MSCBIO 2060 - Current Topics in Computational Biology

Determinants of the rate of protein sequence evolution

Jianzhi Zhang and Jian-Rong Yang
Nature Reviews Genetics 16, 409-429 (2015)
Published online 09 June 2015

PRESENTED BY

RAGHAV PARTHA

Evolutionary rate (R)

- Measure of dynamics of change in a lineage
- Estimated for a pair of species as number of substitutions per site (d) divided by time of divergence (T)
  - d: Fraction of differing amino acid positions, corrected to account hidden substitutions
  - T: irrelevant (constant) if comparing proteins between given set of species

- Historical findings
  - Molecular clock hypothesis – Zuckerkandl & Pauling 1965
  - Neutral theory of evolution – Kimura 1983

Rate determinants

- Neutral theory

$$k = \mu * p$$

k: protein evolutionary rate

μ: rate of mutation

p: proportion of neutral mutations

- p: determined by *functional constraint*, which in turn is inferred from k (Circularity)
- Functional importance, expression level?

Functional importance

- Fitness advantage to the organism
  - More important the protein slower it evolves
  - Measured by fitness reduction upon gene deletion
- Empirical results –
  - Hurst & Smith 1999
    - 175 **essential and non-essential** mouse genes
    - **No significant difference in evolutionary rates**
  - Hirsh & Fraser 2001
    - 500 non-essential genes in yeast
    - **Weak negative correlation between fitness reduction and evolutionary rates**

Hurst, L. D. & Smith, N. G. Do essential genes evolve slowly? Curr. Biol. 9, 747–750 (1999)

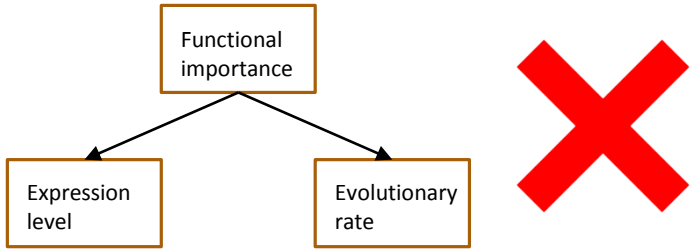
Hirsh, A. E. & Fraser, H. B. Protein dispensability and rate of evolution. Nature 411, 1046–1049 (2001)

Is functional importance irrelevant?

- Possible confounders
  - Laboratory vs natural environment
  - No strong correlation in 400 different laboratory conditions
- Predictive power
  - Two random yeast proteins
    - Slower evolving protein 54% more likely to be more functionally important
  - Two yeast proteins rank-separated by 95% of proteins
    - Slower evolving protein 81% more likely to be more functionally important

