Systems biology and the prospect of a Post-modern Evolutionary Synthesis

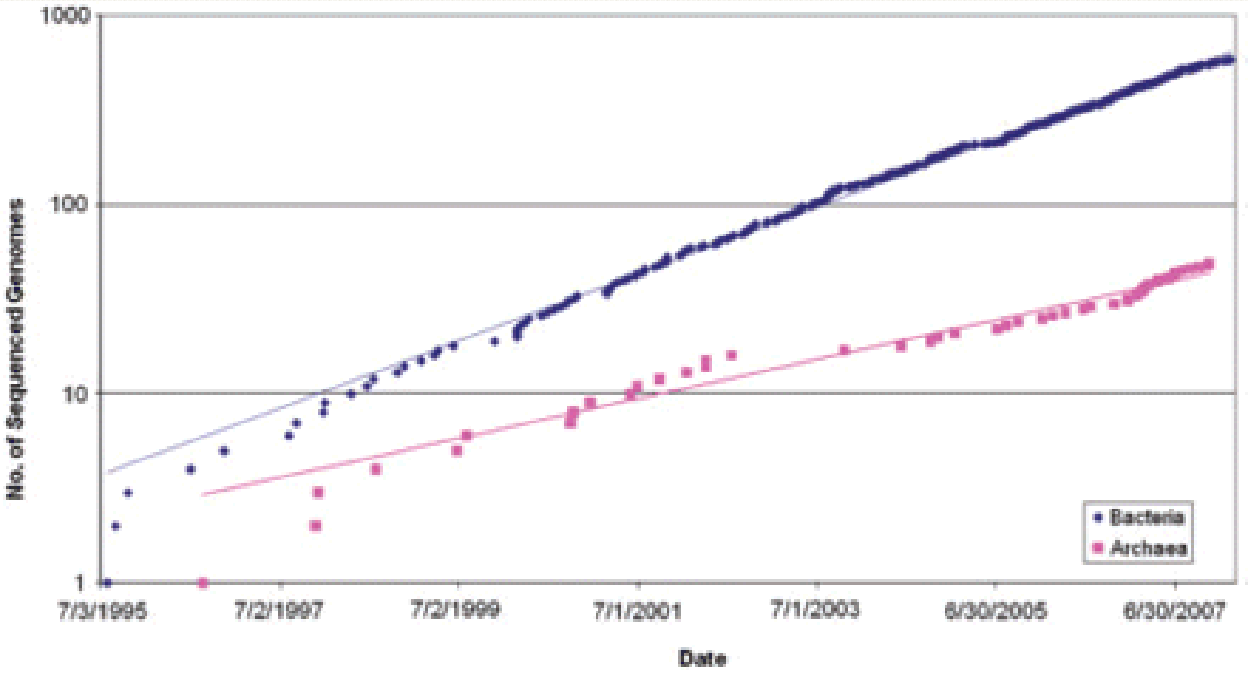
Eugene V. Koonin

National Center for Biotechnology Information, NLM, NIH

Granlibakken, Tahoe City, 10-22-2009

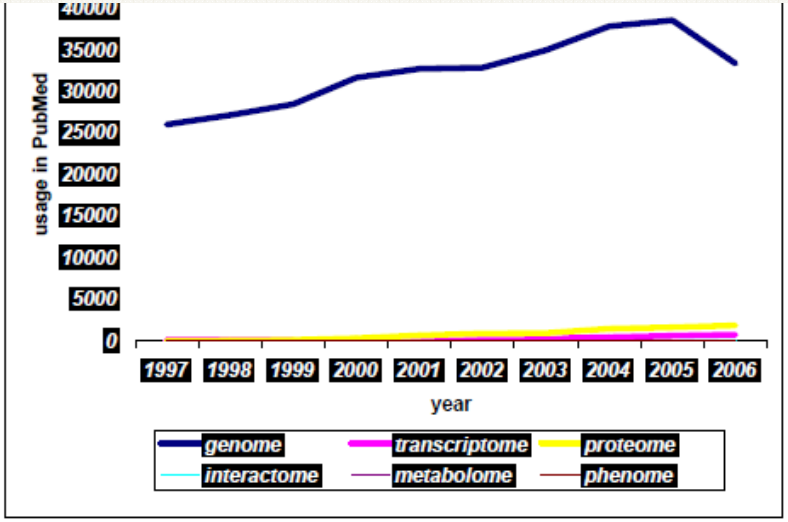
The genomic explosion

Fleischmann et al.
Whole-genome random sequencing and assembly of Haemophilus influenzae Rd.
Science. 1995 Jul 28;269(5223):496-512.

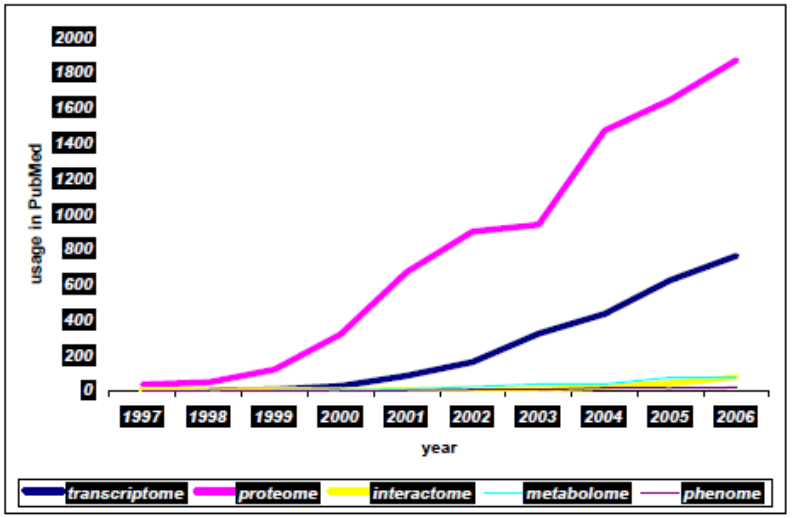


Koonin, Wolf, NAR 2008

The emergence of Systems Biology



(B)



Koonin, Wolf, 2007, in *Evolutionary Genomics and Proteomics*

I fail to see that ours is a better world for the invention of the term 'proteomics' either, especially since it seems to mean different things to almost everyone who is trying to do it. And why on earth do we need 'metabolomics', which doesn't even sound nice, or 'transcriptome' (which is clearly a dense book of American academic records)?

Petsko, 2001, **Homologuephobia.** Genome Biol

So what's the issue with all these new terms (or "exapted" ones, to use a favorite term of Stephen Jay Gould's to indicate something pre-existing that has been recruited for a new function)? Or, for a good measure, with all the mushrooming '-omes' -transcriptome, proteome, metabolome, and even phylome…
Is the world a better place because of them?
-Actually, for what it's worth, I think it is.
-These are not just words, after all: they are ***new memes for the science of a new age***.

Koonin, 2001, An apology for orthologs - or brave new memes. Genome Biol

Kelley & Scott

EMBO reports **9**, 12, 1163–1167 (2008)

The evolution of biology. A shift towards the engineering of prediction-generating tools and away from traditional research practice

The traditional aim of scientific research—most notably in physics—has been to gain a comprehensive understanding of natural phenomena and to generate hypotheses that provide simple, law-like and broad explanations. Contemporary research in bioinformatics is markedly different: bioinformaticians are increasingly generating tools to make accurate predictions for a restricted range of phenomena, irrespective of their simplicity or broader application in science.

…*biologists should not expect evolved biological phenomena to be amenable to theoretical generalization above the low-level principles of chemistry and physics*. *The best possible biological account of an evolved mechanism will probably be immensely messy and offer no cognitively appealing insight into its workings. This conclusion is corroborated by a range of recent findings showing that, far from uncovering simple mechanisms at the foundations of biology, we find an astonishing degree of complexity.*

Genomics and Systems Biology reveal unsuspected evolutionary universals and shed new light on the nature of genome evolution

[from stamp collection to physics?]

Universals of genome evolution

- Ubiquitous power laws:
  - -paralogous family size
  - -network structure
  - -…and more
- Scaling of functional classes of genes with genome size
- **Distribution of evolutionary rates across orthologous gene sets**
- **Structure of correlations between evolutionary and phenomic variables**

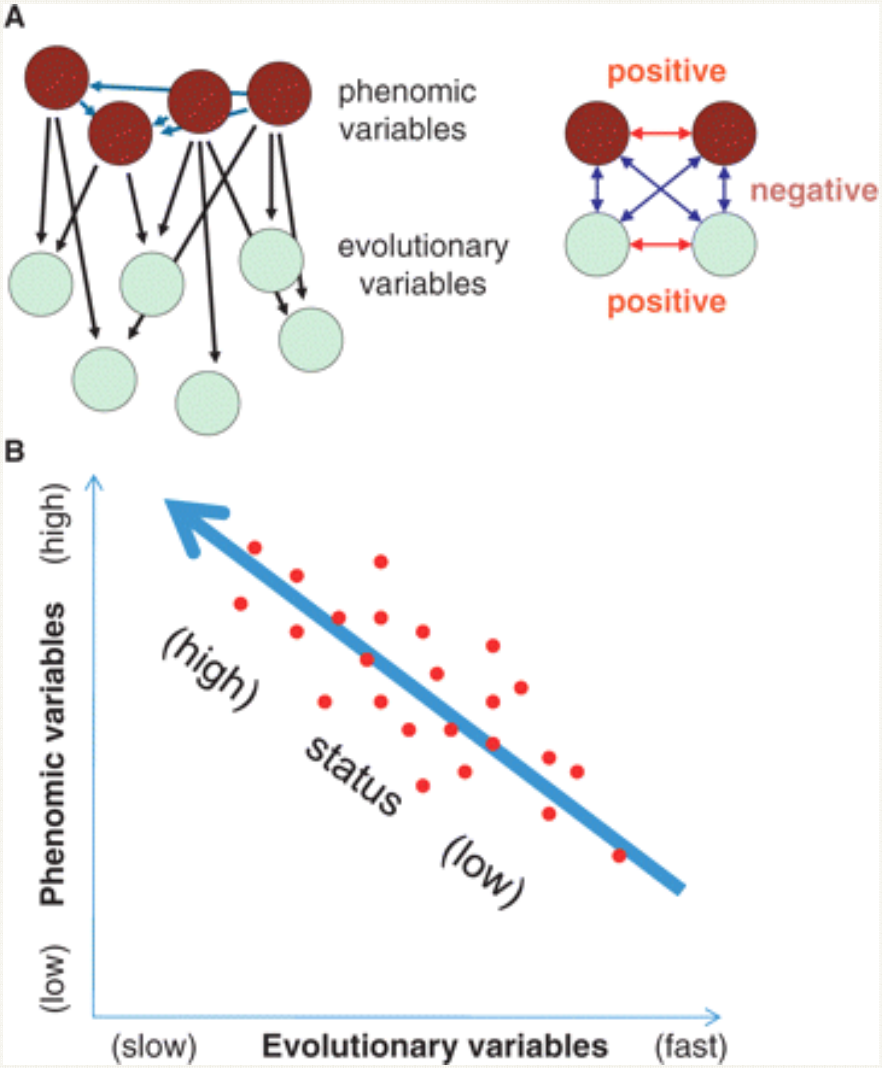
Universal structure of correlations between evolutionary and phenomic variables

	NP	PPI	GI	PGL	**ER**	EL	KE
NP	-						
PPI	0.057	-					
GI	0.060	0.034	-				
PGL	0.000	-0.125	-0.019	-			
ER	-0.070	-0.200	0.034	0.141	-		
EL	0.129	0.199	-0.050	-0.099	**-0.277**	-	
KE	0.027	0.234	-0.048	-0.181	-0.155	0.188	-

The negative correlation between evolutionary rate and expression level is the strongest of all correlations between evolutionary and phenomic variables uncovered by evolutionary systems biology – in particular, much greater than the correlation between gene essentiality and any other variable

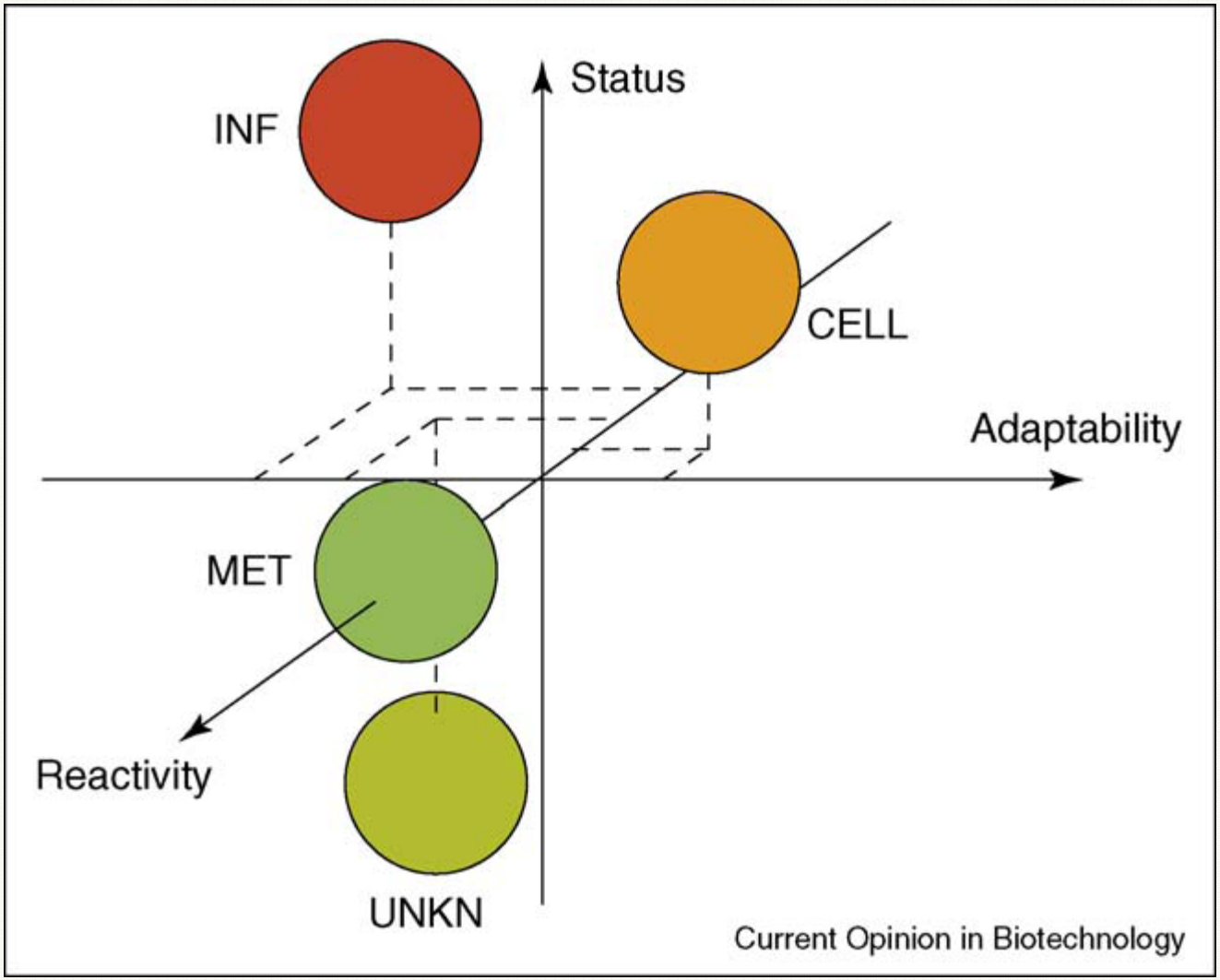
Wolf, Carmel, Koonin, Proc Biol Sci. 2006 Jun 22;273(1593):1507-15

Model of ***gene status***



Wolf, Carmel, Koonin, Proc Biol Sci. 2006 Jun 22;273(1593):1507-15

Functional classes of genes differ in their “status”, “adaptability”, and “reactivity” (the first 3 principal components)

