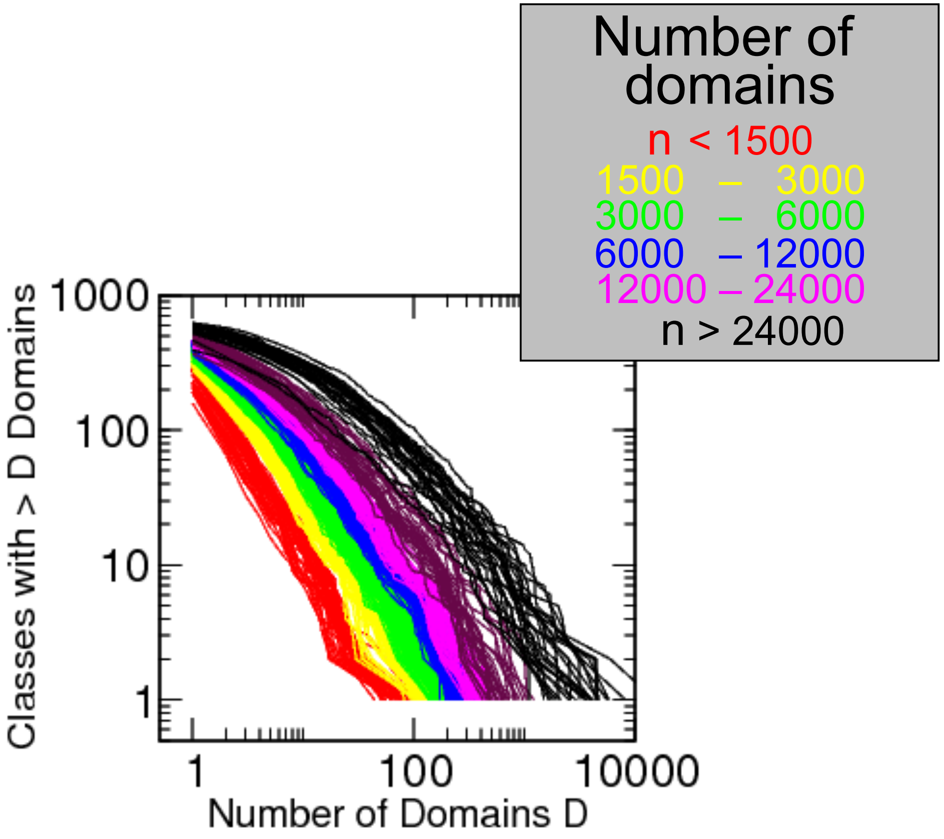
(Dayhoff's Dream)

Scaling Laws for Evolutionary classes



Number of evolutionary families
# classes F
vs genome size n

Population distribution of evolutionary families
class population
cumulative histogram

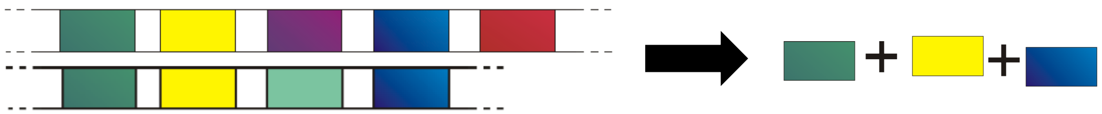


The existence of these scaling laws is
surprising

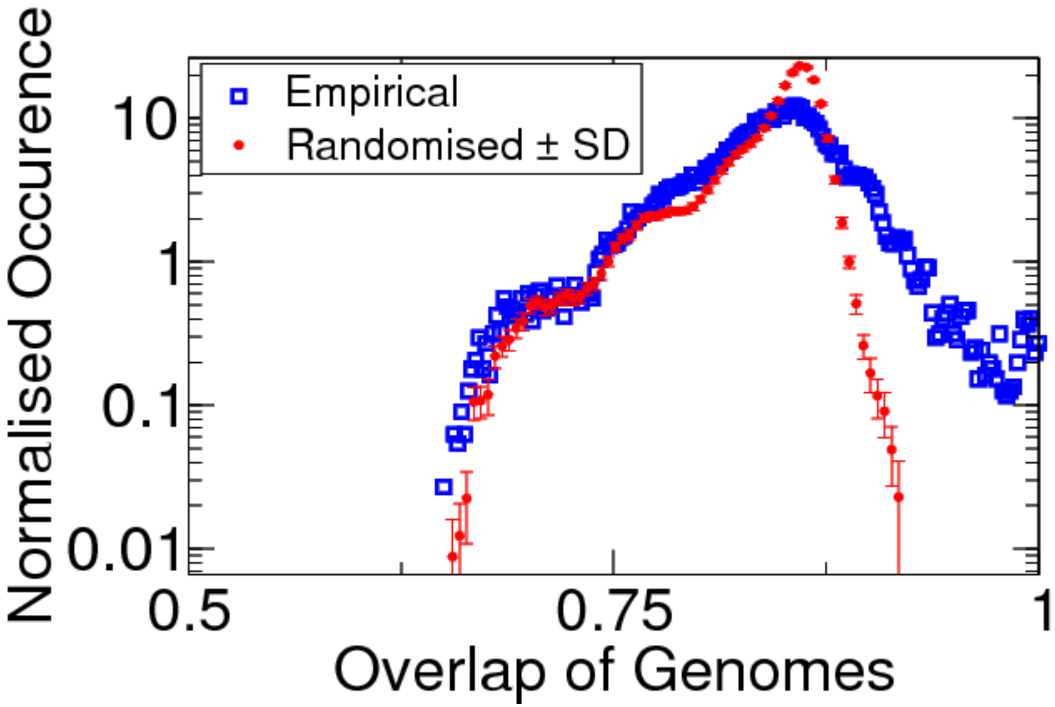
It indicates that domain class partitioning
depends on size
and not on the specific
evolutionary history of a genome

Genome Overlap

Pair of genomes



Common domain class usage



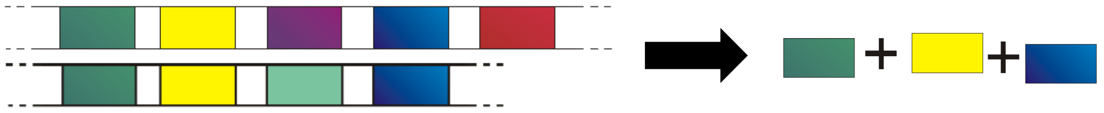
It’s a spin overlap

$$O(g',g'') = \frac{1}{D}\sum_{i=1}^{D}\delta(\sigma_i^{g'},\sigma_i^{g''})$$

$$\sigma_i^g = \begin{cases} 1 & \text{if domain class i is present in genome g} \\ -1 & \text{if domain class i is not present in genome g} \end{cases}$$

