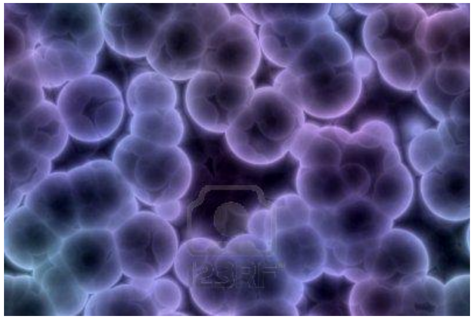
Lecture 3:
Bridging scales: From Biophysics to Populations

1. **Evolution at multiple scales**

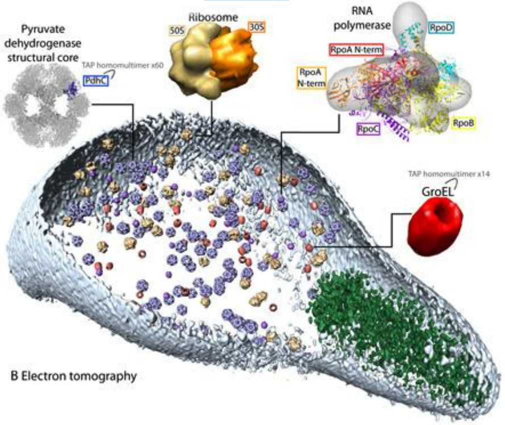
2. **Emerging “universals” from biophysics and genomics**

1. A unified framework of molecular evolution

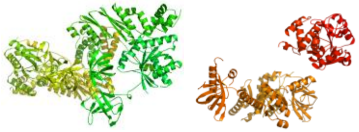
Length and Time Scales in Evolution



POPULATION



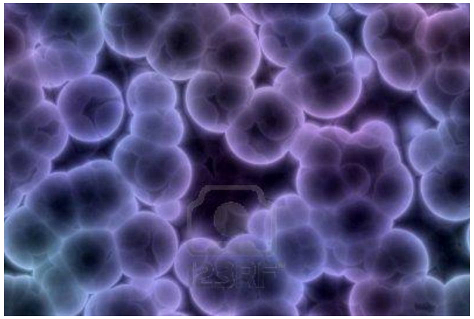
ORGANISM



BIOMOLECULES

ATTGCCATAACGGATGTAAATTGCCATAACGGGCTAA

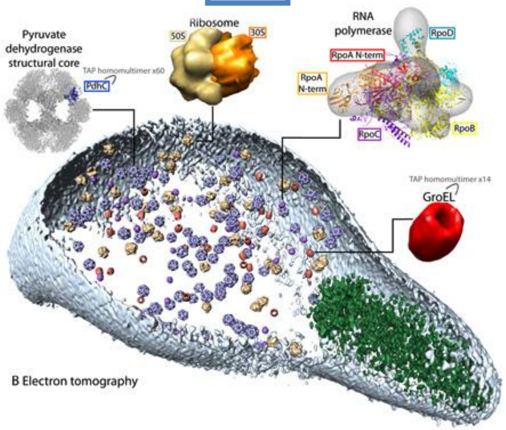
Length and Time Scales in Evolution



POPULATION

Population Genetic/ Ecological variable

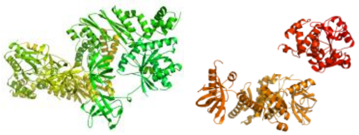
N Population size



ORGANISM

Genomic variables

C protein copy numbers
dN, dS molecular clock rates
PPI protein-protein interaction
μ mutation rate



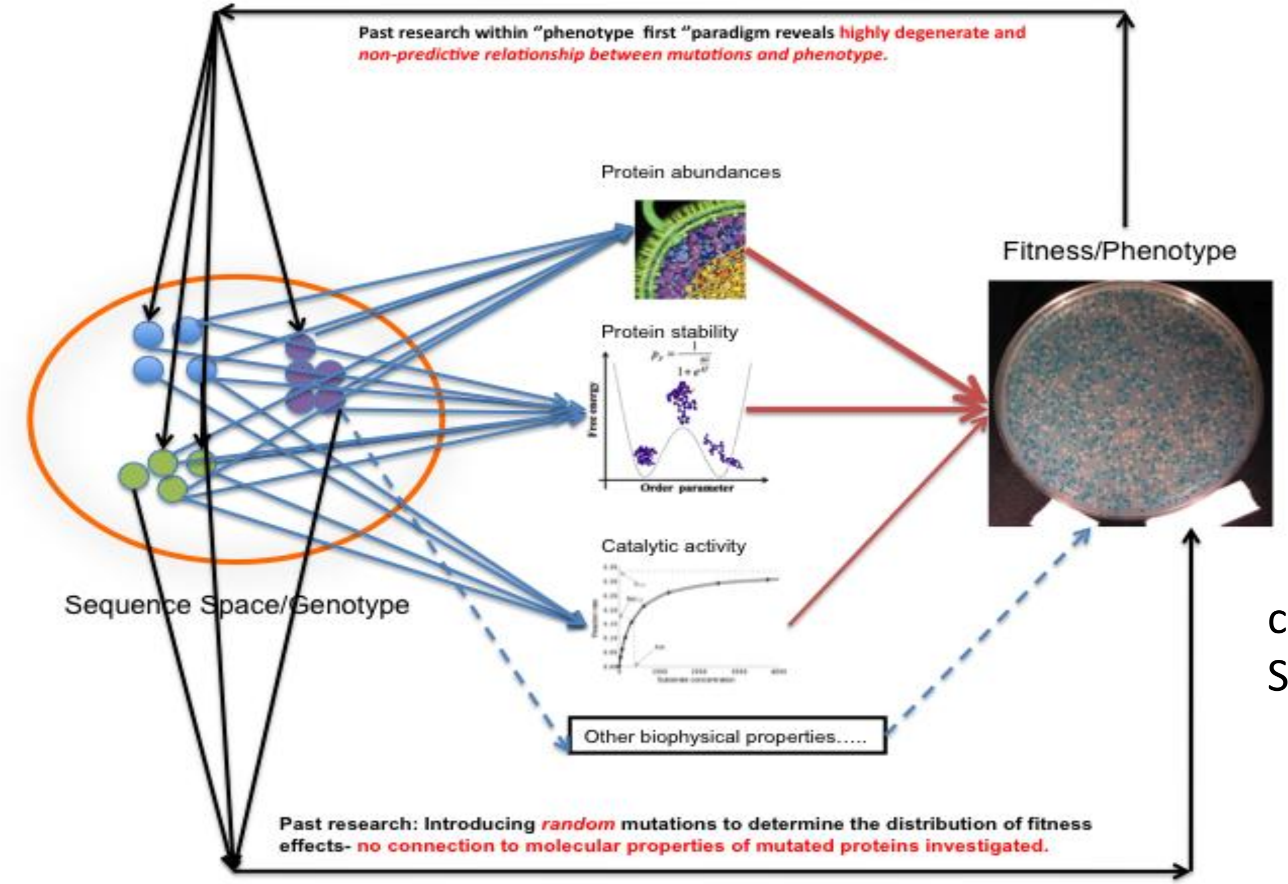
ATTGCCATAACGGATGTAAATTGCCATAACGGGCTAA

BIOMOLECULES

Molecular variables

ΔG Stability
SAS Surface accessible areas
D Fold designability
AA Amino acid composition
ΔG_{pd} Protein-DNA interaction

Hard to Bridge Scales, or Ruggedness of Fitness Landscape..



Past research within "phenotype first "paradigm reveals highly degenerate and *non-predictive relationship between mutations and phenotype.*

Past research: Introducing *random* mutations to determine the distribution of fitness effects- no connection to molecular properties of mutated proteins investigated.

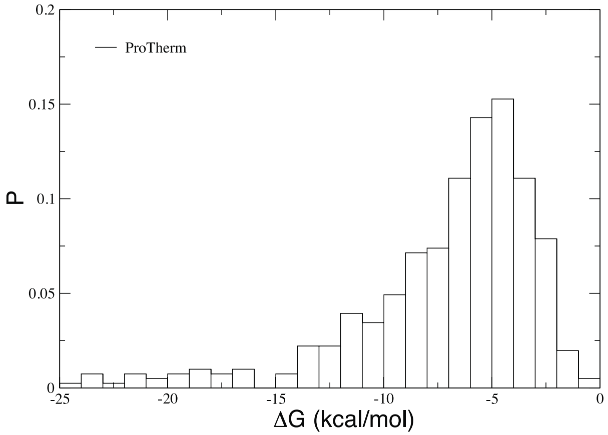
cf: Order parameters in StatMech

Protein Folding Stability

Distribution

Range

Various organisms



-20 – 0 kcal/mol

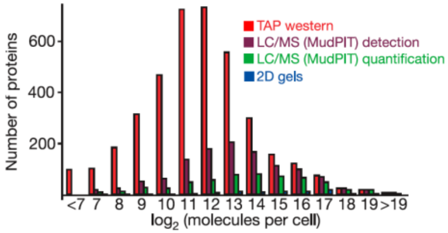
(PROTHERM Database, 2010)

Protein Abundance in the Cell

Distribution

Range

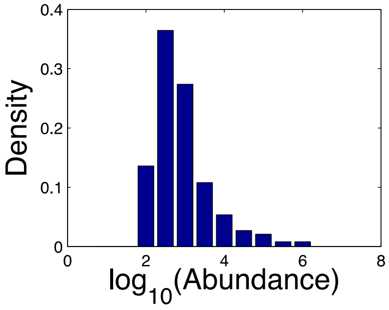
Yeast



~50 - 10^6

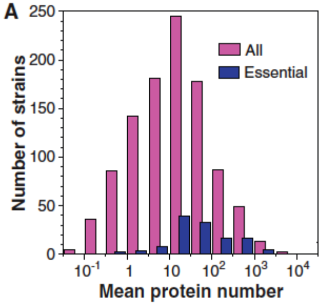
(Ghemmagami S, et al., *Nature* 2003)

E. coli



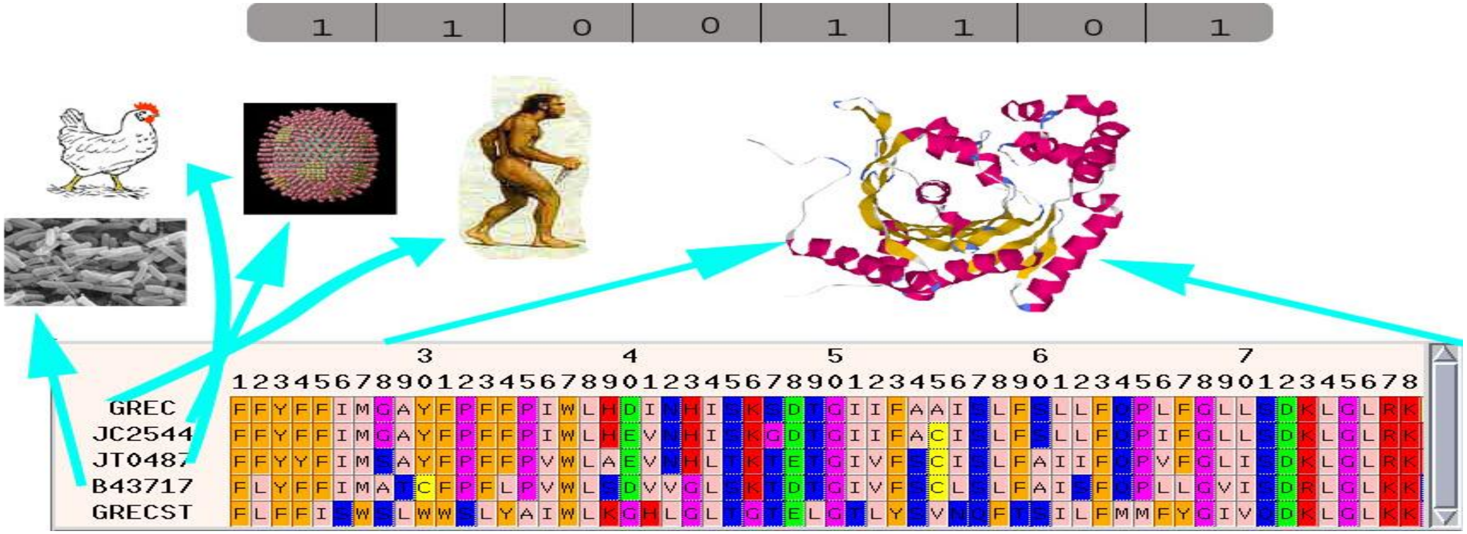
~65 - 10^9

(Ishihama/Frishman, *BMC Genomics* 2008)



(Taniguchi/Xie, *Science* 2010)

Evolutionary rates: Universal Observables in Biology



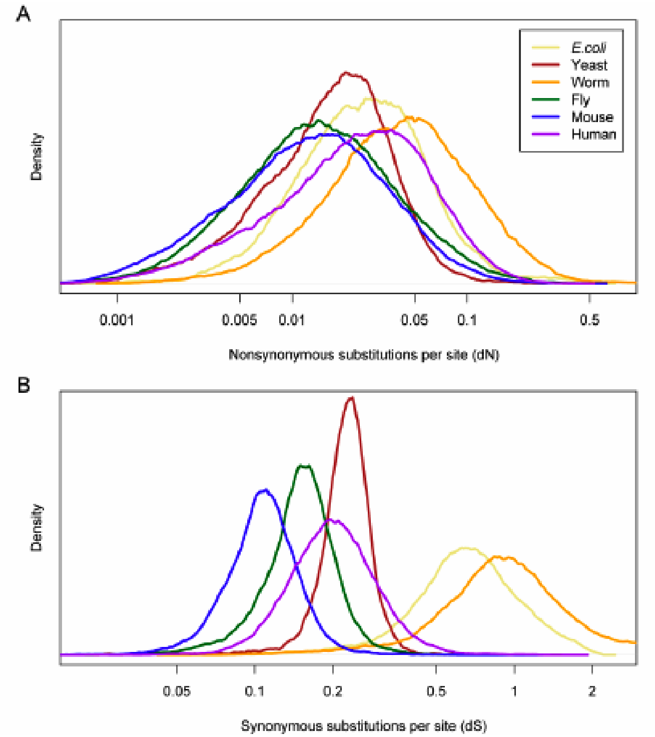
```
                  3         4         5         6         7
        1234567890123456789012345678901234567890123456789012345678
GREC    FFYFFIMGAYFPFFPIWLHDINHISKSDTGIIFAAISLFSLLFQPLFGLLSDKLGLRK
JC2544  FFYFFIMGAYFPFFPIWLHEVNHISKGDTGIIFACISLFSLLFQPIFGLLSDKLGLRK
JT0487  FFYYFIMSAYFPFFPVWLAEVNHLTKTETGIVFSCISLFAIIFQPVFGLISDKLGLRK
B43717  FLYFFIMATCFPFLPVWLSDVVGLSKTDTGIVFSCLSLFAISFQPLLGVISDRLGLKK
GRECST  FLFFISWSLWWSLYAIWLKGHLGLTGTELGTLYSVNQFTSILFMMFYGIVQDKLGLKK
```

The number of differences (''Hamming distance'') between aligned sequences of **orthologous** proteins - number of non-synonymous substitutions Na - is a measure of evolutionary divergence.
Na per unit time (i.e. normalized by number of synonymous substitutions) presents **evolutionary rates.**

Orthologous proteins: Proteins from different species that have same function

Evolutionary Rates:
Some proteins evolve **much** faster than others

Log-normal Distribution-



Range

~2- to 3-orders of magnitude

Pal, et al. *Nat Genetics*, 2003
Drummond DA et al. *Cell* 2008.
Wolf Y, et. al *PNAS* 2009

Overdispersion of molecular clock

Population Size

Distribution

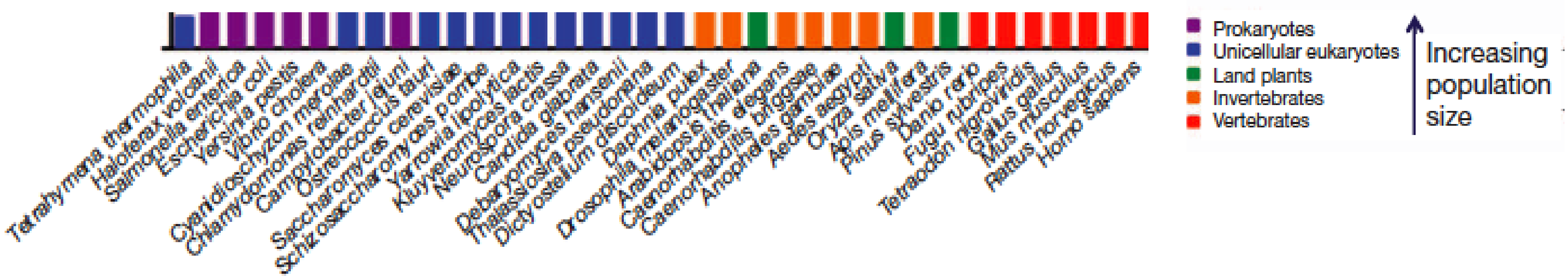
Various organisms

?