Expression level

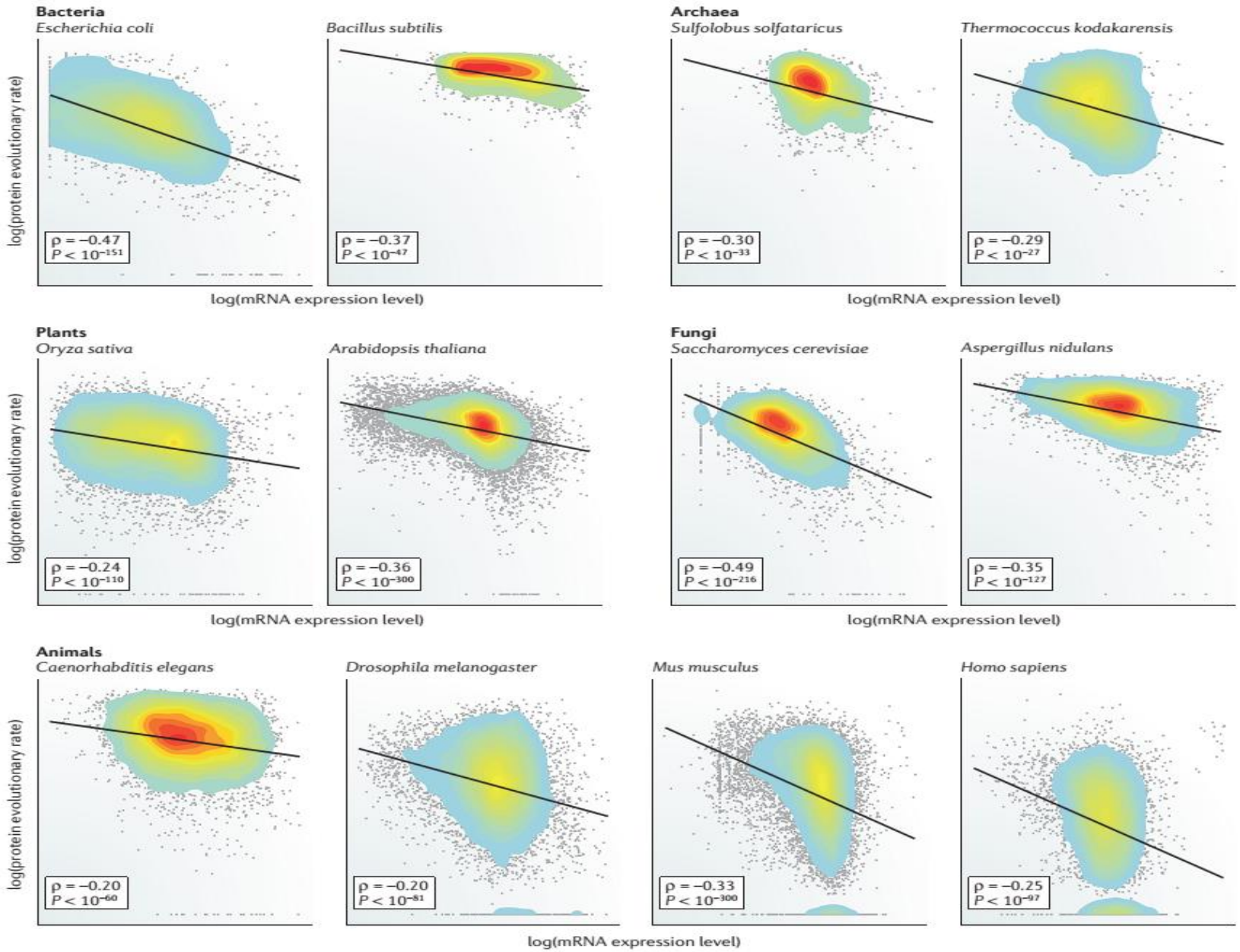
- Pal et al (2001)
  - Strong E-R (Expression-Evolutionary Rate) anticorrelation in yeast
  - True to varying extents across all three domains of life
  - Observed across tissues
- Drummond et al (2005)
  - Stronger signal when mRNA concentrations are used relative to protein concentrations
- Correlation remains after controlling for functional importance



Pal, C., Papp, B. & Hurst, L. D. Highly expressed genes in yeast evolve slowly. Genetics 158, 927–931 (2001)

Drummond, D. A., Bloom, J. D., Adami, C., Wilke, C. O. & Arnold, F. H. Why highly expressed proteins evolve slowly. Proc. Natl Acad. Sci. USA 102, 14338–14343 (2005)

Fig 1.


Bacteria
Escherichia coli
Bacillus subtilis
Archaea
Sulfolobus solfataricus
Thermococcus kodakarensis
ρ = −0.47
P < 10−151
ρ = −0.37
P < 10−47
ρ = −0.30
P < 10−33
ρ = −0.29
P < 10−27
log(protein evolutionary rate)
log(mRNA expression level)
Plants
Oryza sativa
Arabidopsis thaliana
Fungi
Saccharomyces cerevisiae
Aspergillus nidulans
ρ = −0.24
P < 10−110
ρ = −0.36
P < 10−300
ρ = −0.49
P < 10−216
ρ = −0.35
P < 10−127
Animals
Caenorhabditis elegans
Drosophila melanogaster
Mus musculus
Homo sapiens
ρ = −0.20
P < 10−60
ρ = −0.20
P < 10−81
ρ = −0.33
P < 10−300
ρ = −0.25
P < 10−97

Misfolding avoidance hypothesis

- Protein misfolding is cytotoxic and reduces fitness
  - Errors in translation leading to reduced stability
- Higher expressed proteins under stronger pressure to evolve translational robustness
  - Constrains sequence evolution
  - Leads to lower mutation rates and hence E-R anticorrelation
- Other validated predictions for higher expressed proteins
  - Higher folding stability
  - Sites critical for stability more conserved