Genome Clustering by Overlap

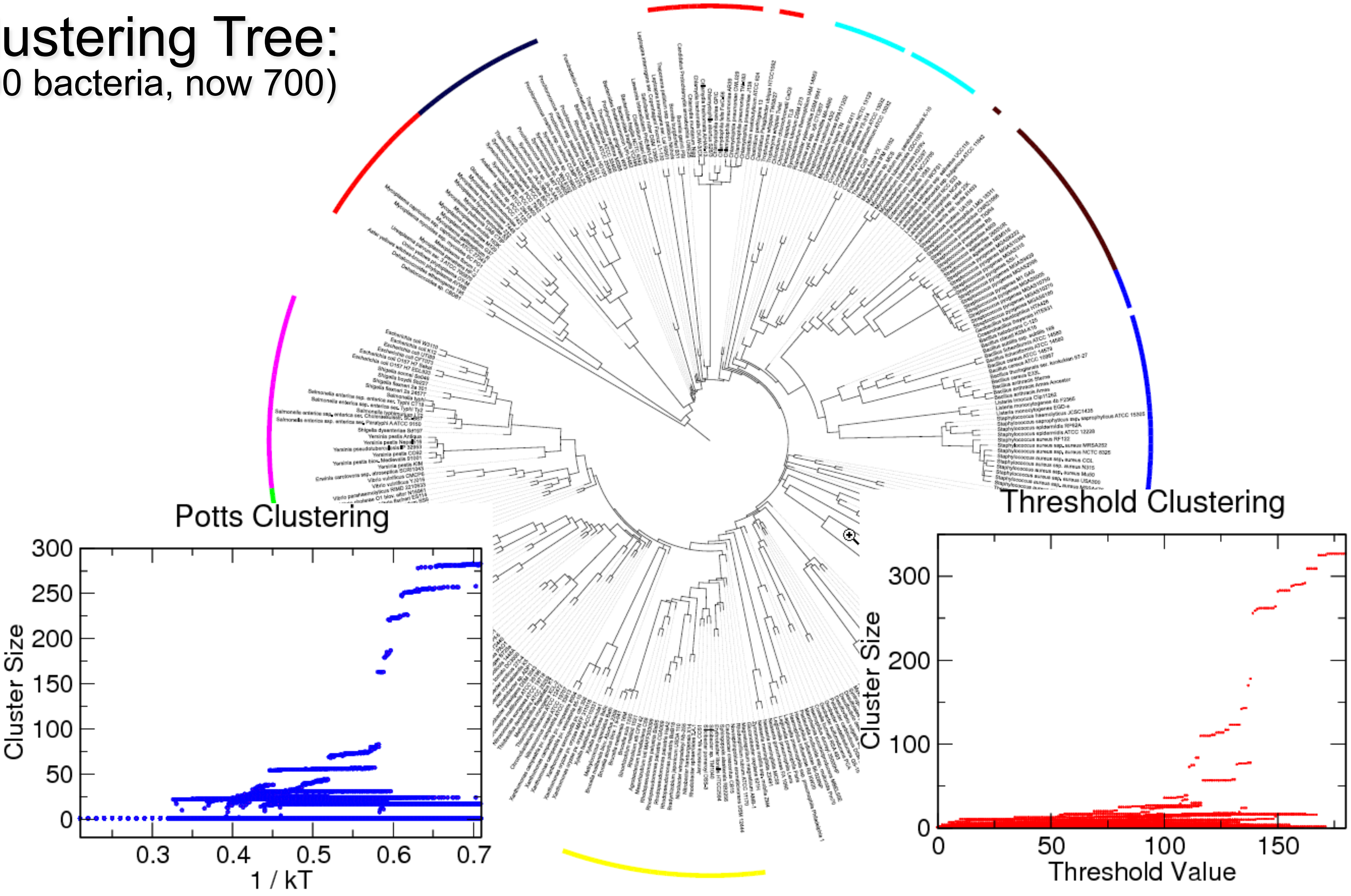
Pair of genomes



Common domain class usage

Clustering Tree:

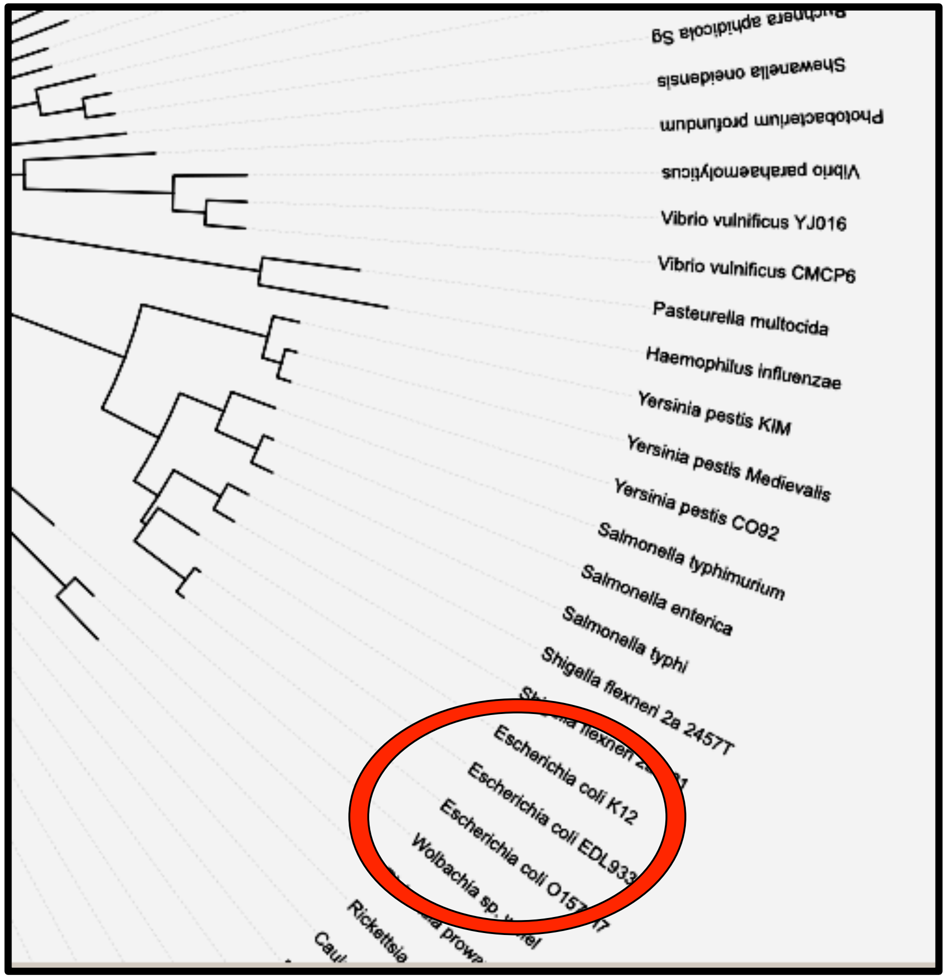
(400 bacteria, now 700)


Potts Clustering

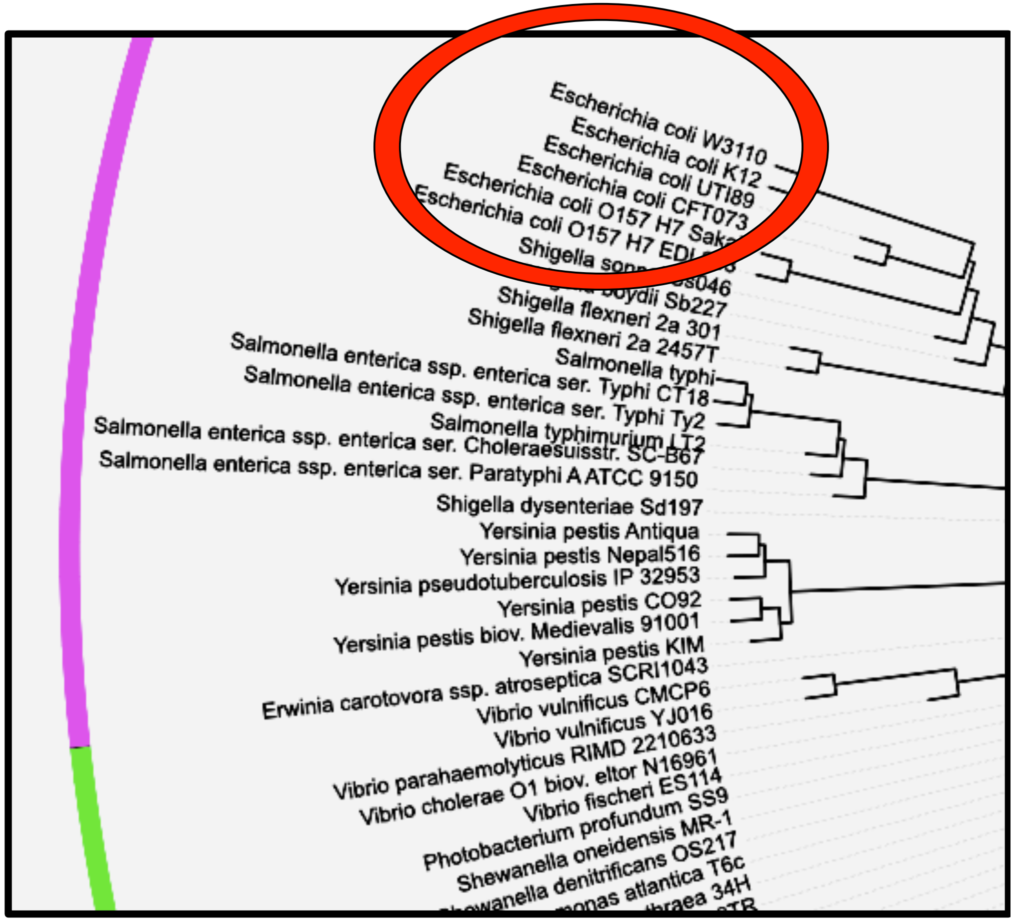
Threshold Clustering

Phylogenetic Tree!