Range

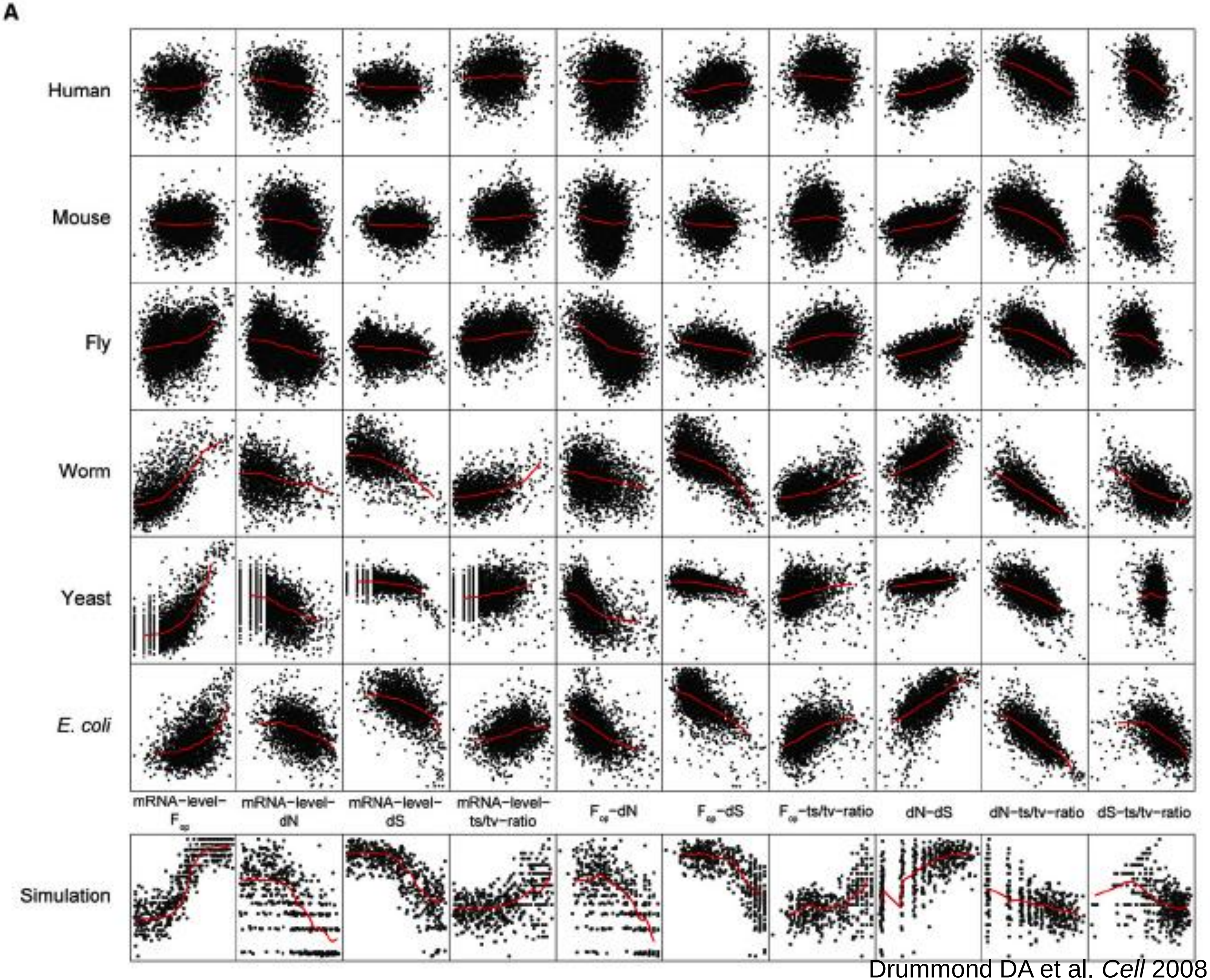
10^8	Prokaryotes
$10^7 – 10^8$	Unicellular Eukaryotes
$\sim 10^5 – 10^6$	Invertebrates
$10^4 – 10^5$	Vertebrates

(Lynch & Connery, *Science* 2003)



(Fernandez & Lynch, *Nature* 2011)

Universal Patterns in Molecular Evolution: Correlation between rates and abundances



Drummond DA et al. *Cell* 2008.

Patterns of molecular evolution

Stability vs. Expression Level

Stability vs. Evolutionary Rate

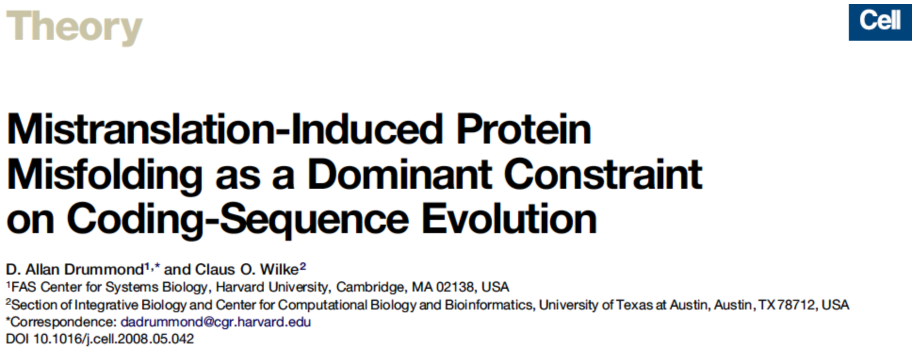
Stability vs. Structure

Expression Level vs. Structure

- **What is the underlying mechanism of evolutionary dynamics?**
- **What is the relative contribution of these variables to evolutionary dynamics?**
- **How do ecological parameters modulate these patterns?**

Emerging constraints in molecular evolution

Selection against misfolding …



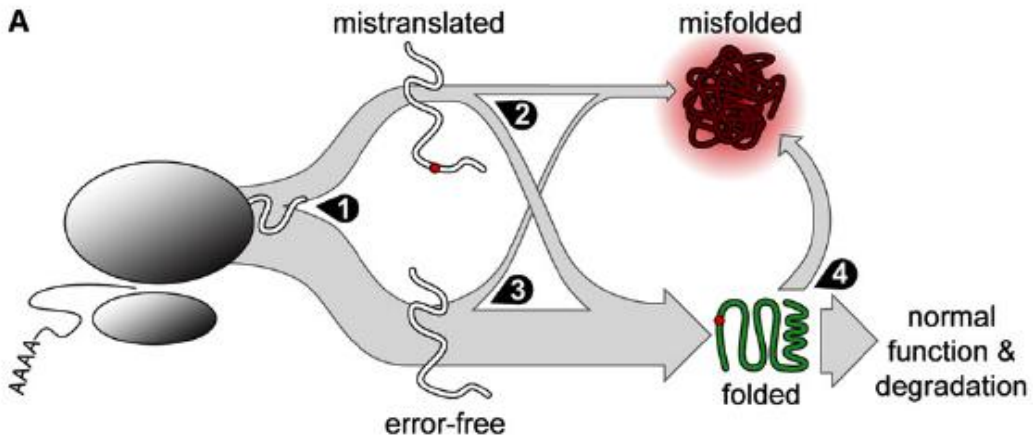
Theory

Cell

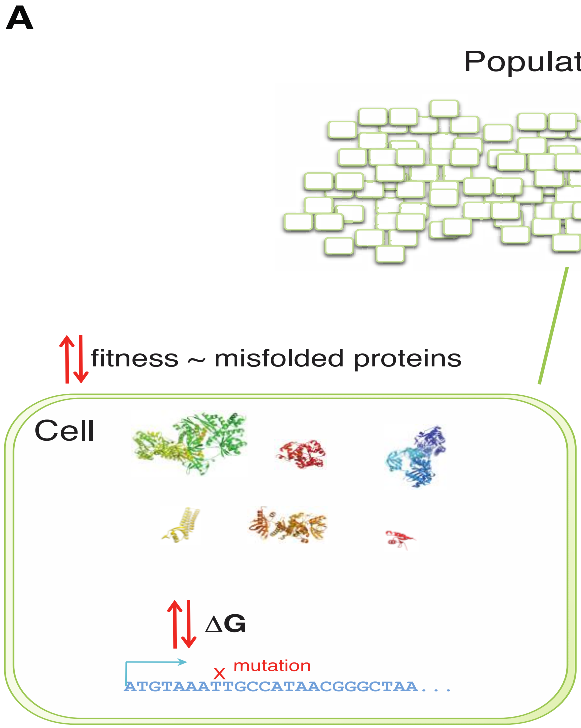
Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution

D. Allan Drummond[1,*] and Claus O. Wilke[2]
[1]FAS Center for Systems Biology, Harvard University, Cambridge, MA 02138, USA
[2]Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas at Austin, Austin, TX 78712, USA
*Correspondence: dadrummond@cgr.harvard.edu

DOI 10.1016/j.cell.2008.05.042



Multiscale Evolutionary Framework



Assumption: Unfolded proteins are toxic.

Fitness ~ 1/(number of unfolded proteins)

Death ~ (number of unfolded proteins)

Misfolded proteins:

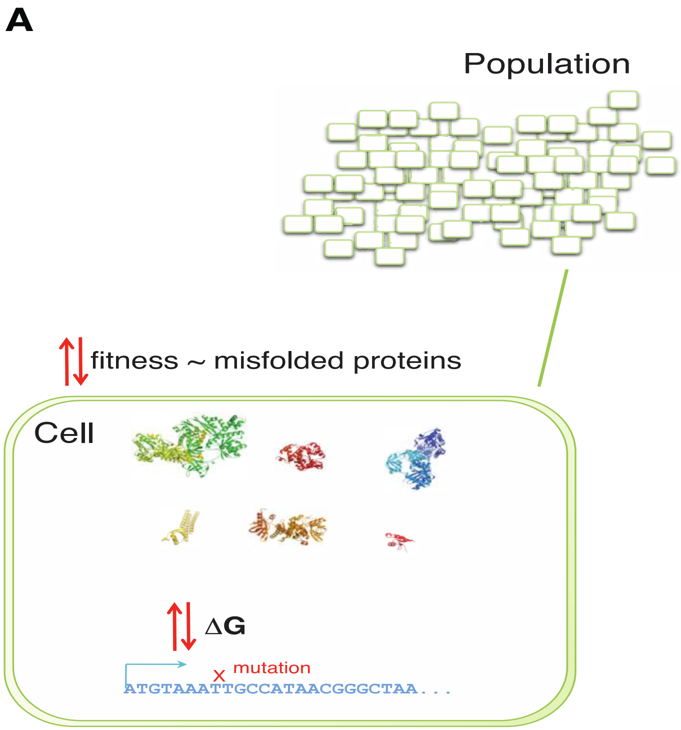
$$d = d_0 \sum_{i=1}^{G} c_i \frac{e^{bDG_i}}{1+e^{bDG_i}},$$

Fitness (Wright-Fisher sense):

$$w_{SS} = \exp(-d) = \exp\left(- d_0 \sum_{i=1}^{G} c_i \frac{e^{bDG_i}}{1+e^{bDG_i}}, \right)$$

(Drummond & Wilke, *Cell* 2008)
(Lobgovsky/Koonin, *PNAS* 2010)

Multiscale Evolutionary Framework



For a given mutation:

$$\Delta G_{k,mut} = \Delta G_{k,wildtype} + \Delta\Delta G$$

Fitness effect:

$$s = \ln w_A - \ln w_B = d_0 c_k \left[\frac{1}{1+e^{-\beta(\Delta G_i + \Delta\Delta G)}} - \frac{1}{1+e^{-\beta \Delta G_i}} \right],$$

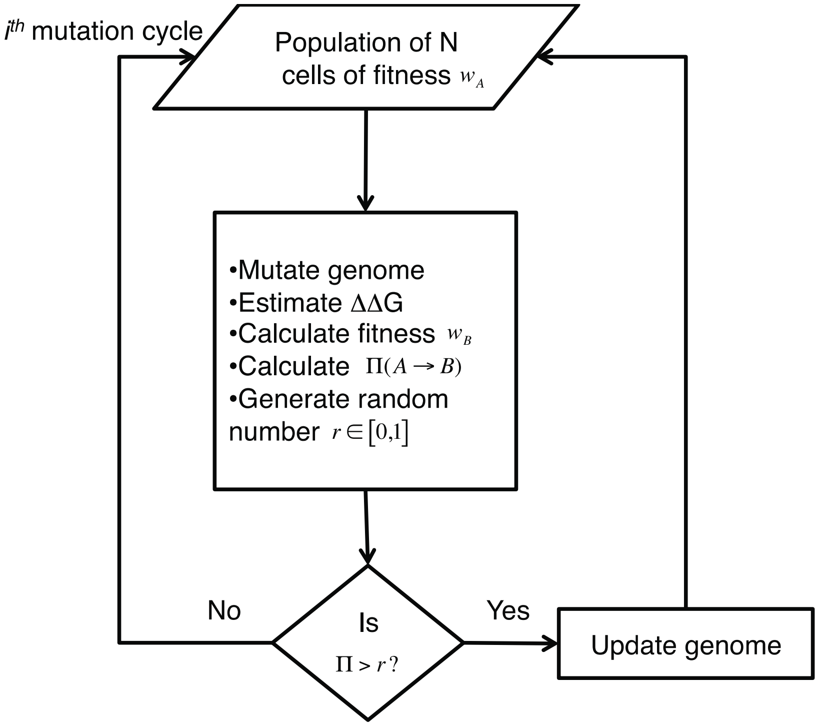
$$s \approx - d_0 c_k e^{\beta \Delta G_k} \left(e^{\beta \Delta\Delta G_k} - 1 \right), \ \ \Delta G_k < -3 \text{ kcal/mol}$$

