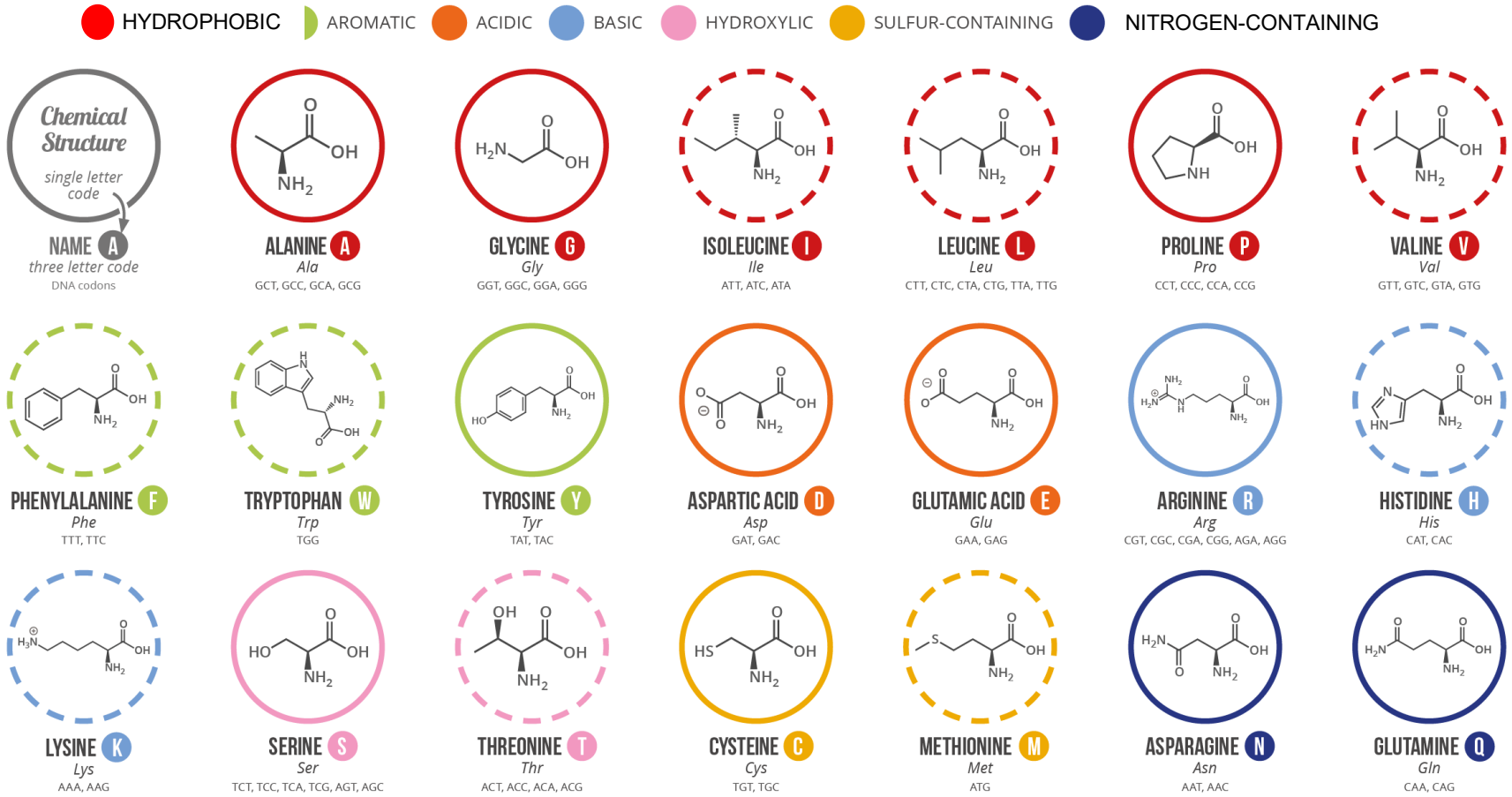
General Issues in Protein Evolution

- The diverse structural / biophysical features of the biological amino acids.
  - Basics of protein structure.
  - Evolutionary order of emergence of amino acids.

- The challenges of protein folding and stability.
  - Determinants and limits to the rate of unassisted folding.
  - The margin of folding stability.

- Constraints on amino-acid sequence evolution.
  - Functional significance.
  - Surface vs. core residues.
  - Expression level.