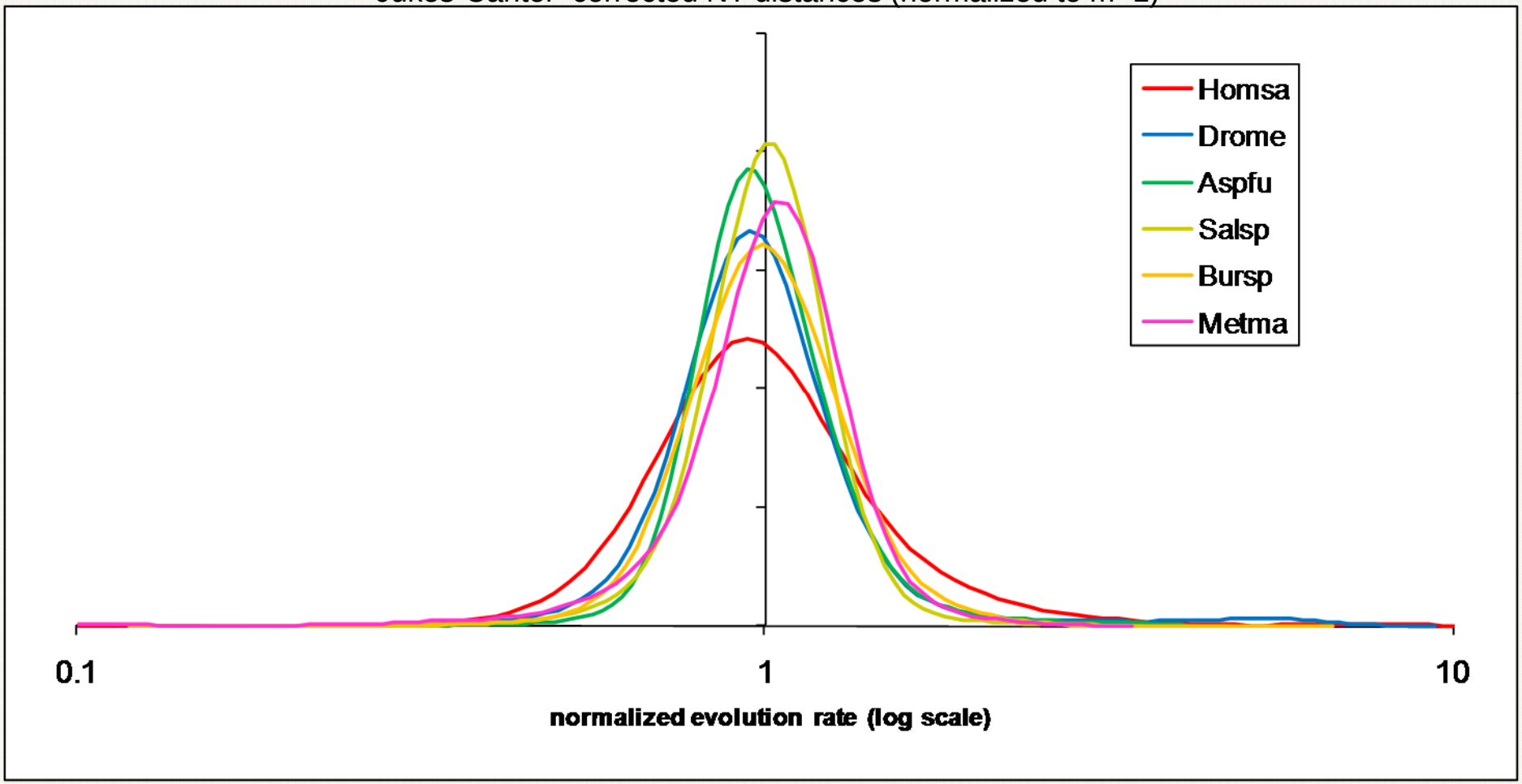


Koonin, Wolf, 2006, Curr. Opin. Biotechnol.

Universals of evolution: distribution of evolutionary rates of genes

Effectively, the same, approximately log-normal distribution in all walks of life, from bacteria to mammals

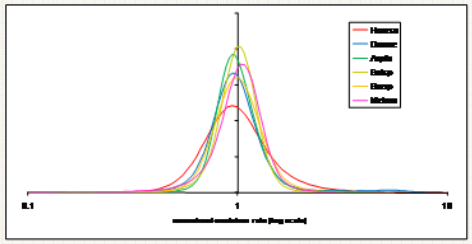
Jukes-Cantor corrected NT distances (normalized to m=1)



Metma - *Methanococcus maripaludis* C5 vs *M. maripaludis* C7 (Euryarchaeota)
Bursp - *Burkholderia cenocepacia* MC0-3 vs *B. vietnamiensis* G4 (Proteobacteria)
Salsp - *Salinispora arenicola* CNS-205 vs *S. tropica* CNB-440 (Actinobacteria)

Wolf et al. 2009 PNAS

The causes of variation of protein evolutionary rate



- The rate of protein evolution varies more than 1000-fold and, for the past 30 years, it was thought that the rate was determined by protein function

MicInerny, 2006, Trends Ecol. Evol. Volume 21, Issue 5

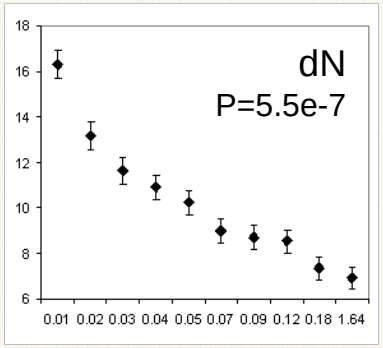
Why do proteins evolve at different rates? Advances in systems biology and genomics have facilitated a move from studying individual proteins to characterizing global cellular factors. Systematic surveys indicate that protein evolution is not determined exclusively by selection on protein structure and function, but is also affected by the genomic position of the encoding genes, their expression patterns, their position in biological networks and possibly their robustness to mistranslation.

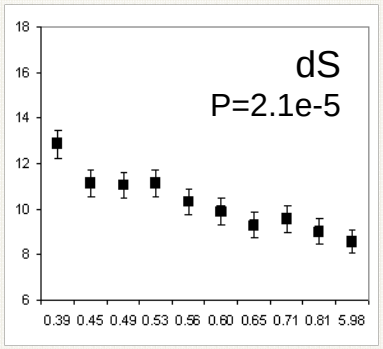
Pal, Papp, Lercher, Nat Rev Genet. 2006 May;7(5):337-48

Universals of evolution: negative correlation between expression and evolutionary rate of genes – highly conserved genes are highly expressed

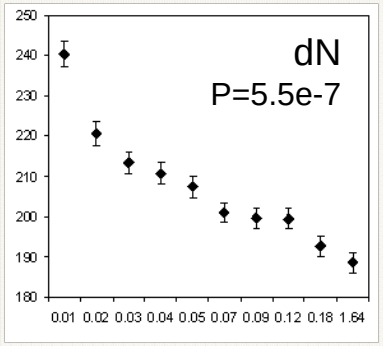
Expression vs. evolutionary rate

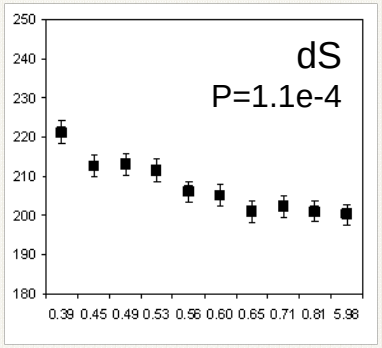
expression breadth





expression level

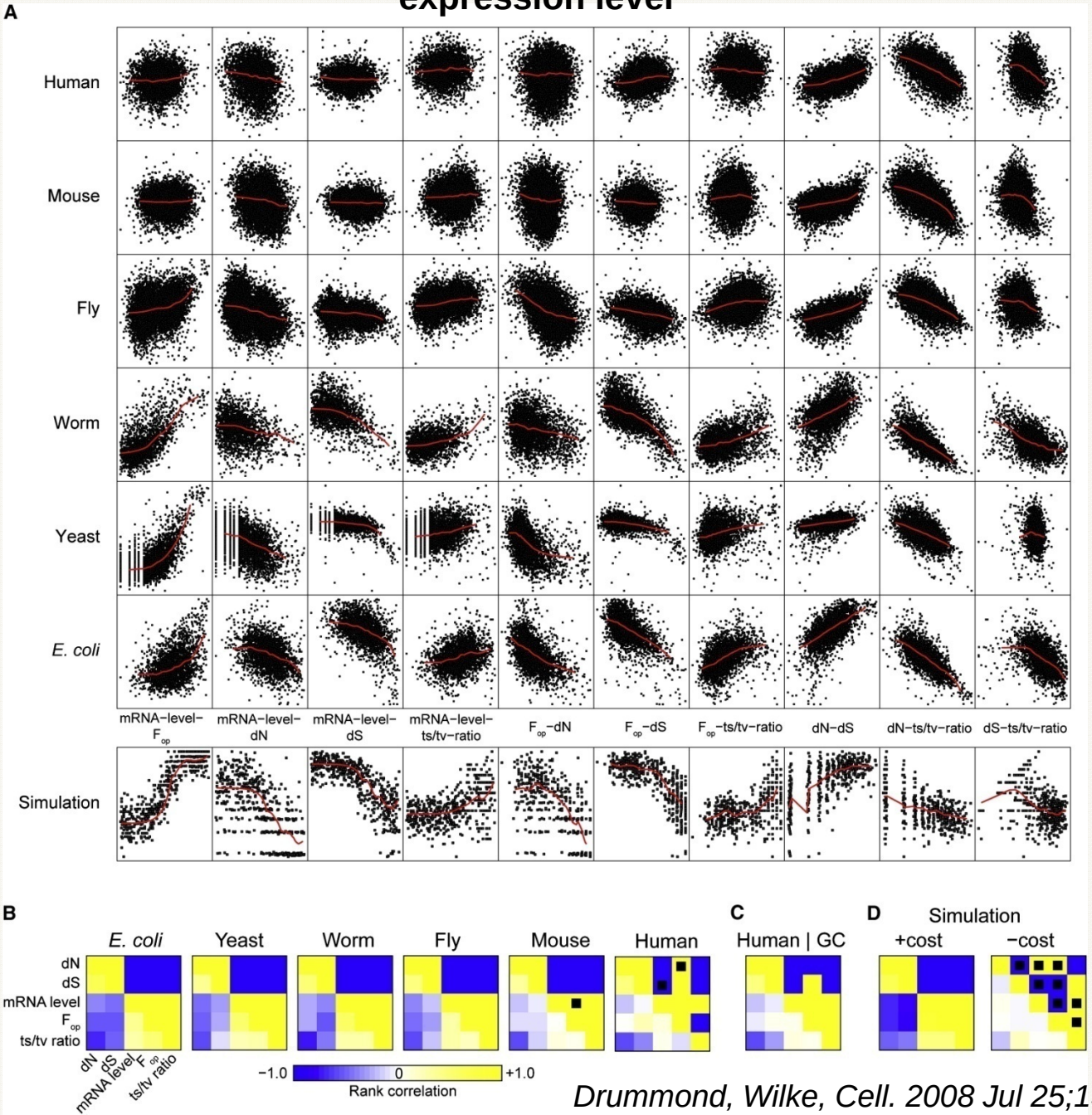




- **Selective constraint on gene sequences is strongly related to the level and breadth of gene expression**

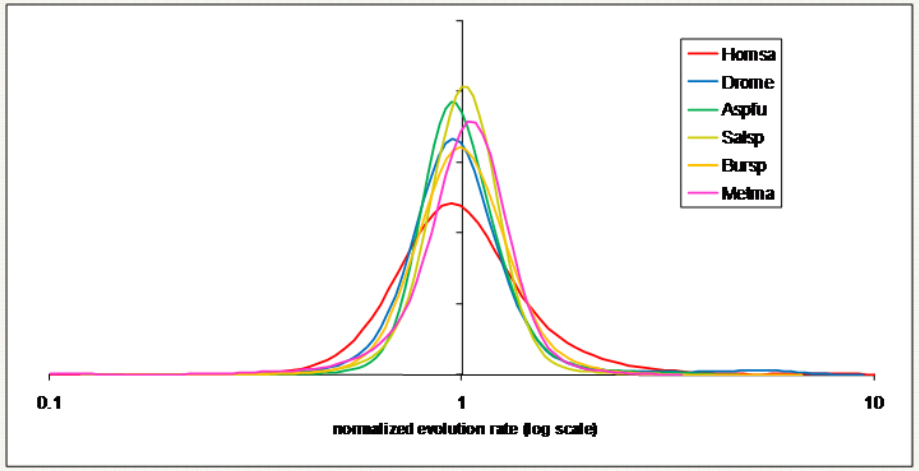
Jordan et al. Mol Biol Evol (2004) 21: 2058

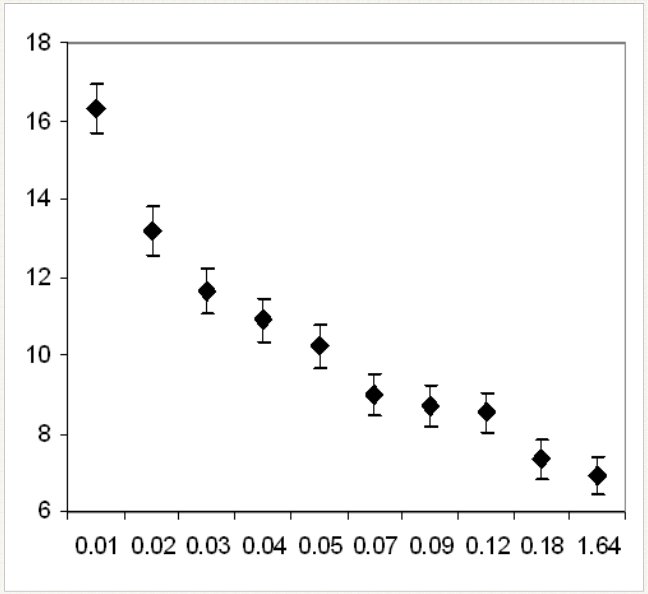
Evolutionary rate of genes is consistently and significantly negatively correlated with expression level



Drummond, Wilke, Cell. 2008 Jul 25;134(2):341-52.

Universals of evolution: distribution of evolutionary rates of genes and the anticorrelation between ER and expression





Universality and simple form of the distribution suggest a simple, generic cause(s) unrelated to unique biological functions

Is the universality and shape of the ER distribution and the ER-expression dependence simple consequences of proteins being heteropolymers in 3D?

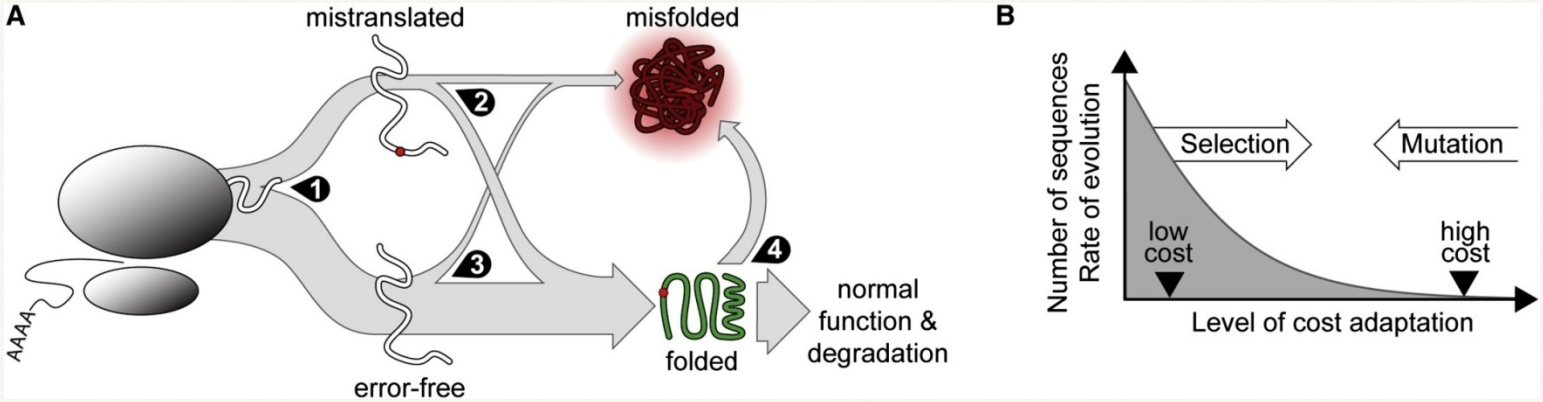
Lobkovsky, Wolf, Koonin, submitted

The MIM hypothesis

•Protein misfolding dominates the cost of mutation and mistranslation to a cell

•Robustness to misfolding - Intrinsic cost – Fitness

•Connect evolutionary rate with expression via misfolding

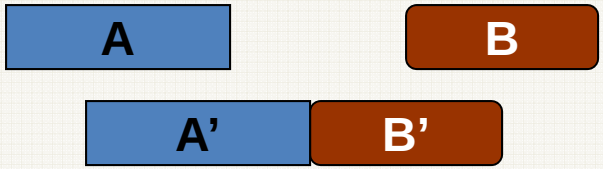


Drummond and Wilke, Cell (2008)

Misfolding probability (robustness to misfolding) of mutants & native sequence – rate of evolution