(Sella & Hirsh, *PNAS* 2005)

Fixation probability (Nμ<<1):

$$\mathrm{P}(A \rightarrow B) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}},$$

Multiscale Evolutionary Framework



For a given mutation:

$$DG_{k,\,mut} = DG_{k,\,wildtype} + DDG$$

Fitness effect:

$$s = \ln w_A - \ln w_B = d_0 c_k \left[\frac{1}{1+e^{-b(DG_i + DDG)}} - \frac{1}{1+e^{-bDG_i}} \right],$$

$$s \approx - d_0 c_k e^{bDG_k}\left(e^{bDDG_k} - 1\right),\ \ DG_k < -3 \text{ kcal/mol}$$

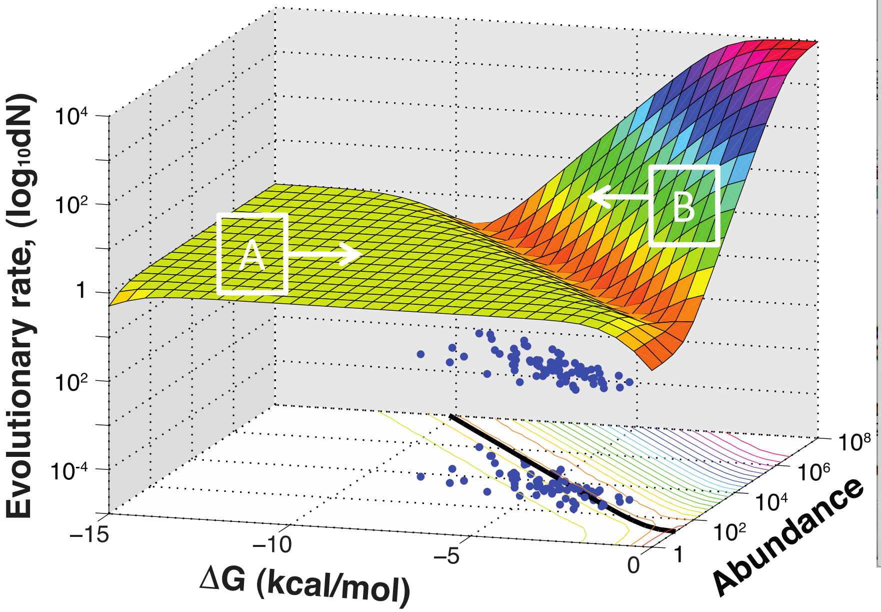
(Sella & Hirsh, *PNAS* 2005)

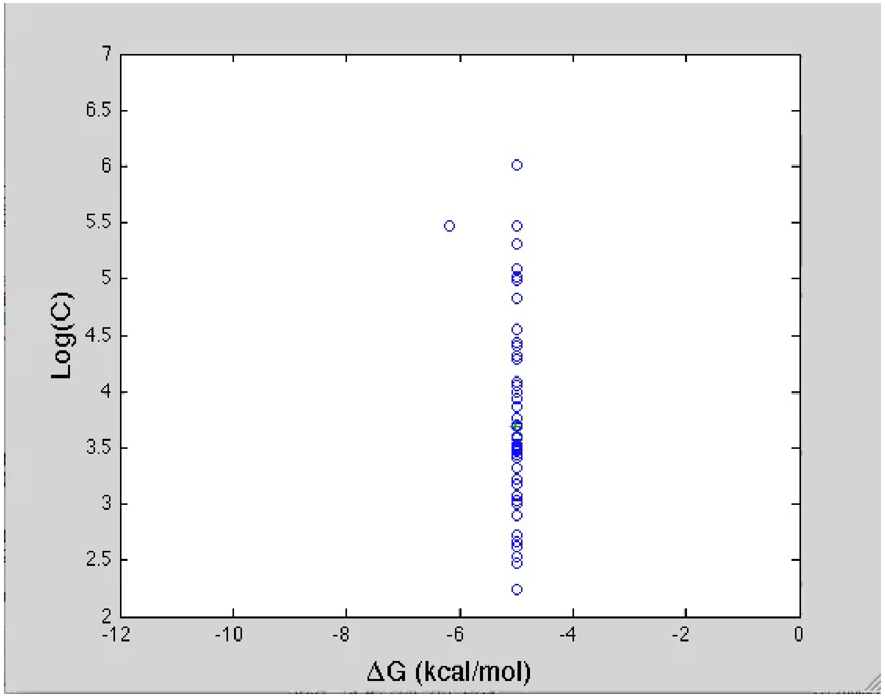
Fixation probability (Nμ<<1):

$$P(A \to B) = \frac{1 - e^{-2s}}{1 - e^{-2Ns}},$$

Average fixation probability: 1/Rate

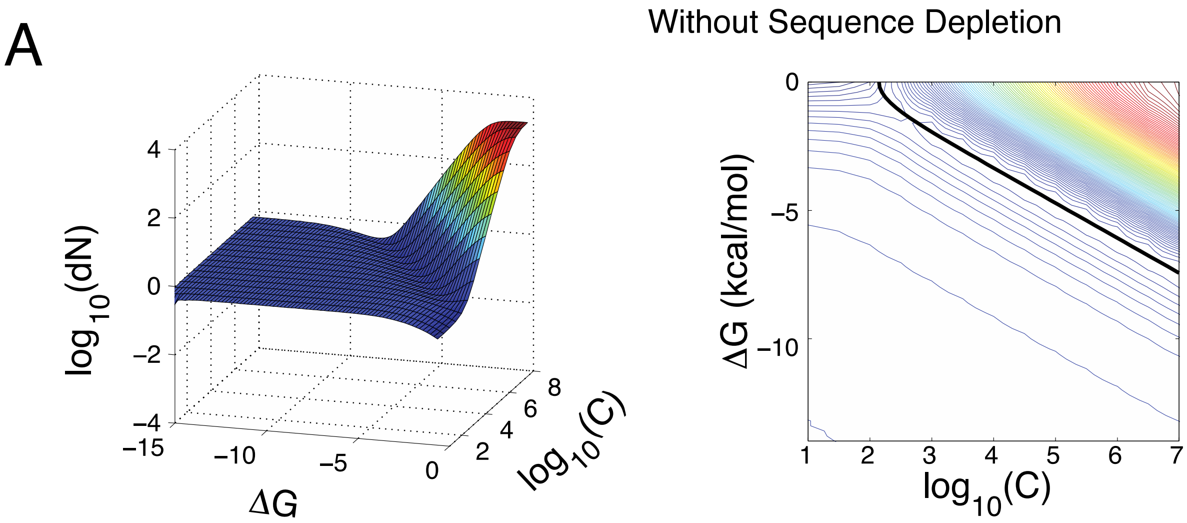
$$\langle \Pi_{i\to f} \rangle = \int_{-\infty}^{-\infty} d(\Delta\Delta G) \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(\Delta\Delta G - \mu)^2}{2\sigma^2} \right) \frac{1-\exp(-2s)}{1-\exp(-2Ns)}$$





Settling in the ‘’gulley’’

Bridging molecular, organismal and population scales: **Highly expressed proteins are more stable and so are proteins in large populations**



Approximation of the minimum:

$$C = \frac{\mu}{\beta\sigma^2}\frac{1}{(N-1)h}\frac{\left(1+e^{-\beta\Delta G}\right)^2}{e^{-\beta\Delta G}}, \quad \beta = 1/k_B T$$

$$\sim \frac{\mu}{\beta\sigma^2}\frac{1}{Nh}e^{-\beta\Delta G}, \quad \Delta G < -3 \text{ kcal/mol}$$

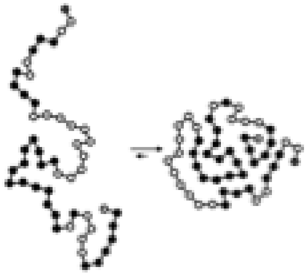
$$\Delta G_k \approx -k_B T\ln N - k_B T\ln C_k - k_B T\ln h - k_B T\ln\left(\frac{1}{k_B T}\frac{\sigma^2}{\mu}\right)$$

Scaling between evolutionary parameters

$$\Delta G_k \approx - k_B T \ln N - k_B T \ln C_k - k_B T \ln h - k_B T \ln\left(\frac{1}{k_B T}\frac{\sigma^2}{\mu}\right)$$

20 kcal/mol	N~10^4-10^8	C~10^1-10^6	h~10^{-6}	σ~1.7 kcal/mol; μ~ 1kcal/mol
	7 kcal/mol	**6 kcal/mol**	**6 kcal/mol**	**1 kcal/mol**

Two-state protein folding paradigm

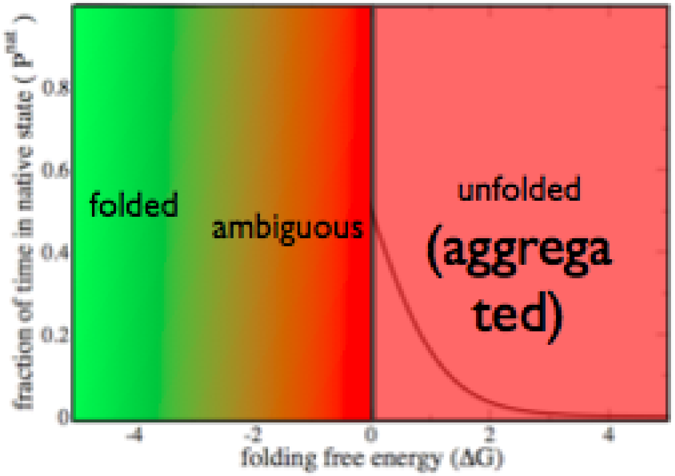


$$P^{nat} = \frac{e^{-G_f/kT}}{e^{-G_f/kT} + e^{-G_u/kT}}$$

$$= \frac{1}{1 + e^{-\Delta G/kT}}$$

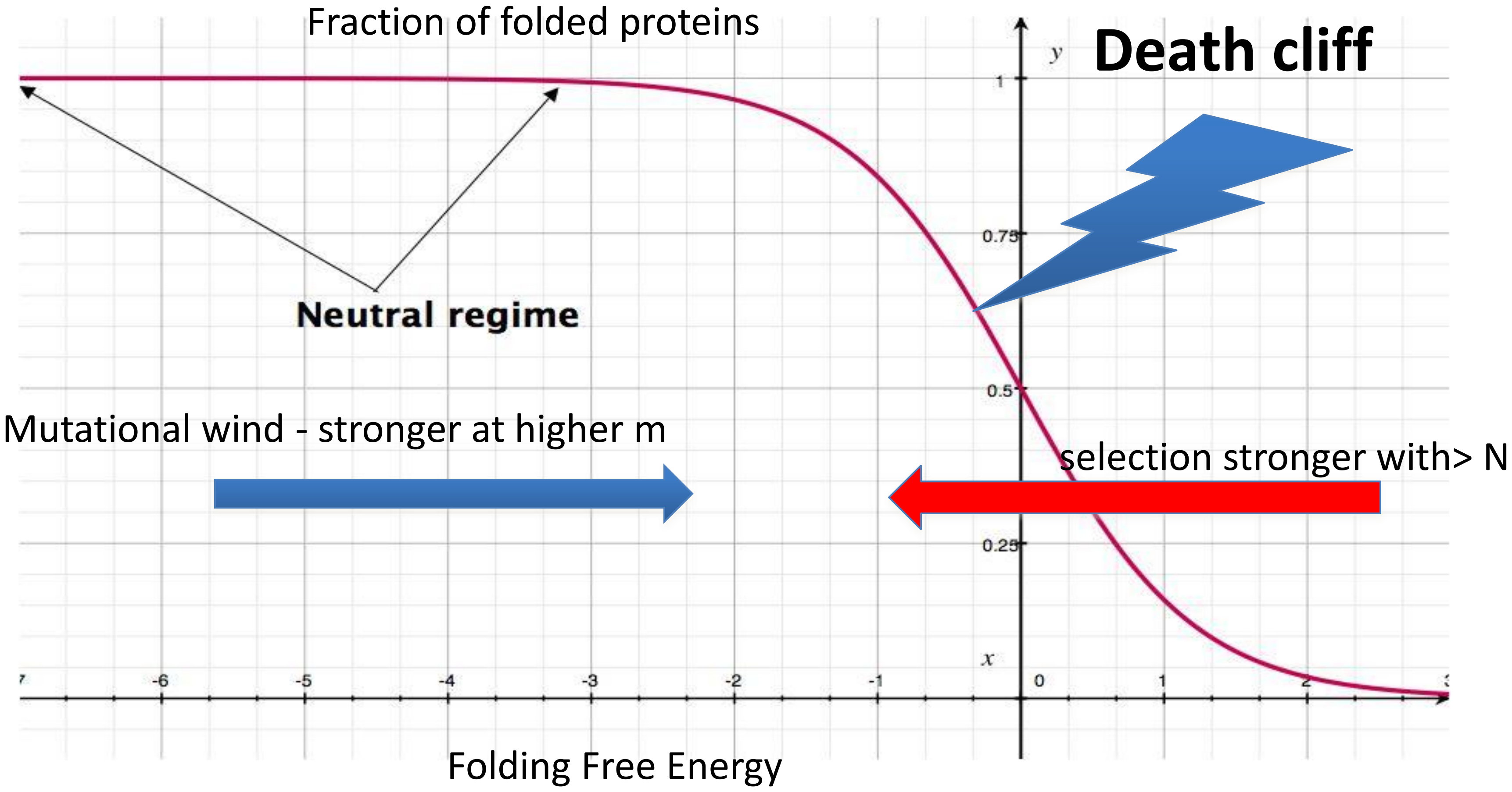
unfolded:
(large entropy, small energy)

folded:
(~ 0 entropy, energy << 0)



ΔG typically ~ -10 to -4 kcal/mole:
"marginally stable"

Evolution of Stability is a Branching (because of cell replication) Random Walk (Biased Diffusional Motion) on the FD Fitness Landscape



Fraction of folded proteins

Death cliff

Neutral regime

Mutational wind - stronger at higher m

selection stronger with> N

Folding Free Energy

Q: How do we know the distribution of step sizes?
A: From experimental ddG statistics!

