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

2584-14

Spring College on the Physics of Complex Systems

26 May - 20 June, 2014

A Genome as a Toolbox:
HGT paradox / Joint partitioning in functions and families

Marco Cosentino Lagomarsino
Université Pierre et Marie Curie
Paris

A Genome as a Toolbox:
HGT paradox /
Joint partitioning in functions and families

June 4th 2014

Spring School, Trieste

Marco Cosentino Lagomarsino

Génophysique / Genomic Physics Group

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

Premise: why study microbes?

“Tout ce qui est vrai pour le Colibacille est vrai pour l'éléphant”
(J. Monod)

TOUT CE QUI EST VRAI POUR LE COLIBACILLE EST VRAI POUR L'ELEPHANT
aquatinte
épreuve d'artiste 1/5

Premise: why study microbes?

Do we really care about the elephant?

Microbes are most of the earth’s biomass

Essential for ecosystems (including our guts)

Biomed (antibiotics)

Hold the key to the origins of life

… and of course we like beer, wine, yogurt, bread …

Premise: why microbial genomics??

The massive amount of sequenced genomes opens new perspectives on microbial

Architecture

Evolution

Adaptation

Ecosystems

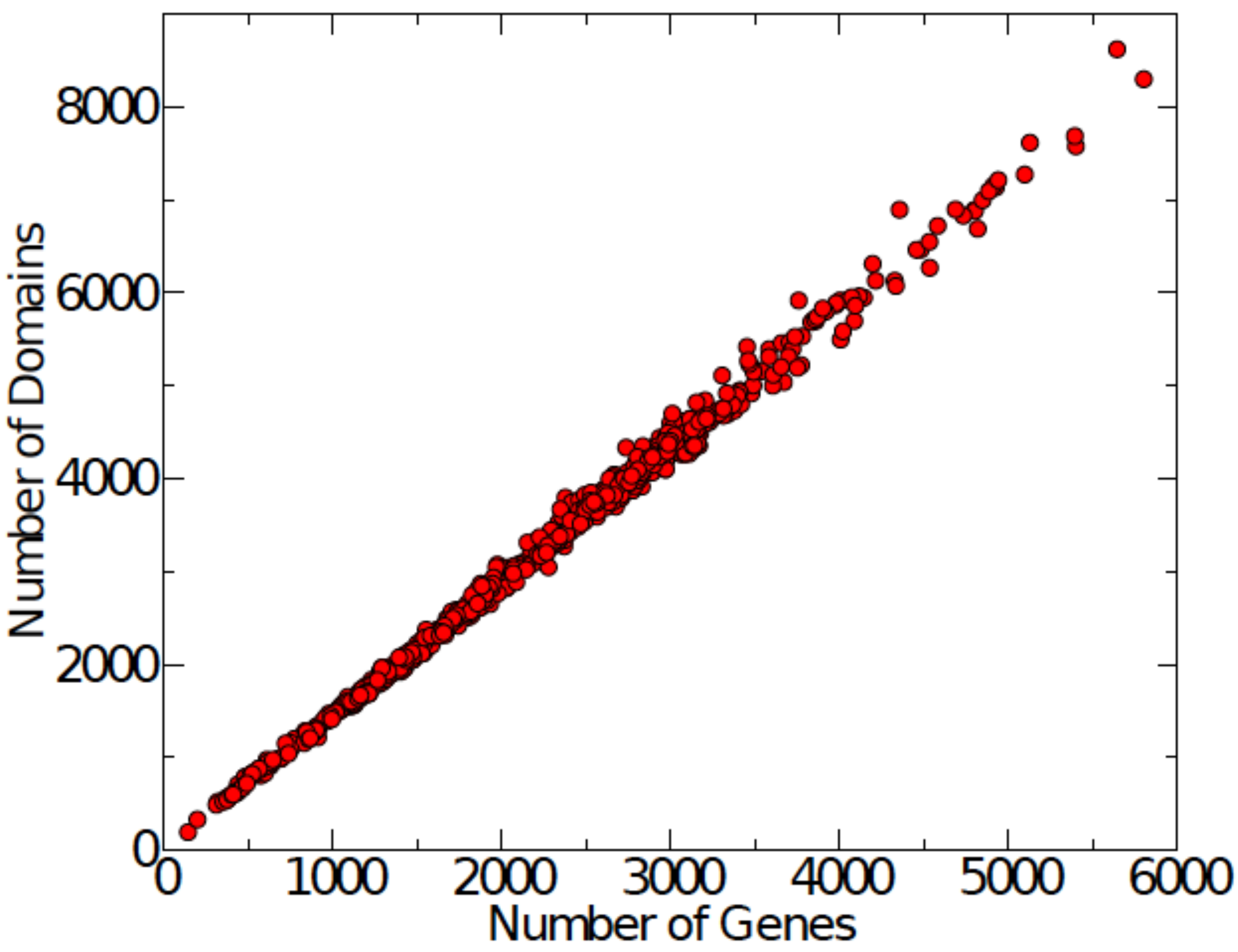
Premise: why with statistical physics???

We know how to build models

Tools are needed to deal with the data
(bioinformatics is mostly data production)

Interesting “exotic” trends, in the perspective of
complex systems theory

The plot that I promised



0) Where we left yesterday ...

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

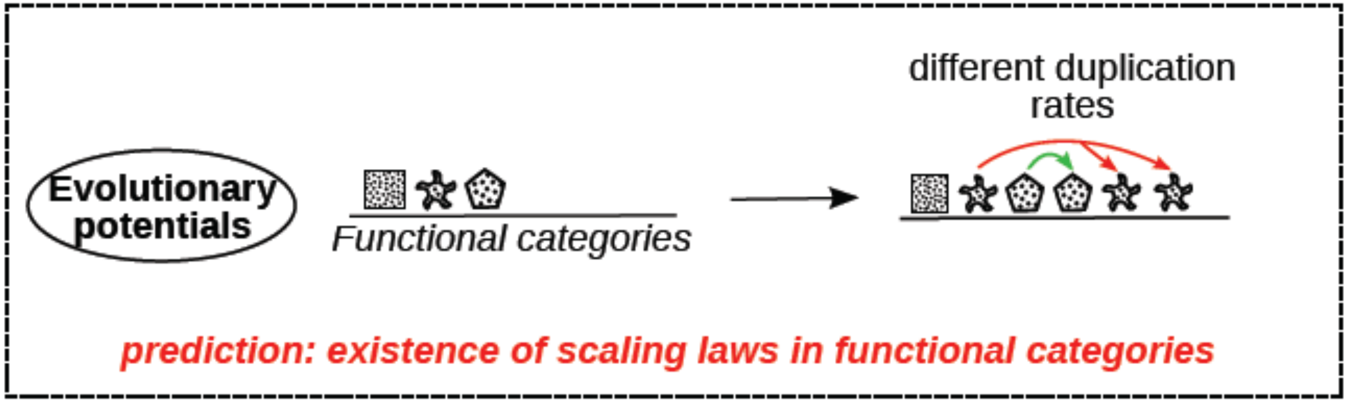
= genome “size”

(related by phylogeny)

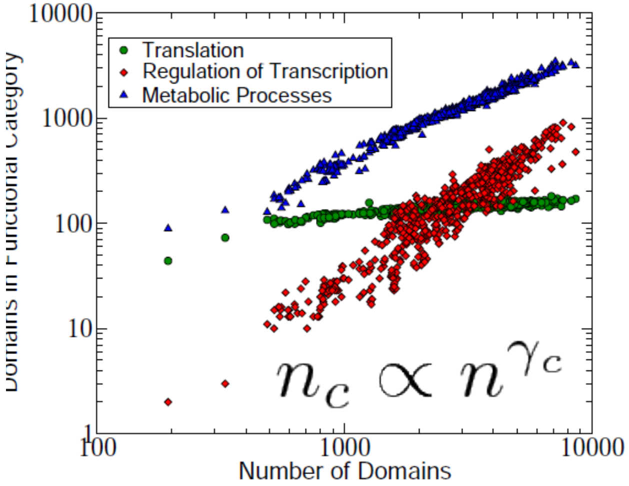
column sum = total family abundance

“Evolutionary Potentials”

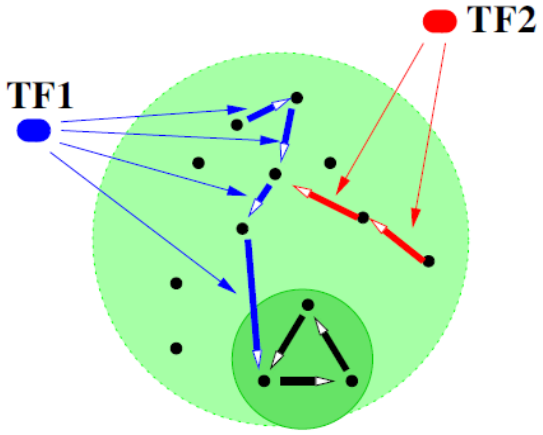
“Preferential Attachment” + Specificity $$\frac{dn_c}{dn} \propto \rho_c \frac{n_c}{n}$$