Key prediction of the MIM hypothesis

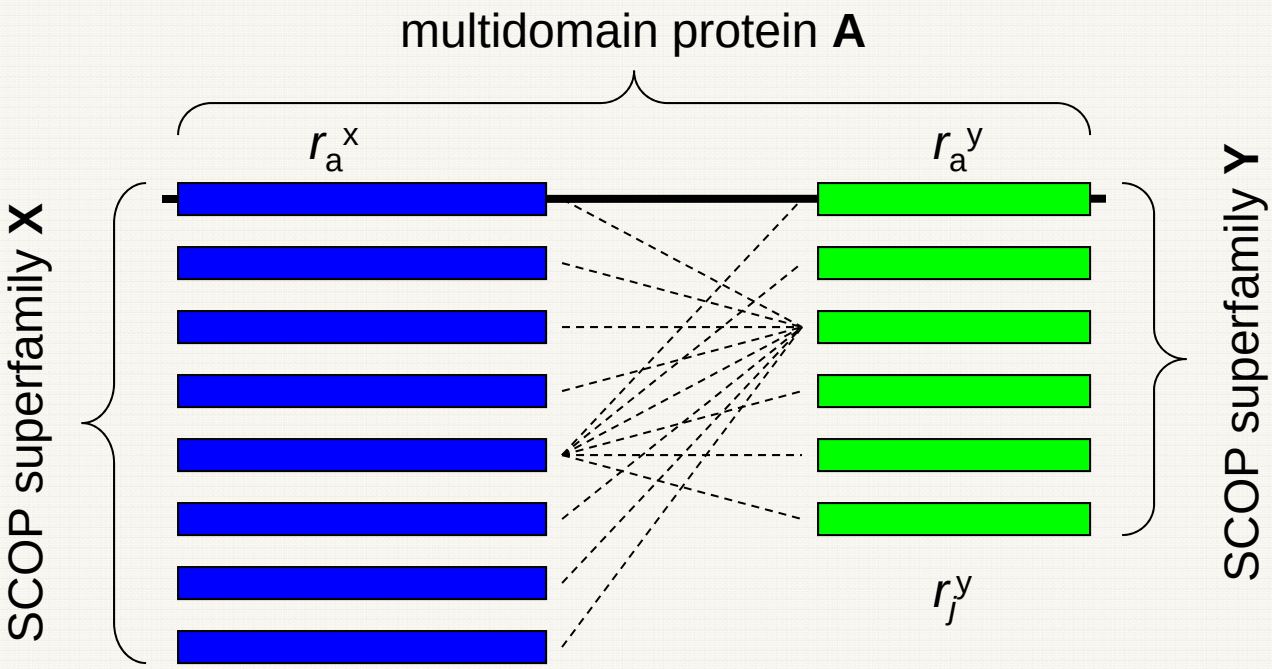
- The mistranslation-induced misfolding hypothesis would predict that, within multidomain proteins, fused domains, on average, should evolve at substantially closer rates than the same domains in different proteins because, within a mutlidomain protein, all domains are translated at the same rate.



$|Ra'-Rb'| << |Ra-Rb|$

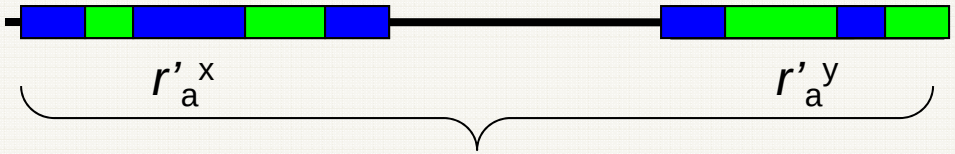
- We performed a comprehensive comparison of the evolutionary rates of mammalian and plant protein domains that are either joined in multidomain proteins or contained in distinct proteins

Wolf, Wolf, Koonin, Biol Direct. 2008 Oct 7;3(1):40



distribution of r_i^x/r_j^y for all possible pairs of domain **X** and domain **Y**

compare r_a^x/r_a^y to the distribution of r_i^x/r_j^y and to $r'_a{}^x/r'_a{}^y$



randomized by alignment permutation to control for sampling error

Pipeline of analysis

Map domains from SCOP/ASTRAL to human and plant proteins
(BLASTP against ASTRAL sequences)

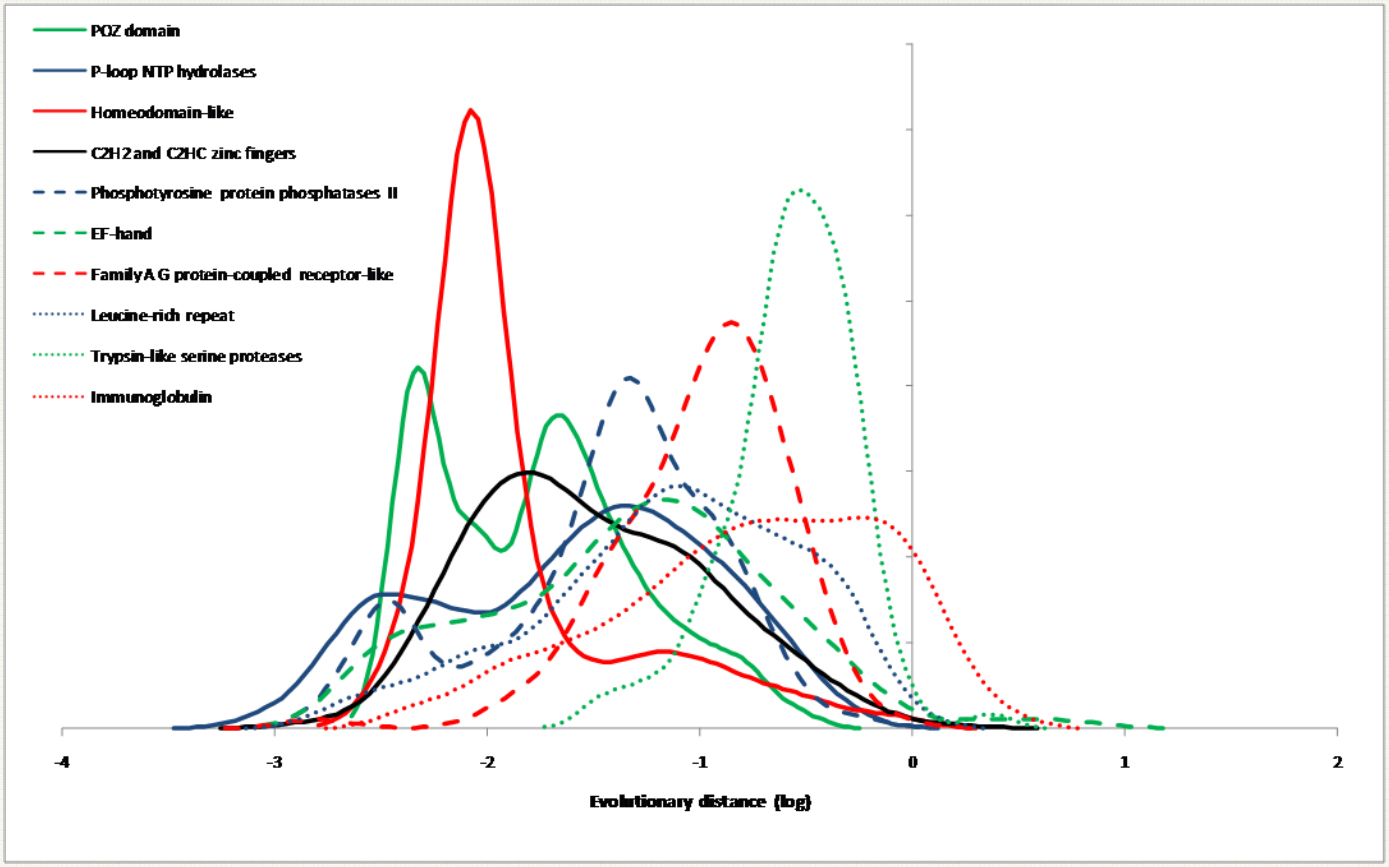
9786 domains in 7404 human proteins
5908 domains in 5118 Arabidopsis proteins

Identify sets of orthologous genes (bidirectional best hits):

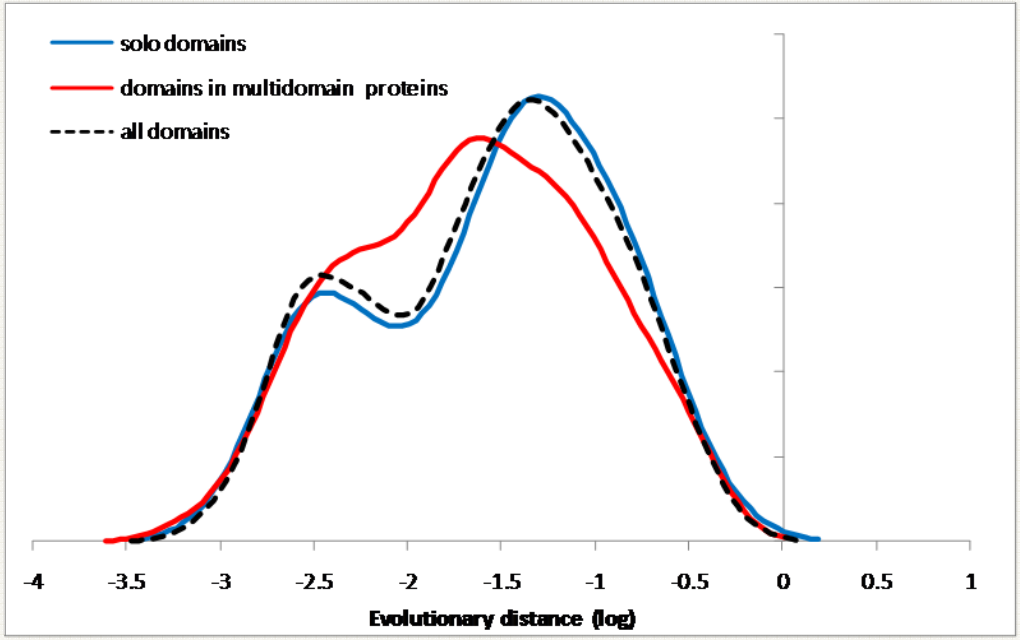
16,603 human-mouse orthologs
15,286 Arabidopsis-poplar orthologs

- Align protein sequences of orthologs (MUSCLE)
- Calculate evolutionary rates (PROTDIST/JTT/Gamma)

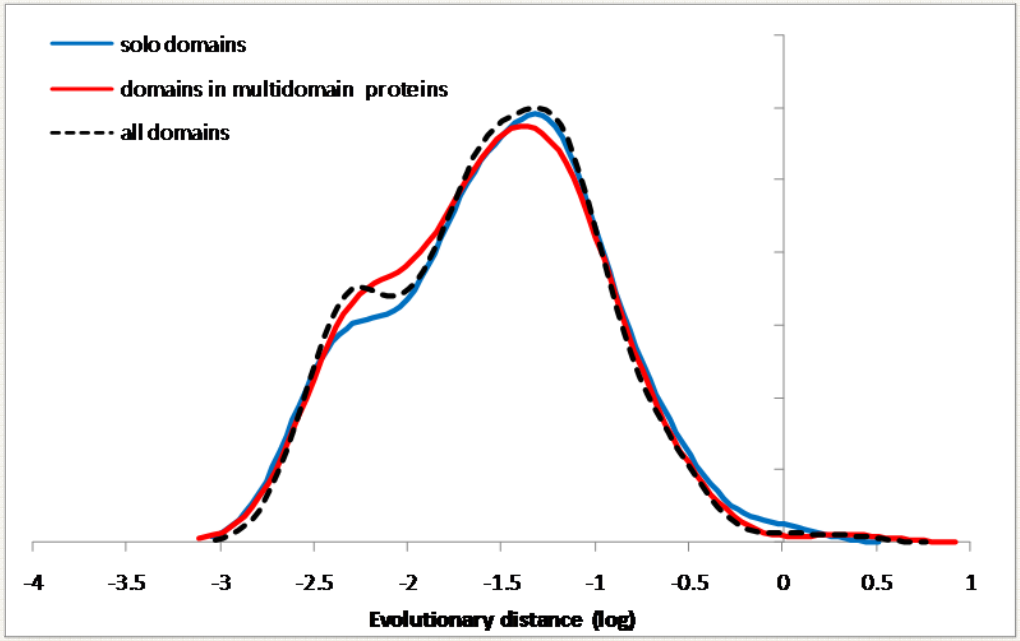
- **Map domains to alignments of orthologs**
- **Derive rates for domains**



Abundant protein domains show widely different evolutionary rate distributions, with the means spanning 3 orders of magnitude



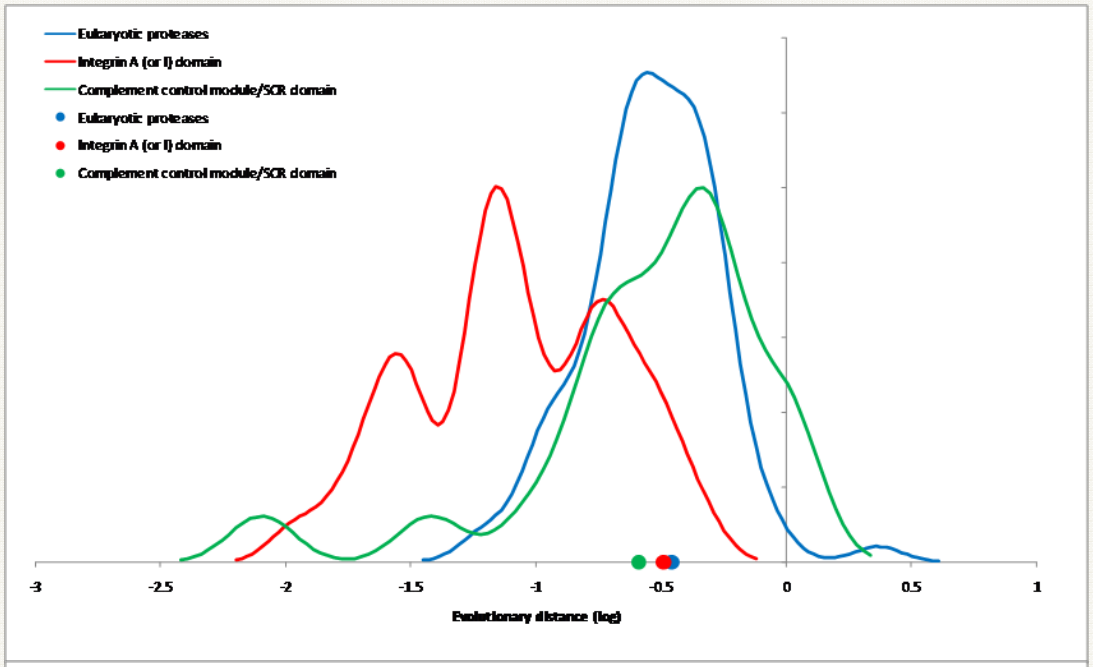
human



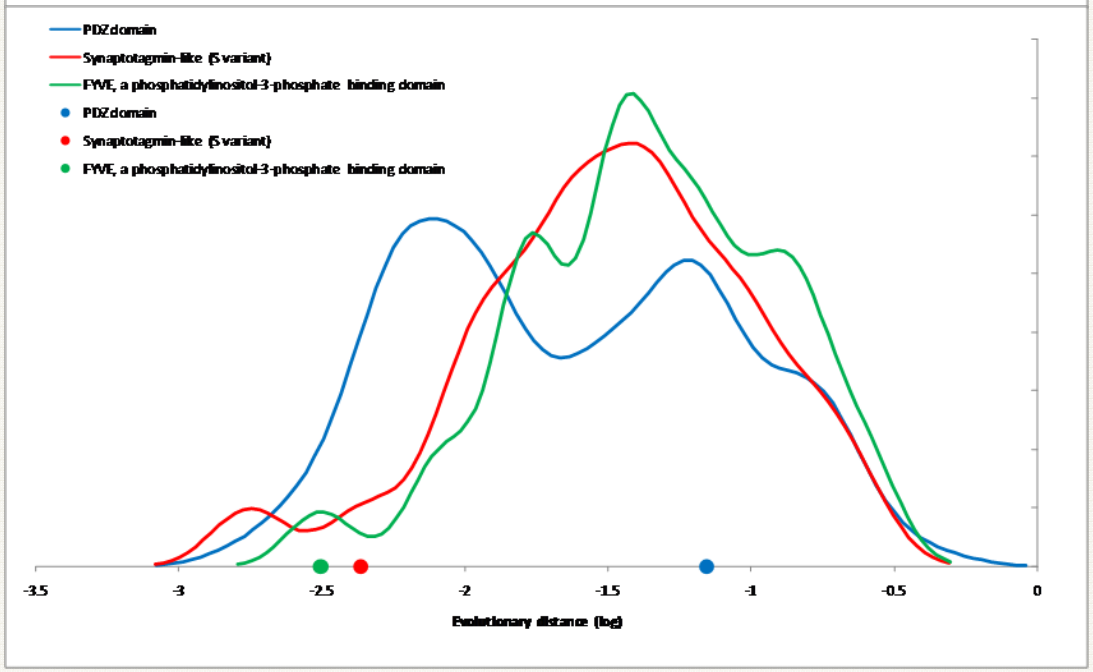
plant

Distributions of evolutionary rates of solo domains and domains within multidomain proteins are very similar