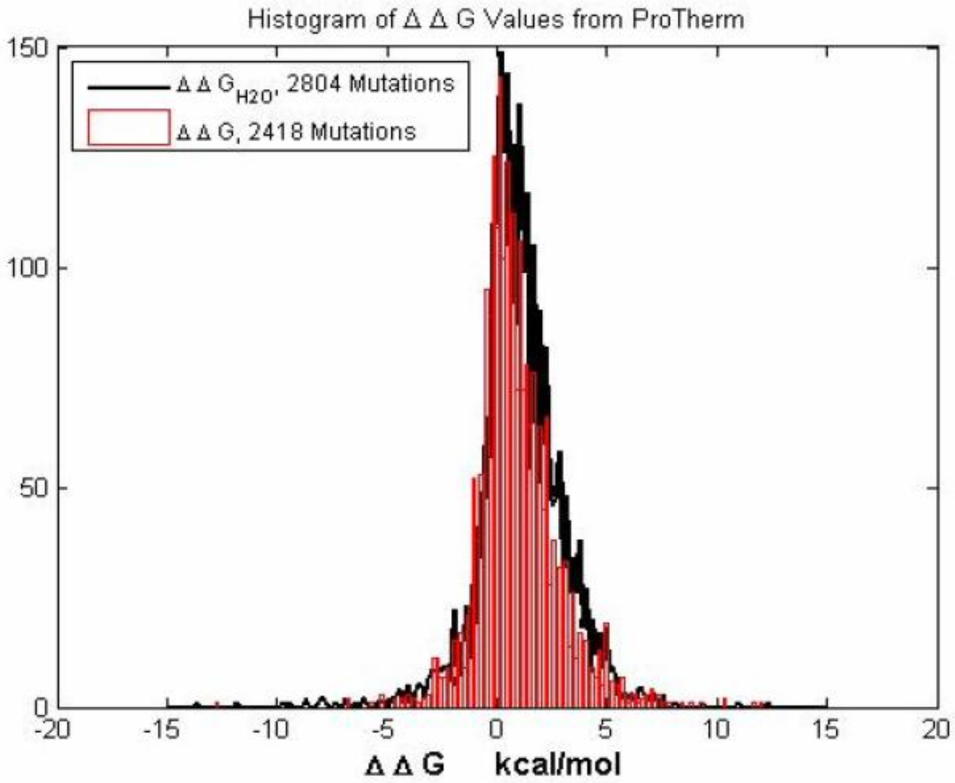
- Asymmetric distribution of ΔΔG; sharper edge on the ΔΔG <0 side, more likely to destabilize protein

•ΔΔG and $ΔΔG_{H2O}$ Shows similar statistics

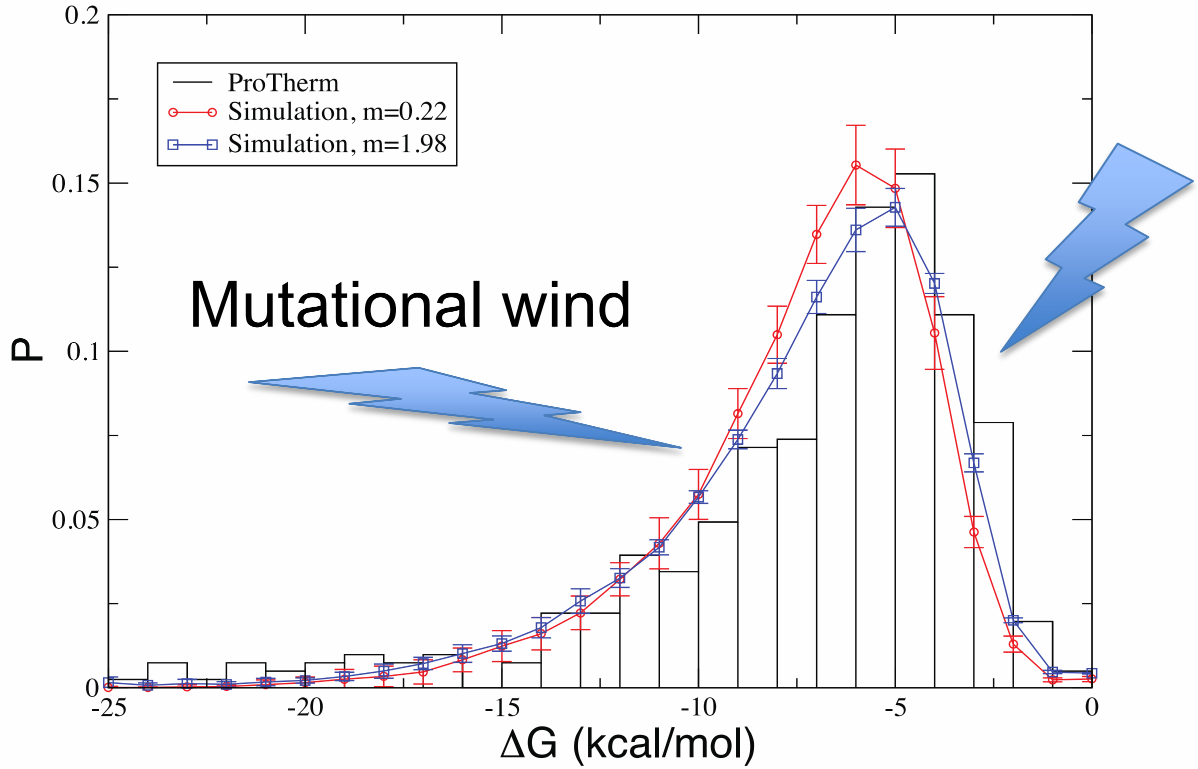
•We could therefore obtain parameters in our random walk model:

•<ddG> =S= 1.1kcal/m

D=<ddg^2> = (3kcal/m)^2



Distribution of protein stabilities : The interplay between protein folding and population genetics



Mutational wind

Cliff hanging

No adjustable parameters

Interplay between mutational wind and ‘’cliff hanging’’:

A.Serohijos, S.Wylie, Peiqiu Chen, K.Zeldovich

Universal Speed Limit on Mutation Rates

~6 Missense mutations Per Proteome Per Replication.

Higher mutation rates ->Population goes extinct

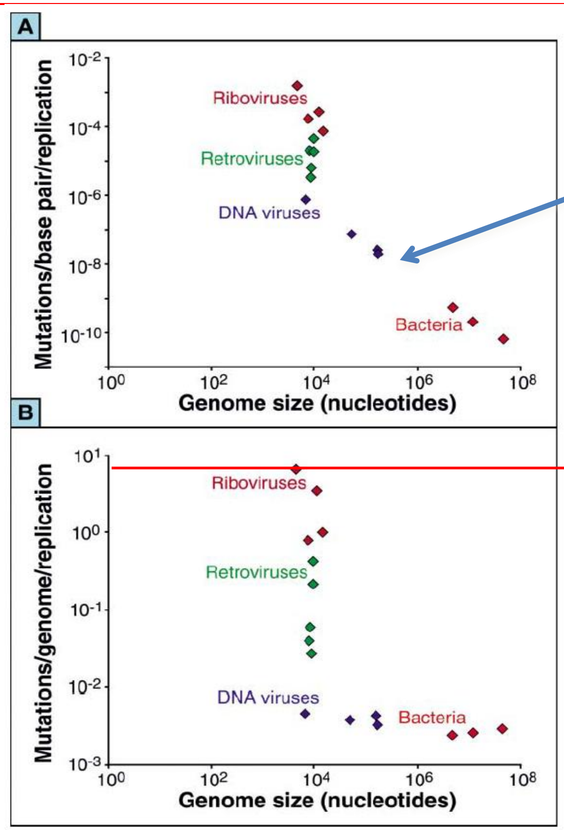
$$\frac{m\mathrm{G}}{b} < \frac{m\mathrm{G}^*}{b} = \frac{2}{\left\{ \frac{s^2}{s^2 + D} + \frac{p^2 \left(s^2 + D \right)}{\left(E_{\min} - E_{\max} \right)^2} \right\}} \approx 6$$

ddG effects of single mutations

# of mutations per genome per replication

K.Zeldovich

RNA world: *At the Edge of the Speed Limit;* **DNA World**: *All organisms are 1000-fold below* (Error correction at work)



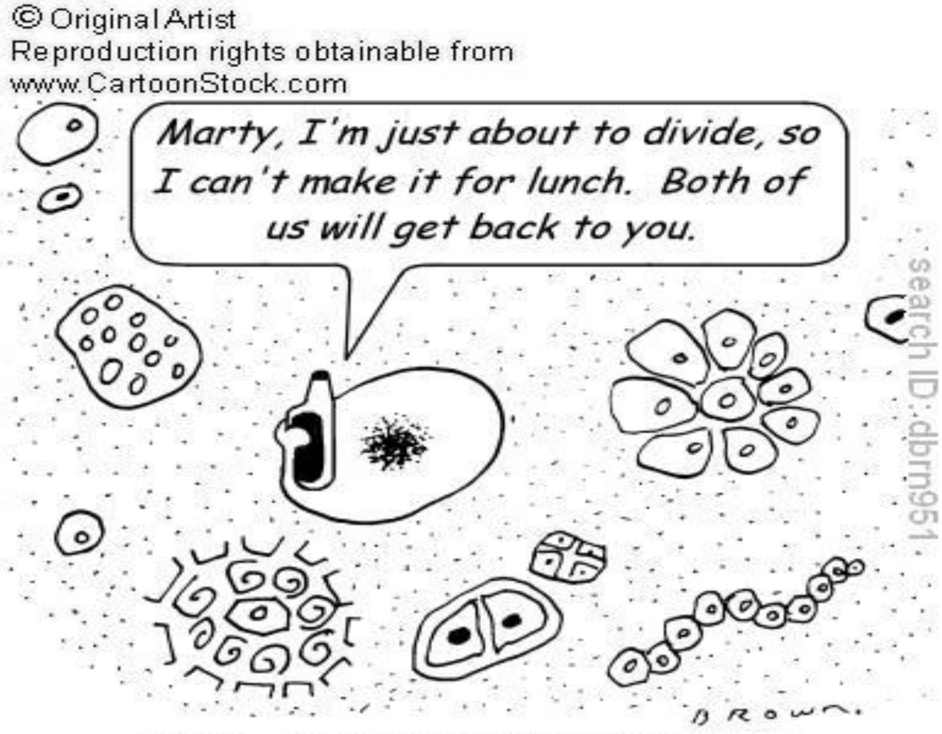
Longer genomes=lower m

$$\frac{2}{\left\{\frac{S^2}{S^2+D}+\frac{\pi^2\left(S^2+D\right)}{\left(\Delta G_{\max}-\Delta G_{\min}\right)^2}\right\}}\approx 6$$

Zeldovich, Chen,ES.PNAS’ 07

Loeb et al, Ann Rev Microbiology, 2005, 58, 183-205

“The dream of every cell is to become two cells.” - F. Jacob -


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com
Marty, I'm just about to divide, so I can't make it for lunch. Both of us will get back to you.
search ID: dbrn951
BROWN.

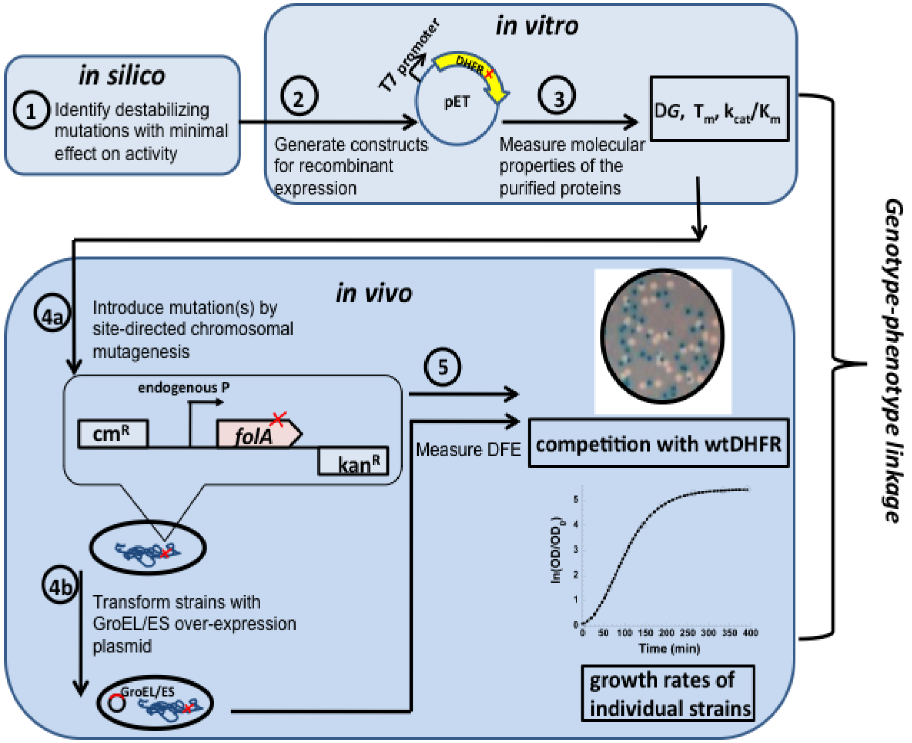
THE CELLULAR PHONE

Cell division rate = fitness
(with caveats –
see below)

How to relate fitness to
Protein Folding? –
That is the question!

Exploring the Genotype-Phenotype relation one mutation at a time:

The experimental setup



Identify destabilizing mutations with minimal effect on activity

Generate constructs for recombinant expression

Measure molecular properties of the purified proteins

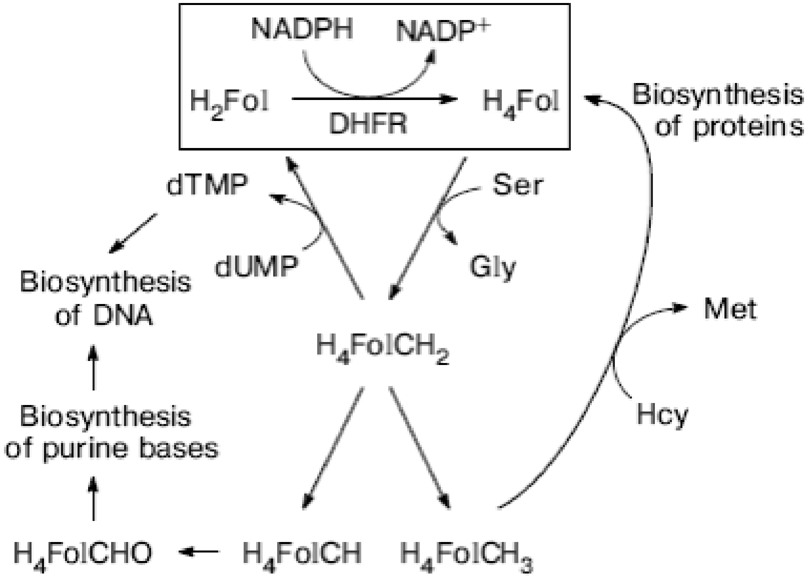
Introduce mutation(s) by site-directed chromosomal mutagenesis