SHOT Prokaryote tree
(gene order + shared orthologs)



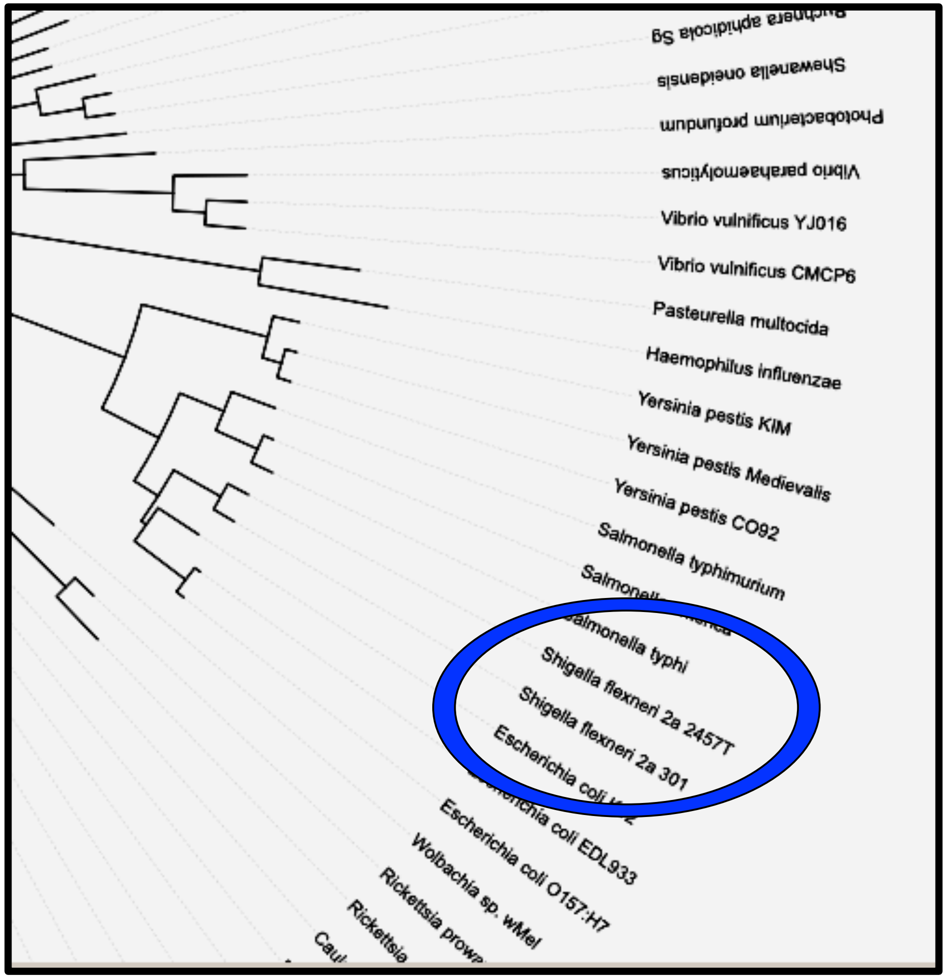
Clusters of Genome Domain Families



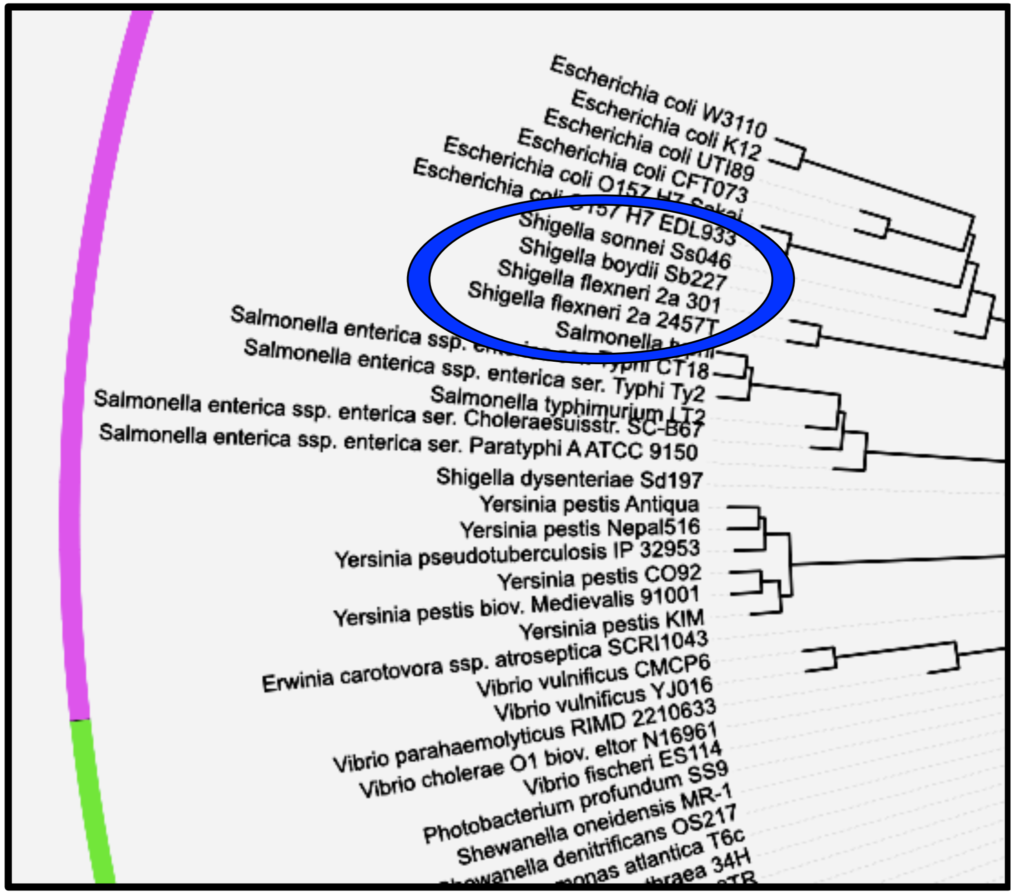
Escherichia coli W3110
Escherichia coli K12
Escherichia coli UTI89
Escherichia coli CFT073
Escherichia coli O157 H7 Sakai
Escherichia coli O157 H7 EDL933
Shigella sonnei Ss046
Shigella boydii Sb227
Shigella flexneri 2a 301
Shigella flexneri 2a 2457T
Salmonella typhi
Salmonella enterica ssp. enterica ser. Typhi CT18
Salmonella enterica ssp. enterica ser. Typhi Ty2
Salmonella typhimurium LT2
Salmonella enterica ssp. enterica ser. Choleraesuisstr. SC-B67
Salmonella enterica ssp. enterica ser. Paratyphi A ATCC 9150
Shigella dysenteriae Sd197
Yersinia pestis Antiqua
Yersinia pestis Nepal516
Yersinia pseudotuberculosis IP 32953
Yersinia pestis CO92
Yersinia pestis biov. Medievalis 91001
Yersinia pestis KIM
Erwinia carotovora ssp. atroseptica SCRI1043
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Vibrio parahaemolyticus RIMD 2210633
Vibrio cholerae O1 biov. eltor N16961
Vibrio fischeri ES114
Photobacterium profundum SS9
Shewanella oneidensis MR-1
Shewanella denitrificans OS217
monas atlantica T6c
hraea 34H

Phylogenetic Tree!