Individual test cases: much variance in rates of solo domains vs multidomain proteins

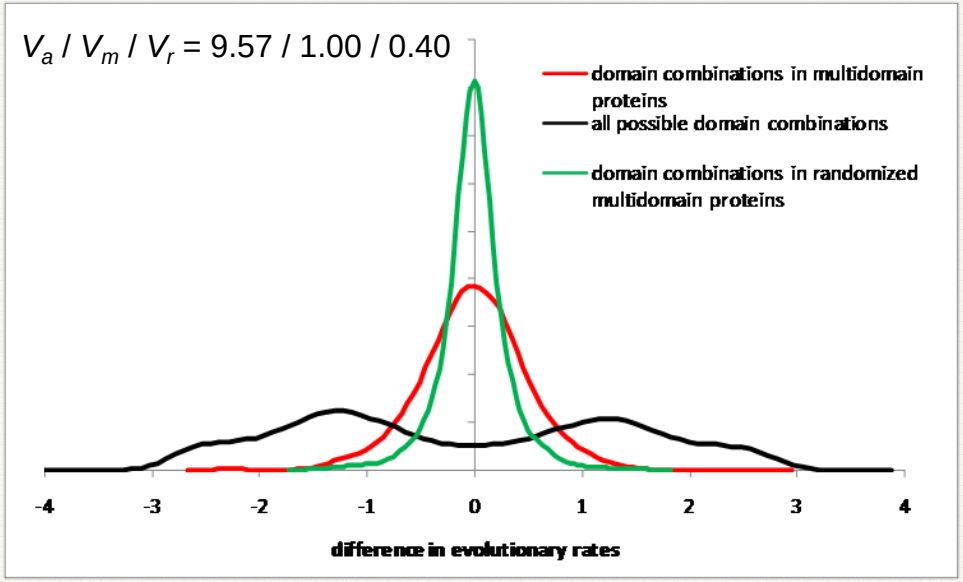


Near complete homogenization

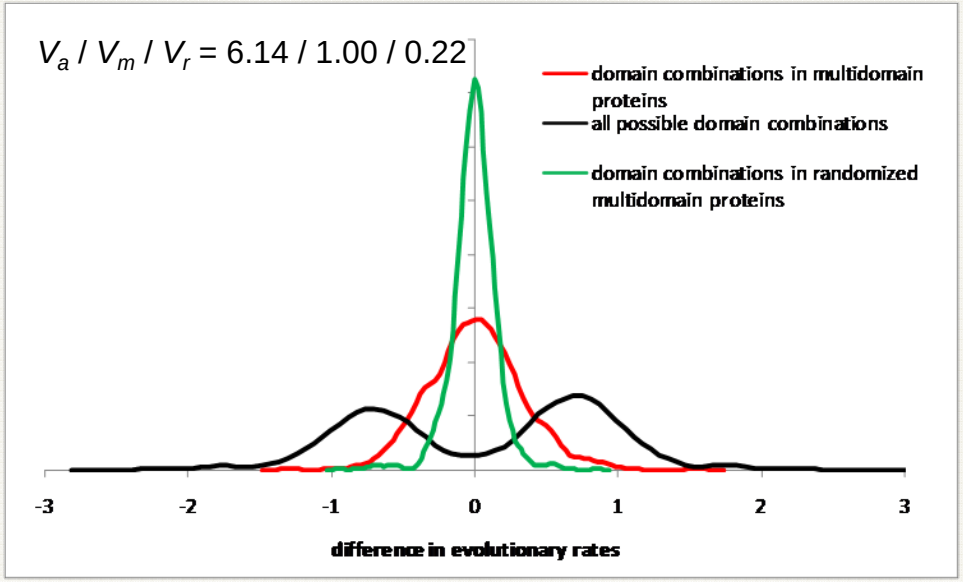


No homogenization whatever…
rather the opposite

Distribution of rate differences (ratios) between domains in multidomain proteins shows much less variance than the distribution of rate ratios between all domains but much more variance than the distribution of ratios between randomized “domains”:
Substantial but far from complete homogenization of domain evolutionary rates



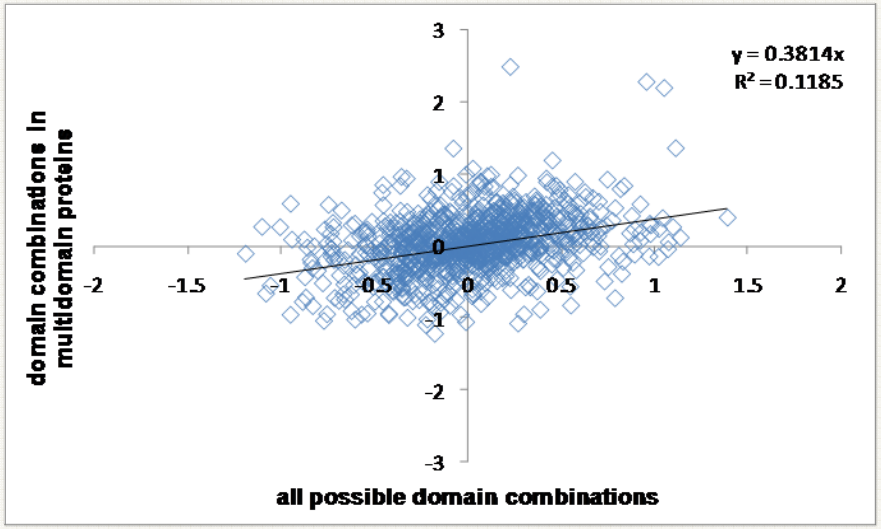
human

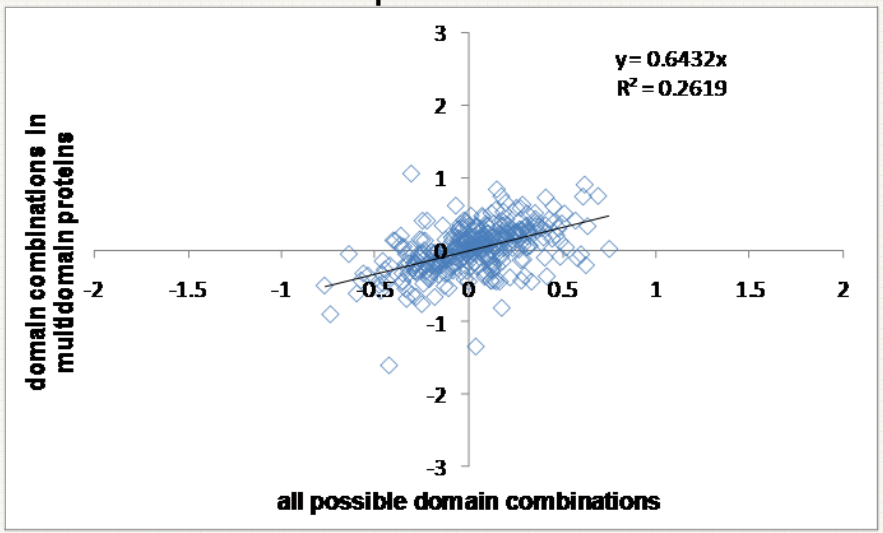


plant

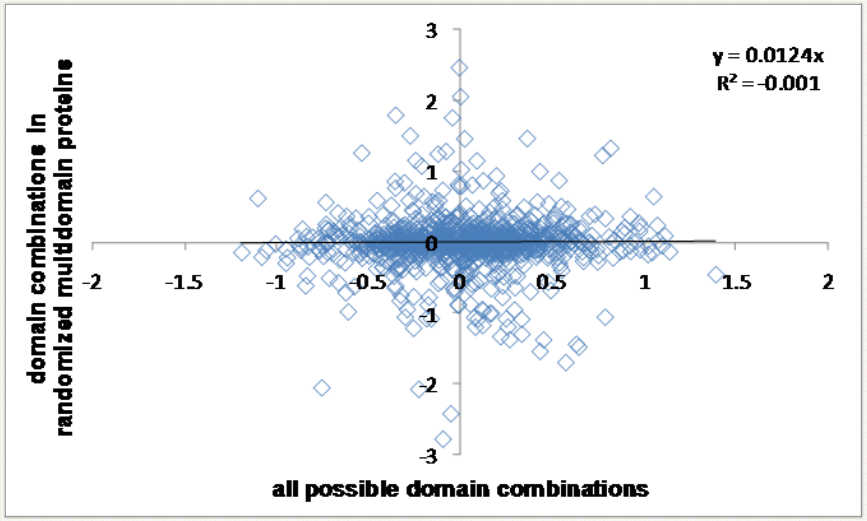
human

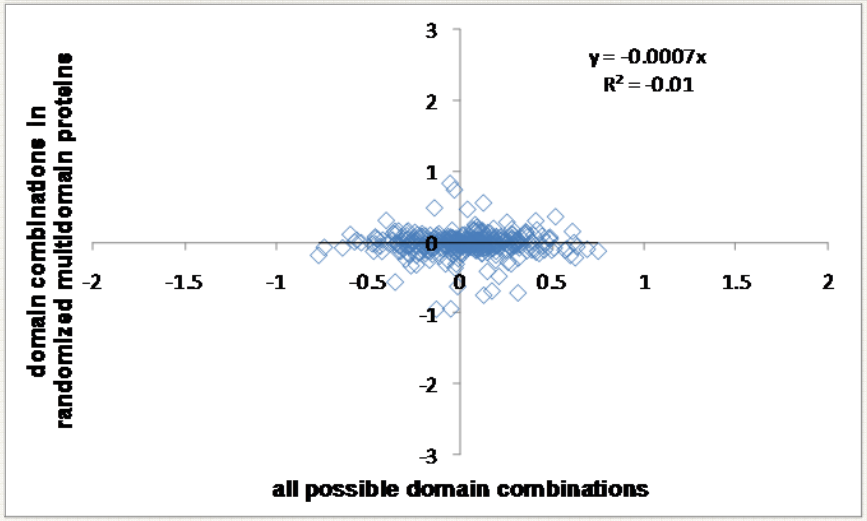
plant





real





rand

•Highly significant positive correlation between rate difference in multidomain proteins and all domain pairs

Example: All domains – 2-fold difference
multidomain – 1.3-fold difference

All domains – 2-fold difference
multidomain – 1.6-fold difference

- Fusion of domains in multidomain proteins results in a substantial but far from complete homogenization of evolution rate
- Expression (rate of translation) and intrinsic constraints exert quantitatively comparable effects on protein sequence evolution
- *Generalized MIM hypothesis: the rate of protein evolution is largely determined by*

  -intrinsic robustness to misfolding that depends on the characteristic stability and designability of the given domain

  -translation rate that serves as an amplifier of the intrinsic fitness cost of misfolding, hence the observed negative correlation between evolution rate and expression

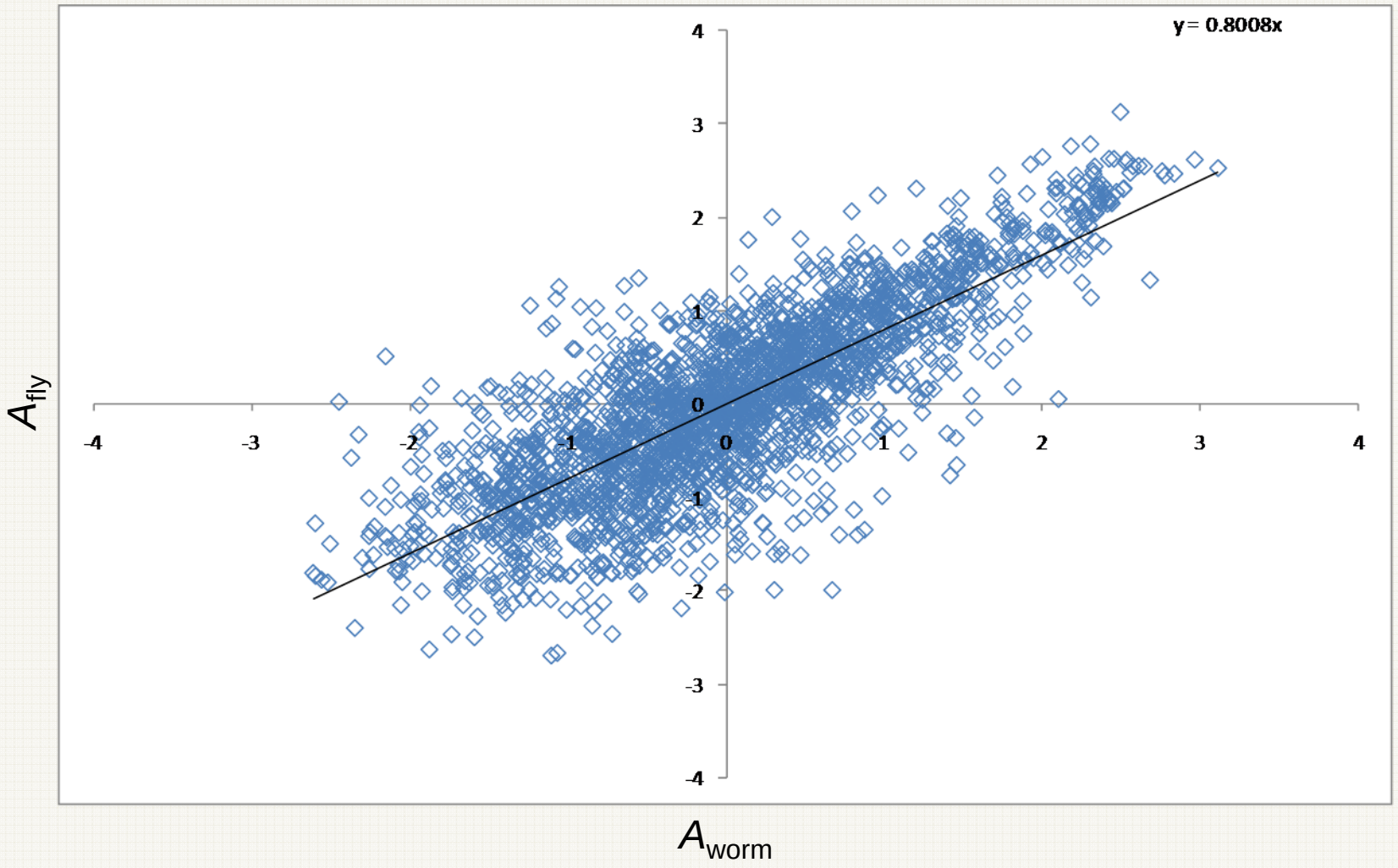
Wolf, Wolf, Koonin, Biol Direct. 2008 Oct 7;3(1):40

Relative contributions of structural-functional constraints and translation rates to the evolution of protein-coding genes: a general misfolding-driven evolution hypothesis

- by analysis of correlations between protein abundances and evolution rates from **2** lineages, the contributions of the two factors can be disentangled