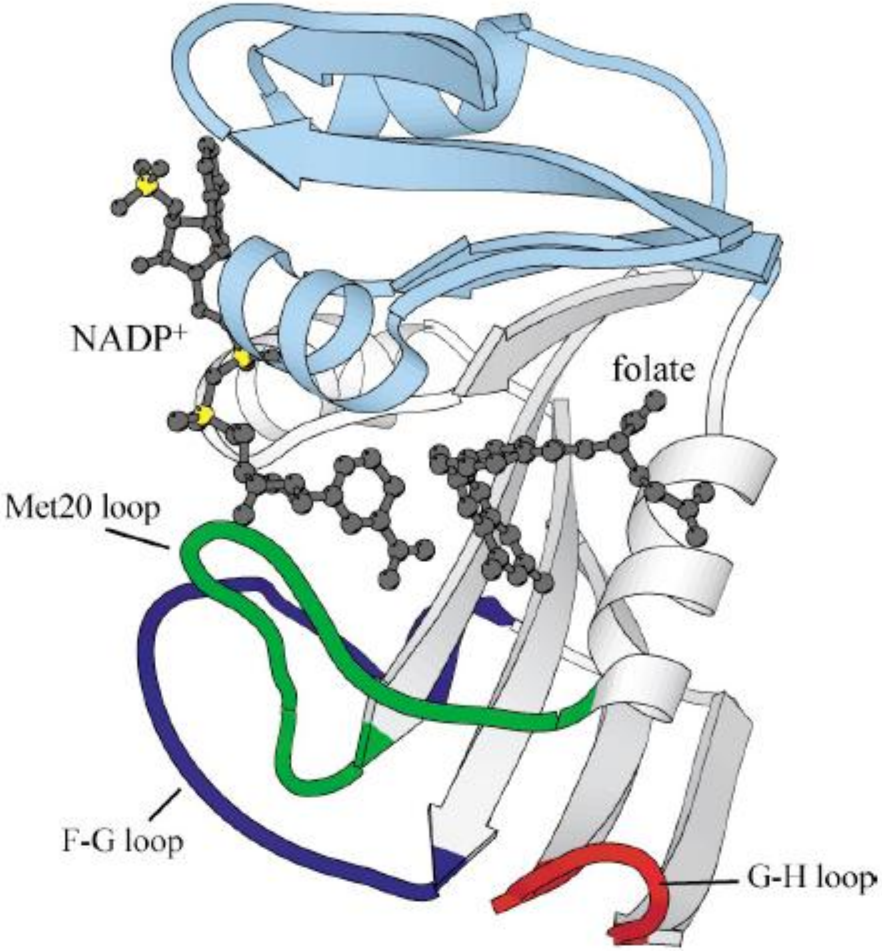
Transform strains with GroEL/ES over-expression plasmid

Measure DFE

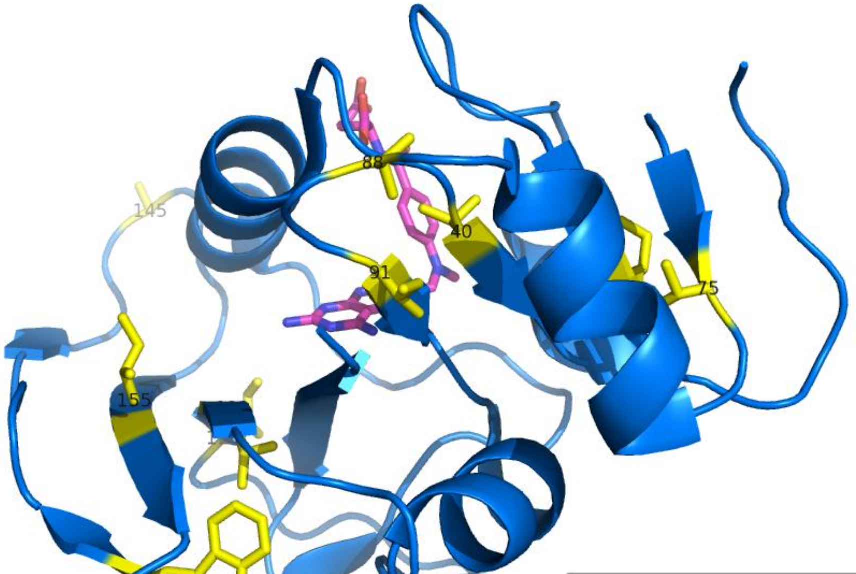
Bottom-up Approach: Introducing mutations of known molecular properties directly on the chromosome.

DHFR Is An Essential Core Metabolism Enzyme

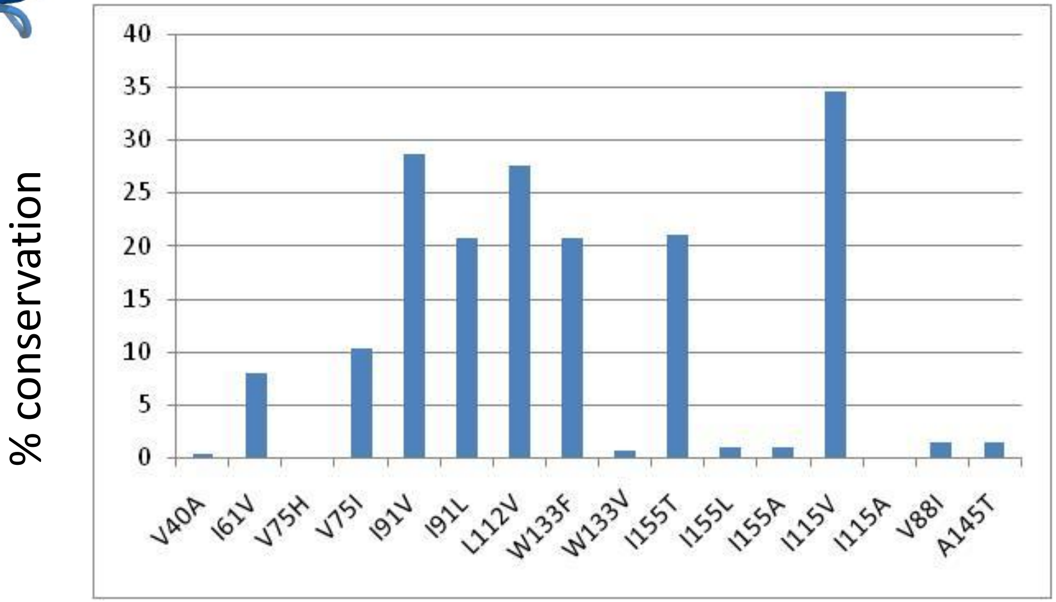




Most mutated residues are deeply buried in the hydrophobic core; half are very conserved evolutionally

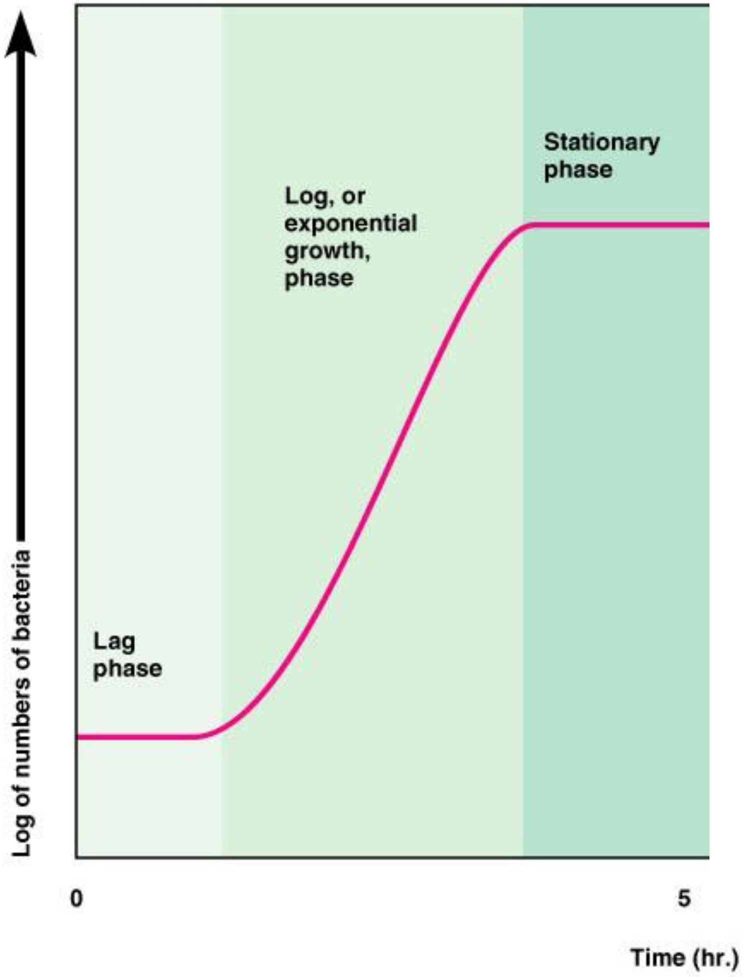


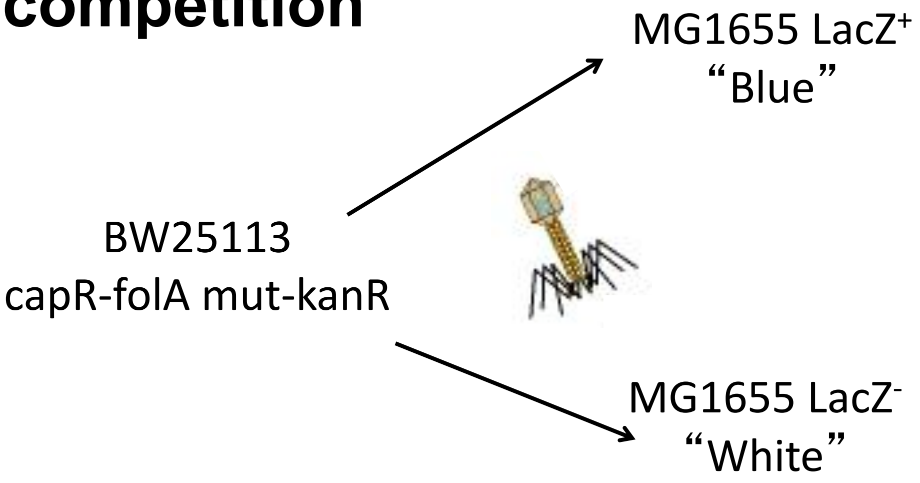
E.*coli*' s DHFR



Mutation	ASA
V40A	0
I61V	0
V75H	0.05
V75I	0.05
I91V	0.07
I91L	0.07
L112V	0
W133F	0.05
W133V	0.05
I155T	0.12
I155L	0.12
I155A	0.12
I115V	0.02
I115A	0.02
V88I	0.34
A145 T	0.89

Bacterial fitness:
Growth rate *vs* competition

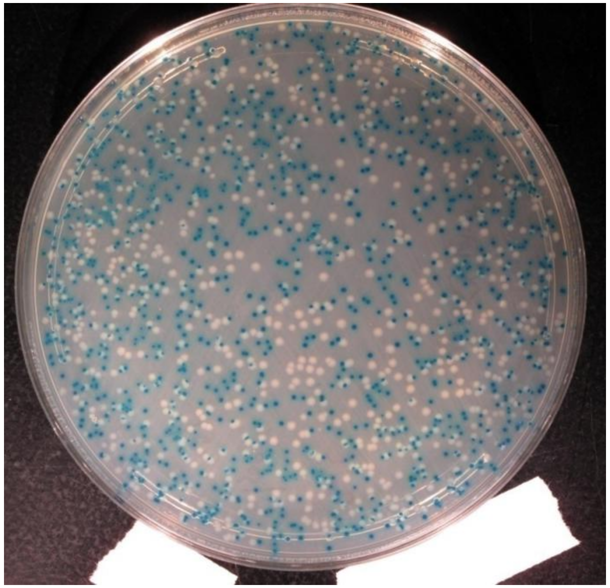




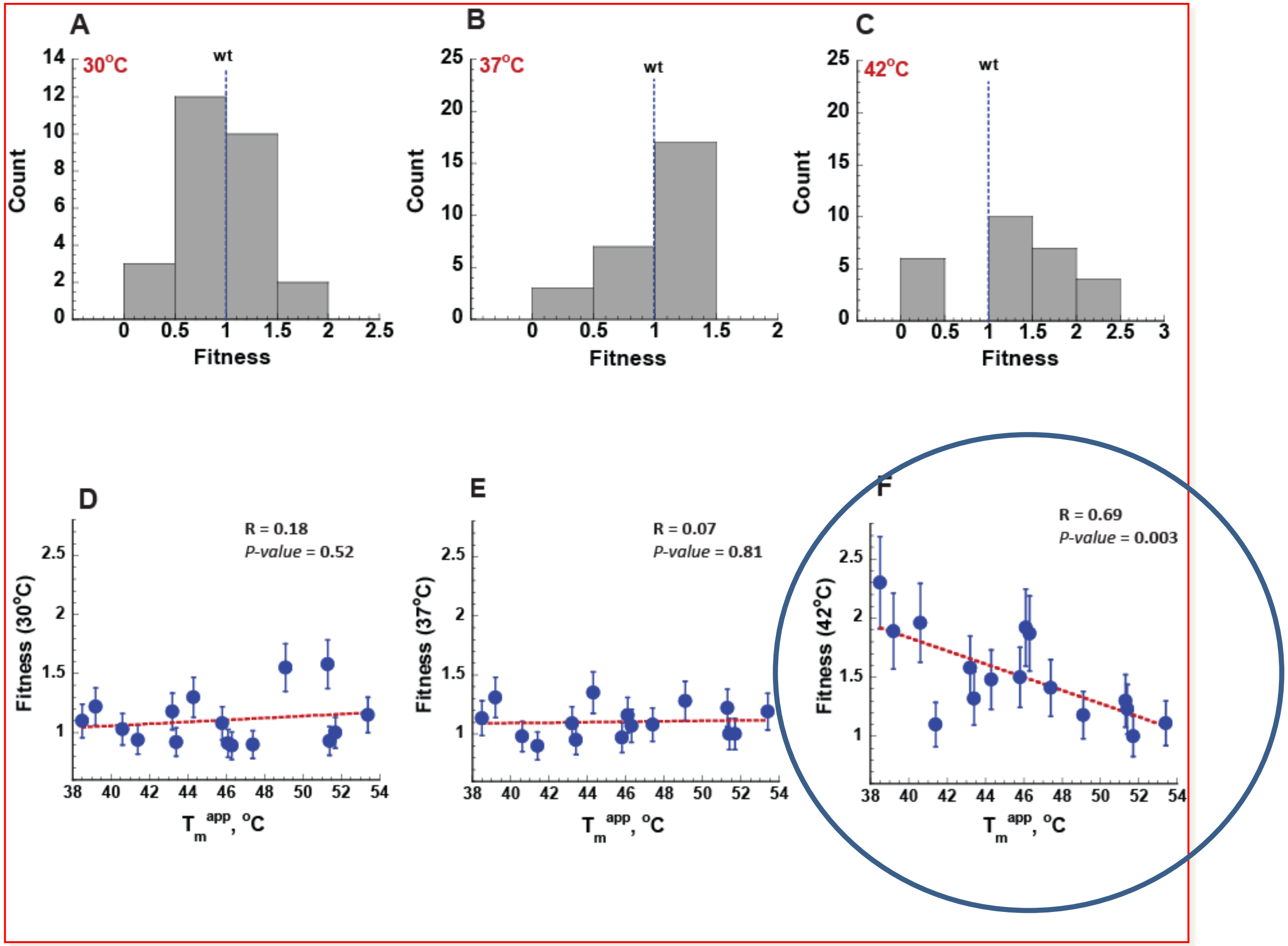
BW25113
capR-folA mut-kanR

MG1655 LacZ^+
“Blue”