SHOT Prokaryote tree



Clusters of Genome Domain Families

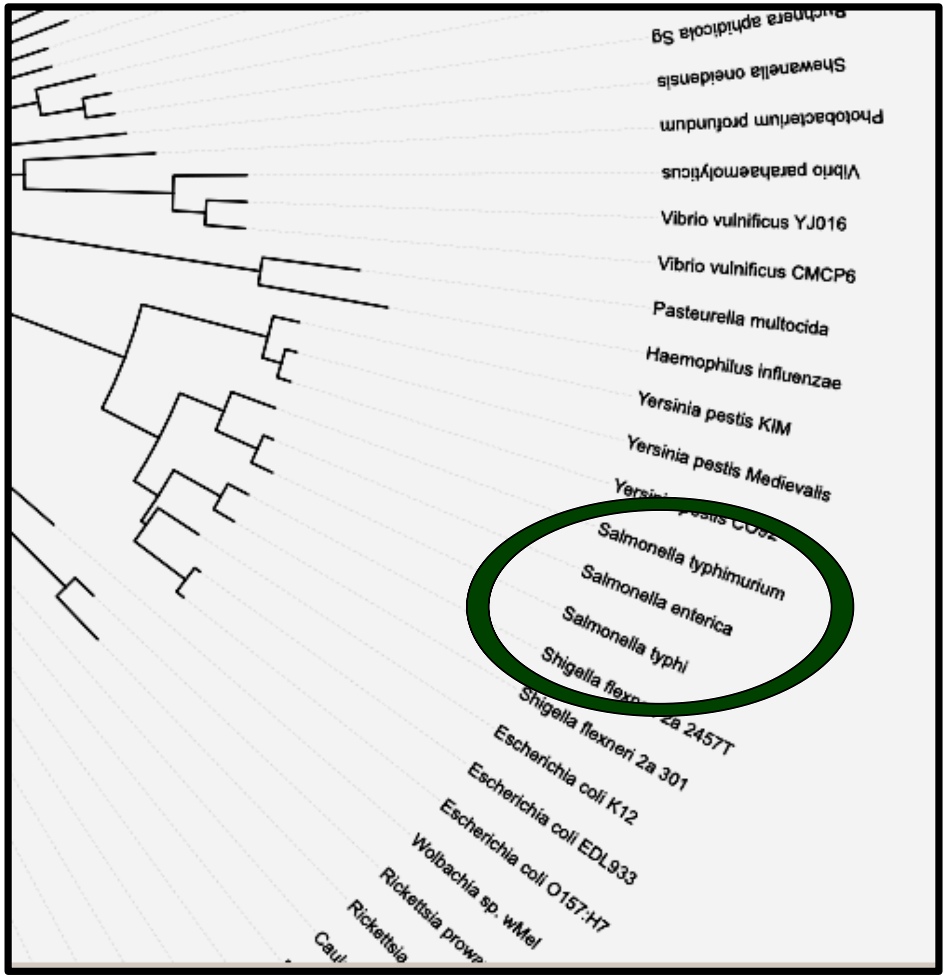


Escherichia coli W3110
Escherichia coli K12
Escherichia coli UTI89
Escherichia coli CFT073
Escherichia coli O157 H7 Sakai
Escherichia coli O157 H7 EDL933
Shigella sonnei Ss046
Shigella boydii Sb227
Shigella flexneri 2a 301
Shigella flexneri 2a 2457T
Salmonella typhi
Salmonella enterica ssp. enterica ser. Typhi CT18
Salmonella enterica ssp. enterica ser. Typhi Ty2
Salmonella typhimurium LT2
Salmonella enterica ssp. enterica ser. Choleraesuisstr. SC-B67
Salmonella enterica ssp. enterica ser. Paratyphi A ATCC 9150
Shigella dysenteriae Sd197
Yersinia pestis Antiqua
Yersinia pestis Nepal516
Yersinia pseudotuberculosis IP 32953
Yersinia pestis CO92
Yersinia pestis biov. Medievalis 91001
Yersinia pestis KIM
Erwinia carotovora ssp. atroseptica SCRI1043
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Vibrio parahaemolyticus RIMD 2210633
Vibrio cholerae O1 biov. eltor N16961
Vibrio fischeri ES114
Photobacterium profundum SS9
Shewanella oneidensis MR-1
Shewanella denitrificans OS217
monas atlantica T6c
hraea 34H

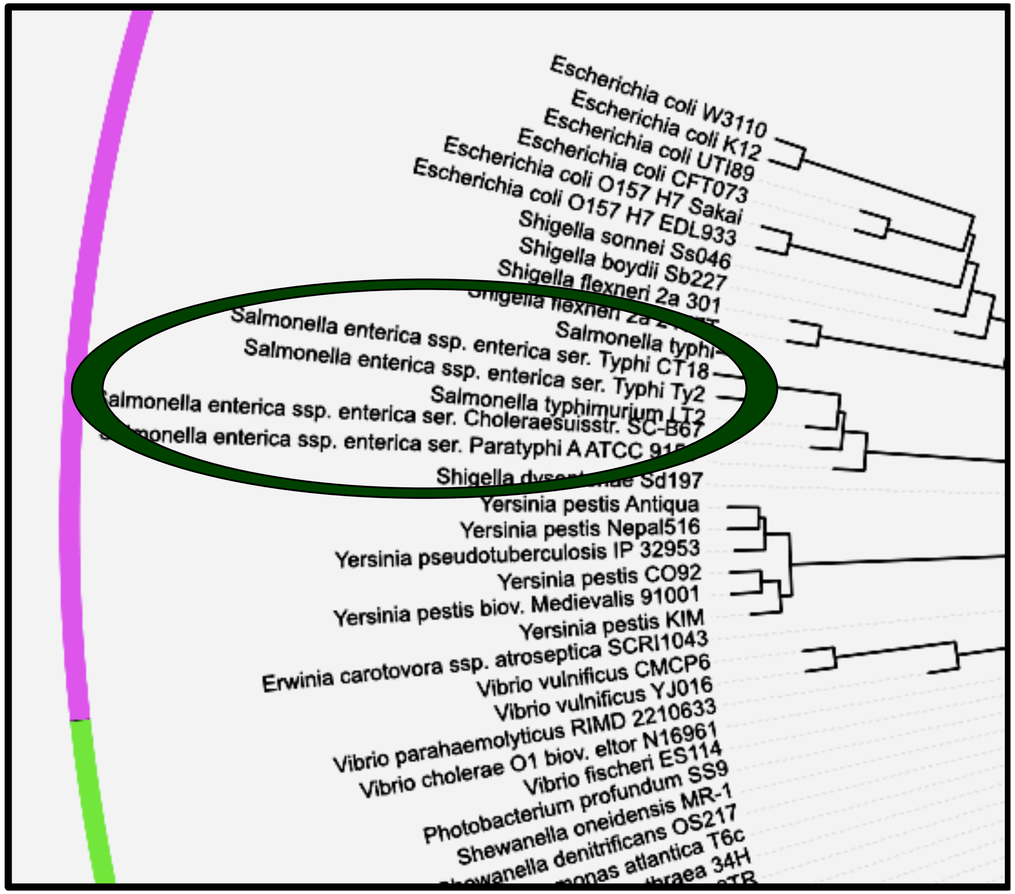
Phylogenetic Tree!