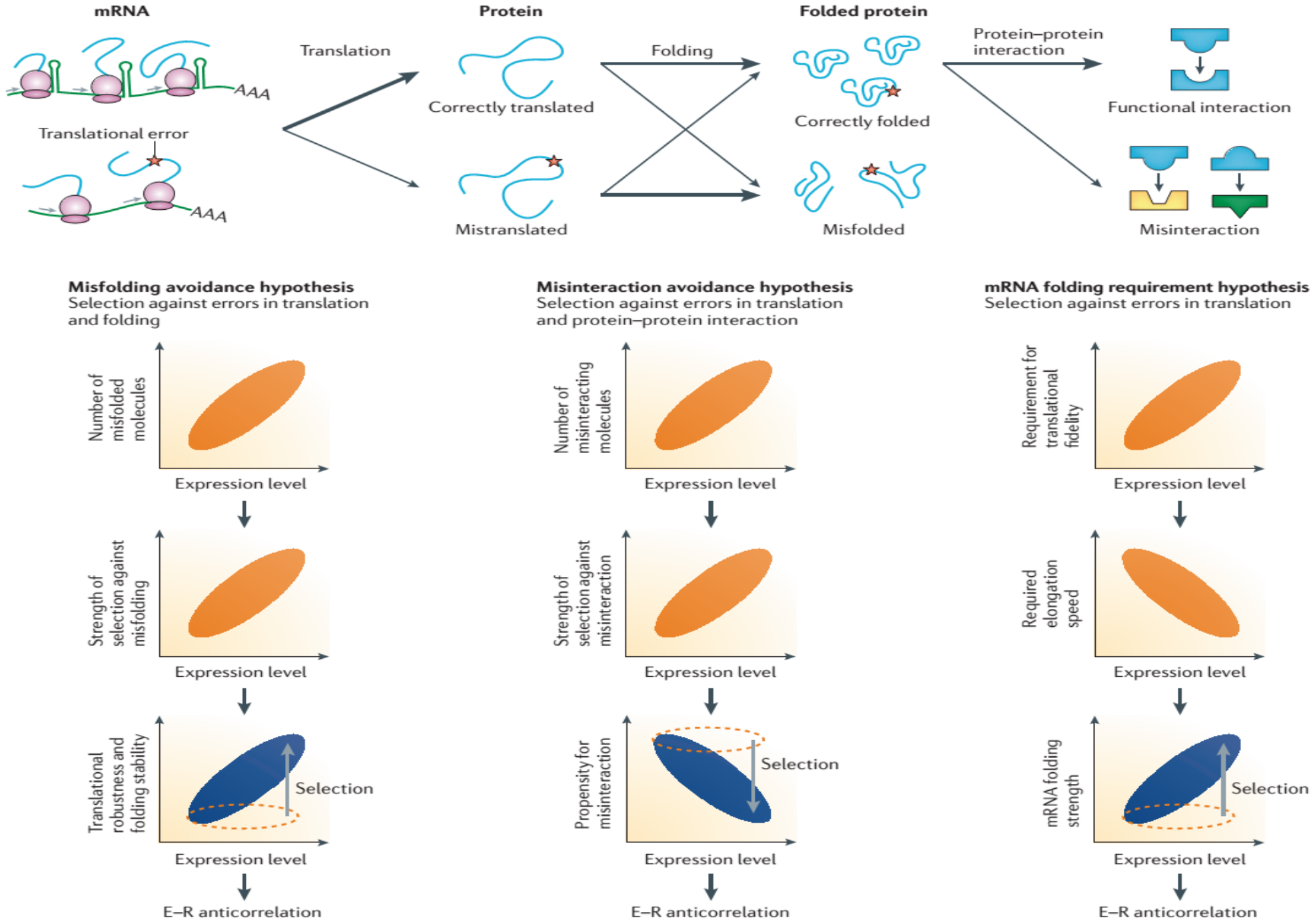
Misinteraction avoidance hypothesis

- Misfolding induced by surface residues is weaker than core residues
  - But, surface residues show E-R anticorrelation as well
- Surface residues critical for interaction
- Number of misinteracting molecules higher for highly expressed proteins
  - Stronger constraint on sequence evolution
  - Leads to lower mutation rates and hence E-R anticorrelation
- Misfolding and misinteraction avoidance hypotheses – complementary insights