MG1655 LacZ^-
“White”

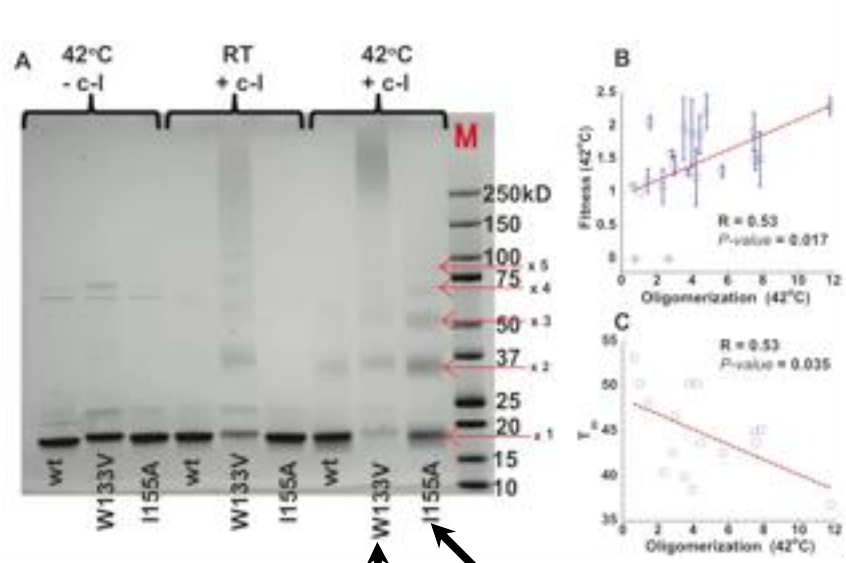


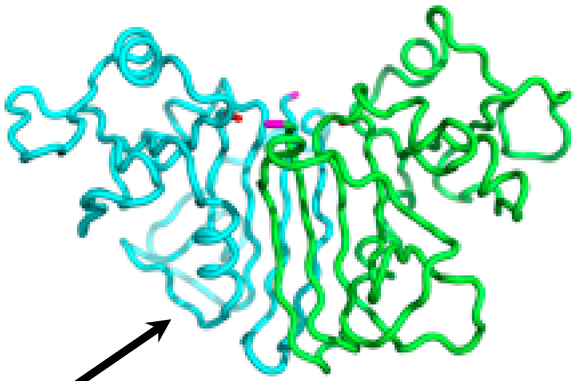
Counter intuitively, for non-lethal mutants fitness is inversely correlated with stability at 42°C!



Why?- see Bershtein, Mu and ES, PNAS, 2012 v.109, pp 4857-62

Soluble oligomerization rescues many mutants from aggregation providing higher fitness to mutant strains

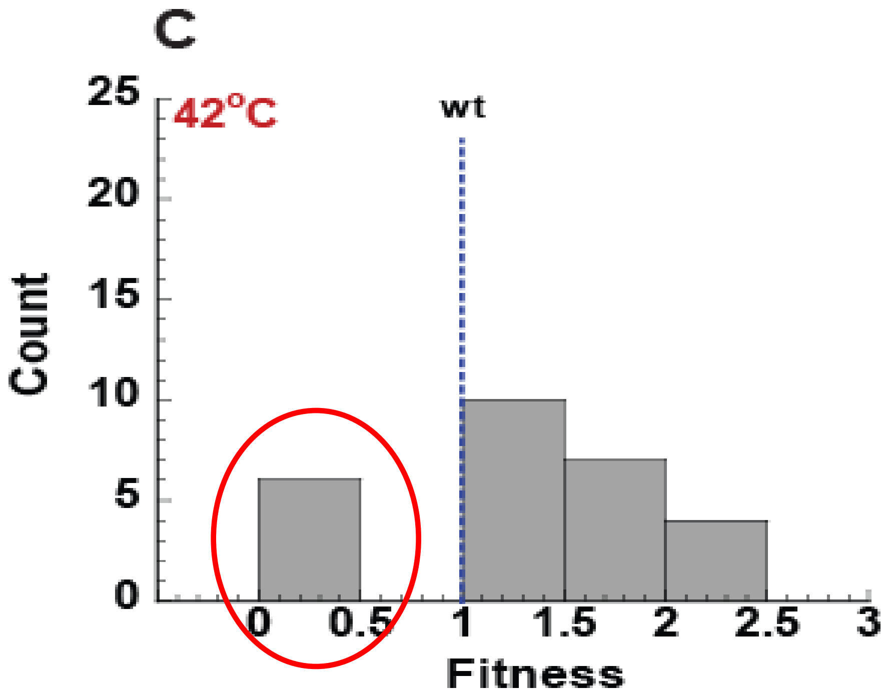




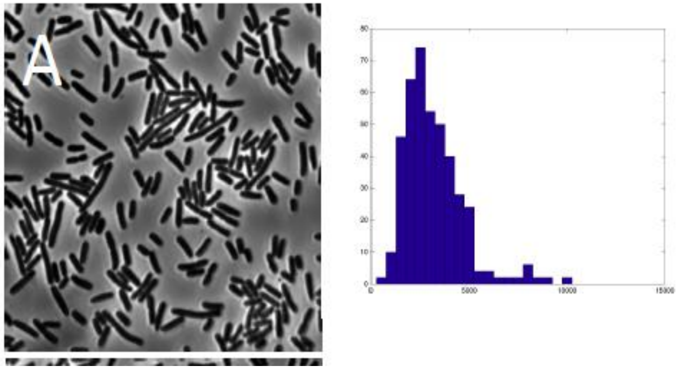
Aggregating mutant

Mutant which oligomerizes

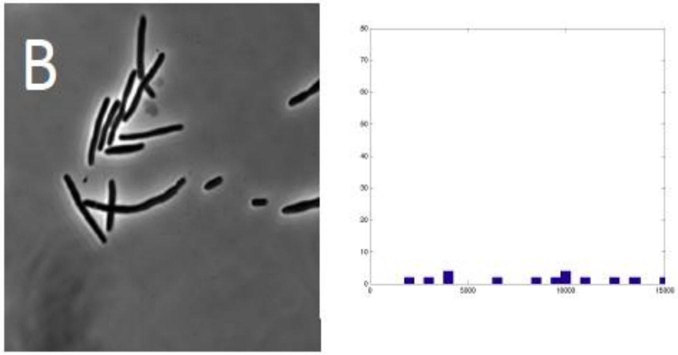
At 42°C 5 out of 27 DHFR mutant strains exhibit slow growth/low fitness



As DHFR activity is diminished the balance between DNA and proteins is shifted: Aggregation prone strains acquire a distinct elongated morphology: a DHFR-deficient phenotype



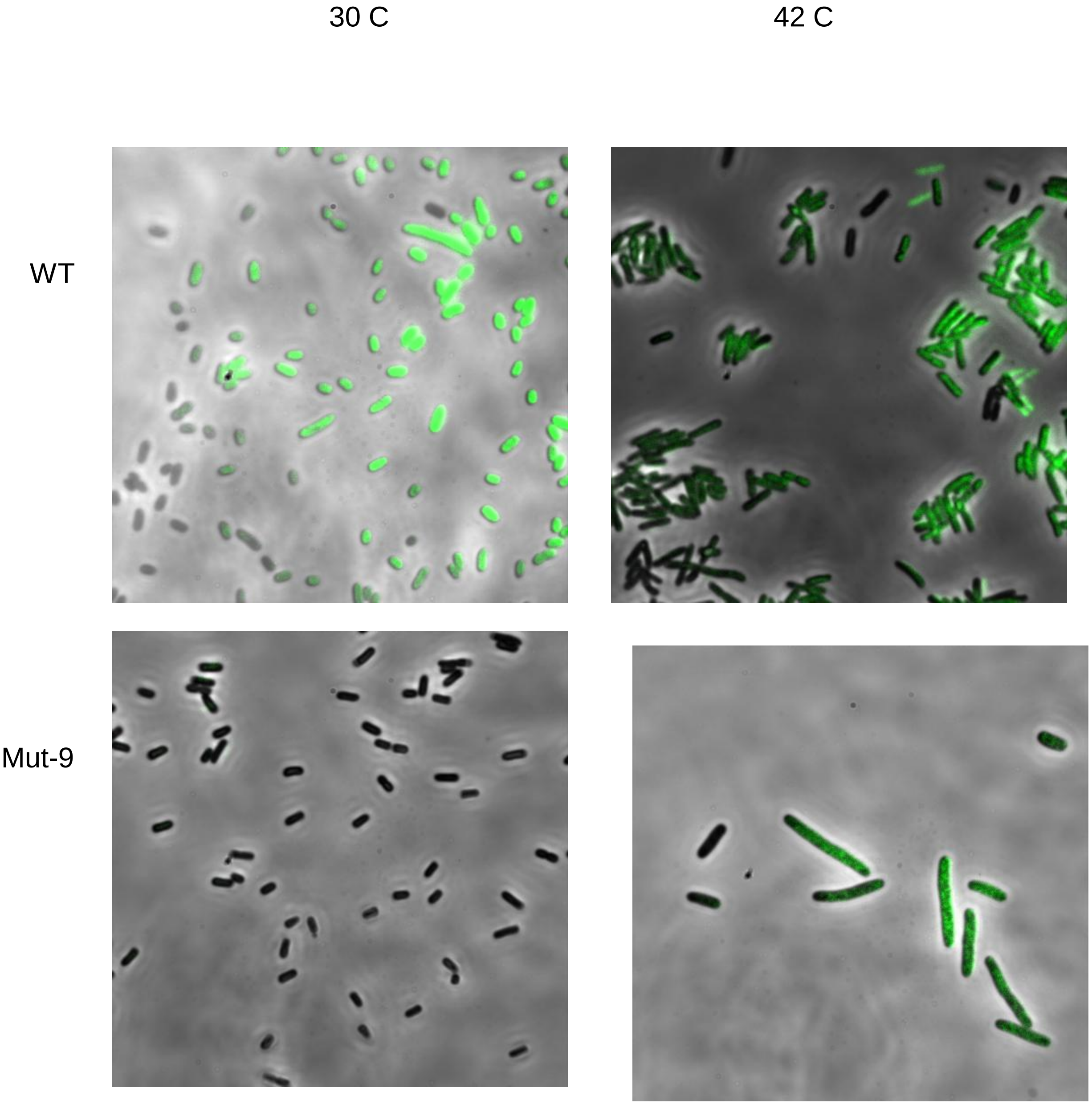
Wt cells



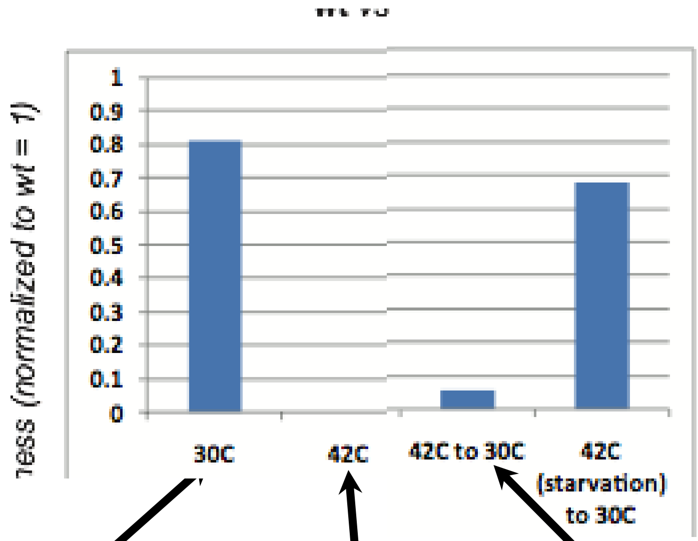
W133V DHFR cells

See also: Kishony et al, Cell, 2010

‘Slow growth mutants of DHFR aggregate in the E.coli cytoplasm: Venus fusion experiment.



Aggregation-prone mutants confer fitness memory on E.coli strains



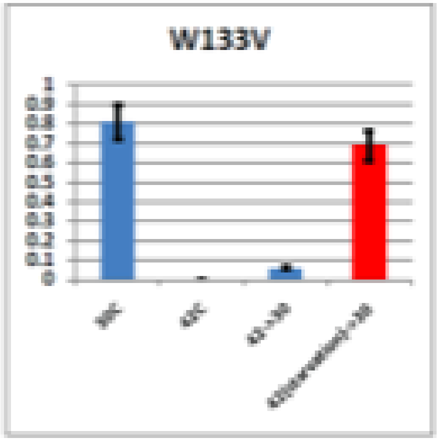
Now:
Before competition at 30C incubate both wt and the mutant at 42 C. Does history of pre-incubation at high T affect fitness at 30C?

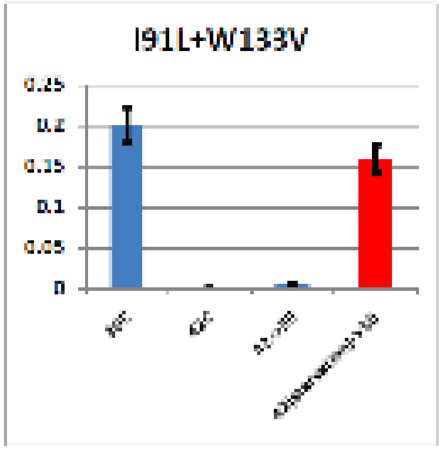
At 30 C the mutant competes well with wt

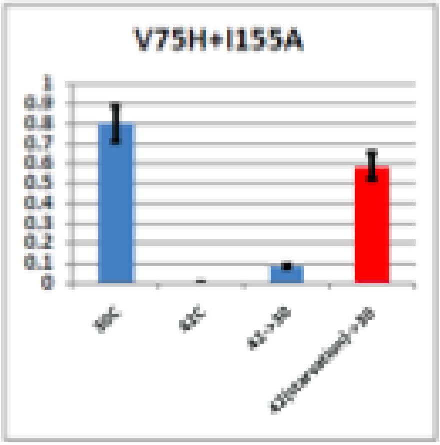
At 42 C the mutant strain does not grow

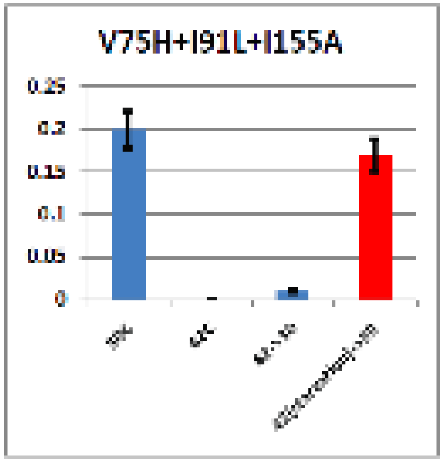
Cells ‘‘memorize’’ past conditions:
after pre-incubation at 42 C fitness at 30 C is lost