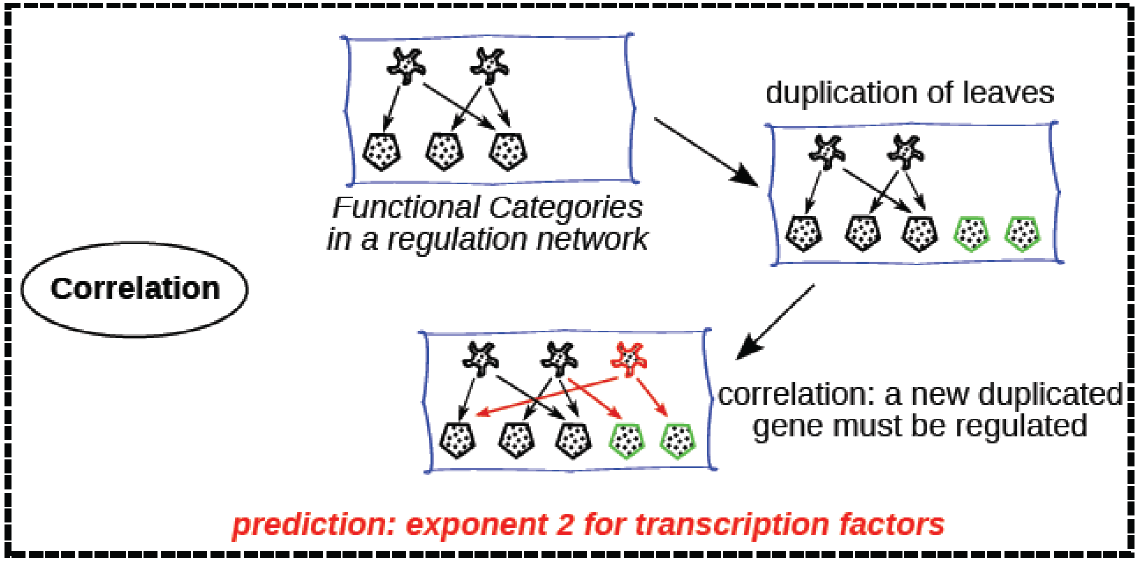
prediction: existence of scaling laws in functional categories



Toolbox model as recipe for coordinated growth



A larger genome gets shorter pathways
TFs control *multiple targets*



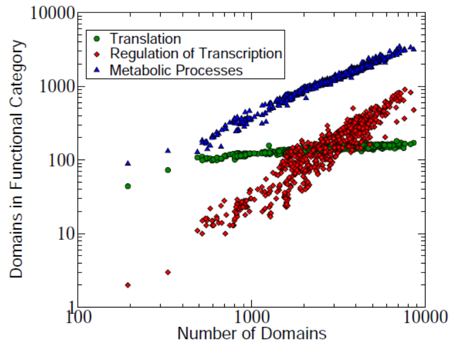
Correlation

Functional Categories in a regulation network

duplication of leaves

correlation: a new duplicated gene must be regulated

prediction: exponent 2 for transcription factors



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

→ quadratic scaling

CRP as minimal model for partitioning into evolutionary families

A

$$p_O + p_N = 1$$

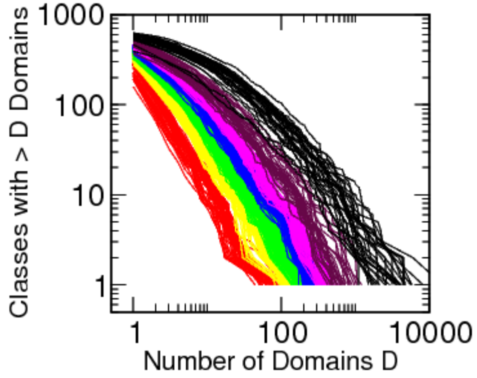
$$p_O = \sum_{i \in \text{classes}} p_O^i$$

normalizations

B

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

(uniform = preferential attachment)

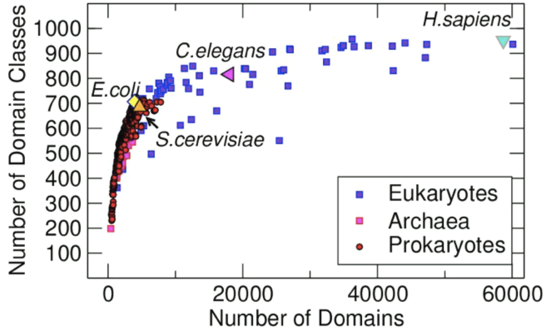


C

p_N *NOT constant in F, n*

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

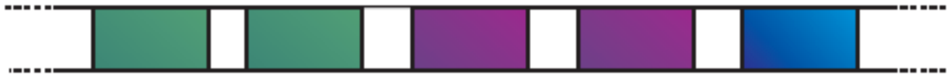
$$p_N \sim \frac{F}{n}$$



Loss does not affect main results
e.g. uniform loss:

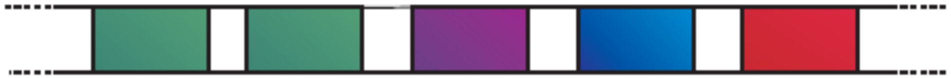
$$\boxed{p_O = (1-\delta)\frac{n - f\alpha}{n+\theta}}$$

i. **Duplication of an existing domain**



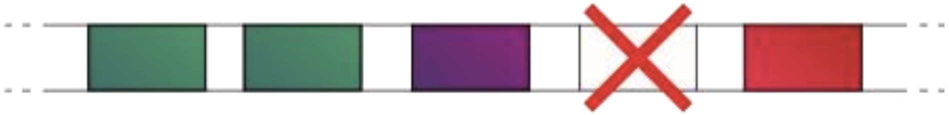
$$\boxed{p_N = (1-\delta)\frac{\theta + f\alpha}{n+\theta}}$$

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



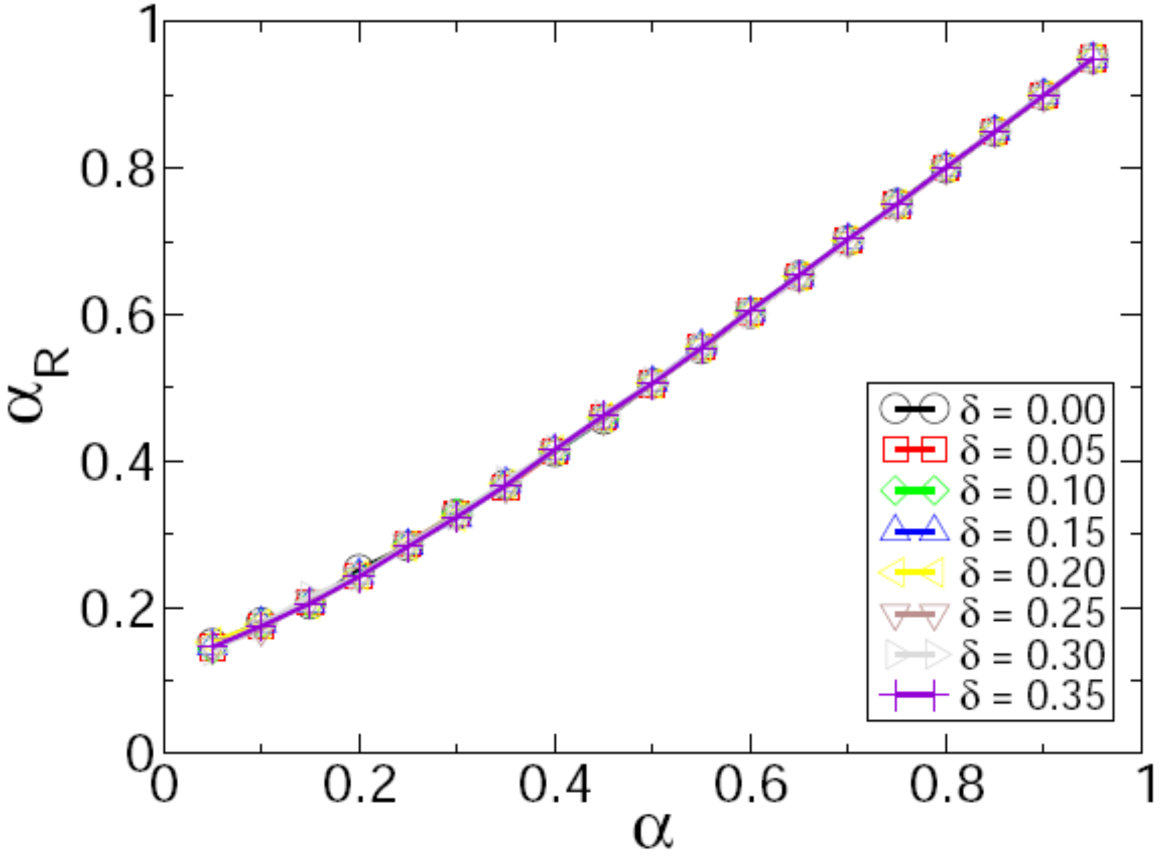
$$\boxed{p_L = \delta}$$

iii.**Loss of an existing domain**



Adding uniform loss does not affect the scaling

Observed Exponent for *F(n)*

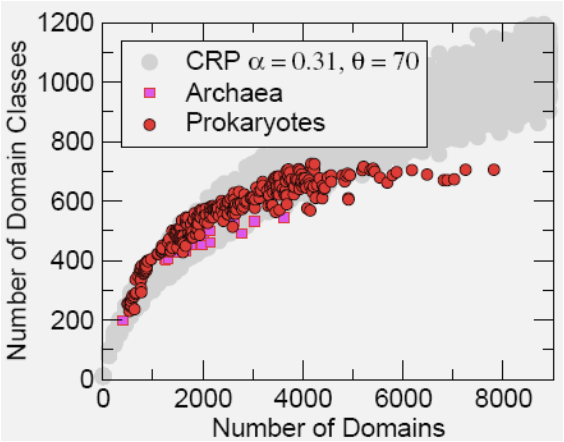


Value of the α parameter

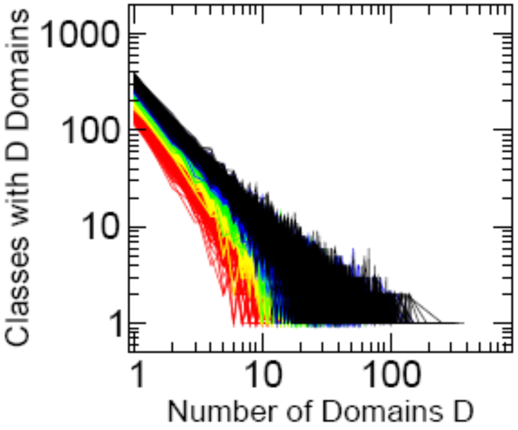
Scaling Results

	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.*	$\sim n^{p_O}$	$= R$	$\sim n^{1-p_O}$	$\sim n$	$\sim j^{-(2+R)}$

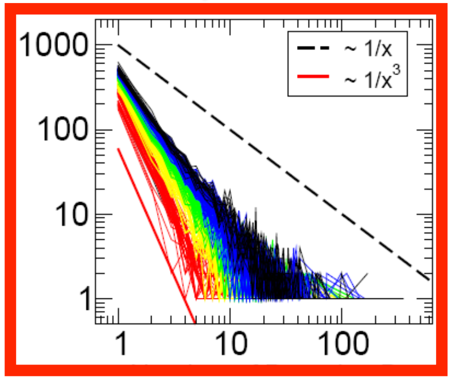
- Agrees with *Universal* Scaling
- θ model fits better *F(n)*
- α model fits better *F(j,n)*



CRP $\alpha = 0.31$



Empirical



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 ...)