Better signal accounting for domain classes that are absent in both genomes when measuring overlap

SHOT Prokaryote tree

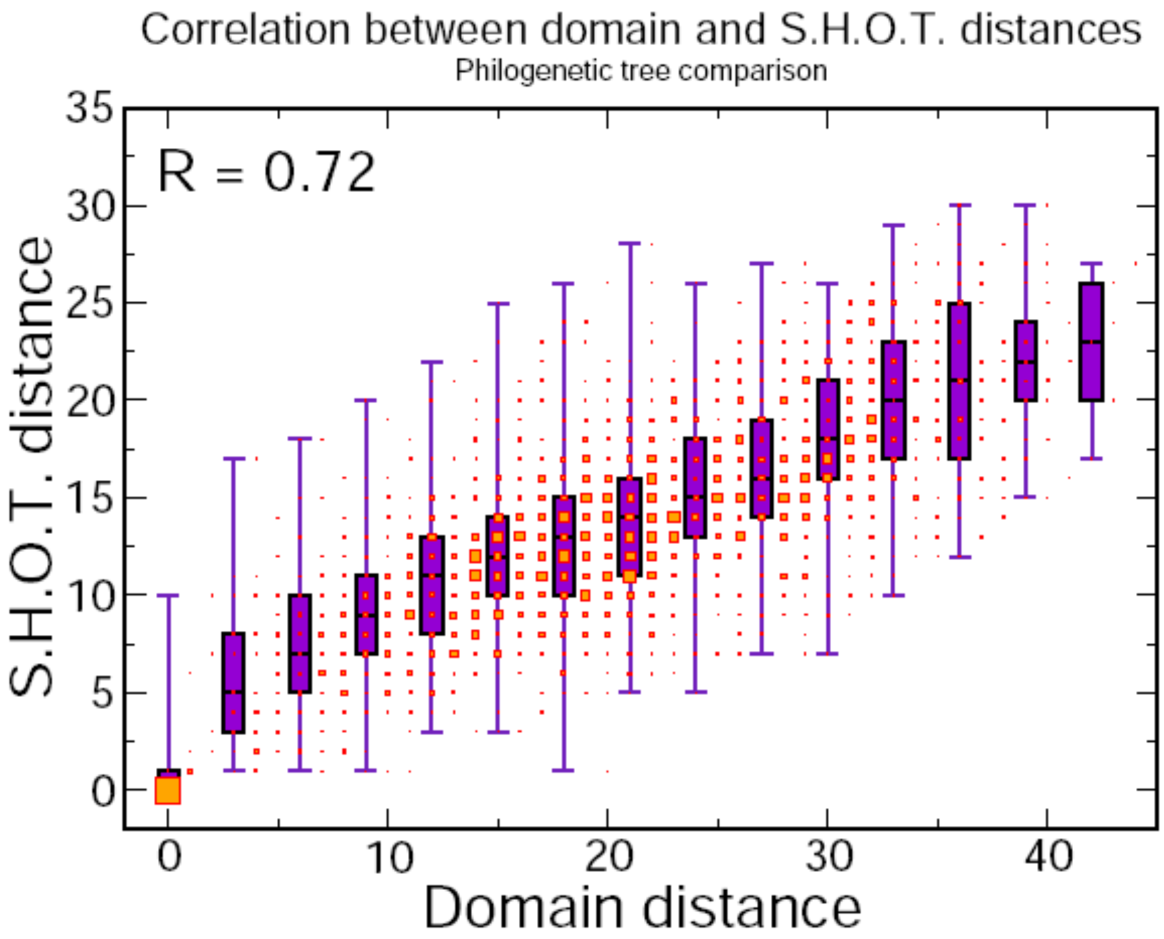


Clusters of Genome Domain Families


Escherichia coli W3110
Escherichia coli K12
Escherichia coli UTI89
Escherichia coli CFT073
Escherichia coli O157 H7 Sakai
Escherichia coli O157 H7 EDL933
Shigella sonnei Ss046
Shigella boydii Sb227
Shigella flexneri 2a 301
Shigella flexneri 2a 2457T
Salmonella typhi
Salmonella enterica ssp. enterica ser. Typhi CT18
Salmonella enterica ssp. enterica ser. Typhi Ty2
Salmonella typhimurium LT2
Salmonella enterica ssp. enterica ser. Choleraesuisstr. SC-B67
Salmonella enterica ssp. enterica ser. Paratyphi A ATCC 915
Shigella dysenteriae Sd197
Yersinia pestis Antiqua
Yersinia pestis Nepal516
Yersinia pseudotuberculosis IP 32953
Yersinia pestis CO92
Yersinia pestis biov. Medievalis 91001
Yersinia pestis KIM
Erwinia carotovora ssp. atroseptica SCRI1043
Vibrio vulnificus CMCP6
Vibrio vulnificus YJ016
Vibrio parahaemolyticus RIMD 2210633
Vibrio cholerae O1 biov. eltor N16961
Vibrio fischeri ES114
Photobacterium profundum SS9
Shewanella oneidensis MR-1
Shewanella denitrificans OS217
monas atlantica T6c
thraea 34H

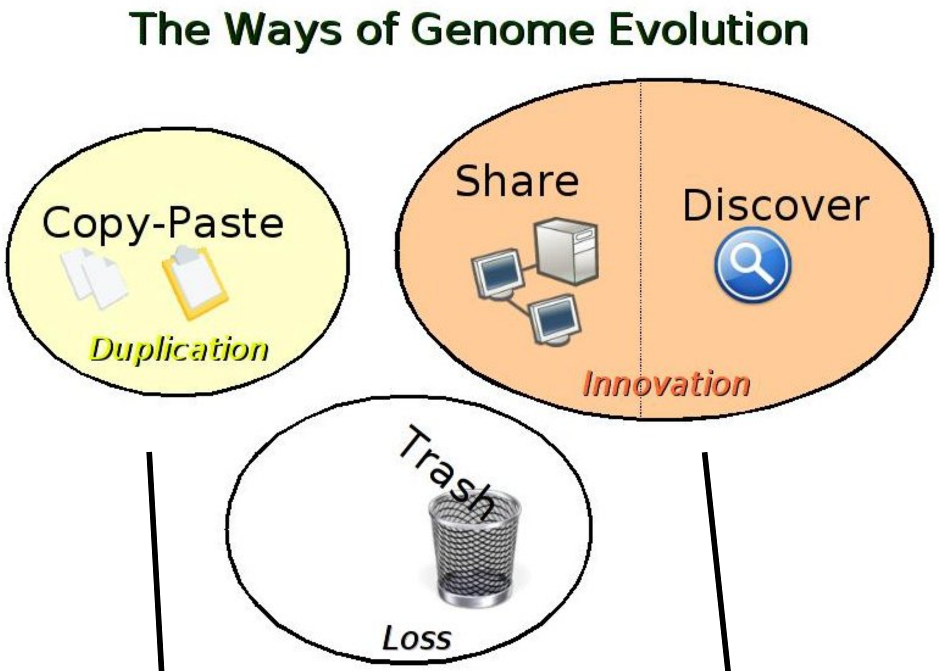
Clustering genomes by domain class overlap gives a phylogenetic Tree