The Twenty Natural Biological Amino Acids

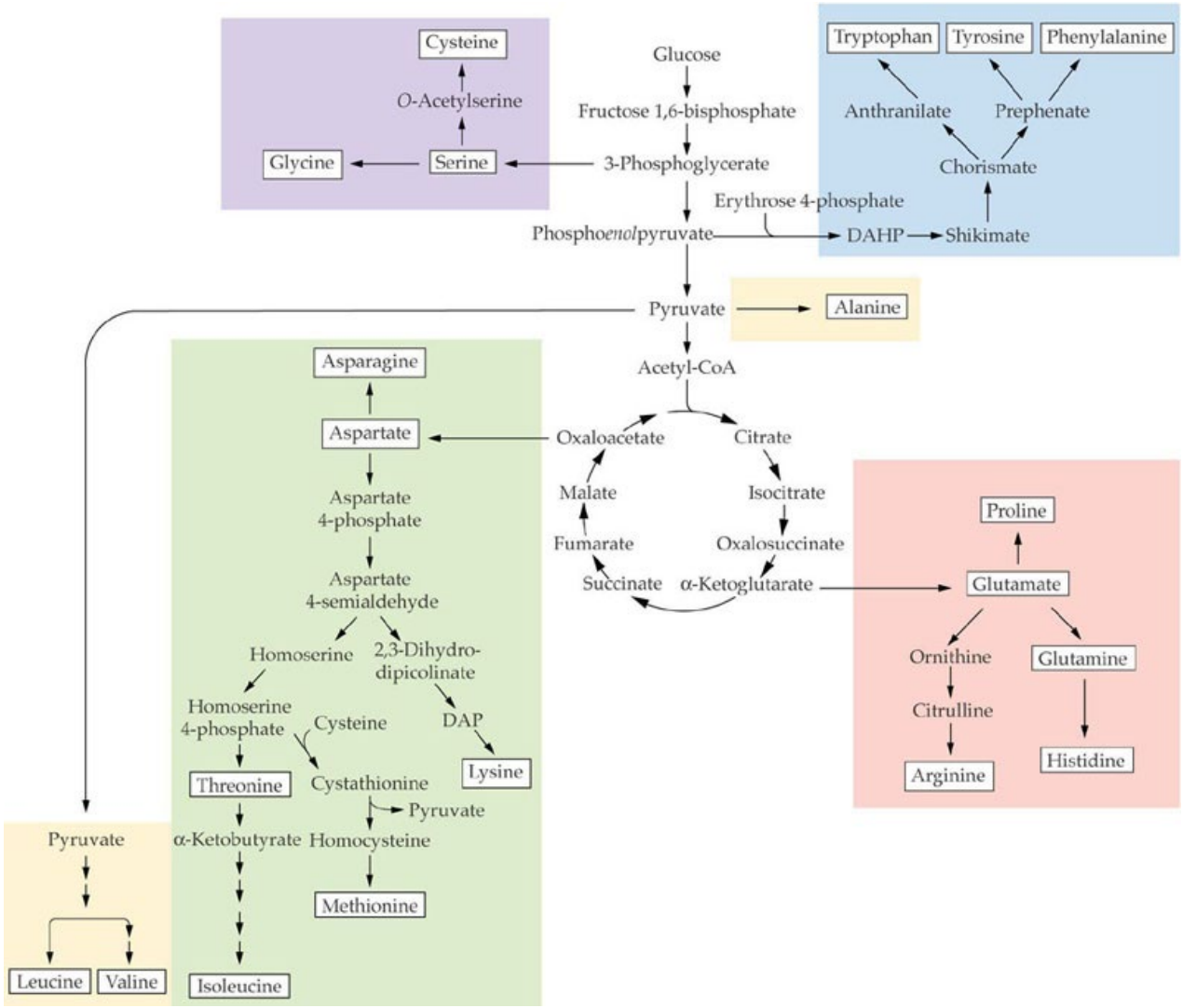


Diverse Physical / Chemical Properties of Amino Acids: dictates where they are deployed in proteins, and the consequences of mutations