1) “HGT paradox”

HGT in Bacteria

Recent genomic studies in Bacteria suggest that
most new genes are the result of horizontal transfer
rather than duplication

Is innovation affected
by the universe of accessible genes?

Expansion-innovation model with HGT from finite universe of families

(HGT family expansion rate)

$$\frac{dn_i}{dt} = \dot{n}_i = n_i + \gamma$$

(family expansion with pref. attachment = time scale)

$$\dot{F} = \gamma(D - F)$$

(HGT innovation rate)

Expansion-innovation model with finite universe

$$\dot{n} = \sum_{i=1}^{F} \dot{n}_i + \dot{F} = n + \gamma D$$

Total growth in size per *dt*

Expansion-innovation model with finite universe

using $\frac{dX}{dn} = \frac{dX}{dt} \bigg/ \frac{dn}{dt}$

$$\frac{dn_i}{dn} = \frac{n_i + \gamma}{n + \gamma D}$$

$$\frac{dF}{dn} = \frac{\gamma(D - F)}{n + \gamma D}$$

We are back to the same type of model…

set $\alpha = -\gamma \qquad \theta = \gamma D$

$$p^{(i)}_{\text{old}} = \frac{n_i - \alpha}{n + \theta}$$

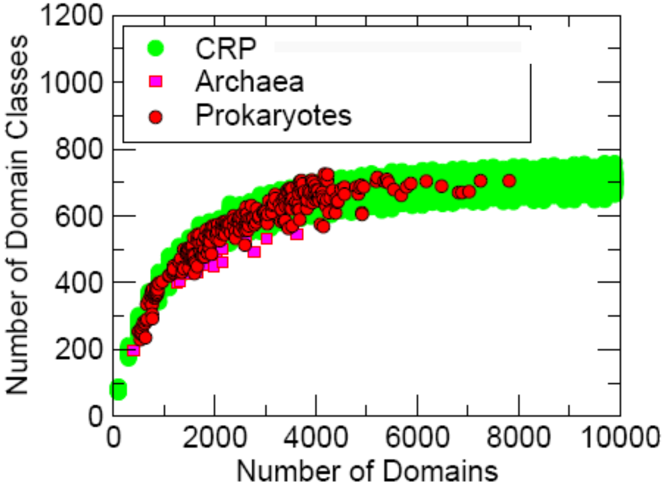
$$p_{\text{new}} = \frac{\theta + \alpha F}{n + \theta}$$

One gets a CRP with *negative* α

Can be analyzed by mean-field and simulation
(as usual)

Models with finite universe gives the best fit with data

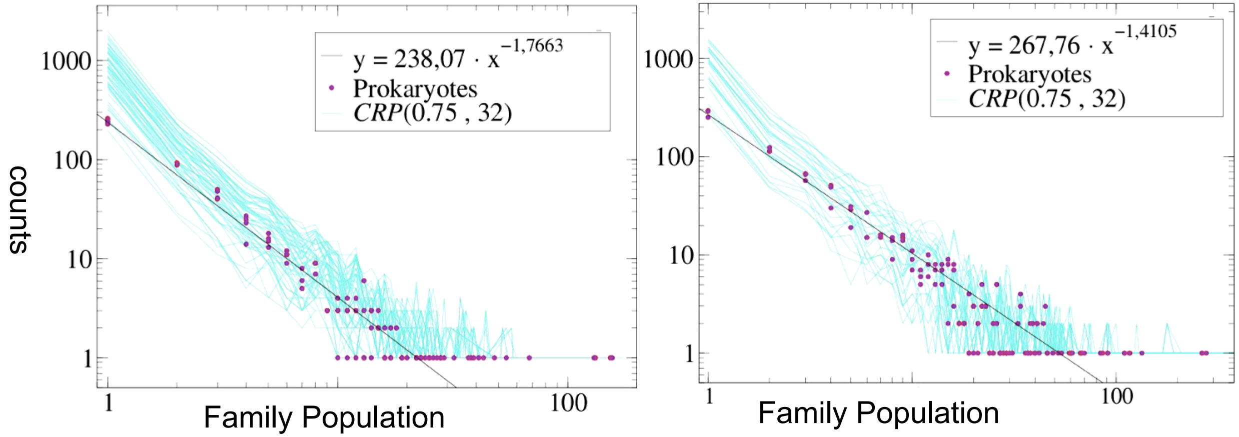
domain classes F
vs domains n

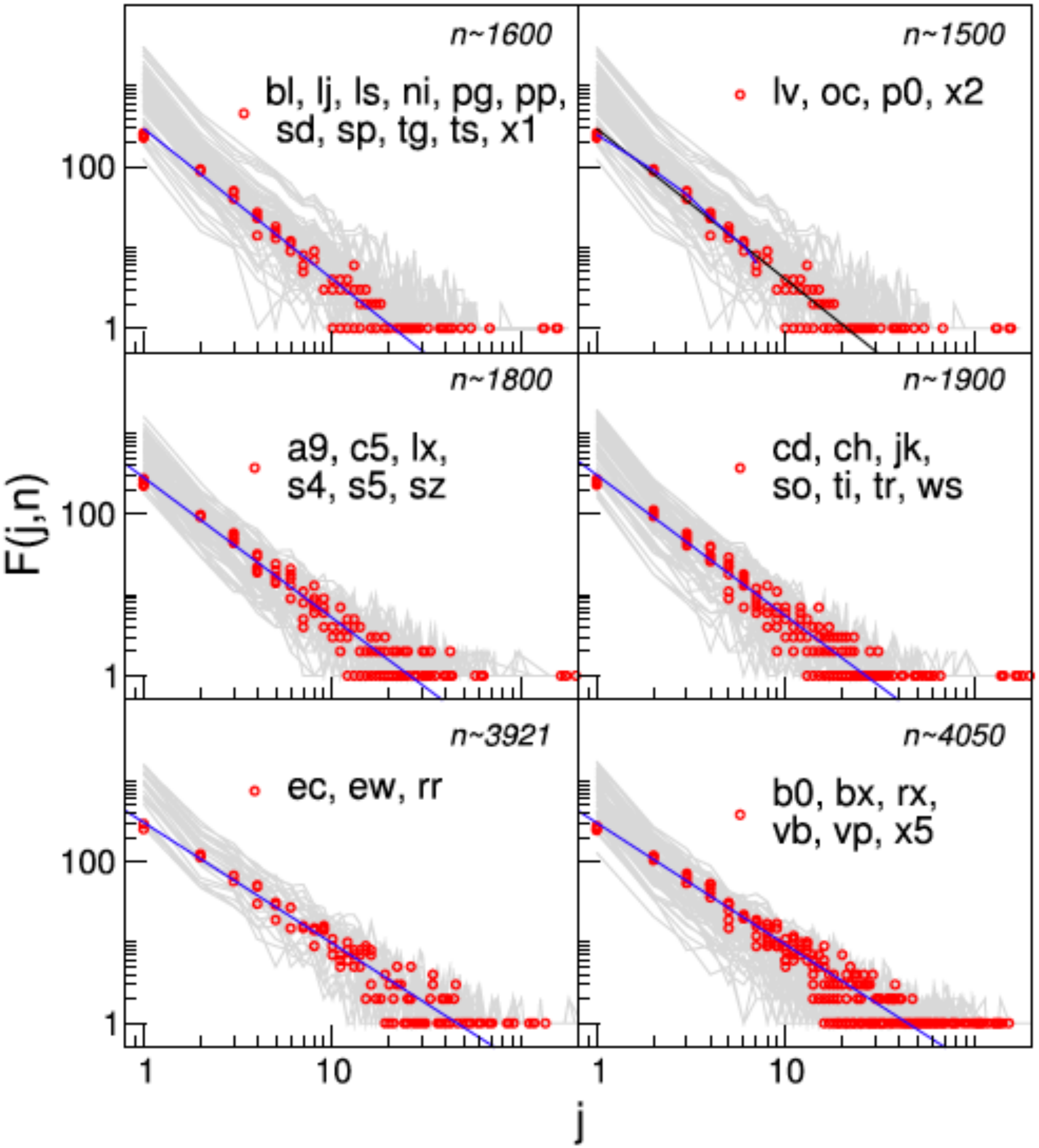


domain family histogram

n ~ 1500

n ~ 3921




n~1600
bl, lj, ls, ni, pg, pp, sd, sp, tg, ts, x1
n~1500
lv, oc, p0, x2
n~1800
a9, c5, lx, s4, s5, sz
n~1900
cd, ch, jk, so, ti, tr, ws
n~3921
ec, ew, rr
n~4050
b0, bx, rx, vb, vp, x5
F(j,n)
100
1
j
10

Exercise:

Compare a CRP with negative α (see previous slides)

with a model with *positive* α and a finite universe.

In the latter model, one has the mean-field equation:

$$\frac{dF}{dn} = \frac{\alpha F + \theta}{n + \theta} \frac{D - F}{D}$$

Show that the mean-field dynamics of $f = F/D$

Is not the same in the two kinds of models

The HGT Paradox in Bacteria /data

Recent genomic studies in Bacteria suggest that most new genes are the result of horizontal transfer rather than duplication

Two questions

For duplications-deletions, it is natural that family expansion rates are proportional to family size but is this the case for HGT? (and why?)

Does HGT affect the universe of accessible genes?