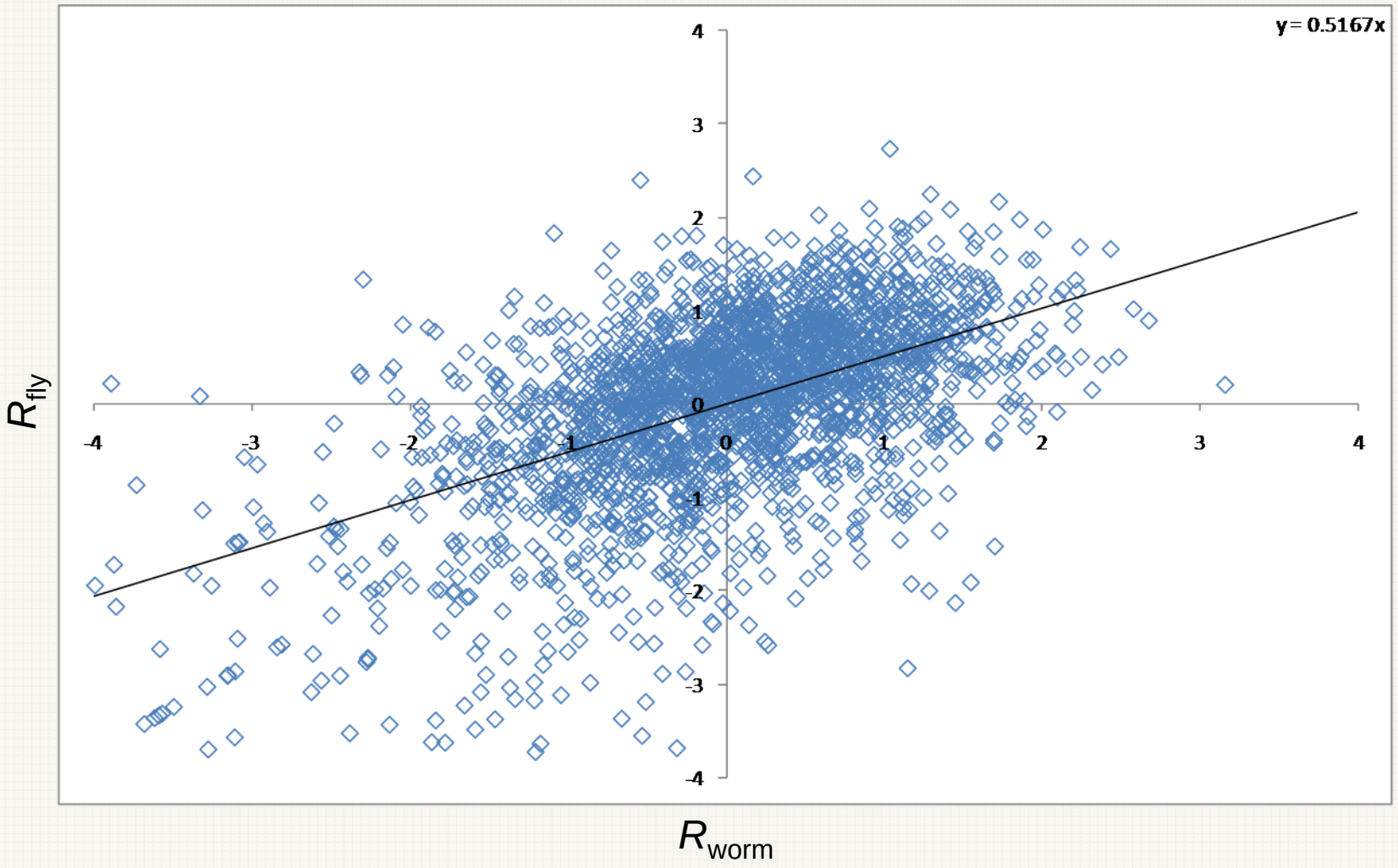
Wolf, Gopich, Lipman, Koonin, submitted

Striking correlation between protein abundances in distant animals (fly and nematode)

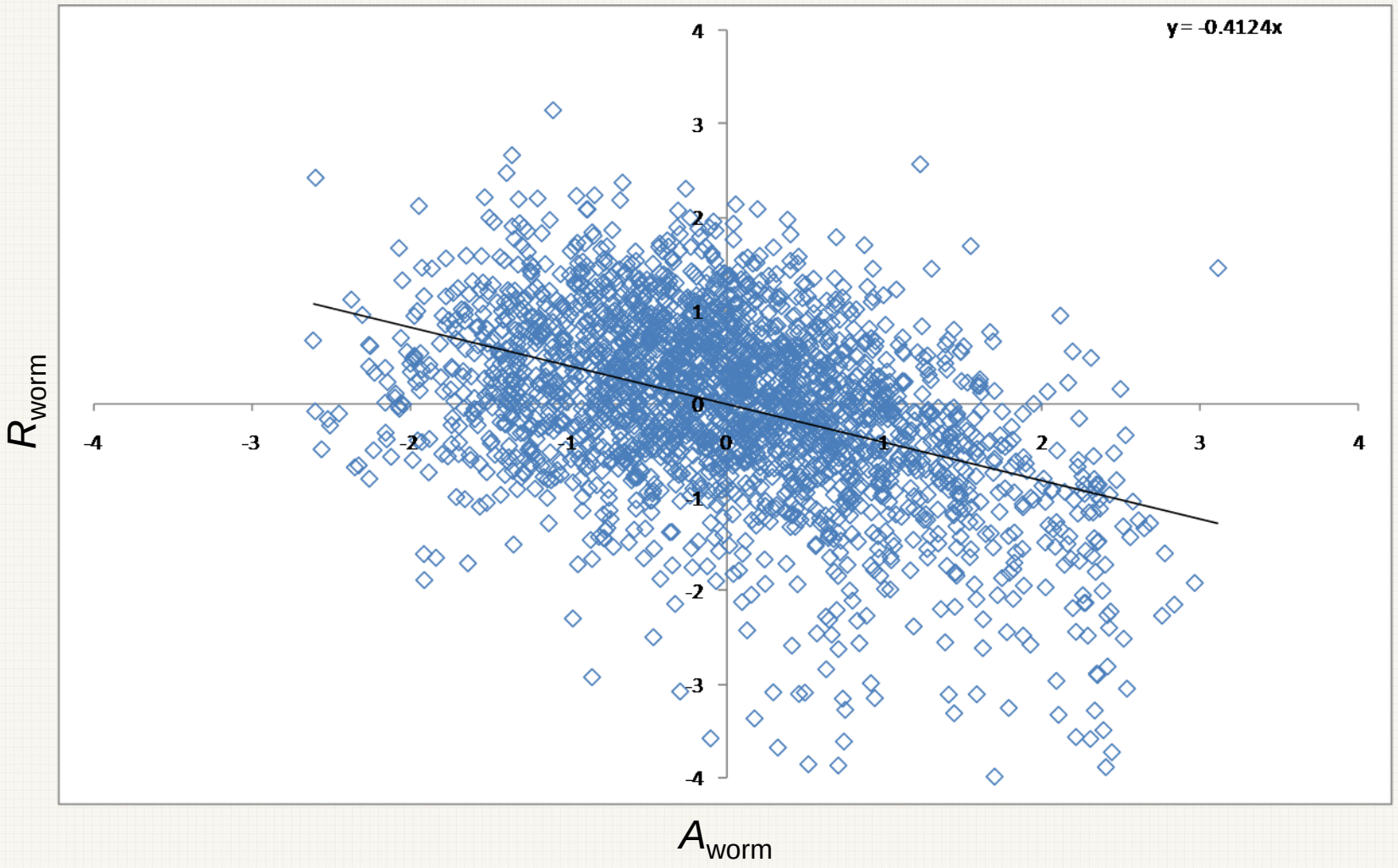


Data from **Schrimpf et al. PLOS Biol. 2009**

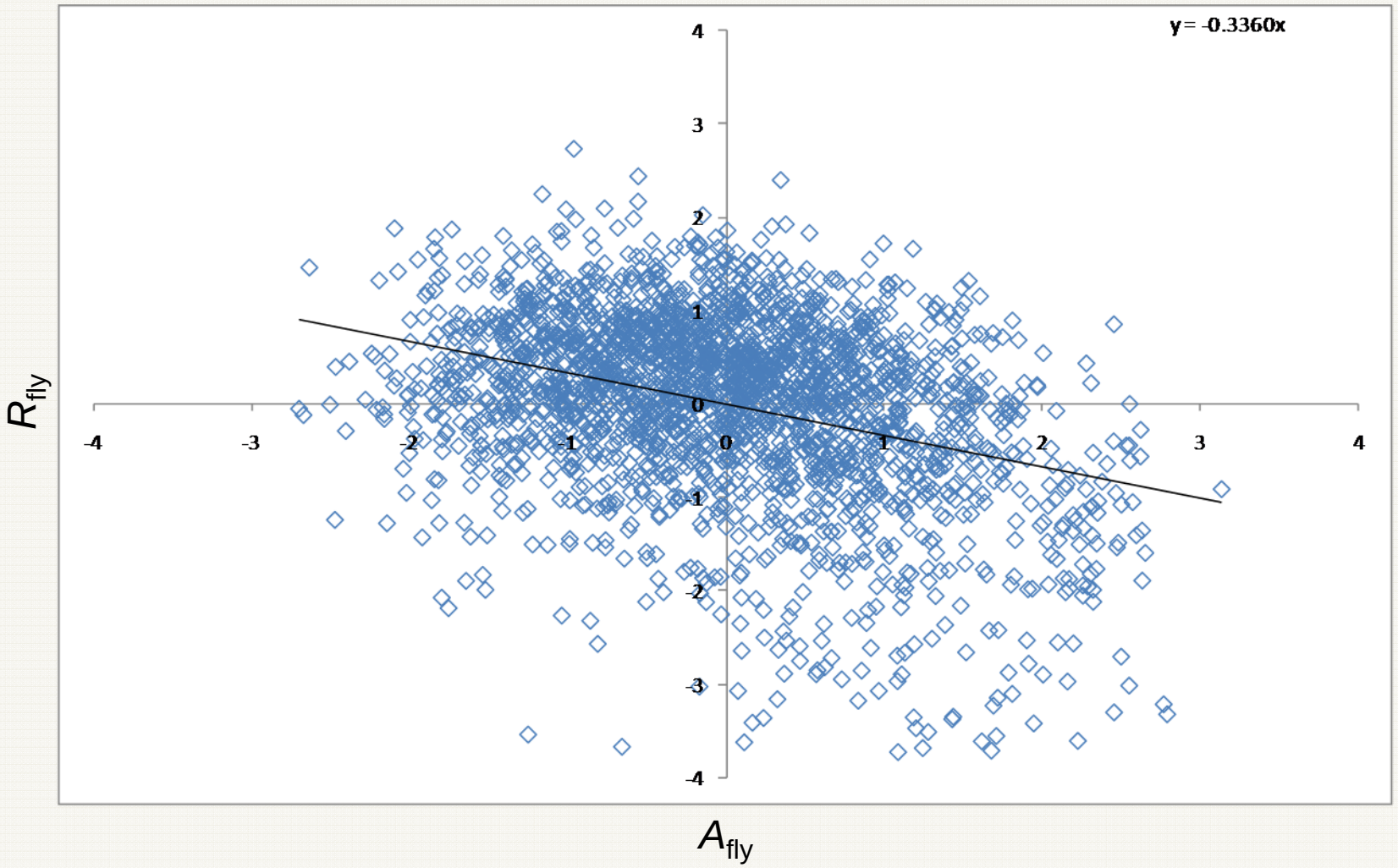
Somewhat less striking but highly significant correlation between evolutionary rates in the fly and nematode lineages



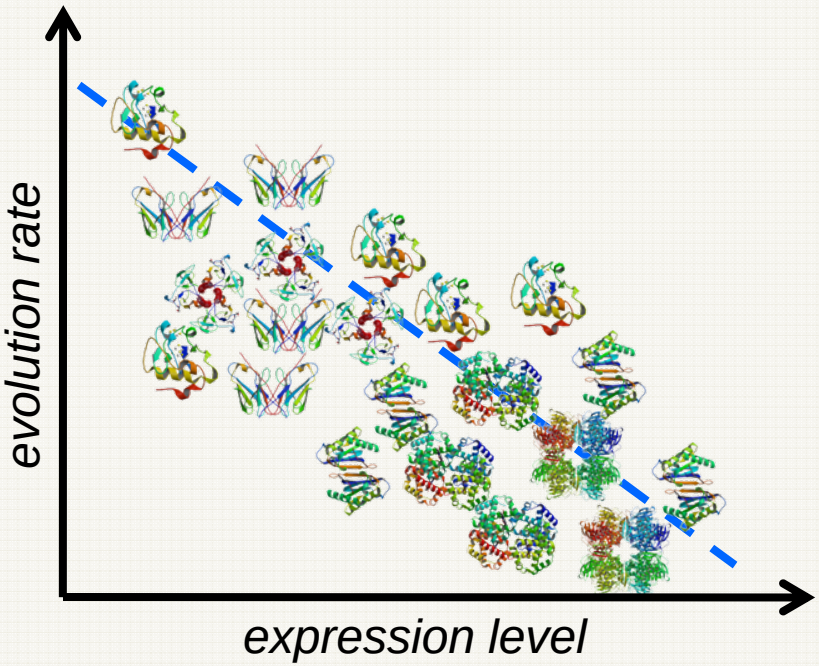
Universal anti-correlation between rate and abundance (nematode)



Universal anti-correlation between rate and abundance (fly)

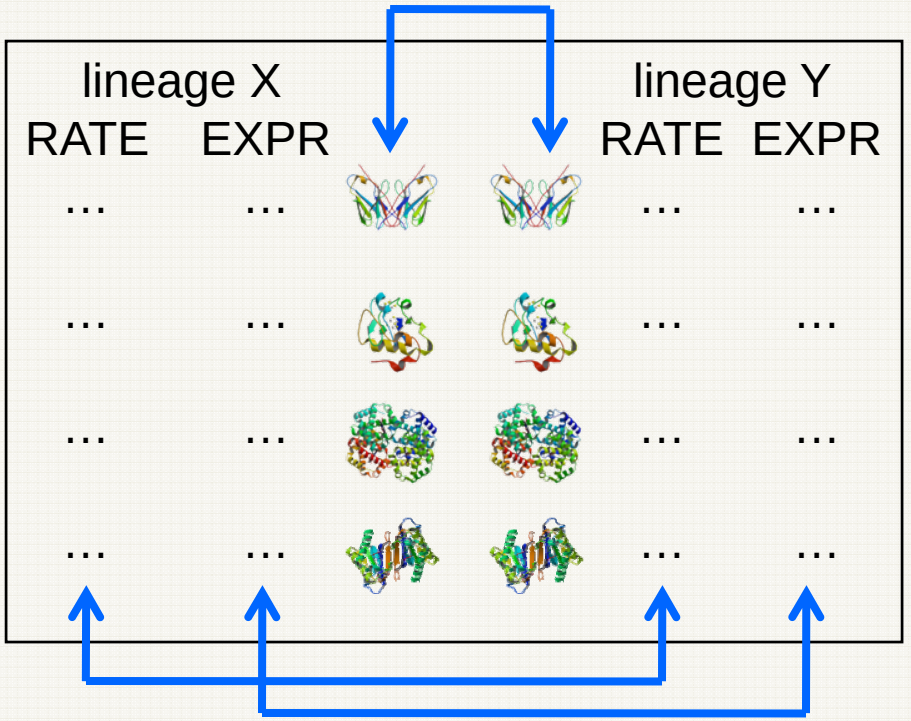


Correlation between expression level and evolution rate…



…conflates the pure effect of expression with the effect of intrinsic structural and functional constraints

orthologs have the *same* structure and function…



…so the difference between their evolution rates is determined by the difference between their expression levels