Memory effect is due to an active process: Incubation at starvation erases it

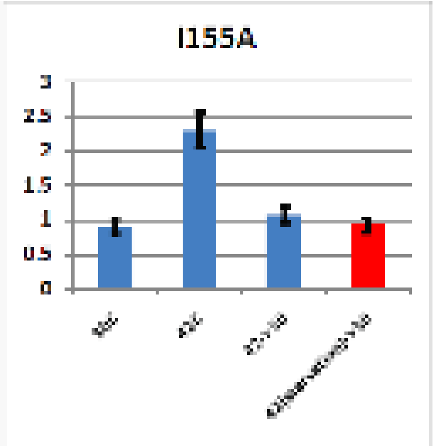




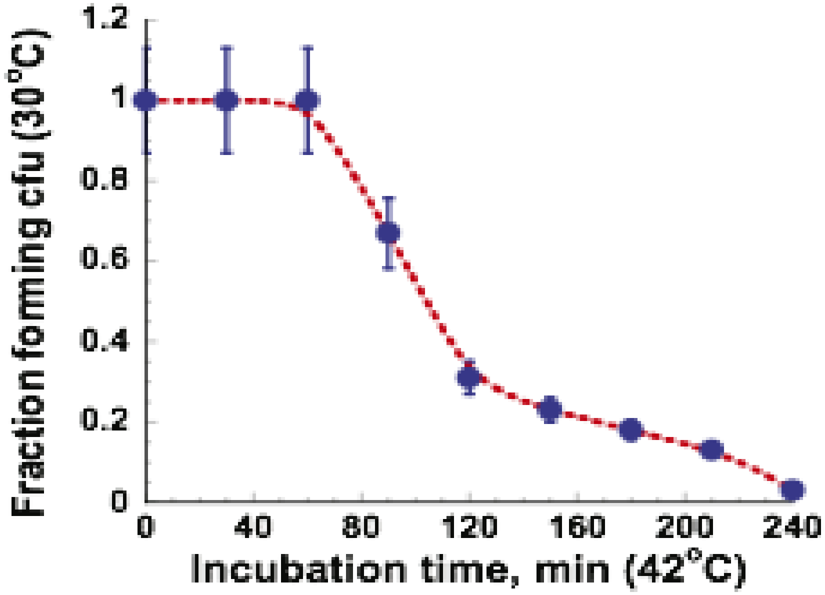




Text



Lag in Fitness Memory : Hint at nucleation of aggregation/amyloid formation of DHFR mutants



Fitness of the mutants strain at 30C depends on how long cells were pre-incubated at 42 C

Q: Why are DHFR aggregates ‘’toxic”?

A: Because aggregates float in the cytoplasm interfering with other cellular processes ‘’non-specifically”.

B: Because aggregates sequester newly made DHFR depleting cells of an essential enzyme

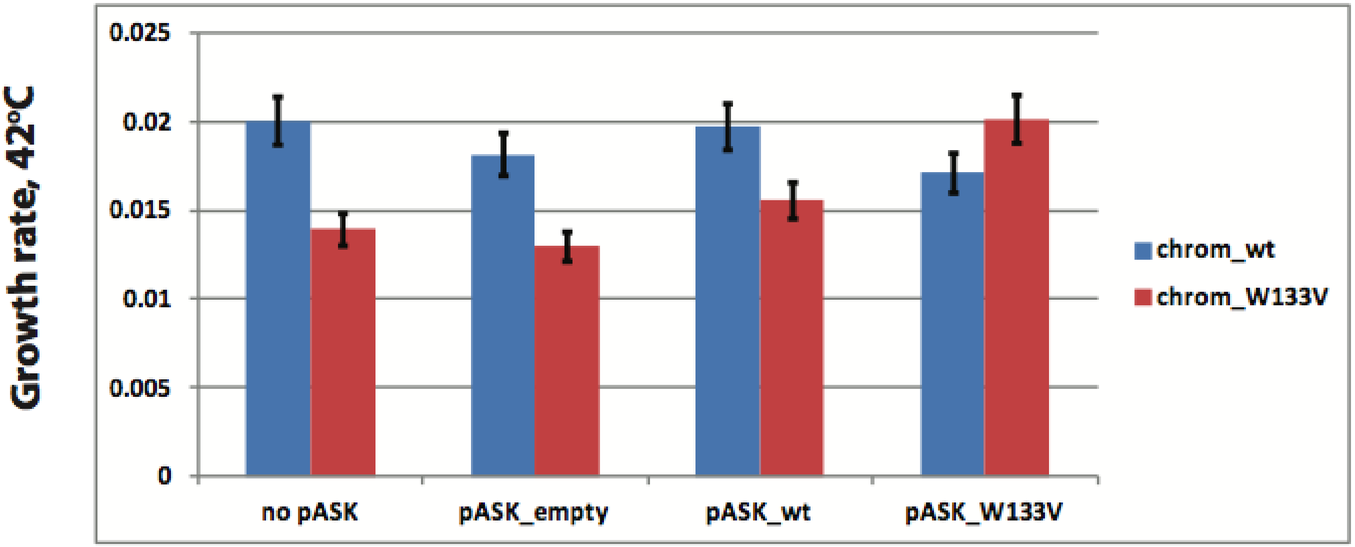
How to answer this question experimentally: ‘’flood” the cells with aggregation-prone DHFR from a plasmid..

If (A) is correct - cells will die.

If (B) is the answer cells will improve fitness.

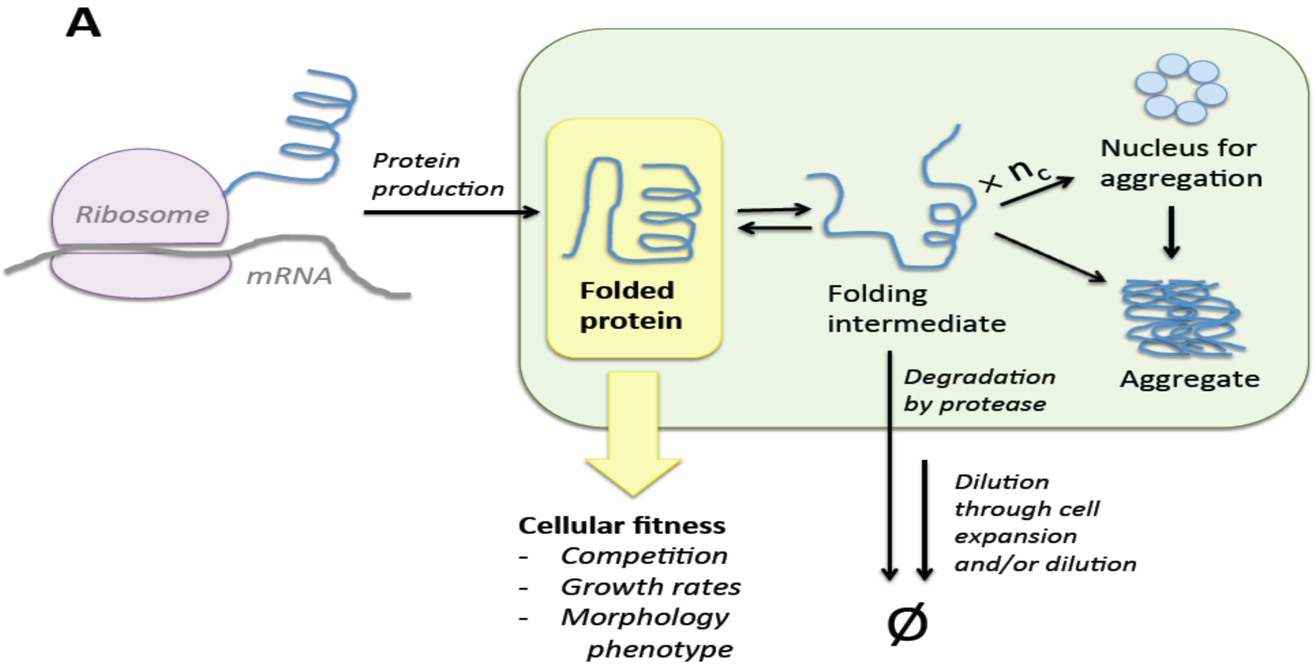
And the answer is…..

B: aggregates sequester newly made DHFR depleting cells of an essential enzyme



Massive overexpression of W133V or I91L+W133V DHFR improves fitness

Theory: a dynamic equilibrium



Physics/Biology Foundations of the model:

1) The cytoplasm is an active medium where concentrations are determined by steady state fluxes rather than Boltzmann equilibrium
2) Aggregates are formed via nucleation, from the Molten-Globule state which competes with cell division.
3) ''Aggregation toxicity'' is due to sequestration of newly folded proteins into aggregates

Equations, shmequations…

$$\frac{dF}{dt} = \mathrm{Pr} - k_u F + k_f U - k_{dil} F$$

$$\frac{dU}{dt} = k_u F - k_f U - n_c k_n U^{n_c} - k_a A U - k_{dil} U$$

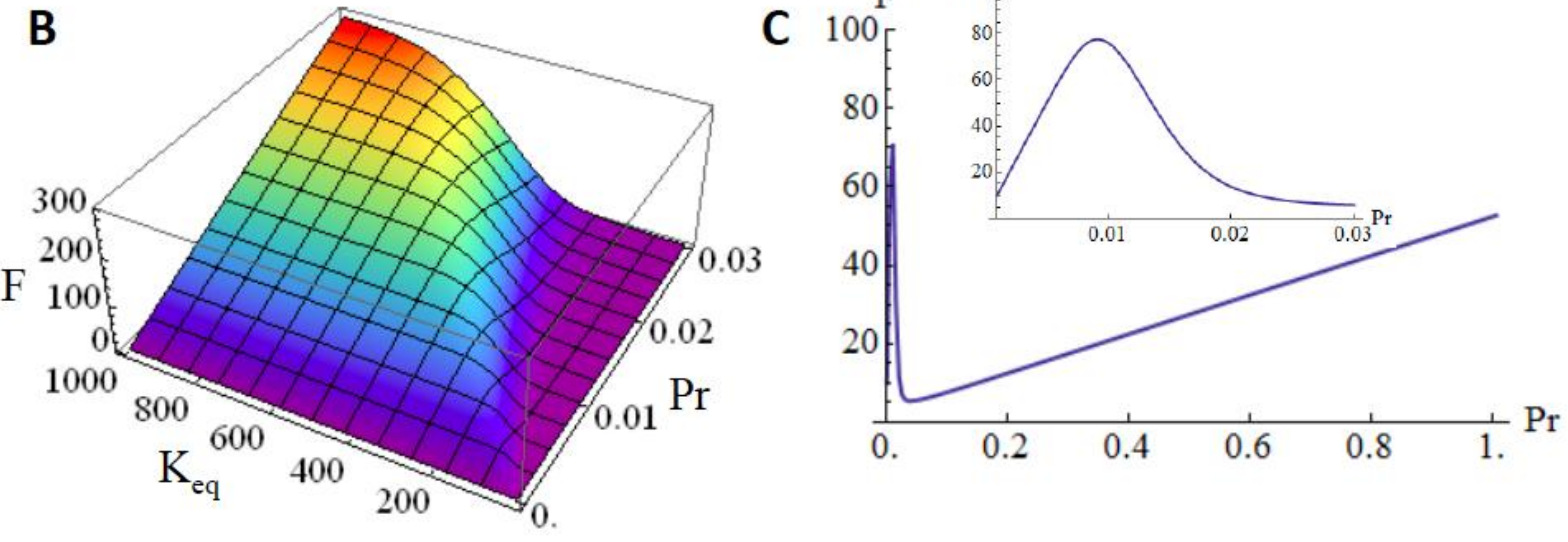
$$\frac{dNuc}{dt} = k_n U^{n_c} - k_a U Nuc - k_{dil} Nuc$$

$$\frac{dA}{dt} = k_a U Nuc - k_{dil} A$$

Jaie Woodard

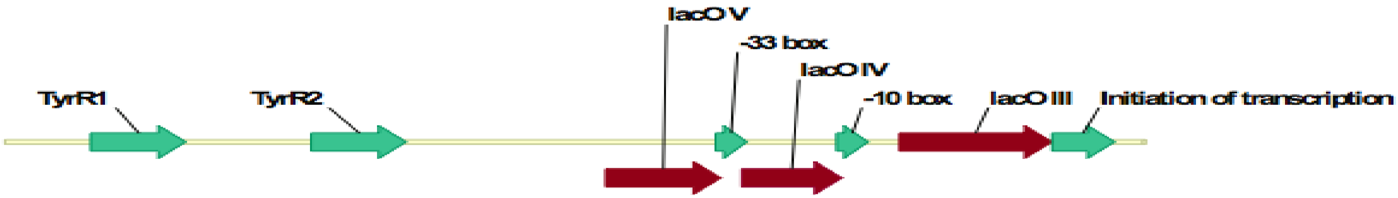
Theory Predicts: Less is more.

Downregulation of the production of an aggregation prone mutant would increase concentration of folded protein and rescue fitness

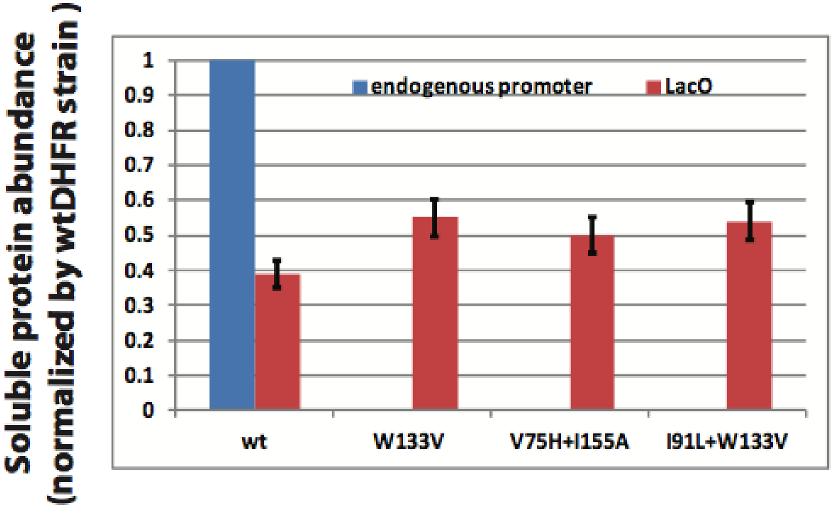


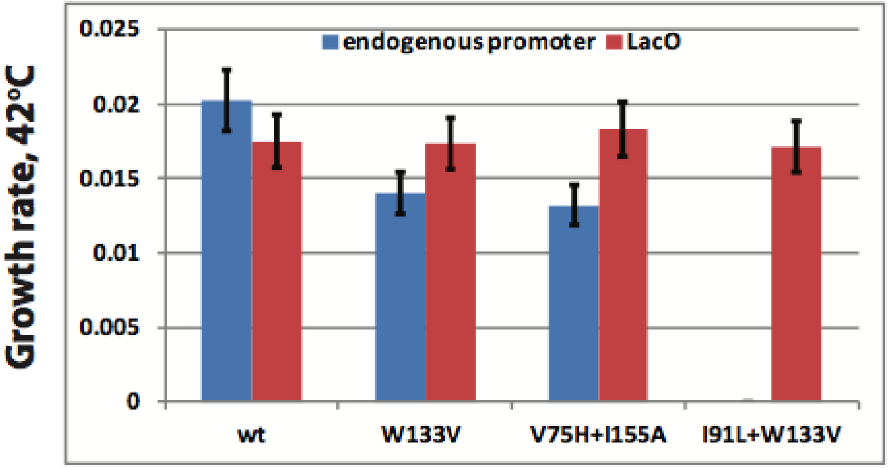
Testing the ‘’less is more” prediction: experiment #1