Study on data

Obtain HGTs

(Lercher data set on 21 genomes)
(HGT-DB database, 959 genomes)

See where they expand and innovate in terms of Domain families

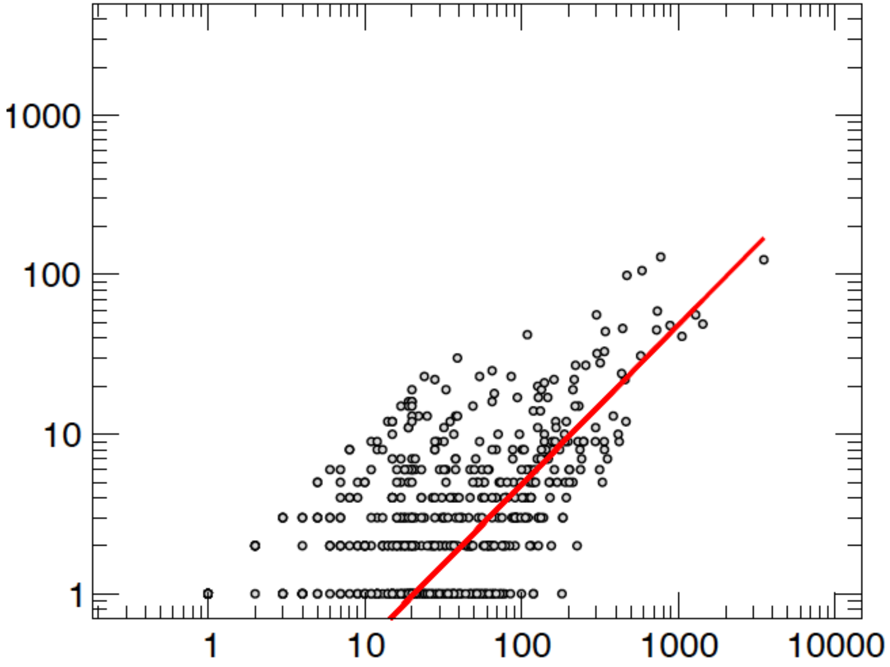
(SUPERFAMILY)
(PFAM)

(Grassi et al 2012)