Reference Phylogenetic tree

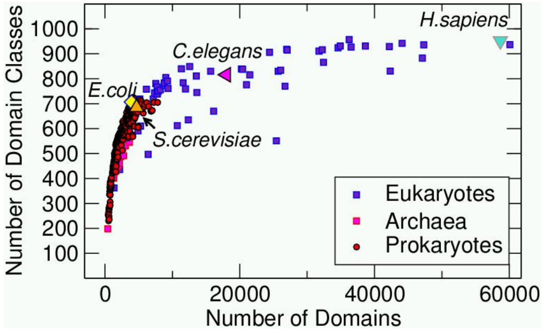


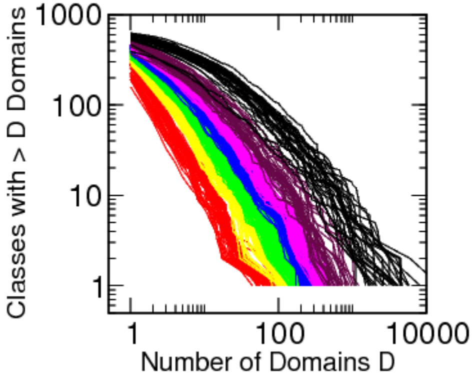
Clusters of Genome Domain Families

Duplication / Innovation / Loss Model



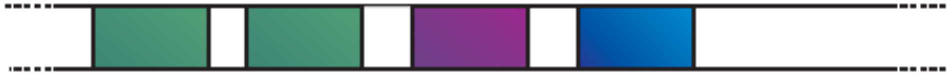
p(old) *p(loss)* *p(new)*



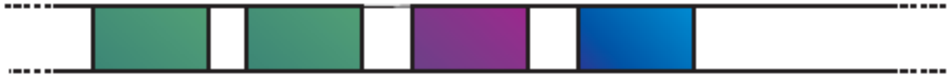


Duplication / Innovation Model

i. Duplication of an existing domain

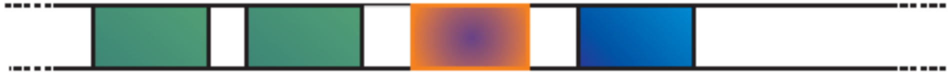


ii. Innovation, genesis or transfer of a domain

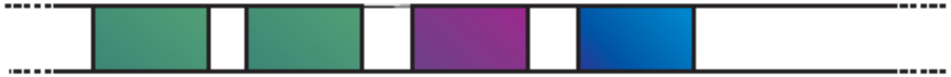


Duplication / Innovation Model

i. **Duplication of an existing domain**

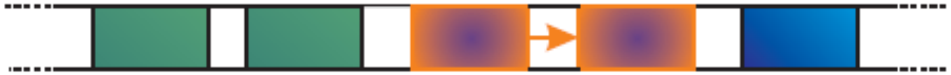


ii. **Innovation, genesis or transfer of a domain**

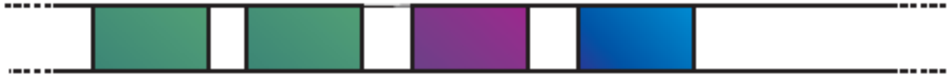


Duplication / Innovation Model

i. **Duplication of an existing domain**

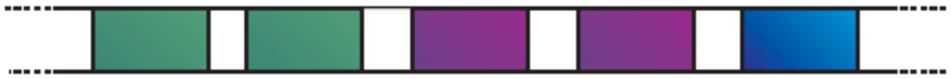


ii. **Innovation, genesis or transfer of a domain**

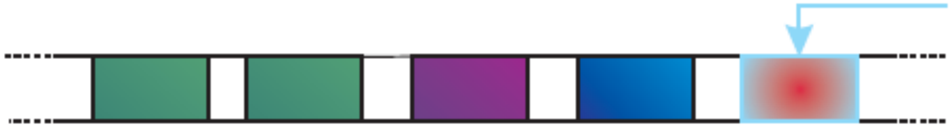


Duplication / Innovation Model

i. Duplication of an existing domain

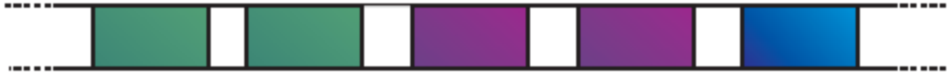


ii. Innovation, genesis or transfer of a domain

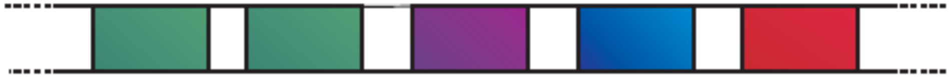


Duplication / Innovation Model

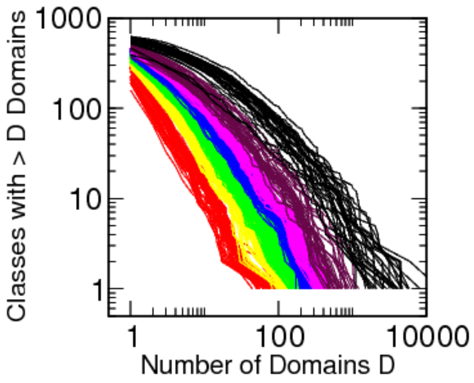
i. **Duplication of an existing domain**



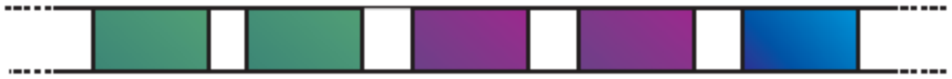
ii. **Innovation, genesis or transfer of a domain**



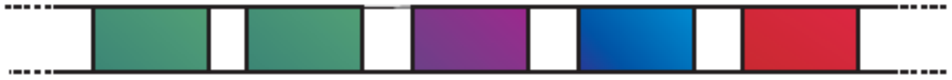
Duplication / Innovation Model

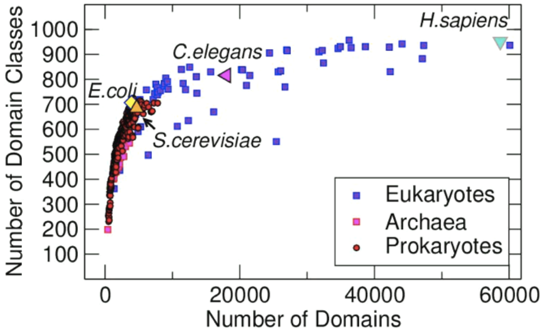


i. **Duplication of an existing domain**



ii. **Innovation, genesis or transfer of a domain**





Requirements

A

$$p_O + p_N = 1$$

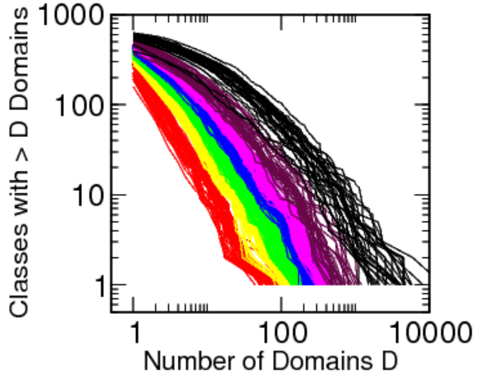
$$p_O = \sum_{i \in \text{classes}} p_O^i$$

normalizations

B

$$p_O^i \sim \frac{n_i}{n}$$

(uniform = preferential attachment)

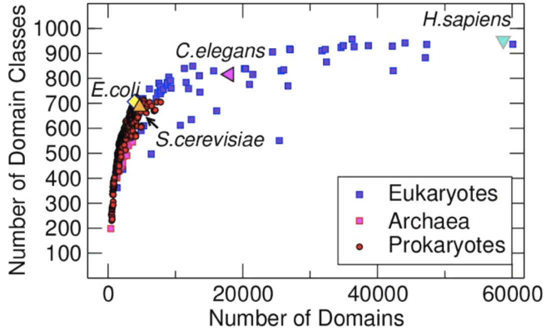


C

p_N *NOT constant in F, n*

$$p_N \sim \frac{1}{n}$$

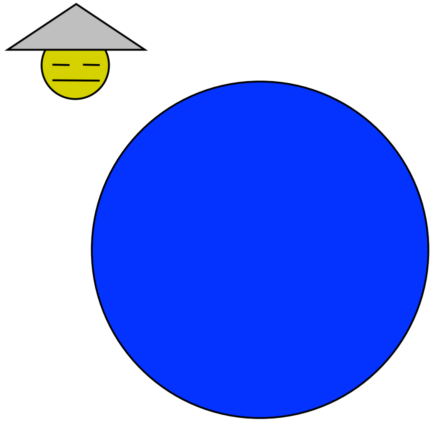
$$p_N \sim \frac{F}{n}$$



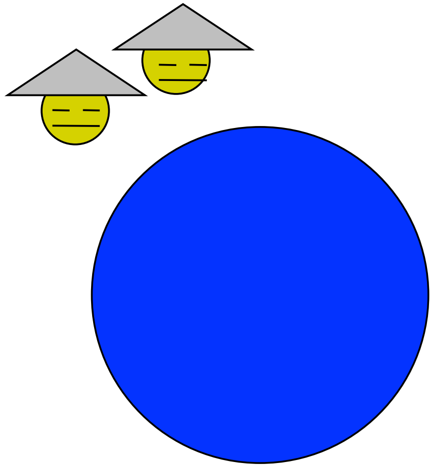
Simplest Case

$$\boxed{p_O = \frac{n-f\alpha}{n+\theta}} \quad \boxed{p_N = \frac{\theta+f\alpha}{n+\theta}} \quad \begin{array}{l} \theta > -\alpha \\ 0 \leq \alpha \leq 1 \end{array}$$