Downregulation of the DHFR transcription by placing it under an IPTG-controllable promoter indeed restores fitness and folded abundance



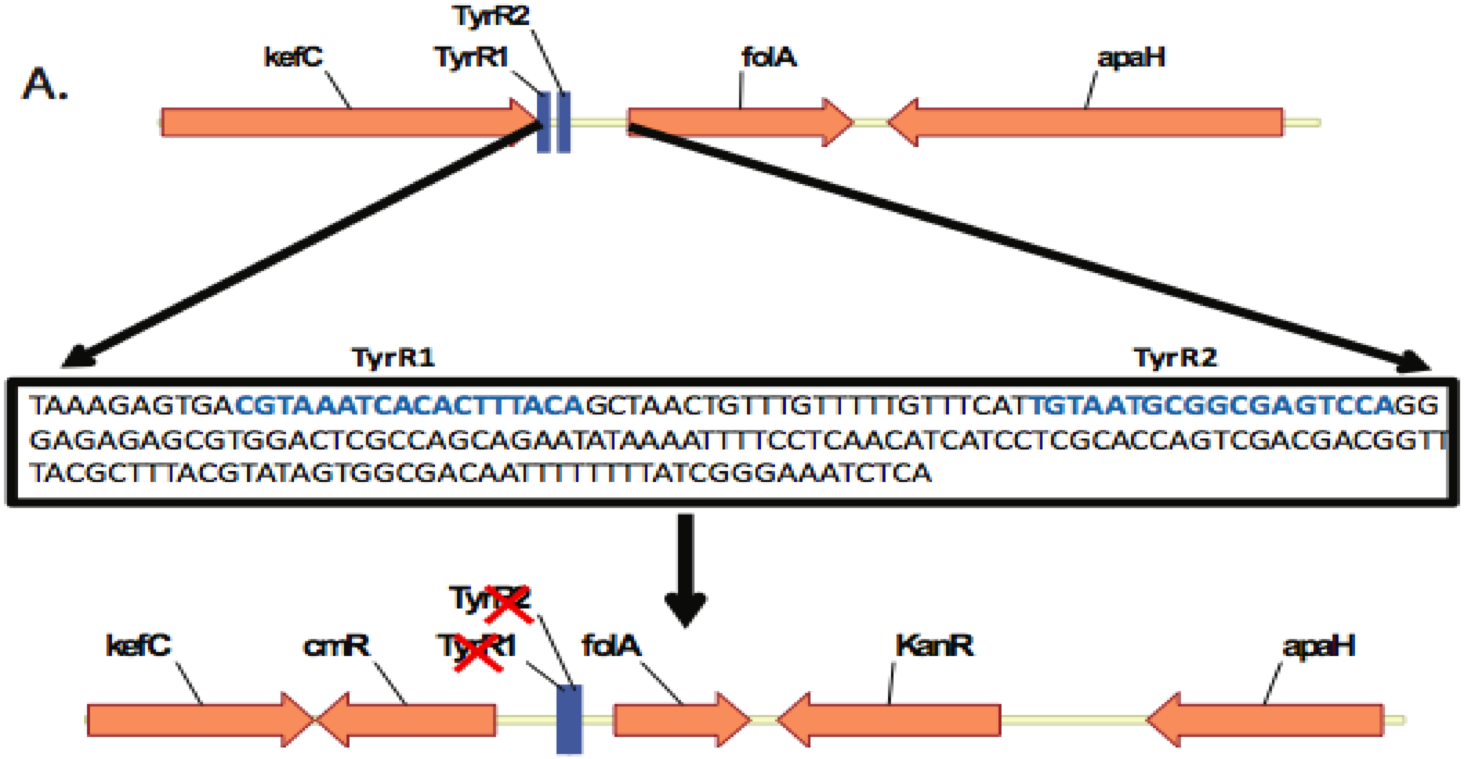
ATCGATTAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTAAT
GCGGCGAGTCCAGGGAGAGAGCGTGGACTCGCCAGCAGAATATAAAATTT**AATTGTGAGC**
GGATAACAATTTCGAC**TTGTGAGCGGATAACAATT**ATAGTGGCGAC**AATTGTGAGCGGATA**
ACAATTTCACACAGATCGGGAAATCTCAT





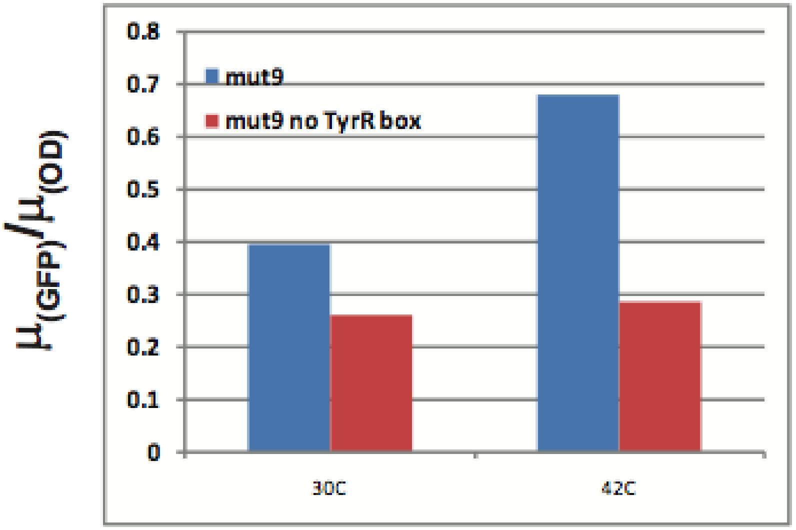
Testing the ‘’less is more” prediction: experiment #2

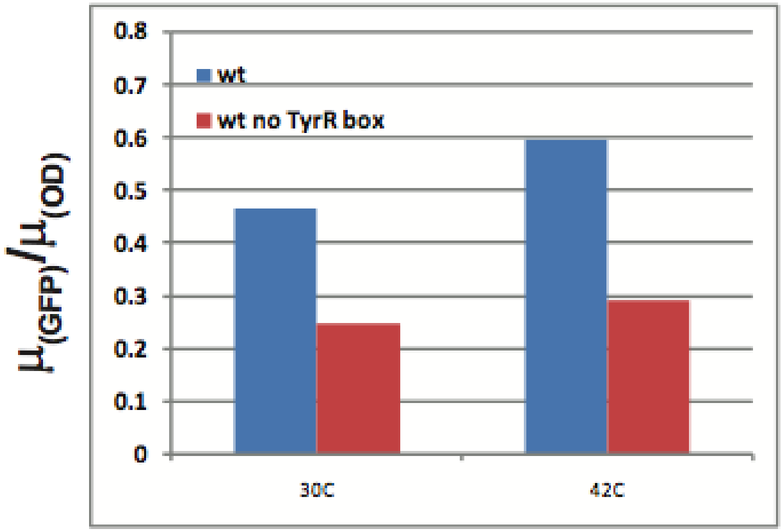
2) **Delete the tyrR box- a genetic element, which controls DHFR expression via a positive feedback loop.**



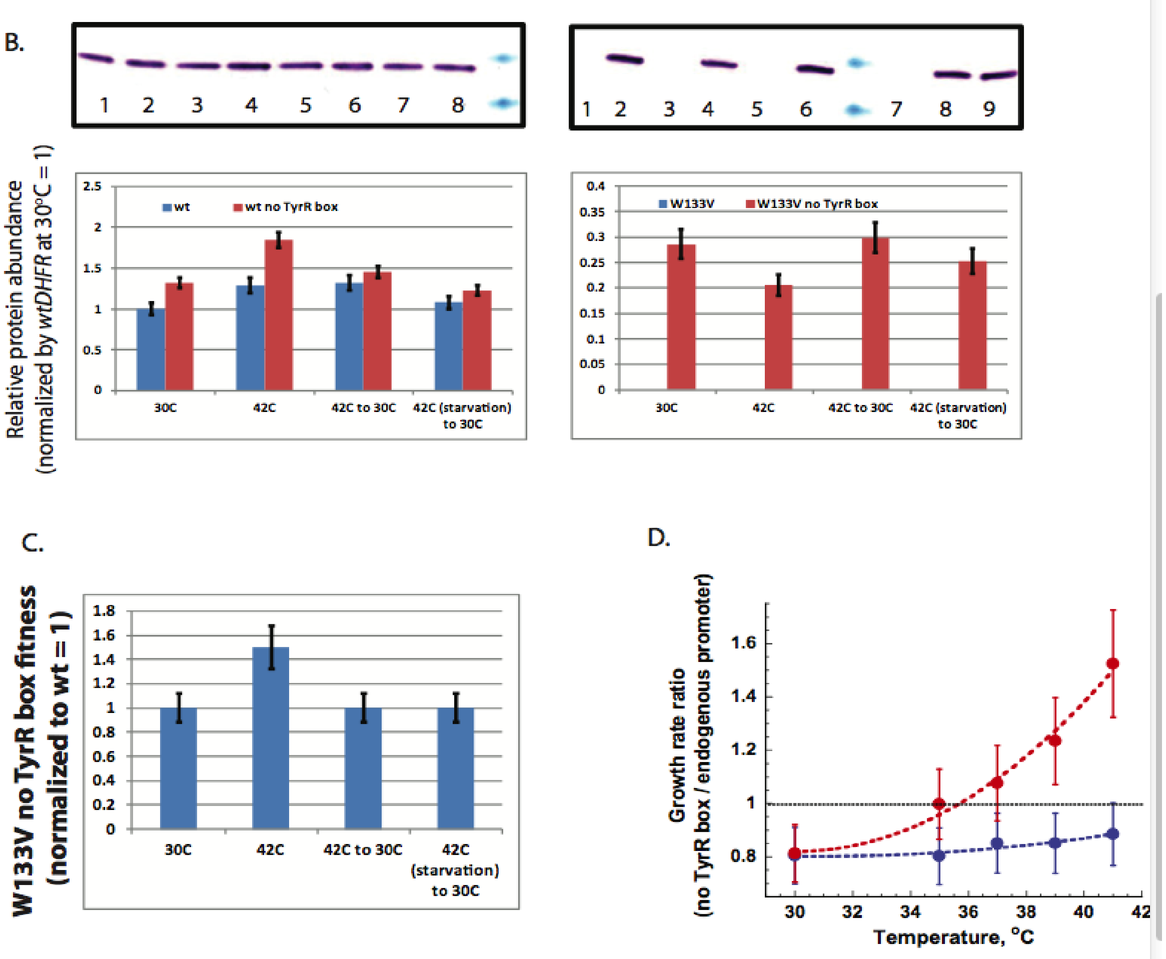
TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTAATGCGGCGAGTCCAGG
GAGAGAGCGTGGACTCGCCAGCAGAATATAAAATTTTCCTCAACATCATCCTCGCACCAGTCGACGACGGTT
TACGCTTTACGTATAGTGGCGACAATTTTTTTTATCGGGAAATCTCA

Real-time qPCR and GFP under DHFR promoter indeed confirm that tyrR controls DHFR expression through positive feedback: deletion of tyrR box decreases production

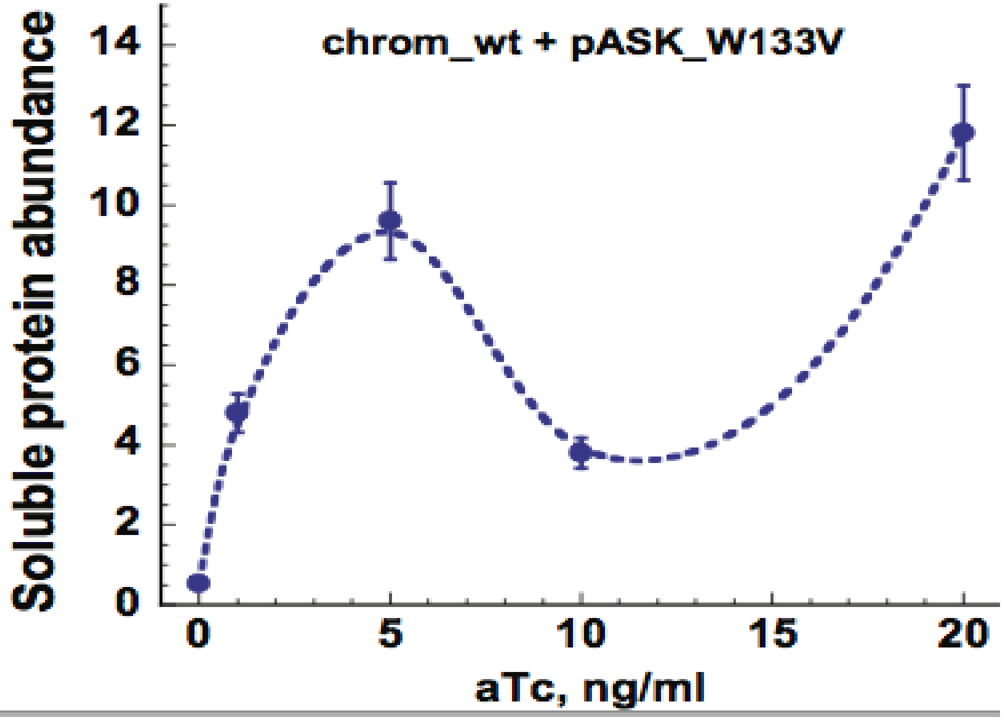


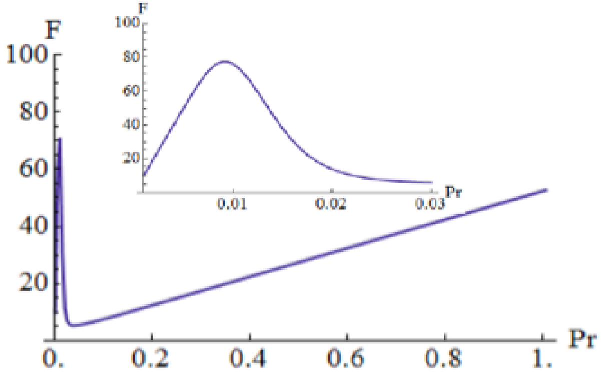


At the same time tyrR deletion hugely increases abundance of folded protein in cytoplasm and rescues fitness:



Finally 3rd experiment: add mutant strains from the aTc controllable plasmid and see how the amount of folded protein depends on production.





Acknowledgements:

Shimon Bershtein
Mu Wanmeng
Jaie Woodard
Jiabin Xu
Konstantin Zeldovich (now at UMASS)
Adrian Serohijos
Scott Wylie
Ryan Lee
Abhishek Chintapalli
Elston He
George Whitesides
Xiaowei Zhuang
Roy Kishony
Sunny Xie

$$$ NIH, HFSF, DARPA

THINK BIG!!!