Comparison of protein abundances and evolution rates in two lineages can help disentangle contributions of *structural-functional constraints* and *translation rates*

evolution rate

translation rate

lineage X → $R_{X,i} = \beta S_i + \alpha_X T_{X,i} + E_{X,i}$

lineage Y → $R_{Y,i} = \beta S_i + \alpha_Y T_{Y,i} + E_{Y,i}$

structural and functional constraints

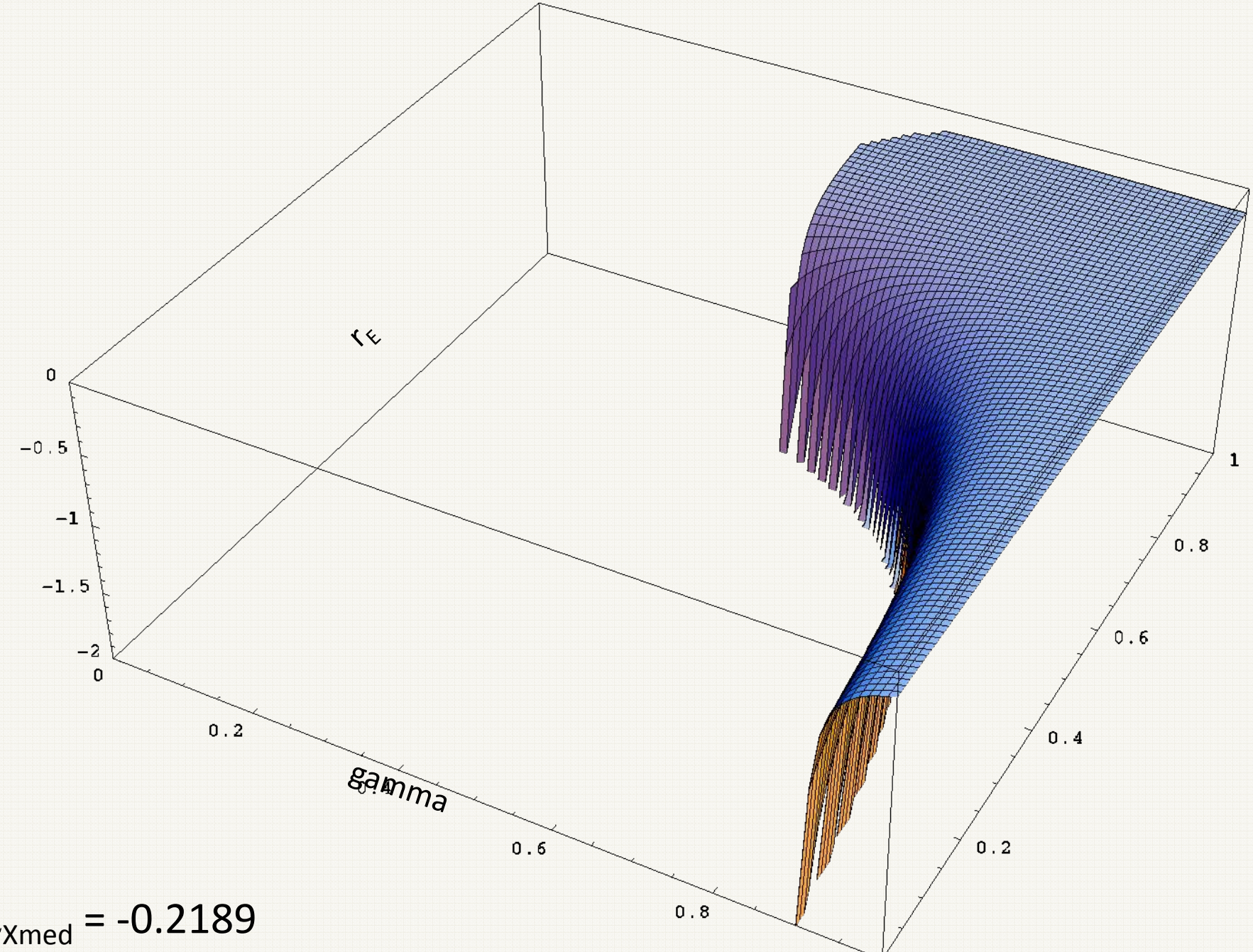
random effects

[here be algebra…]

$$\alpha_X = \frac{r_{RAXX} - r_{RAYX} + (r_{RAYY} - r_{RAXY})\langle T_{X,i} T_{Y,i} \rangle}{\gamma(1 - \langle T_{X,i} T_{Y,i} \rangle^2)}$$

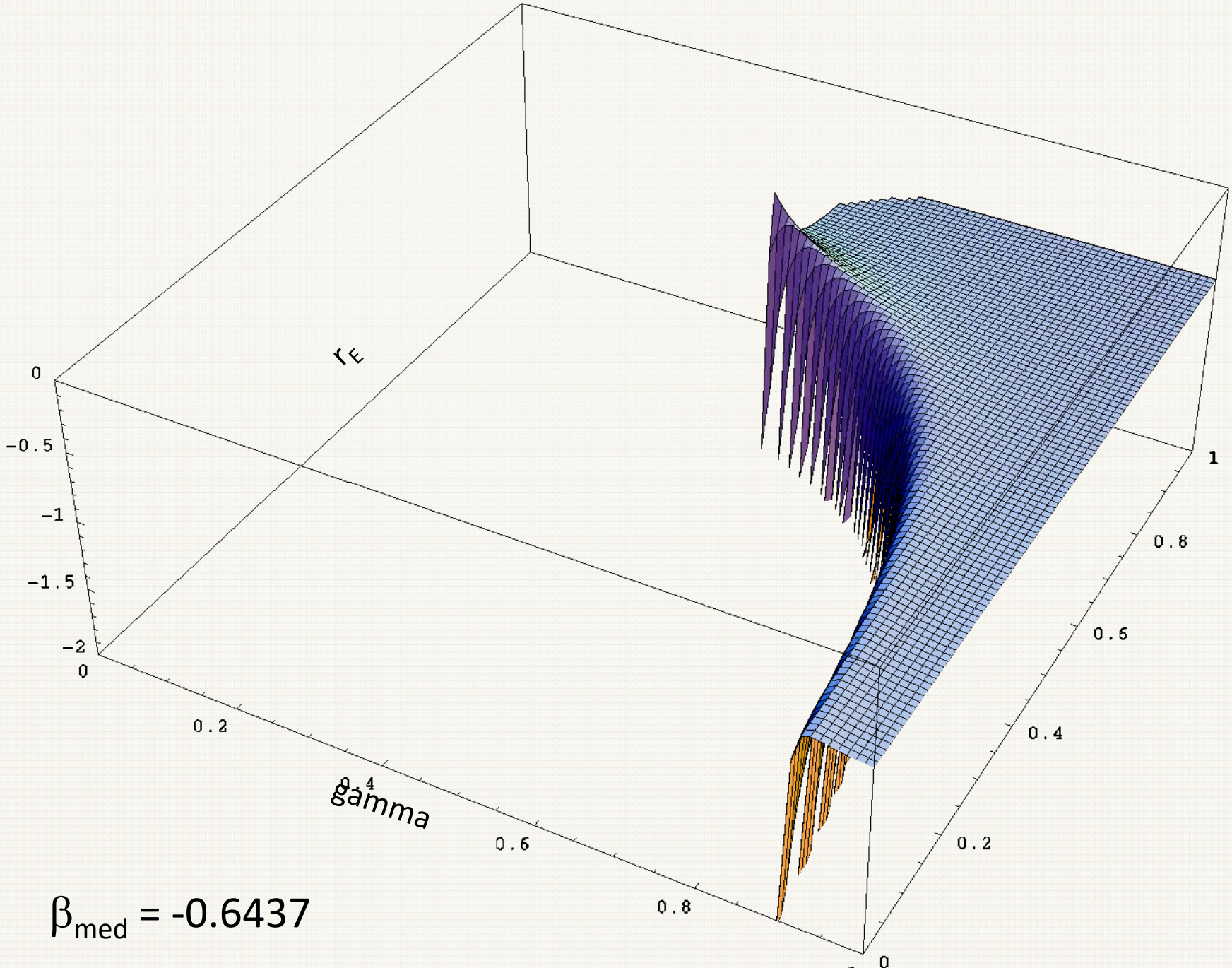
$$\alpha_Y = \frac{r_{RAYY} - r_{RAXY} + (r_{RAXX} - r_{RAYX})\langle T_{X,i} T_{Y,i} \rangle}{\gamma(1 - \langle T_{X,i} T_{Y,i} \rangle^2)}$$

α_{X-} - translation rate contribution - dependence on model parameters



α_{Xmed} = -0.2189

β – intrinsic constraint contribution – dependence on model parameters



β_{med} = -0.6437

Model Solution:

$$\beta/\alpha = 3..5$$

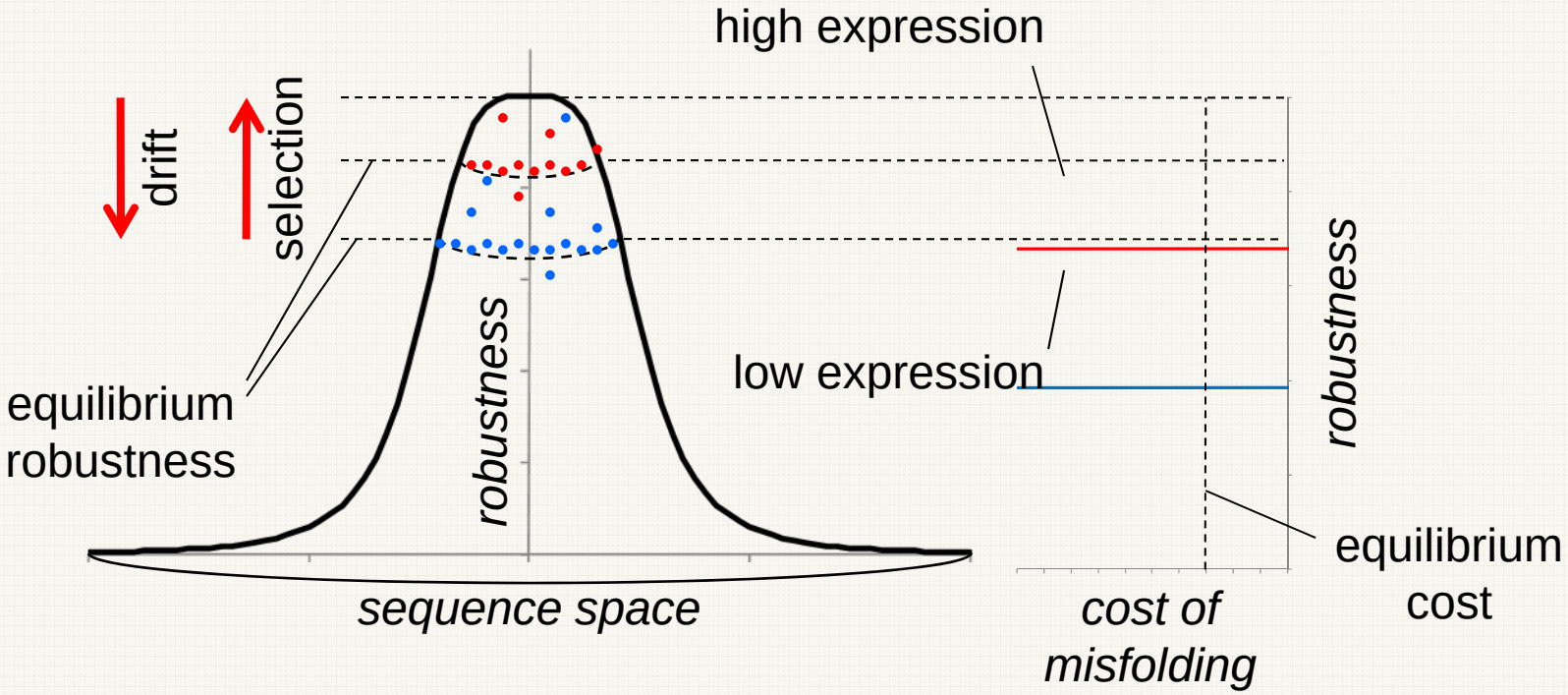
- The contribution of intrinsic ***structural-functional constraints (that determine robustness to misfolding)*** to the protein evolution rate is several-fold greater than the contribution of ***translation rate***

Implications for robustnes/fitness landscape:

Size of neutral network is

- **inversely proportional to misfolding robustness of native sequence**
- **directly proportional to the rate of evolution**
- **narrow, steep peaks for highly robust, strongly expressed proteins**

-slow evolution



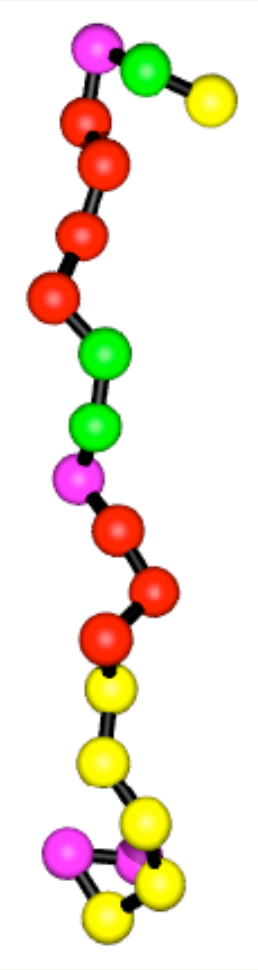
Approximating protein evolution with a simple model of folding

sequence → misfolding probability → fitness → evolution rate

- "Proteins:" hetero-polymers identified by their sequences
- Construct a simple kinetic model of "protein" folding
- Fold a particular "protein" a large number of times (low T quench)
- "Native" folded conformation: most often reached
- Compute the correct folding probability (CFP) as the fraction of folding events that reached the "native" conformation
- Fitness = - # of misfolded copies produced to reach a required abundance
- For each distinct native structure compute **the fold network**: all sequences connected by point substitutions which have a substantial probability to fold to this particular structure
- For each network compute the fixation probability for all transitions between member sequences
- **Evolution: Markov process on the fold network**

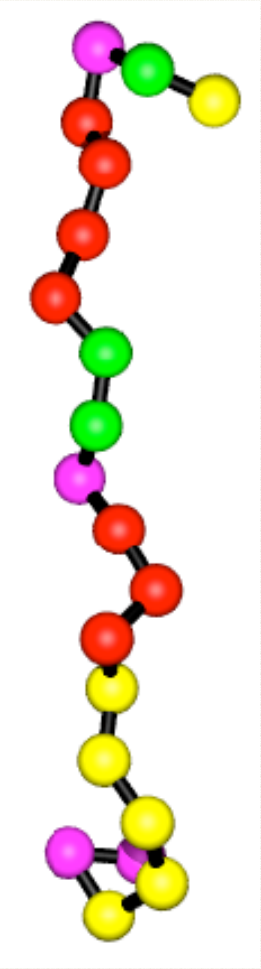
Evolution Rate: **Time averaged transition rate for the Markov process on the fold network**

Model “protein:” flexible chain with 4 monomer types



- Monomers H, P, +, - represent persistence length runs of amino-acids
- Distance between nearest neighbors is fixed bond angles are unrestricted
- Pairwise interactions: Lennard-Jones, screened Coulomb, parameters depend on the pair in question
- Brownian dynamics: over-damped, non-hydrodynamic, implicit solvent

Correct Folding Probability (CFP)



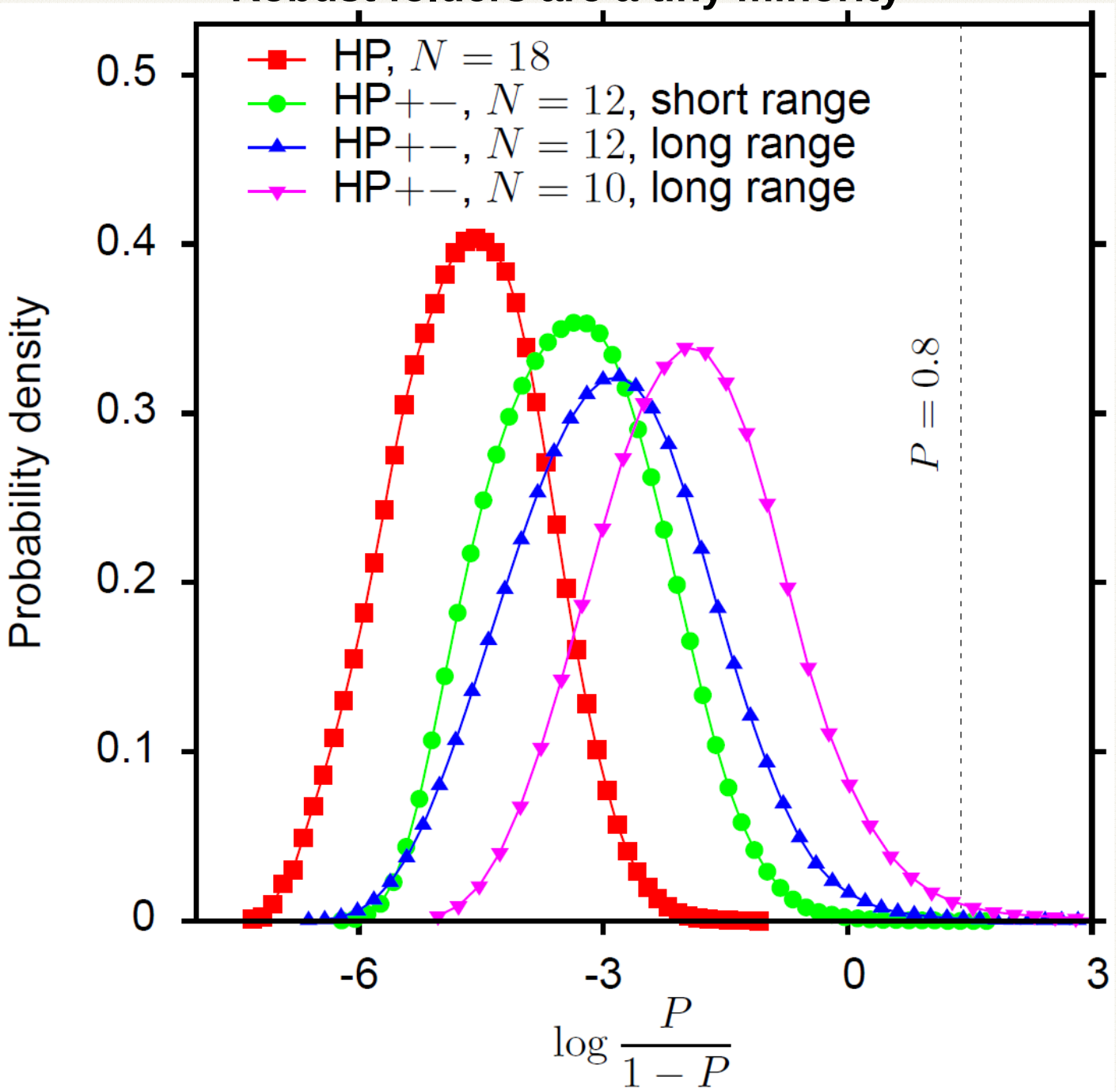
- Equilibrate at $T > T_{fold}$
- Rapid quench to $T < T_{fold}$
- Find the exact energy minimum conformation
- Repeat a large number of times and compare folded structures via 3D alignment
- Identify “native conformation” as one reached most often
- CFP: fraction of folding experiments that reached the “native conformation”