mRNA folding requirement hypothesis

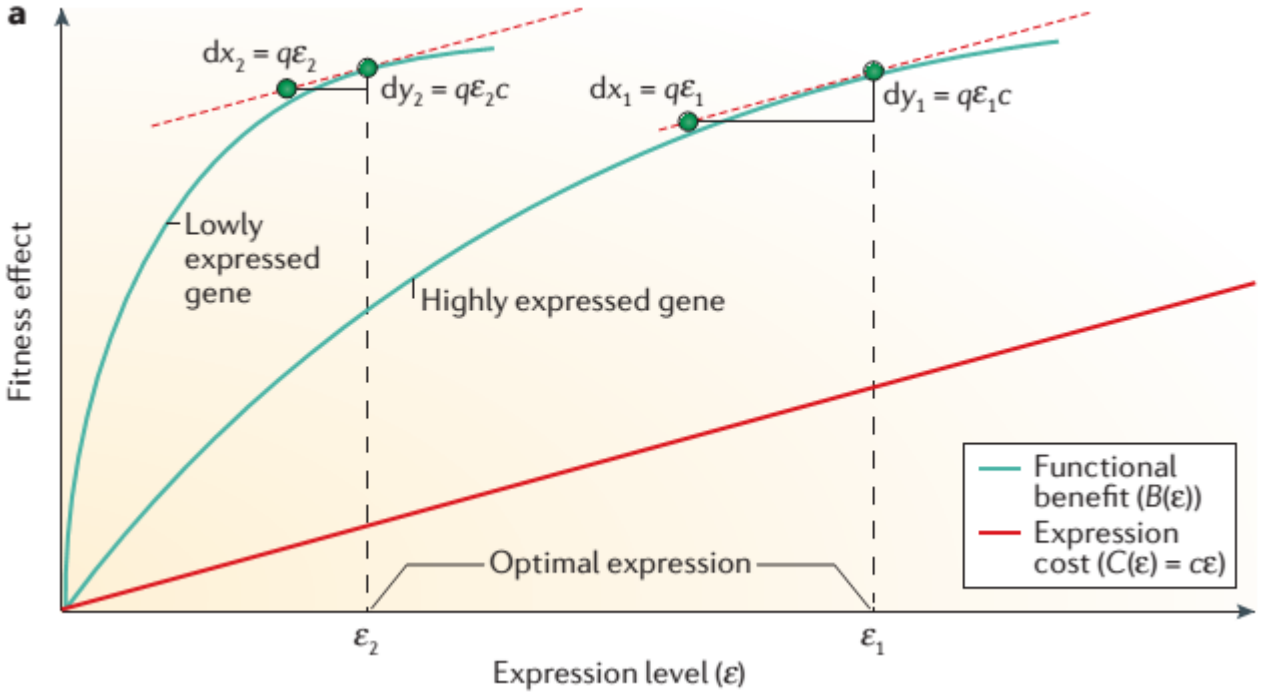
- mRNAs of highly expressed genes have stronger folding
  - Not a product sequence level differences
  - Random mutations more harmful
  - Lower substitution rate (confirmed in yeast)
- Why stronger mRNA folding?
  - Stronger the folding, slower the ribosome elongation
  - Higher translational fidelity

Fig 2.


mRNA
Translation
Protein
Folding
Folded protein
Protein–protein interaction
Translational error
Correctly translated
Mistranslated
Correctly folded
Misfolded
Functional interaction
Misinteraction
Misfolding avoidance hypothesis
Selection against errors in translation and folding
Misinteraction avoidance hypothesis
Selection against errors in translation and protein–protein interaction
mRNA folding requirement hypothesis
Selection against errors in translation
Number of misfolded molecules
Number of misinteracting molecules
Requirement for translational fidelity
Strength of selection against misfolding
Strength of selection against misinteraction
Required elongation speed
Translational robustness and folding stability
Propensity for misinteraction
mRNA folding strength
Expression level
Selection
E–R anticorrelation

Expression cost hypothesis

- C : Cost, B: Benefit, ϵ: abundance/expression
- Optimal abundance: $B'(\epsilon) = C'(\epsilon)$
- Mutation that decreases activity by a fraction q => loss of qϵ molecules