A. Family expansion rates by HGT Are (roughly) proportional to family size

Detailed study of 21 genomes in the *E.coli* clade

Measured number of horizontal transfers

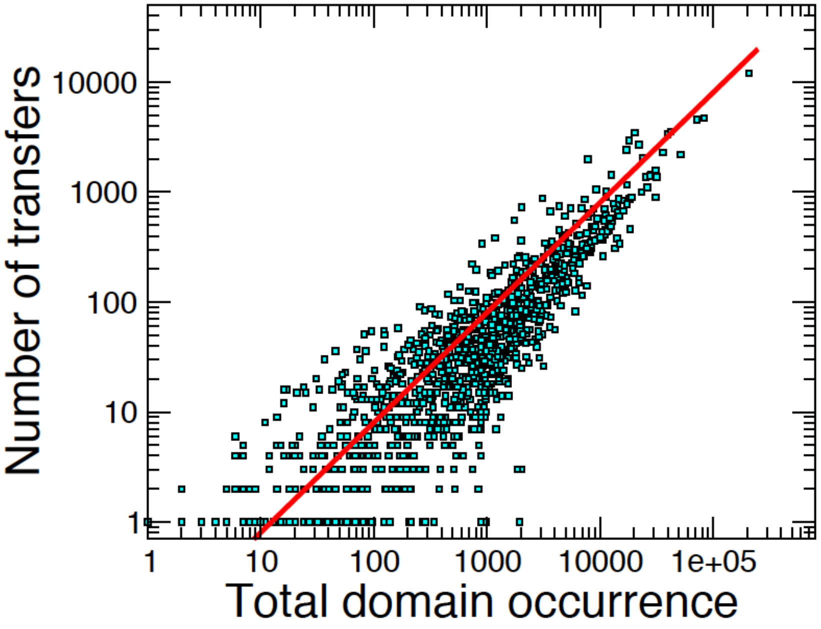


Family size

A. Family expansion rates by HGT are proportional to family size

Systematic data on HGT from 959 bacterial genomes

Measured number of horizontal transfers

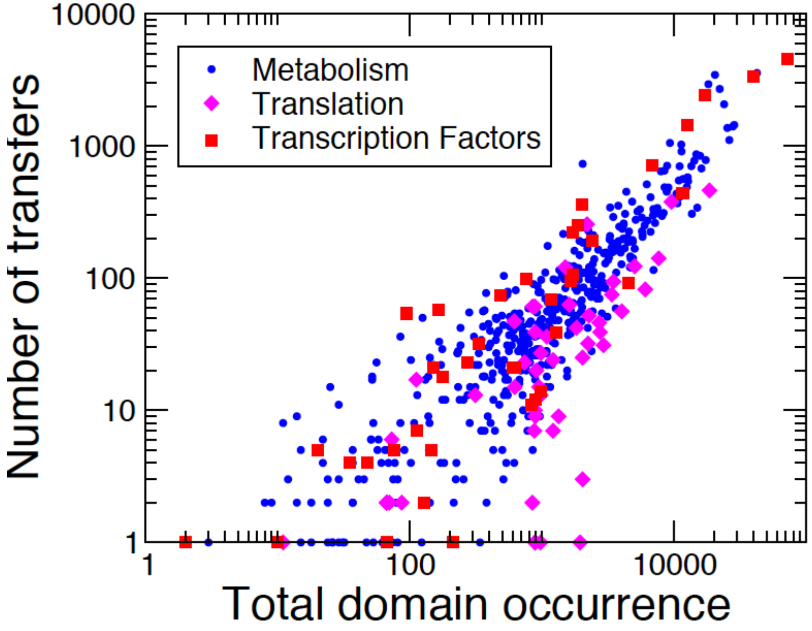


Family size

A. Family expansion rates by HGT are proportional to family size

Not dependent on functional category

Measured number of horizontal transfers

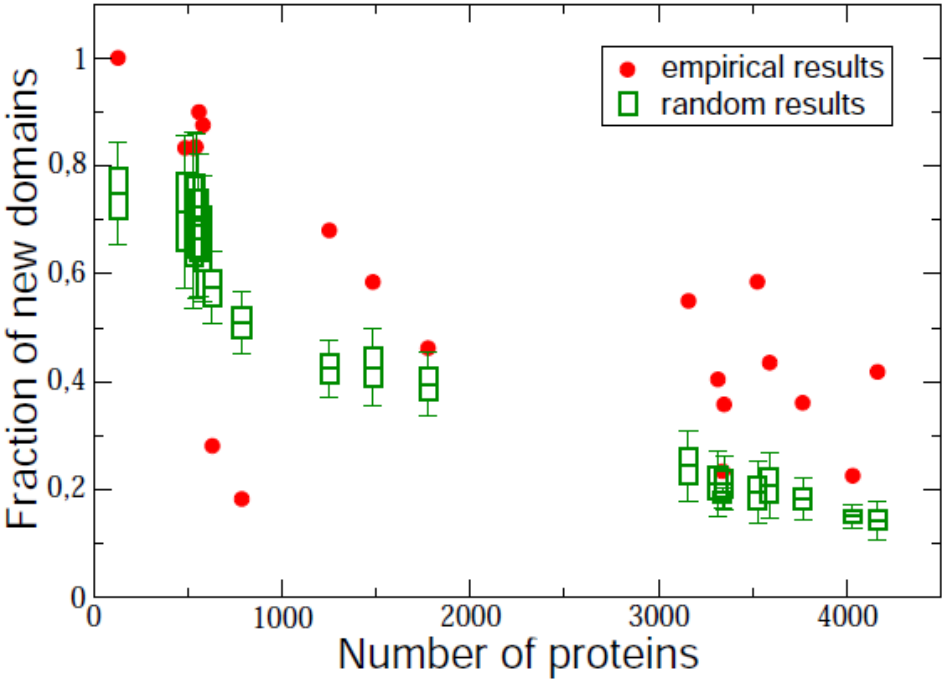


Family size

B. Novel domains acquired by horizontal transfer are compatible with extraction from a finite universe

Detailed study of 21 genomes in the *E.coli* clade

Measured probability of horizontal transfers carrying new domains

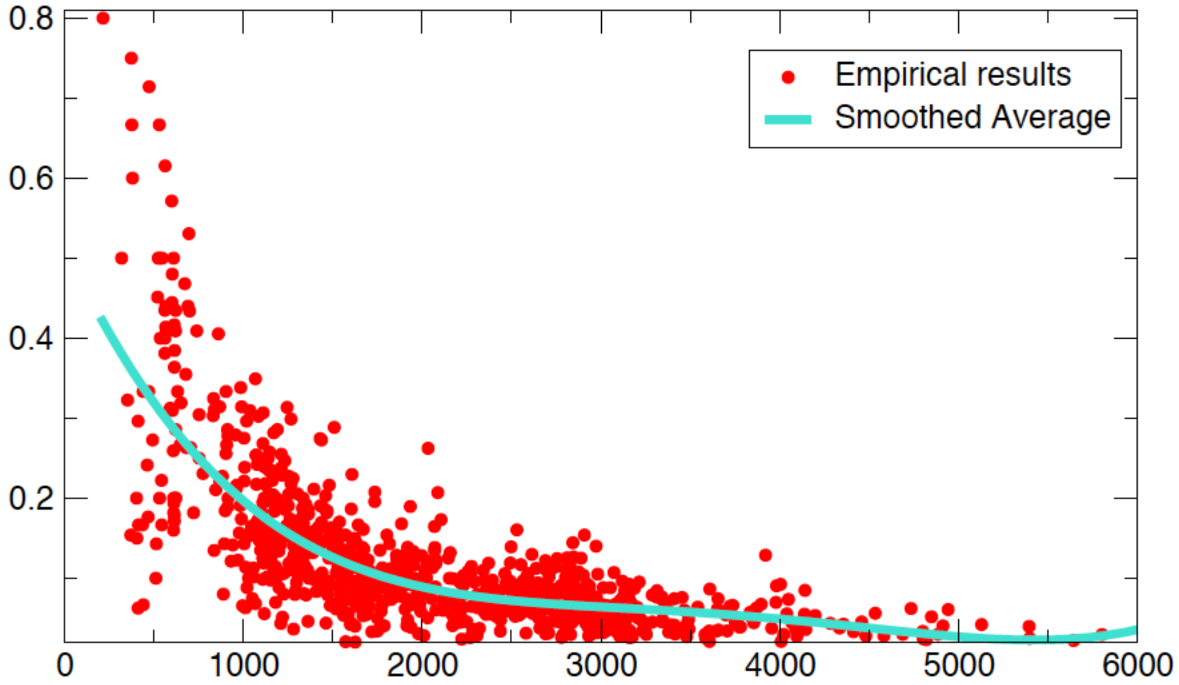


Genome size in proteins

B. Novel domains acquired by horizontal transfer are compatible with extraction from a finite universe

Systematic data on HGT from 959 bacterial genomes

Measured probability of horizontal transfers carrying new domains

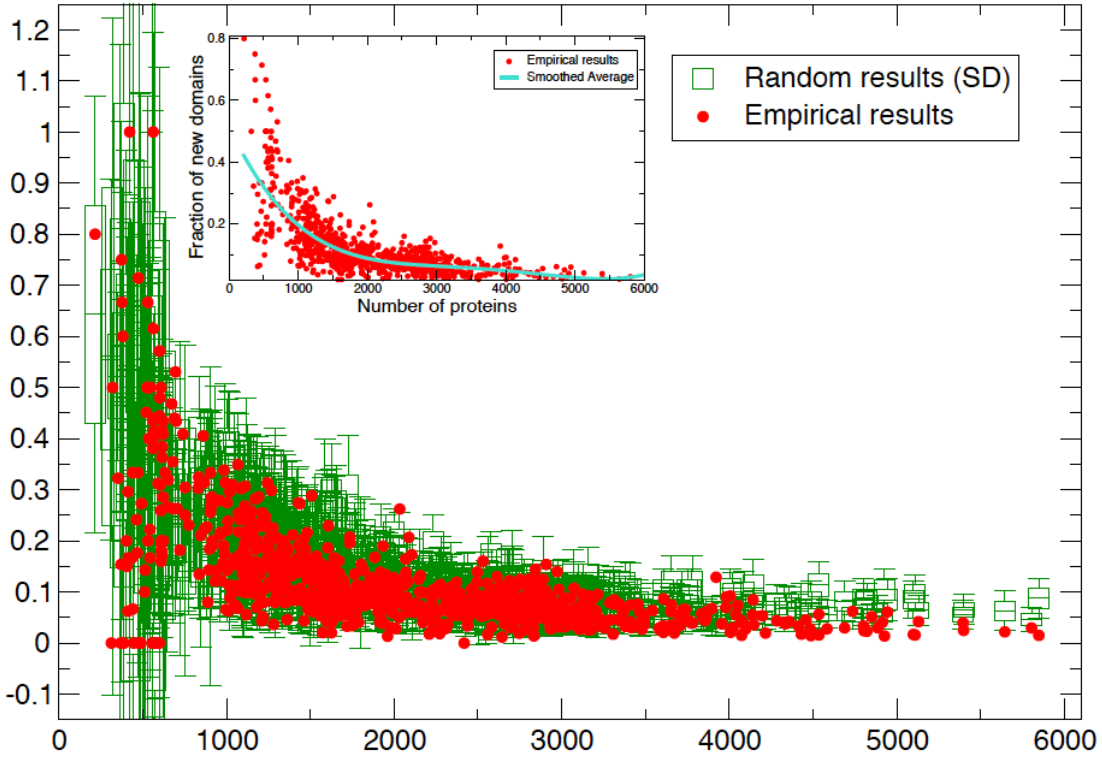


Genome size in proteins

B. Novel domains acquired by horizontal transfer are compatible with extraction from a finite universe

Systematic data on HGT from 959 bacterial genomes

Measured probability of horizontal transfers carrying new domains



Genome size in proteins

randomization = random re-assignment of horizontally transferred genes to receiving genomes

2) Joint partitioning of a genome
into functional and evolutionary classes

(Monod marries Dayhoff)

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

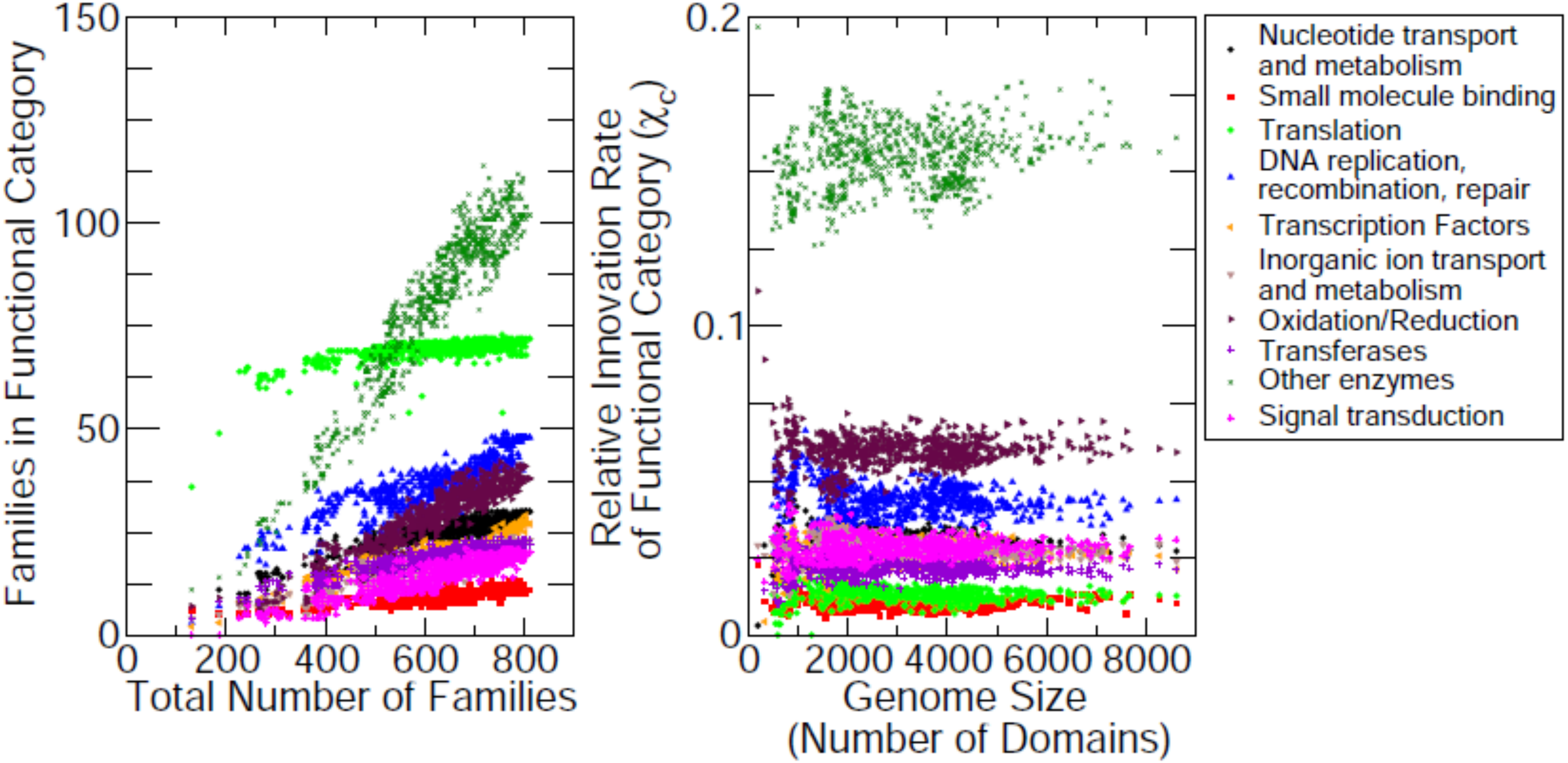
row sum

= genome “size”

(related by phylogeny)

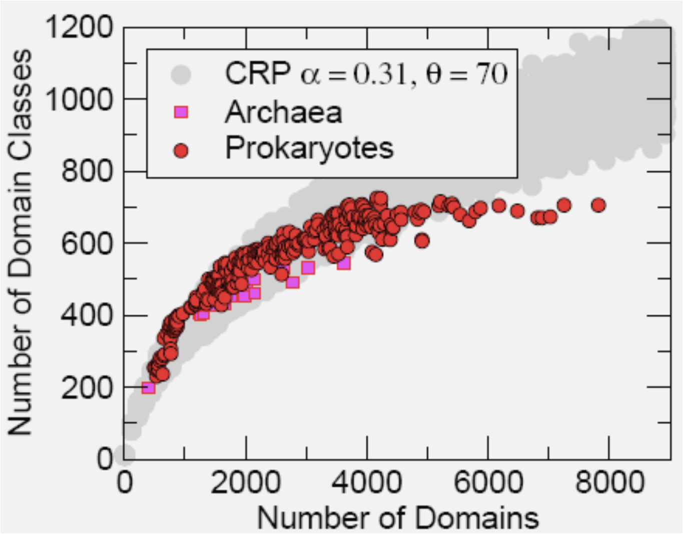
column sum = total family abundance

New “law”: the number of evolutionary families belonging to a functional category grows linearly with a category-dependent coefficient