Fitness cost for misfolded proteins

- Compute the total probability of misfolding P_m *;* include correctly translated and mistranslated sequences with appropriate probabilities
- To reach an abundance level A of correctly folded proteins, the cell must make

  $N_m = A\, P_m/(1 - P_m)$ misfolded copies
- Postulate: Fitness = $-N_m$
- Two parameters: required abundance *A*, per-monomer mistranslation probability *r*

Robust folders are a tiny minority



Evolution:
Markov process on the fold network

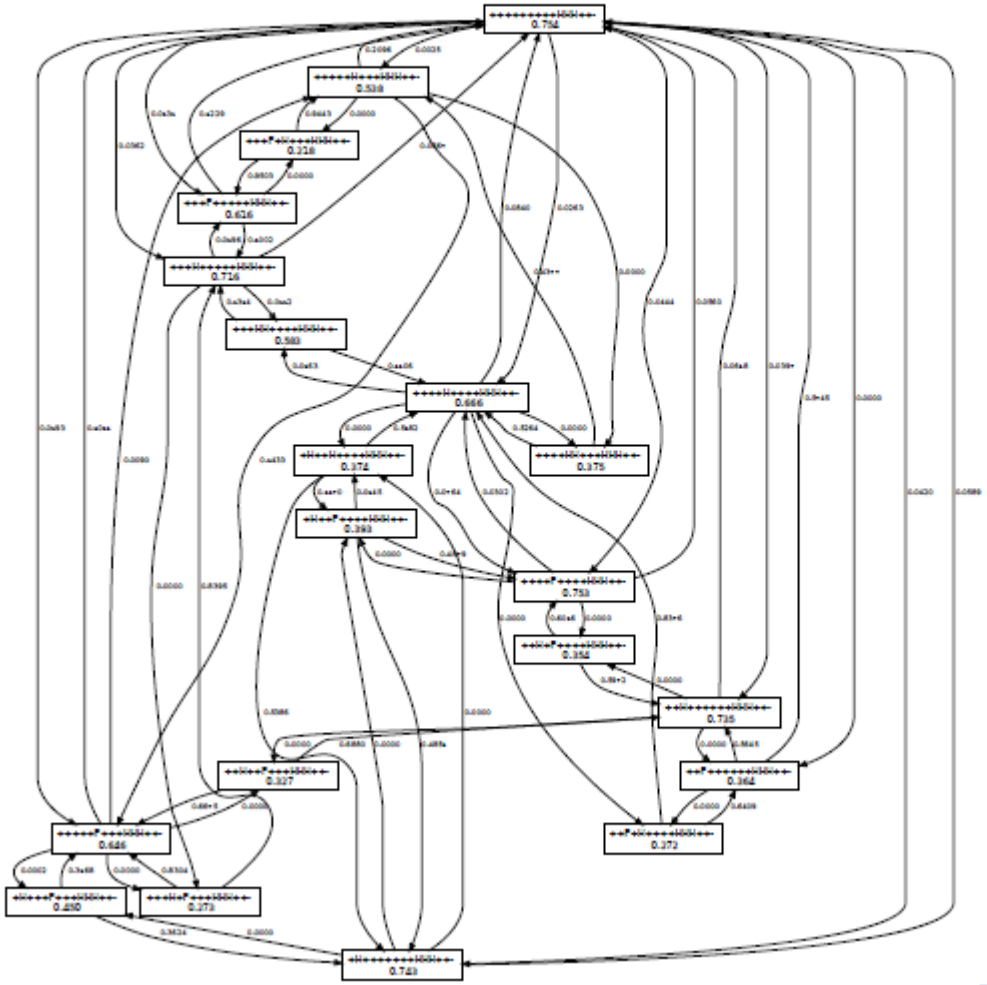
- Population dynamics (Kimura) + fitness concept ➔ fixation probabilities for transitions between the members of the fold network
- Parameters: fitness ($-N_m$) gaps (s), effective population size N_e

$$Pfix = 1/2Ne \text{ for } s=0$$

$$Pfix = 1-2^{-2s}/1-e^{-4Nes} \text{ for } s\neq 0$$

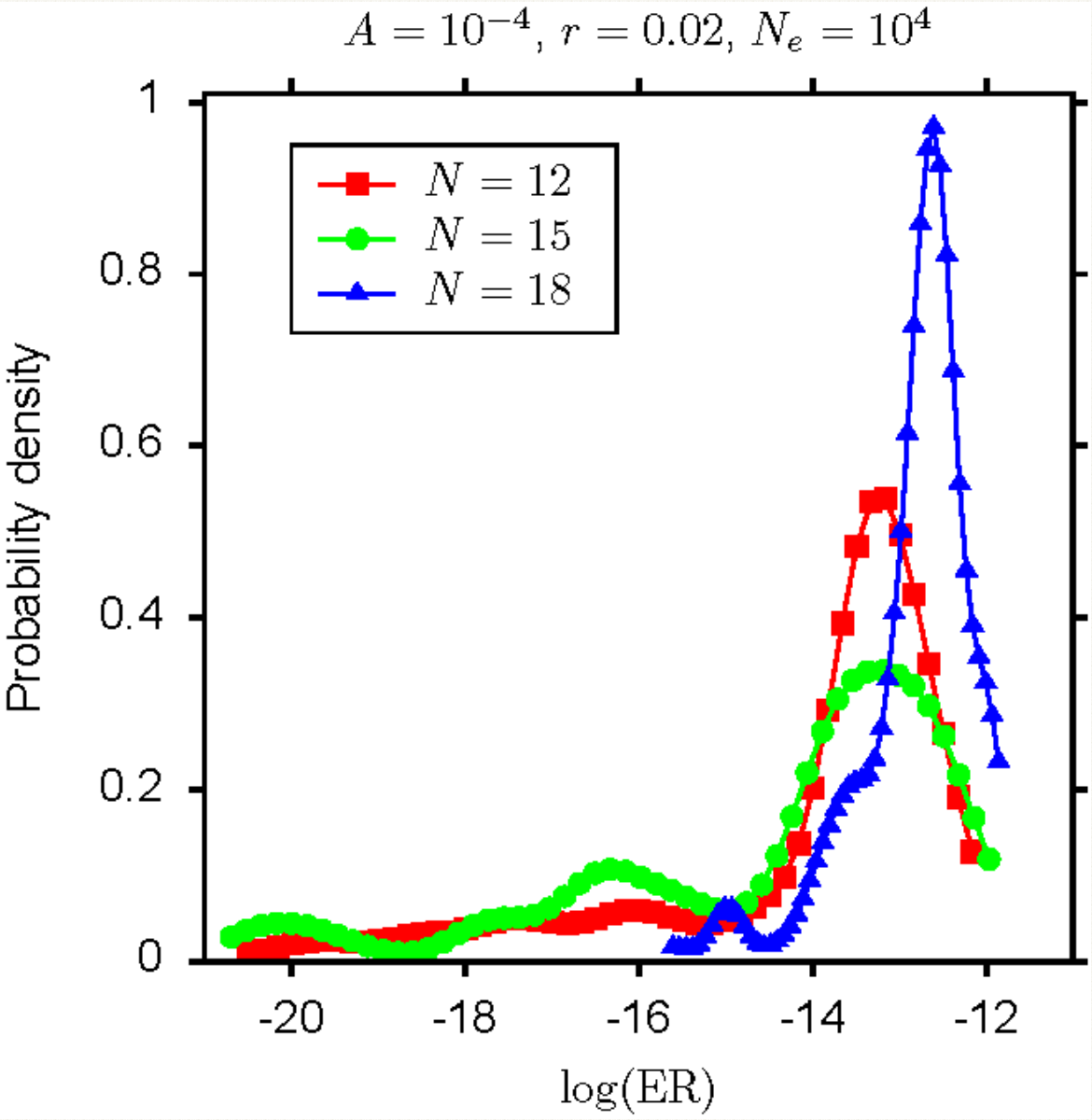
- Instantaneous evolution rate: fixation probability of a random mutant – broad variability among sequences
- Fold-averaged evolution rate: time average of the transition rate of the Markov process on the fold network

Fixation probability of mutants

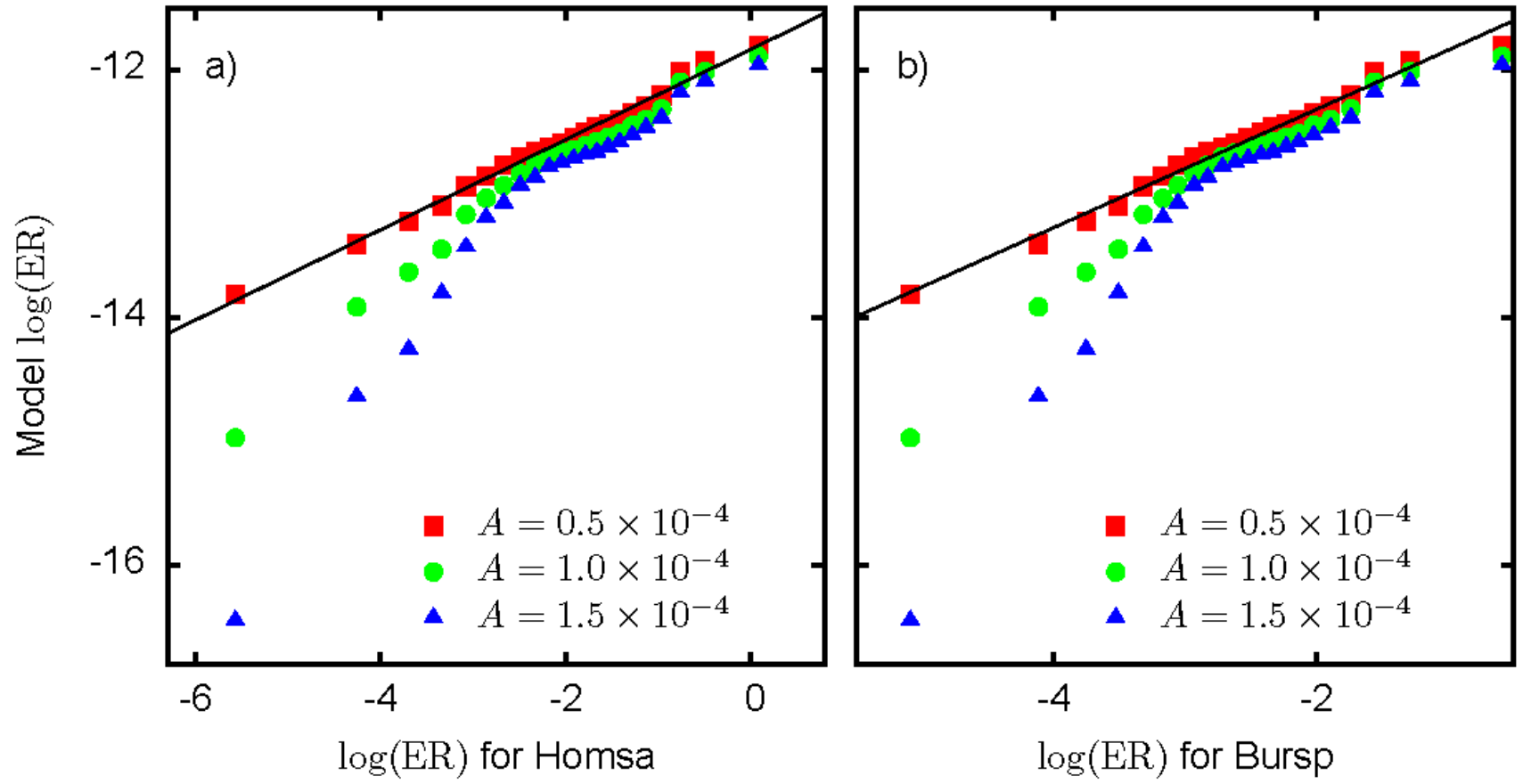


Markov process on directed graph

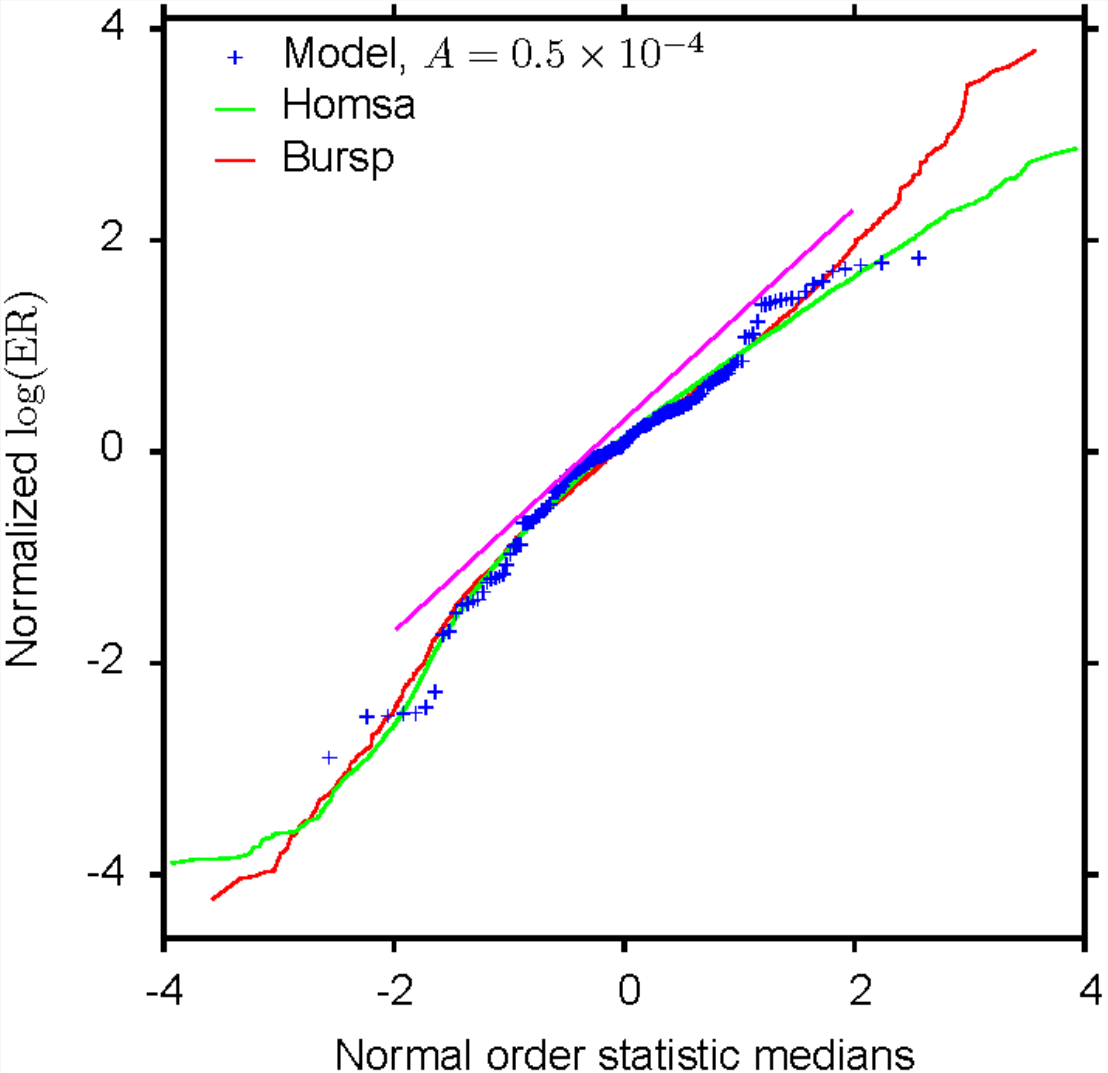
Evolution rate distribution: log-normal with a heavy tail at low rates