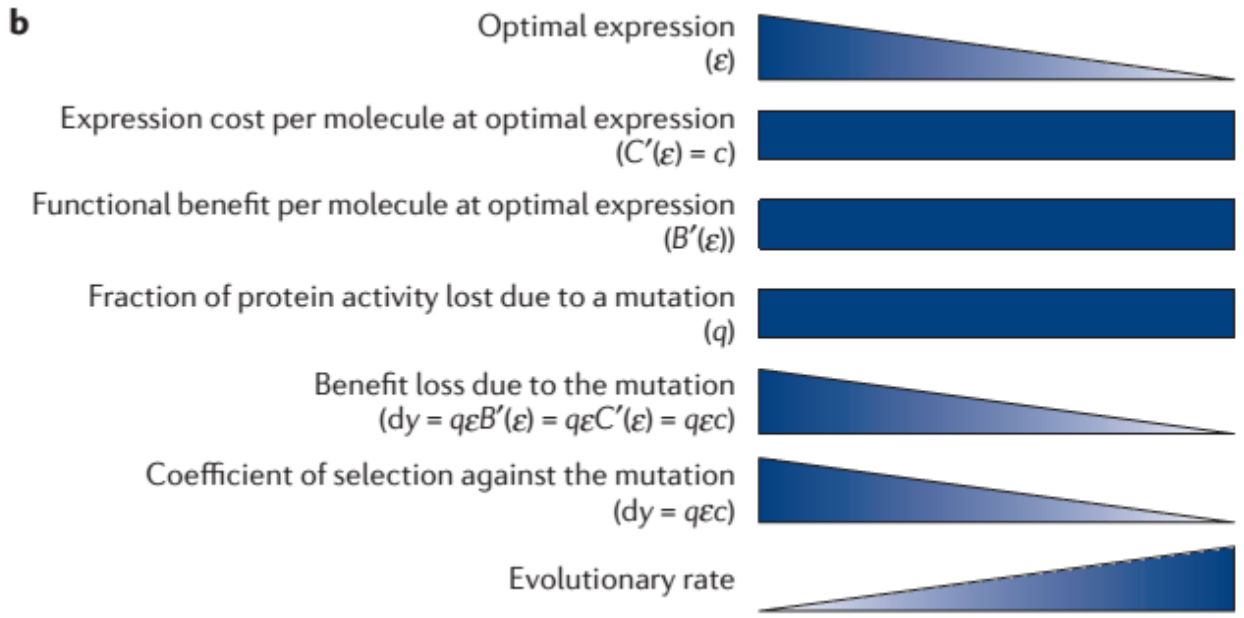


Fig 3.

Expression cost hypothesis

- Lacking empirical evidence
  - How to quantify expression cost?
  - How does cost scale with protein length?
  - What if cost per molecule is not constant for proteins?

Table 1 | **Correlates of the protein evolutionary rate**

Correlates	Properties of faster evolving proteins	Organisms	Refs
Gene expression level	Lower expressions	Bacteria, archaea and eukaryotes	9,14,34, 37
Functional importance	Lower importance and higher dispensability	Yeast and mammals	27,30–32
Expression breadth among tissues	Lower expression breadth and higher tissue specificity	Mammals	32,100
Expression timing in development	Expression in late embryogenesis and adulthood	Zebrafish	101
Promoter and gene body methylation	Higher levels of promoter methylation but lower gene body methylation levels	Mammals	102
Chaperone targeting	Higher levels of chaperone targeting	Bacteria and eukaryotes	103,104
Protein subcellular localization	Higher tendency to be extracellular	Yeast and mammals	105
Codon usage bias	Weaker codon usage bias	Bacteria and eukaryotes	14
Distance from the origin of replication	Larger distance	Bacteria and archaea	106, 107,108
Pleiotropy	Lower pleiotropy	Eukaryotes	57
Protein–protein interaction network properties	Lower connectivity, closeness and betweenness	Eukaryotes	58,109 ,110
Metabolic network property	Lower flux and connectivity	Yeast	111
Regulatory network properties	Higher centrality	Yeast	112
Targeting by microRNAs	Fewer types of targeting microRNA	Mammals	55
Gene compactness	Shorter introns and untranslated regions	Mammals	32
Protein length	Longer proteins	Yeast and mammals	113,114
mRNA folding	Weaker mRNA folding	Bacteria and eukaryotes	12
GC content	Lower GC content	Mammals	113
Domain structure	Lower density of domains	Animals and plants	54
Protein disordered regions	More-disordered regions	Bacteria and eukaryotes	115
Protein structural designability	Higher inter-residue contact density and higher fraction of buried sites	Yeast	114
Protein conformational diversity	Lower conformational diversity	Mammals	116

Broader implications

- Functional constraint
  - Functional importance only a minor determinant
  - Creation of toxicity more important
- Misfolding/misinteraction in genetic diseases
- Integrative approach to study multiple factors constraining evolutionary rates