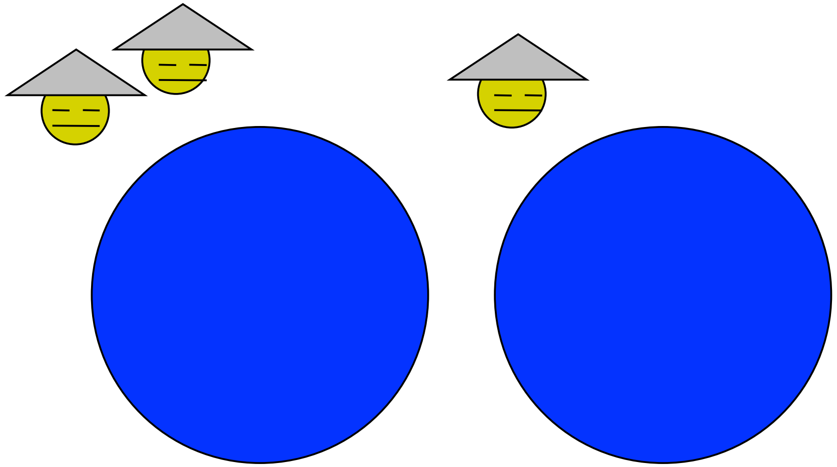
Chinese Restaurant Process



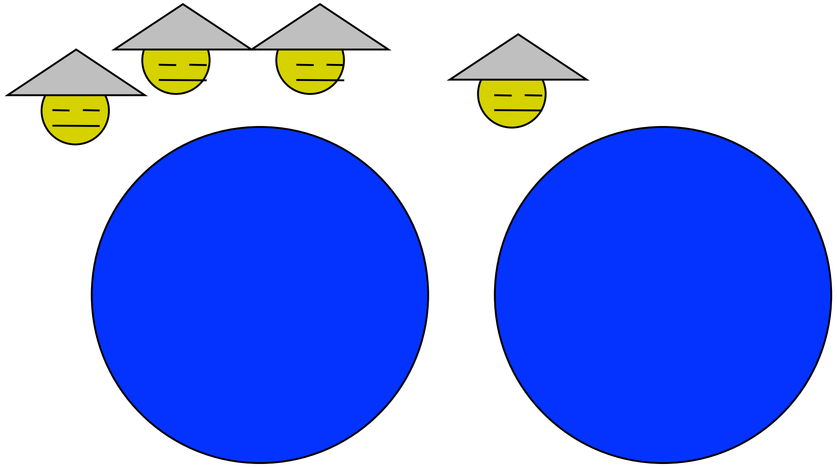
Chinese Restaurant Process



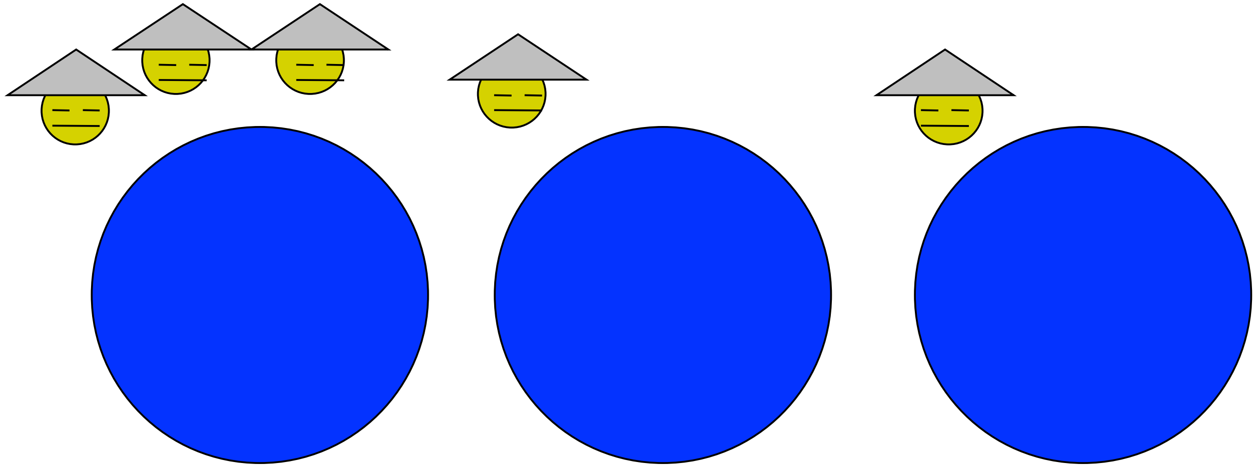
Chinese Restaurant Process



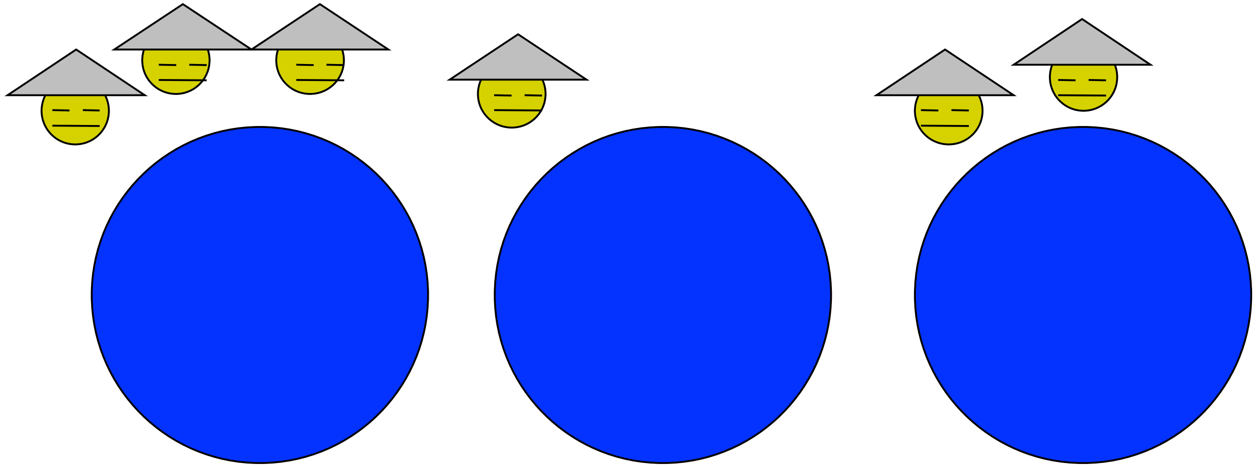
Chinese Restaurant Process



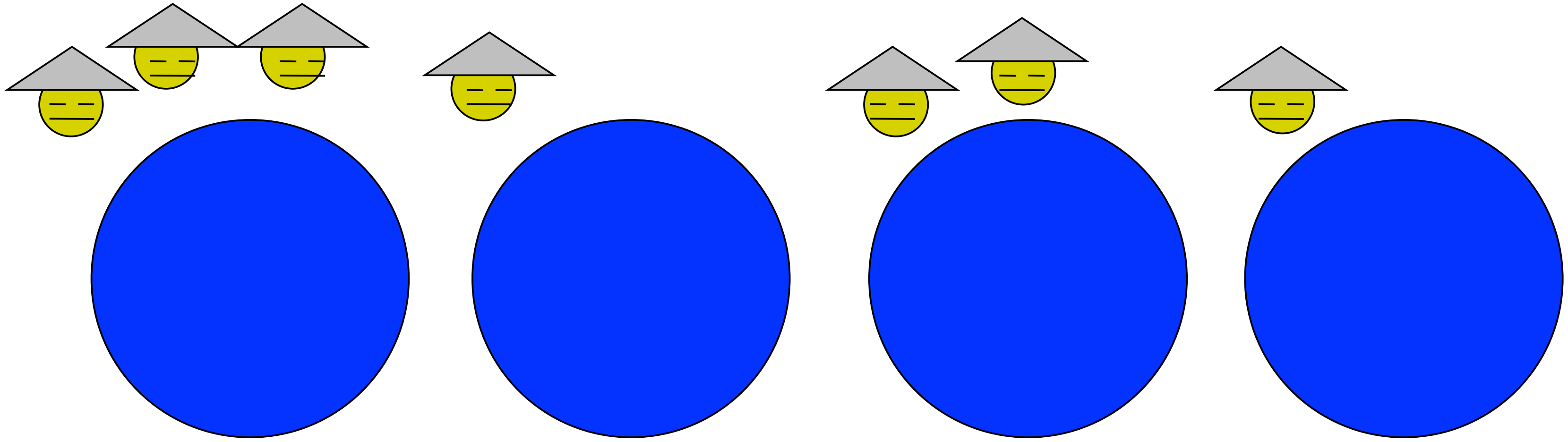
Chinese Restaurant Process



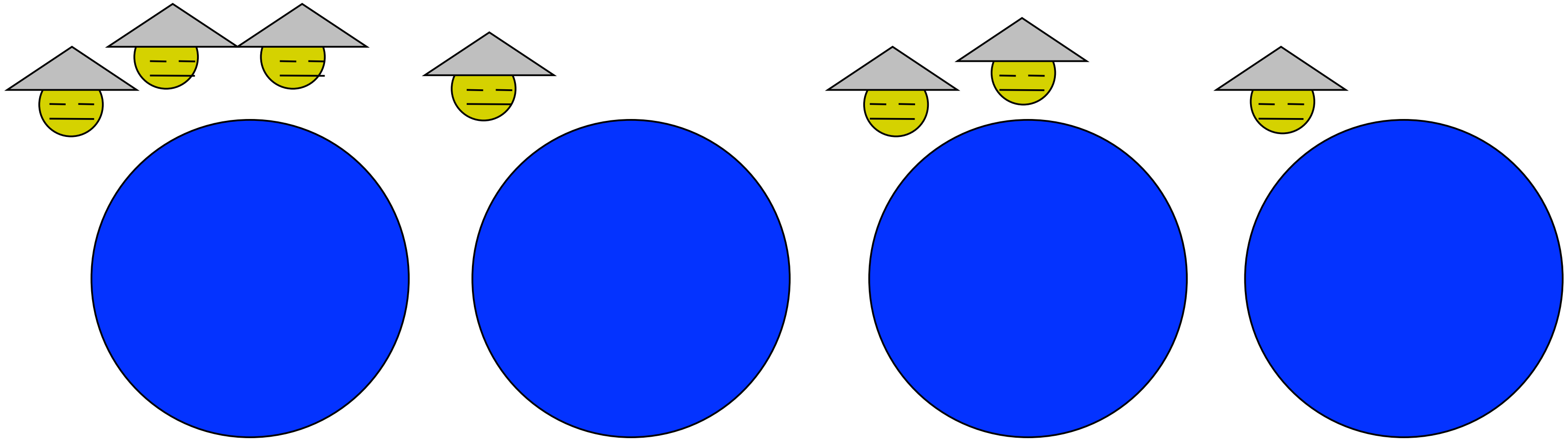
Chinese Restaurant Process



Chinese Restaurant Process



Chinese Restaurant Process



Exercise: write a simulation of this process

$$\boxed{p_O = \frac{n-f\alpha}{n+\theta}} \quad \boxed{p_N = \frac{\theta+f\alpha}{n+\theta}} \quad \begin{array}{c} \theta > -\alpha \\ 0 \leq \alpha \leq 1 \end{array}$$

Plot some realizations of
f(n) (#families)
f(j,n) (#families with j members)

Mean-field for families

$$\boxed{p_O = \frac{n - f\alpha}{n + \theta}} \qquad \boxed{p_N = \frac{\theta + f\alpha}{n + \theta}} \qquad \begin{array}{l} \theta > -\alpha \\ 0 \leq \alpha \leq 1 \end{array}$$

$$\boxed{p_O^{(i)} = \frac{n_i - \alpha}{n + \theta}}$$

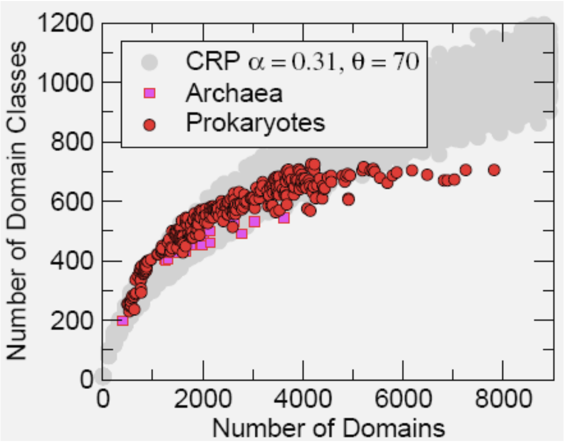
$$\frac{d\langle n_i \rangle}{dn} = p_O^{(i)}(\langle n_i \rangle)$$

$$\frac{d\langle f \rangle}{dn} = p_N$$

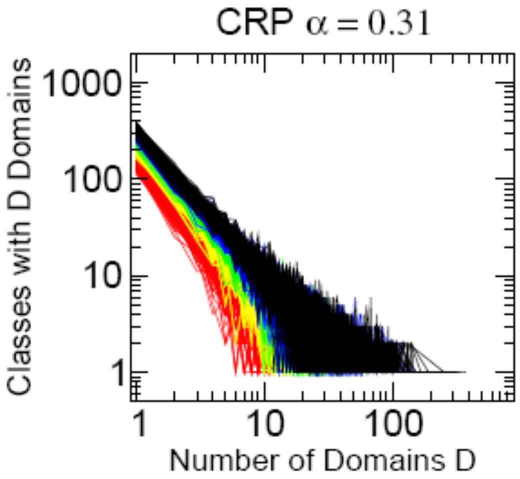
More in the afternoon…

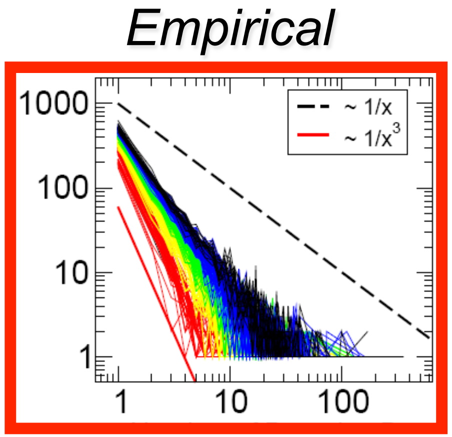
Scaling Results

	K_i	$\frac{p_N}{p_O}$	$\frac{p_N}{p_O^i}$	$F(n)$	$F(j,n)/F(n)$