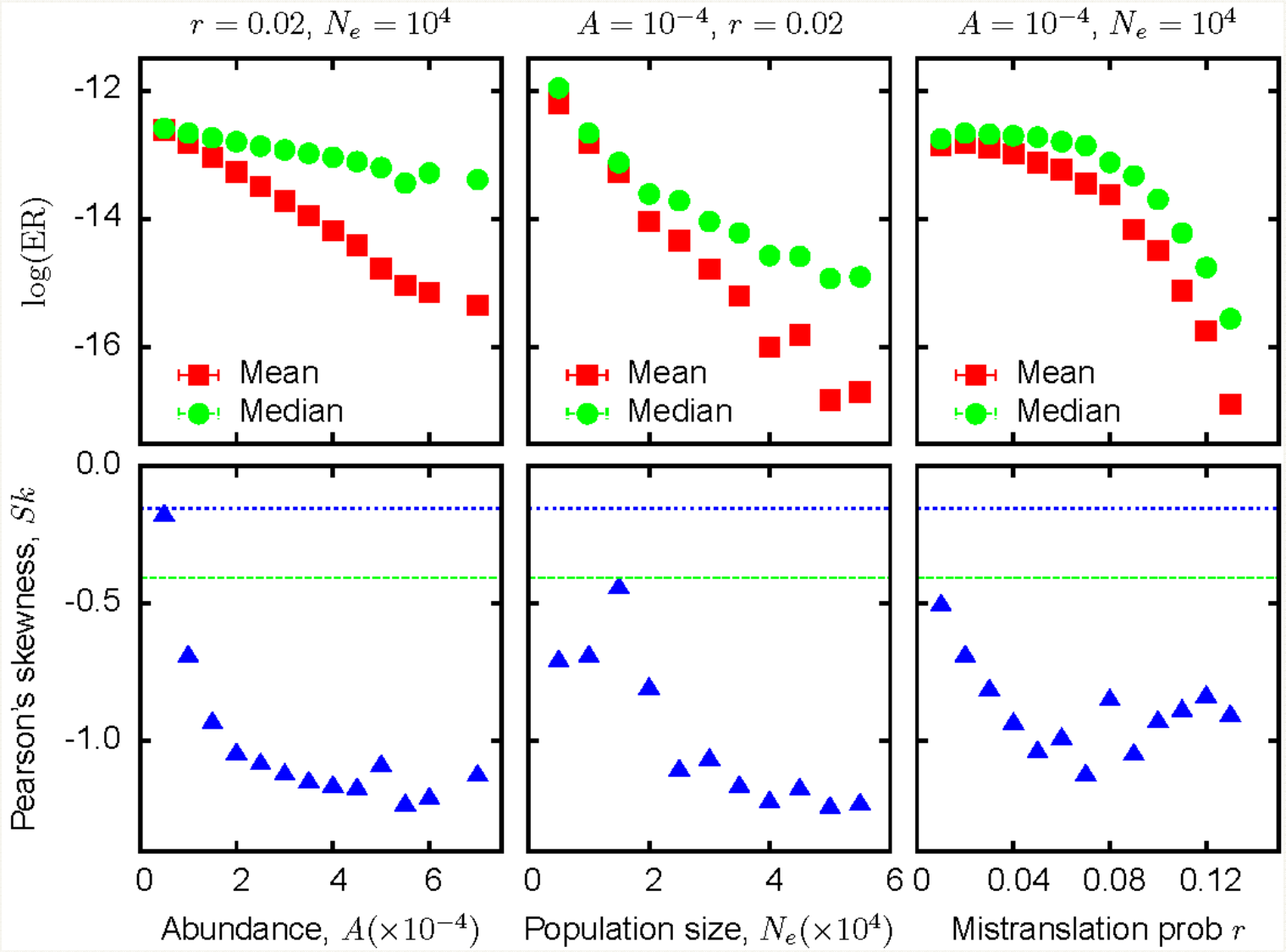
Evolution rate distributions given by the model closely fit the empirical distribution



Observed and model-derived distribution similarly deviate from the Gaussian



The model reproduces the other dependencies of evolution rate



Sources of misfolding probability P_{mf}: correctly translated vs. mistranslated

- Sequence length N
- Correct folding probability p of the correctly translated protein
- Probability r for a mistranslated protein to fold correctly
- Mistranslation probability p_{mt} per monomer

Estimate p_{mt}^0 for which the two sources of mistranslation probability are comparable

$$\boxed{p_{\mathrm{mt}}^0 = 1 - \left(\frac{1-r}{2-r-p}\right)^{1/N}}$$

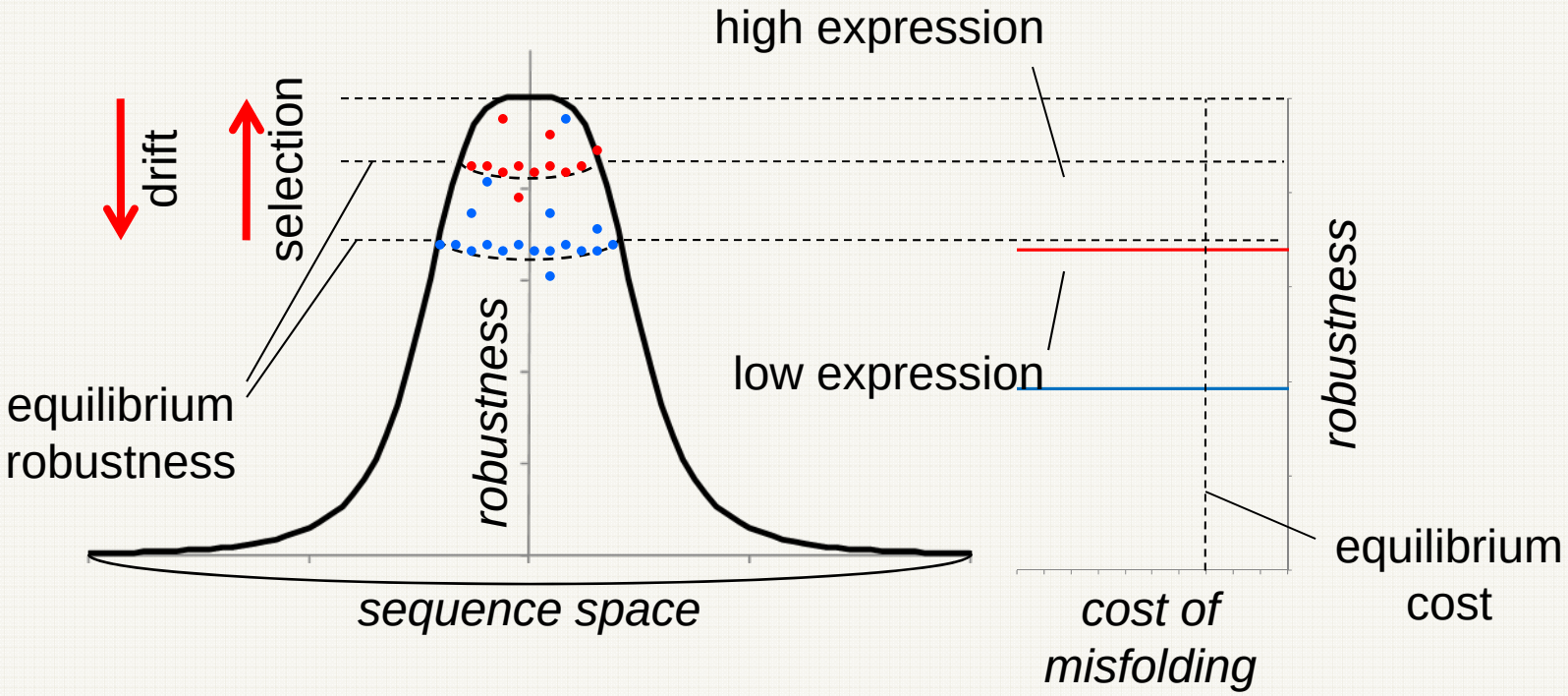
Example: $p = 0.95,\ r = 0.5,\ N = 200 \quad \Rightarrow \quad p_{\mathrm{mt}}^0 = 0.00047$

Measured p_{mt} ~10^{-4}-10^{-5}

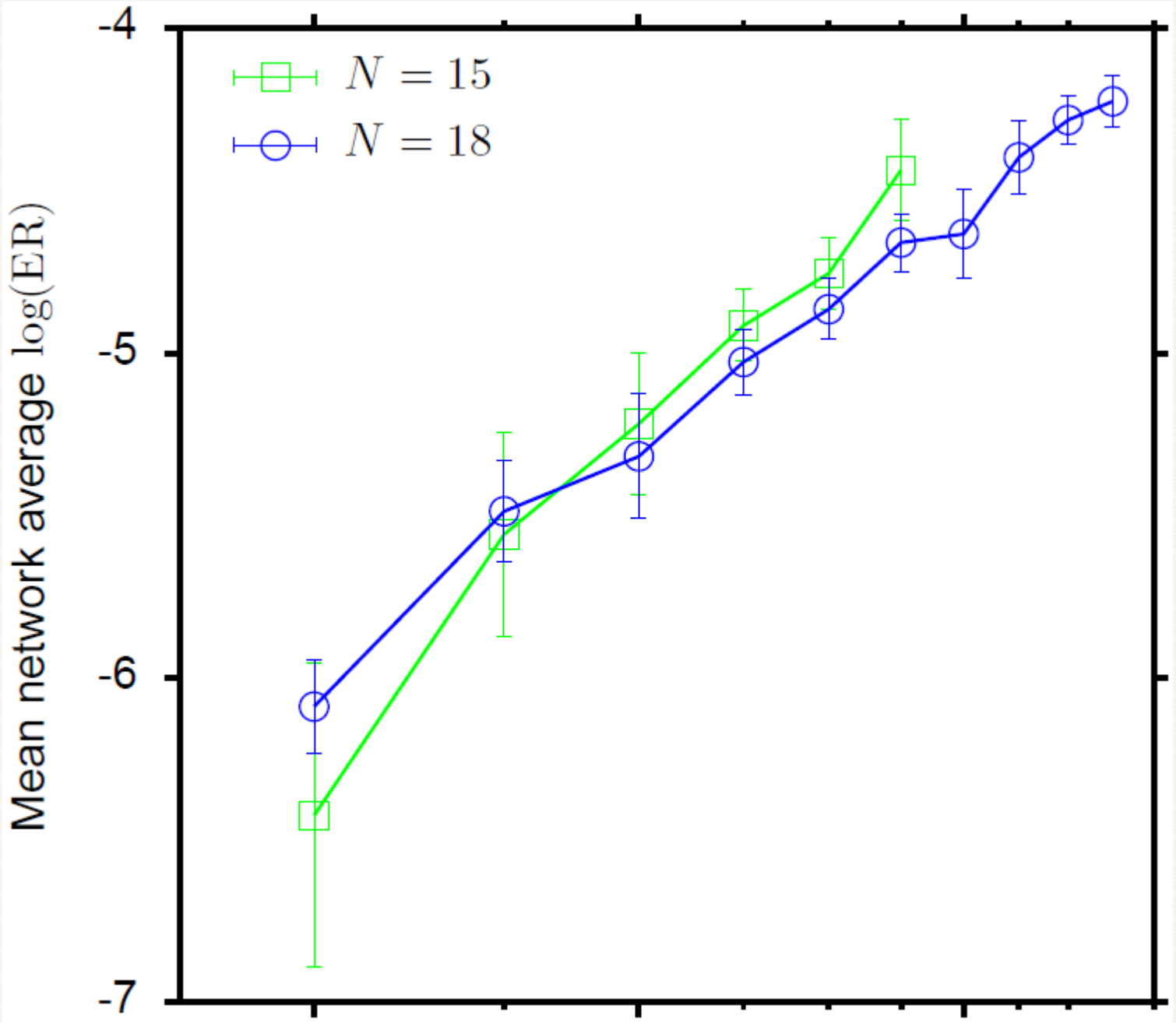
Misfolding of the native sequence could be more important than mistranslation

Implications for robustnes/fitness landscape:

- **rate of evolution is proportional to the size of neutral network**
- **narrow, steep peaks for highly robust, strongly expressed**
- **slow evolving proteins**



In the model framework, ER is indeed proportional to network diameter



A simple model of protein folding combined with the MIM hypothesis quantitatively reproduces

- **universal distribution of evolution rates**
- **dependence of evolution rate on expression**
- **dependence of evolution rate on *Ne***

General hypothesis of misfolding-driven evolution of proteins

•principal determinant of evolution rate – ***neutral network size***

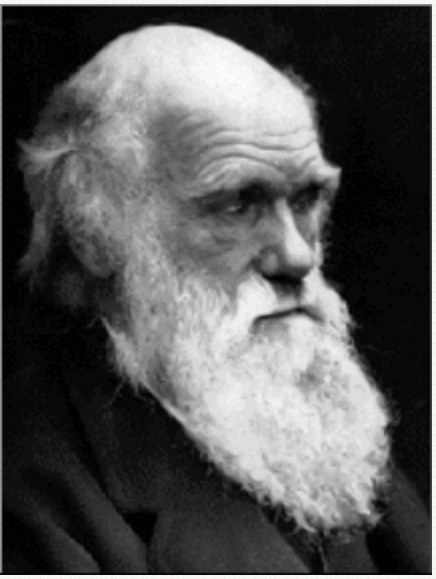
•***neutral network size*** is inversely proportional to ***misfolding robustness of native sequence***

•***misfolding robustness*** is determined by ***intrinsic structural-functional constraints***

-critical for highly expressed proteins

-amplified by ***mistranslation***

On the nature of evolution



"…as natural selection works solely by and for the good of each being, all corporeal and mental endowments will tend to progress towards perfection."
Charles Darwin, 1859, *Origin of Species*, ch. 14

"…natural selection does not work as an engineer works.
It works like a tinkerer – a tinkerer who does not know exactly what he is going to produce but uses whatever he finds around him whether it be pieces of string, fragments of wood, or old cardboards…"

Francois Jacob. *Evolution and tinkering*.
Science. 1977 Jun 10;196(4295):1161-6

Conjectures and refutations

•**Comparative genomics and systems biology reveal several surprising universals of genome evolution such as the distribution of evolution rates across genes, the distribution of paralogous gene family size, connections between expression and evolution rate…and more**

•The existence of these universals calls for simple, general models of evolutionary processes akin to those used in statistical physics (eg, birth and death processes), and at least in some cases, such models seem to explain the observed universal patterns

•These general evolutionary models do not explicitly include selection suggesting that the basic processes underlying evolution are non-adaptive

•Where evidence of selection is seen, the subject of selection is often not a specific function of gene/protein but rather robustness to malfunction (MIM hypothesis) or, more generally, maintenance of the basic organization of genome/cell

Koonin, *Darwinian evolution in the light of genomics*. NAR 2009

Conjectures and refutations

- The quest for simple, law-like regularities underlying evolution might not be futile
- The applicability of simple models to the basic processes of evolution does not contradict Jacob's tinkering metaphor: the specific outcomes of these regular processes are all contingency, and result of (adaptive) tinkering
- In general terms, evolutionary biology is not that different from physics/astrophysics/cosmology: the basic processes obey law-like patterns but the specific outcomes are determined by chance + adaptation, and are unpredictable – within the applicable constraints
- Adaptation "takes advantage" of the inherent stochasticity of the evolutionary process

So what about a Post-Modern Synthesis of Evolutionary Biology in the light of Evolutionary Genomics and Systems Biology?

The old order (Modern Synthesis) is no longer…

- Pan-selectionism gone
- Gradualism gone
- Tree of Life gone – giving way to Forest of Life

However, a new order could be in sight

- **coming from the results of evolutionary genomics/systems biology**
- **resembling physics more than stamp collection**

Acknowledgments

- Yuri I. Wolf
- David J. Lipman
- Irina Gopich (NIDDK)
- Alexander Lobkovsky

Previous relevant work:

Nick Grishin (currently UT Southwestern-Dallas)
Liran Carmel (currently Jerusalem University)
I. King Jordan (currently Georgia Tech)
Georgy Karev
Pavel Novichkov (currently LBL)
Igor Rogozin
Dmitry Krylov
Roman Tatusov

Essential discussion:

Allan Drummond (Harvard)
Claus Wilke (UT-Austin)
Sergei Maslov (Brookhaven)
Scott Roy
Josh Cherry
W. Ford Doolittle (Dalhousie)
Michael Lynch (U. Indiana)
Bill Martin (U. Duesseldorf)
Valerian Dolja (Oregon State U)
Tania Senkevich (NIAID)

Indispensable data:

Manuel Weiss (U. Zurich)
Sabine Schrimpf (U. Zurich)
Cristian von Mering (U. Zurich)