$$f_c = A_c + \chi_c f$$

To sum up: we have counts for



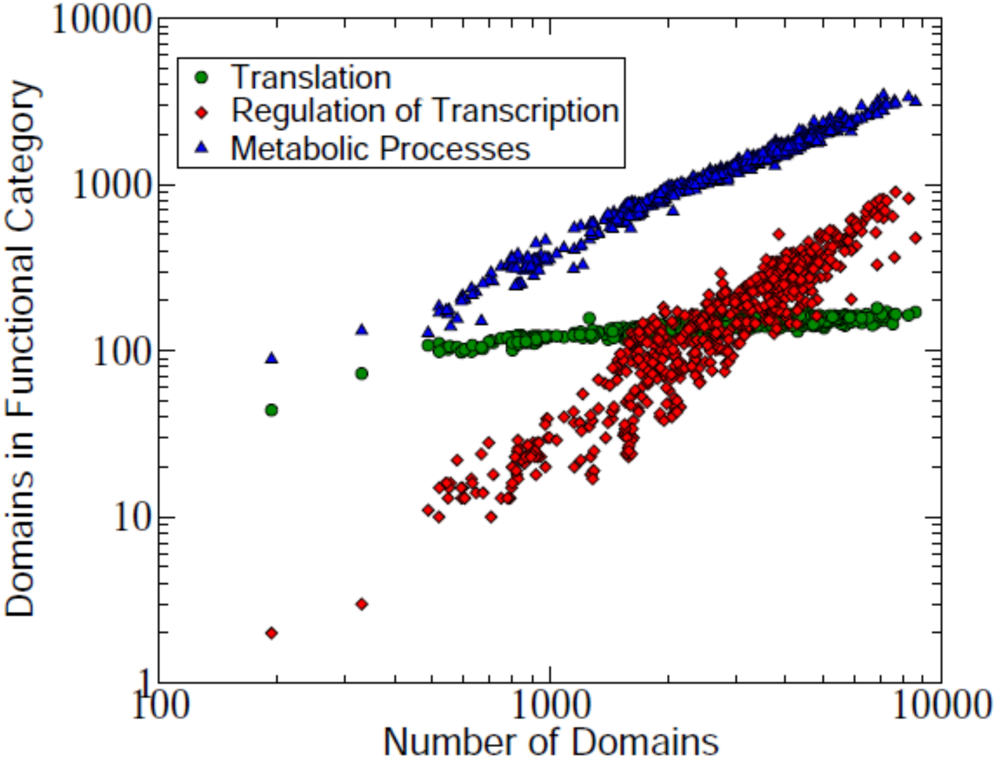
Evolutionary Classes:

- common behavior reproduced by class-expansion / innovation

Functional Categories:

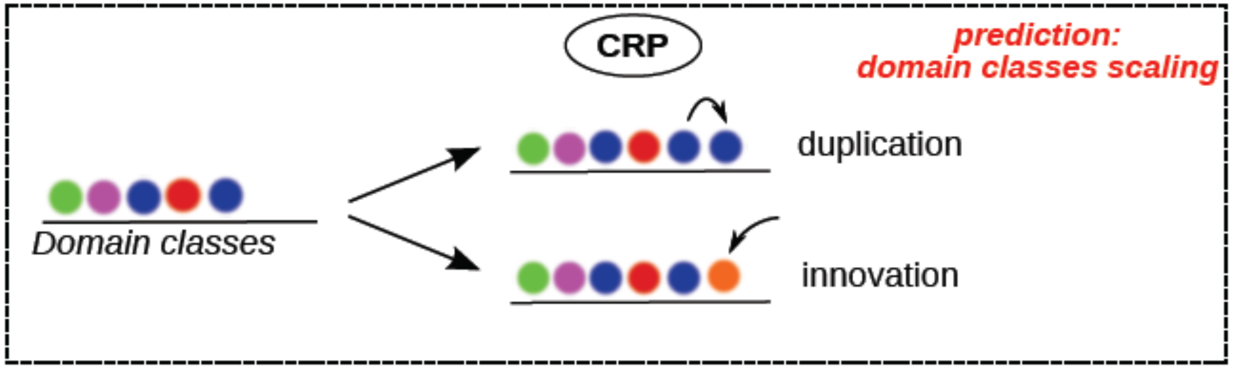
- Grow like Power-laws

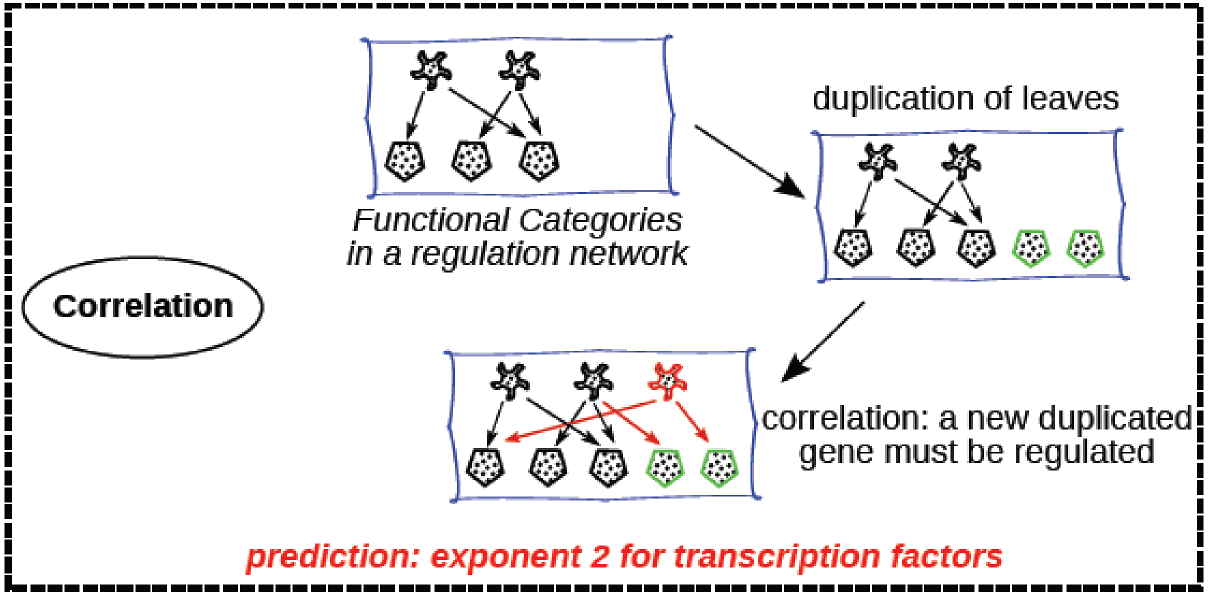
- Exponent ~two for transcription factors



Can a common model describe them ?

combine CRP with functional growth models





CRP with correlated family expansion

$$p_O^i = \frac{\sum_{j=1}^{f} a_{i,j} n_j - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta}$$

$a_{i,j} \longrightarrow$ creating members in family *i*
Is affected by the population of family *j*

we want couplings $a_{i,j}$ to describe dependencies between functional categories

Choice we put couplings only in family expansion

CRP with correlated family expansion

$$p_N = \frac{\alpha f + \theta}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta}$$

Choice we put couplings only
in family expansion

One needs to describe innovation at the function level

We set $p_N^{(c)} = \chi_c p_N$

a newly added family belongs to category *c* with probability χ_c.

In mean field:

$$\partial_n f_c = \chi_c p_N$$

proportionality law for categories $f_c = A_c + \chi_c f$

Mean-field equations

$$\partial_n n_i = p_O^i$$

$$\partial_n f = p_N$$

$$\partial_n f_c = \chi_c p_N$$

$$\partial_n n_c = \partial_n \sum_{i\in c} n_i = \sum_{i\in c} \partial_n n_i + \partial_n f_c = \sum_{i\in c} p_O^i + \chi_c p_N$$

Different correlated recipes are possible

Simplest case, two functional classes:
TFs and Targets (Metabolic Enzymes)

$$\frac{n_{met}}{U}\frac{n_i}{n_{TF}}$$

(Pure Toolbox Model)

$$a_{i,j} =$$

If i is a TF and *j* a leaf.
(= 0 otherwise, TFs are slaved by Targets)

$$\delta_{i,j} + b_{i,j}$$

$$b_{i,j} = n_i / n_{met}$$

(Allows for intrinsic growth of TF classes, at equal rates, Generalizable to arbitrary exponents)

Different correlated recipes are possible

Pure Toolbox Model

Both variants give

$$n_{TF} \sim n_{met}^2$$

in mean field

Allows for intrinsic growth of TF classes, at equal rates, Generalizable to arbitrary exponents

Toolbox recipe

(i) We restate the toolbox

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

as

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

This rescaling leaves invariant $\dfrac{\Delta n_{TF}}{\Delta n_{met}}$

Toolbox recipe

(ii) We impose

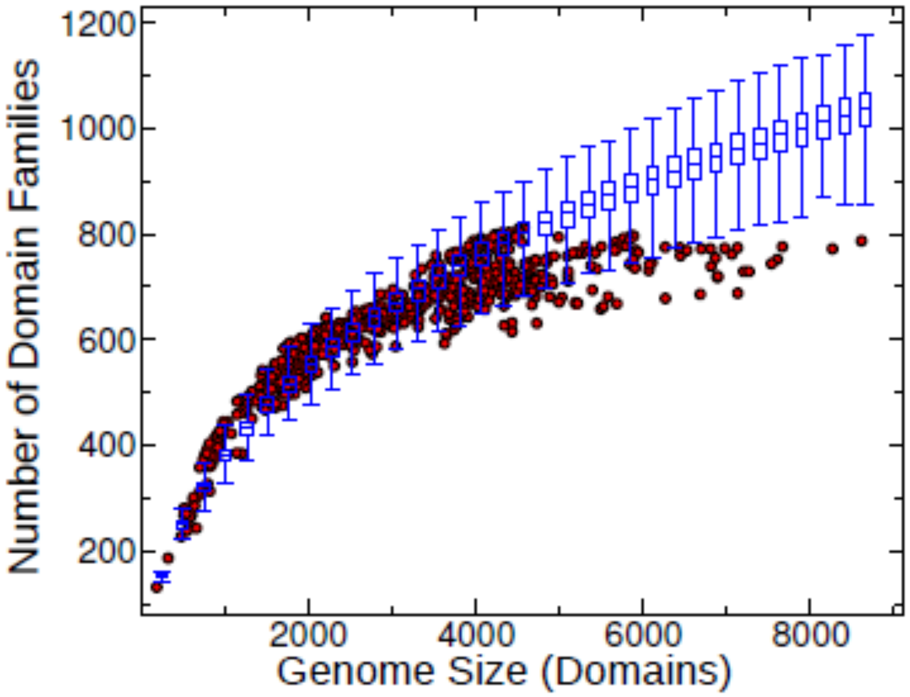
$$\begin{cases} p_O^{met} := \sum_{i \in met} p_O^i = \dfrac{n_{met} - \alpha f_{met}}{C(n)} \\ p_O^{TF} := \sum_{i \in TF} p_O^i = \frac{\frac{n_{met}}{U} n_{met} - \alpha f_{TF}}{C(n)} \end{cases}$$