CRP $\alpha = 0$	$\sim n$	$\sim n^{-1}$	$\sim n^{-1}$	$\sim \log(n)$	$\sim \frac{\theta}{j}$
CRP $\alpha > 0$	$\sim n$	$\sim n^{\alpha-1}$	$\sim n^{\alpha-1}$	$\sim n^{\alpha}$	$\sim j^{-(1+\alpha)}$
Qian *et al.*	$\sim n^{p_O}$	$= R$	$\sim n^{1-p_O}$	$\sim n$	$\sim j^{-(2+R)}$



- Agrees with *Universal* Scaling
- θ model fits better *F(n)*
- α model fits better *F(j,n)*





The Scaling of the Innovation Rate Poses a Biological Question

Data and model:
innovation is less likely than duplication with increasing size

WHY?

- **Neutral or adaptive trend ?**
- Small number of shapes in nature ?
- Role of effective population size ?

Other hypothesis:

- Increased difficulty of “wiring” new functions into increasingly complex interaction networks:

***dF* new folds require *dn* new genes for incorporation**
OPTIMIZATION PROBLEM
***dn* is a function of *n* (the *size* of the problem)**
(exponential, polynomial ...)

The Scaling of the Innovation Rate Poses a Biological Question

Data and model:
innovation is less likely than duplication with increasing size

WHY?

TOMORROW! ☺

Conclusions

- Evolutionary potentials rationalize exponents for functional categories
- Toolbox model gives a proportional recipe for transcriptional regulation vs metabolism
- Duplication-innovation processes rationalize the partitioning of a typical genome into evolutionary families