and

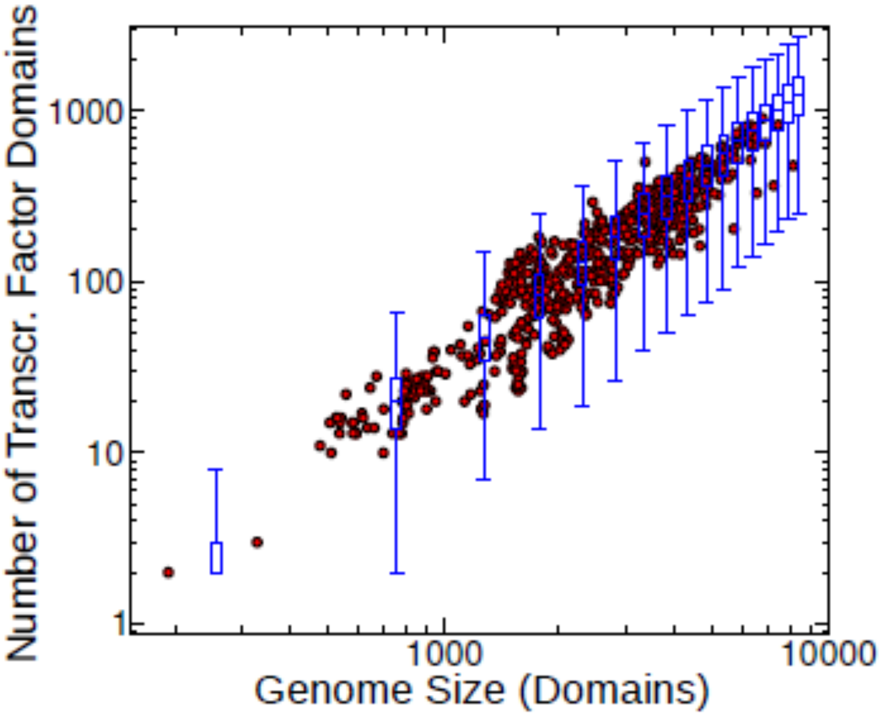
$$\begin{cases} p_O^i = \dfrac{\sum_{j \in met} \frac{n_{met}}{U} \frac{n_i}{n_{TF}} n_j - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} & \text{if } i \in TF \\ p_O^i = \dfrac{n_i - \alpha}{\sum_{i,j=1}^{f} a_{i,j} n_j + \theta} & \text{if } i \in met \end{cases}$$

Metabolic families grow on their own / TFs follow

CRP with correlated duplication agrees well with empirical data (both variants)



Evolutionary Classes



Functional Categories

Power-law with exponent

ζ ~ 1.6 for transcription factors

[1.6 explained as finite-size effect]

Non-trivial prediction:
domain class histograms for transcription factors

From mean-field master equation:

Targets

$$P(d) \sim \left(\frac{1}{d}\right)^{1+\alpha}$$

Restricted to TFs

$$P(d)_{TF} \sim \left(\frac{1}{d}\right)^{1+\frac{\alpha}{2}}$$

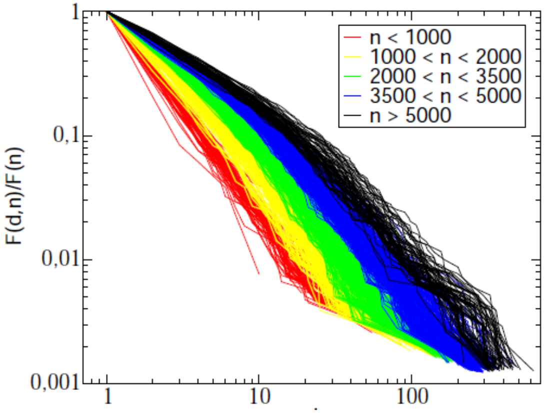
Corrected by category exponent!!

In general, the histogram restricted to a functional category is expected to scale as:

$$P(d)_c \sim \left(\frac{1}{d}\right)^{1+\beta_c}$$

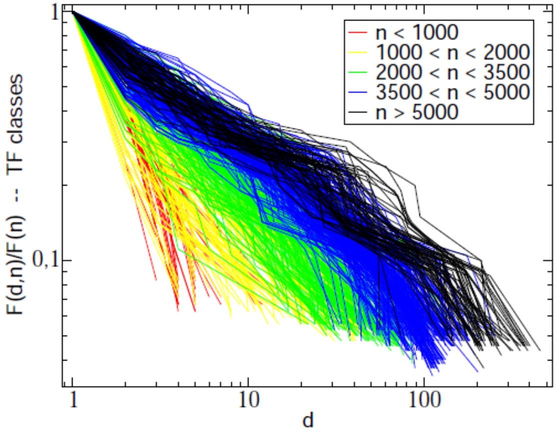
where $\beta_c = \alpha / \zeta_c$

Empirical data follow the predicted trend



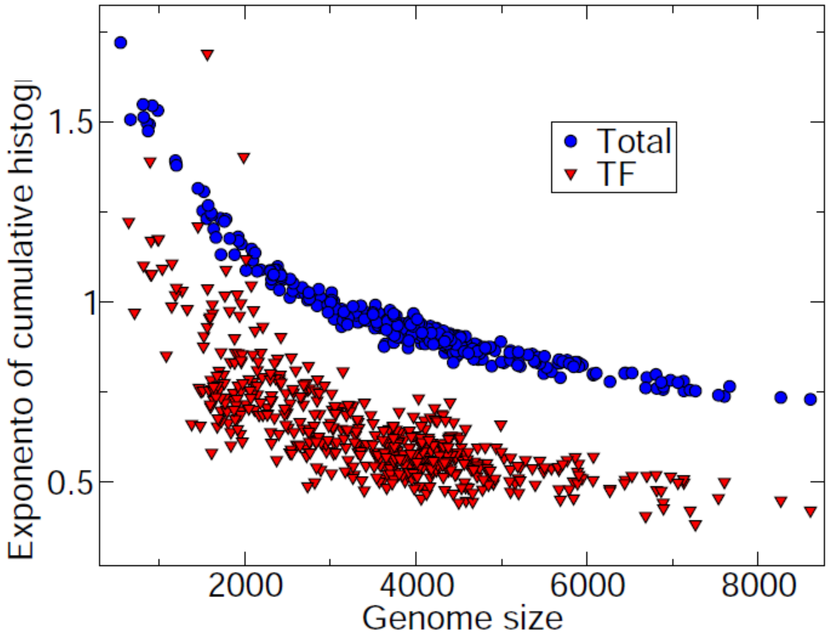
Total

TFs



$$P(d) \sim \left(\frac{1}{d}\right)^{1+\alpha}$$

$$P(d)_{TF} \sim \left(\frac{1}{d}\right)^{1+\frac{\alpha}{2}}$$

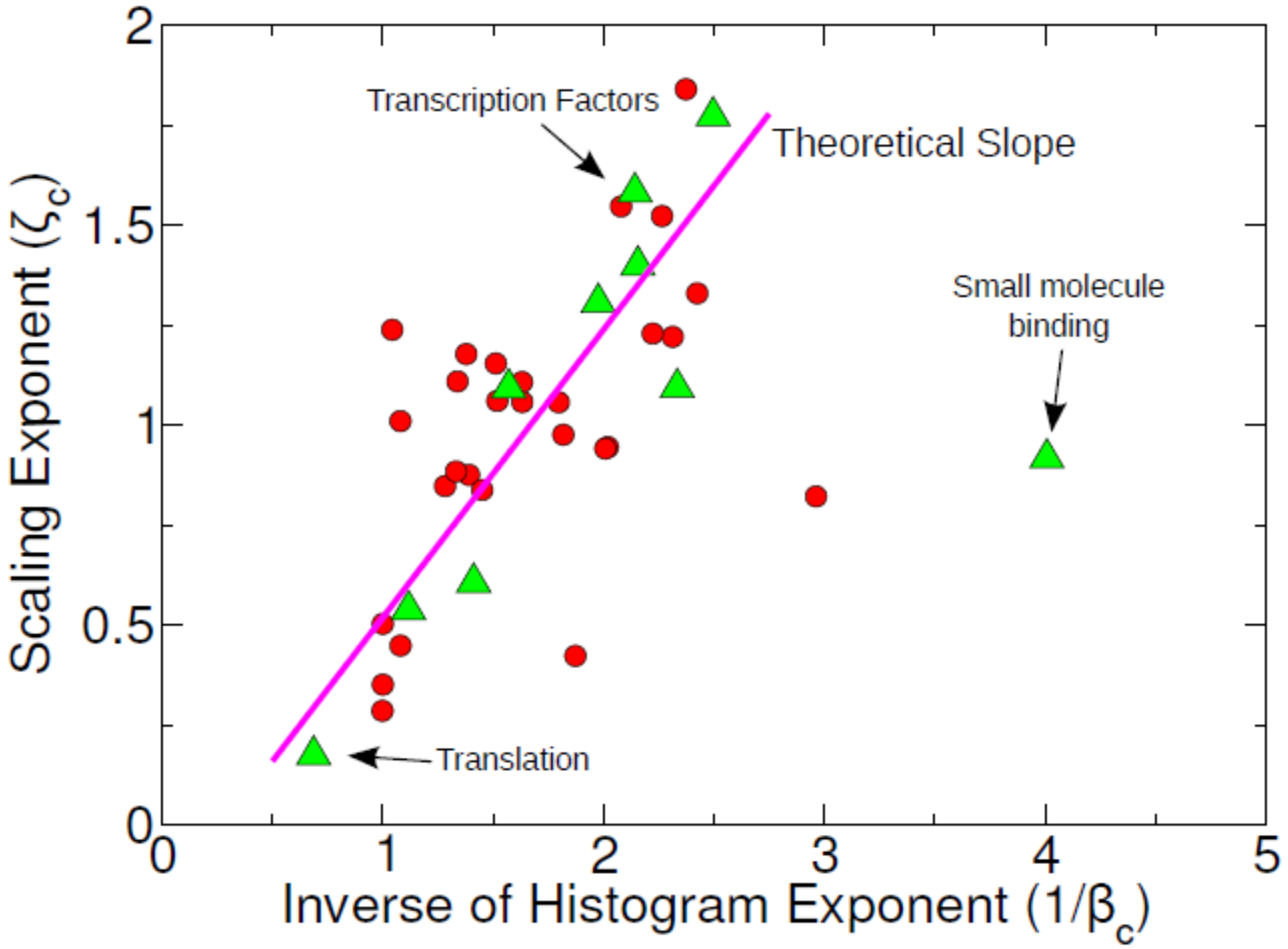


Total

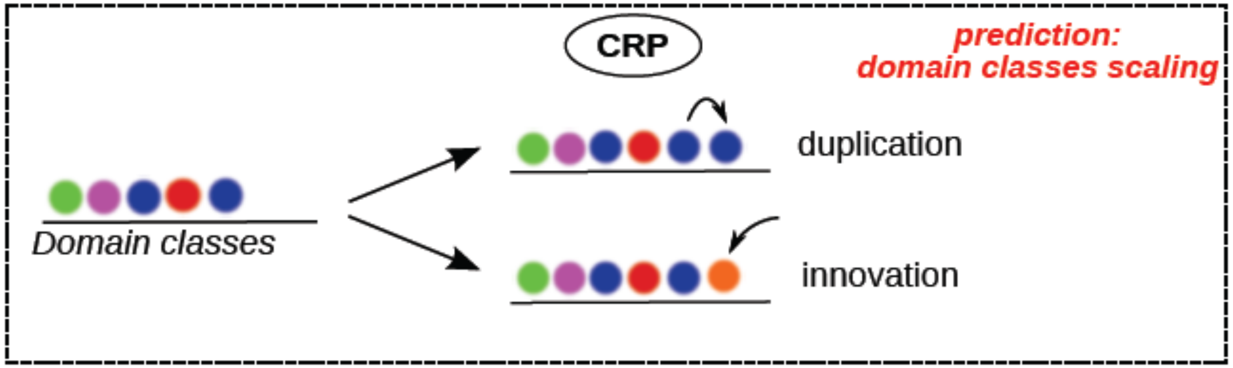
TFs

Empirical data follow the predicted trend

Valid for many categories $\beta_c = \alpha/\zeta_c$



CRP with evolutionary potentials – also possible



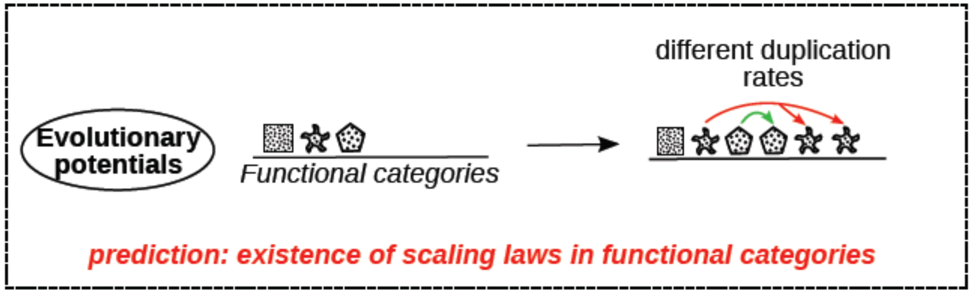
CRP

prediction: domain classes scaling

Domain classes

duplication

innovation

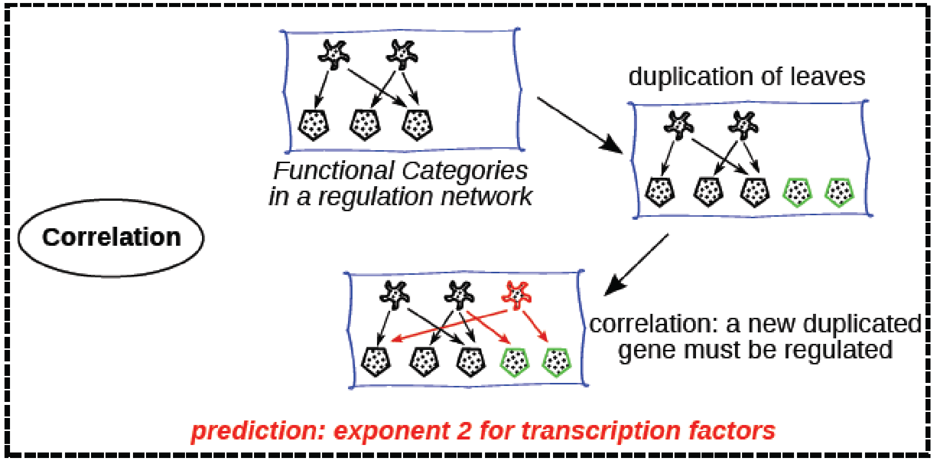


Evolutionary potentials

Functional categories

different duplication rates

prediction: existence of scaling laws in 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

CRP with evolutionary potentials

Insert evolutionary potentials in family expansion moves:

$$p_O^i = \frac{\rho_{c(i)} n_i - \alpha}{\sum_{j=1}^{f} \rho_{c(j)} n_j + \theta}$$

Giving per-function rates:

$$p_O^c := \sum_{i \in c} p_O^i = \frac{\rho_c n_c - \alpha f_c}{\sum_{j=1}^{f} \rho_{c(j)} n_j + \theta}$$

As usual:

$$p_N = \frac{\alpha f + \theta}{\sum_{j=1}^{f} \rho_{c(j)} n_j + \theta} \qquad p_N^c := \chi_c p_N$$

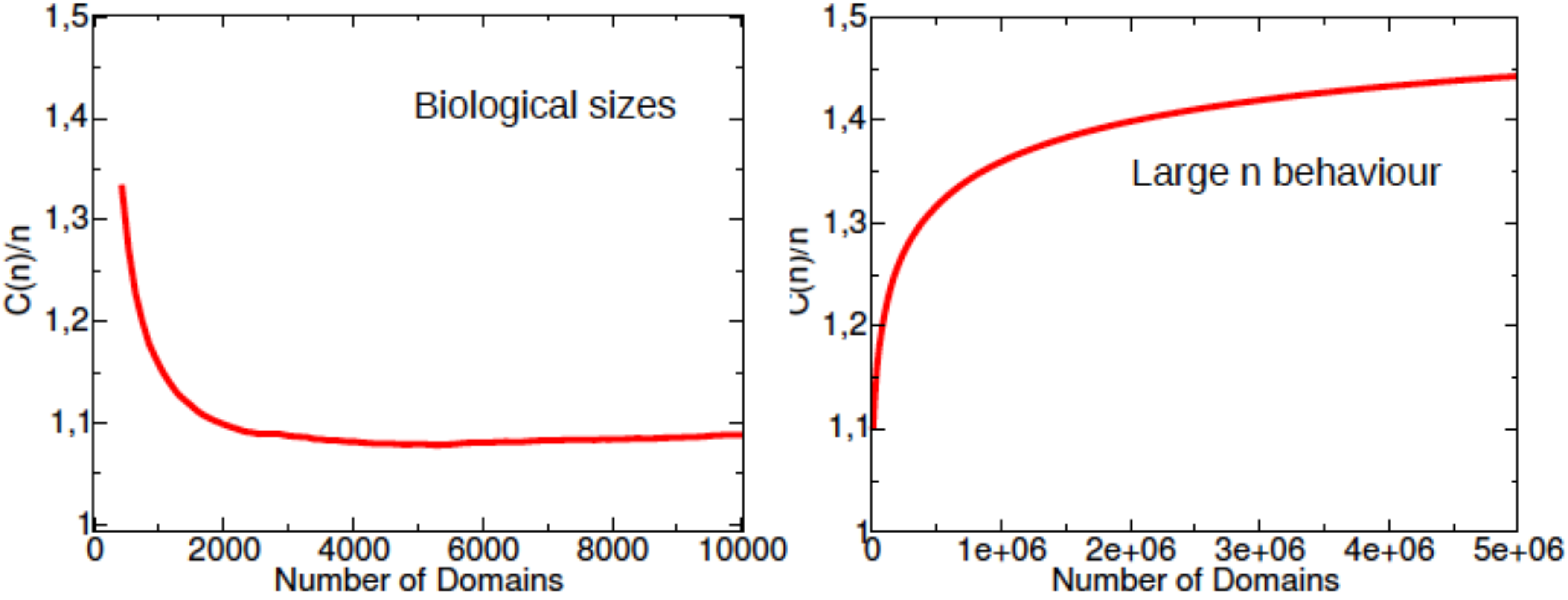
CRP with evolutionary potentials

$$\partial_n n_c = \frac{\rho_c n_c + \theta \chi_c}{C(n)}$$

If *C(n) ~ n*
these are the usual
Evolutionary potentials

$$C(n) \simeq \sum_i \rho_i n_i$$

Model with 3 categories (met, TF ,others)



CRP with evolutionary potentials

Problems:

→ Does not give large-*n* power-law

→ cannot easily give exponents > 1 (as for TFs)

A common description of homology classes and functional scaling laws in terms of evolutionary potentials is possible but not entirely convincing

Conclusions

- Effectively finite universe for innovation

- Nontrivial predictions from joint partitioning into functional and evolutionary classes