Amino acid (abbrv.)	Polarity	Charge	MW	Hydropathy	Interface	GC	Cost
Alanine (Ala, A)	nonpolar	0	89	2.11	0.01	0.83	13
Arginine (Arg, R)	polar	+	174	-4.32	-0.09	0.83	22
Asparagine (Asn, N)	polar	0	132	-4.88	-0.27	0.17	12
Aspartic acid (Asp, D)	polar	−	133	-3.29	-0.75	0.50	10
Cysteine (Cys, C)	nonpolar	0	121	1.53	1.04	0.50	25
Glutamic acid (Glu, E)	polar	−	147	-2.26	-0.79	0.50	11
Glutamine (Gln, Q)	polar	0	146	-4.07	-0.41	0.50	12
Glycine (Gly, G)	nonpolar	0	75	0.20	-0.18	0.83	14
Histidine (His, H)	polar	+	155	-3.49	0.12	0.50	33
Isoleucine (Ile, I)	nonpolar	0	131	4.24	1.11	0.11	30
Leucine (Leu, L)	nonpolar	0	131	4.24	0.91	0.38	32
Lysine (Lys, K)	polar	+	146	-0.27	-1.18	0.17	28
Methionine (Met, M)	nonpolar	0	149	1.91	1.01	0.33	30
Phenylalanine (Phe, F)	nonpolar	0	165	2.64	1.27	0.17	59
Proline (Pro, P)	nonpolar	0	115	3.75	-0.18	0.83	16
Serine (Ser, S)	polar	0	105	-2.82	0.14	0.50	14
Threonine (Thr, T)	polar	0	119	-1.83	0.10	0.50	15
Tryptophan (Trp, W)	nonpolar	0	204	1.83	0.79	0.68	76
Tyrosine (Tyr, Y)	polar	0	181	-0.31	0.88	0.17	55
Valine (Val, V)	nonpolar	0	117	4.09	0.76	0.50	26

- Polar side-chains “prefer” water.
- Hydrophobic residues – key to protein folding, membrane proteins.
- Two cysteines can produce covalent disulfide bonds.
- Basic residues – important for binding to nucleic acids.

Near Universal Biosynthetic Pathways for Amino Acids



Davis (1999) Hypothesis for the Serial Additions of Amino Acids to the Proteome

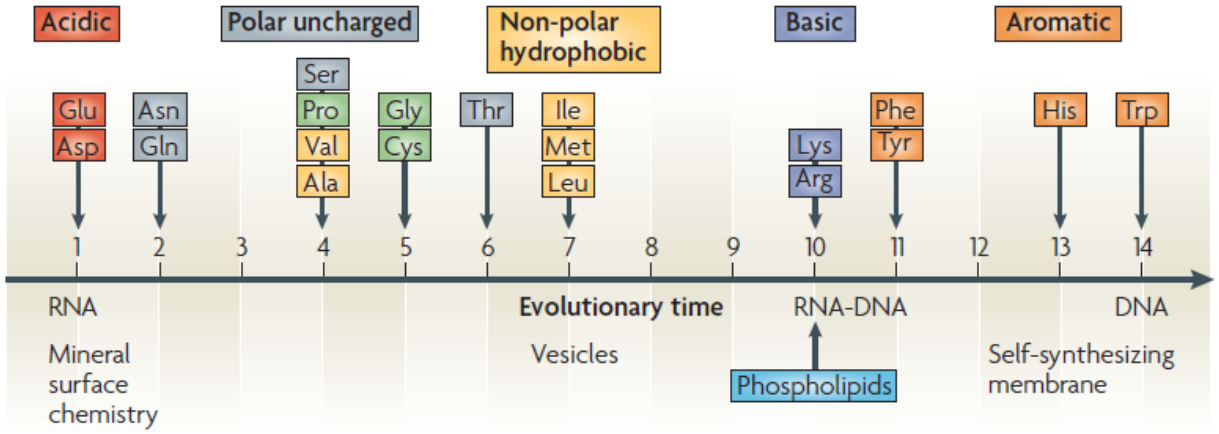


Figure 1 | **Biosynthetic pathways for amino acids, phospholipids and central metabolism.** It has been proposed by Davis[22,26] and many others that primitive biochemical pathways existed before a genetic-based system. These include the reductive carbon cycle (the equivalent of the tricarboxylic acid (TCA) cycle working in reverse), the reductive pentose pathway and the central trunk (proposed to be a remnant of the formose cycle), which together make up the central biochemical pathway (CBP). Davis proposes that the emergence of genetically encoded amino acids correlates with the number of chemical reactions from the CBP that are required to generate each amino acid (evolutionary steps 1–14). Acidic amino acids (Glu and Asp) are close to the CBP (1–2 reactions), whereas aromatic residues (such as Trp) require up to 14 steps to be synthesized and may have appeared later. The hydrophobic amino acids that could associate with membranes required four steps. The synthesis of phospholipids requires 10 steps and, therefore, self-synthesizing membranes might only have arisen after this point. Davis provides detailed analysis to argue that the evolutionary appearance of the different amino acids correlates well with the emergence of their corresponding triplet codes; this implies co-evolution of biochemistry and the genetic code, an idea extensively championed by Wong[60]. Davis also identified an 11-amino acid sequence in the FtsZ–tubulin family that he mapped to his evolutionary stage 7.5, which could support a role for this protein in cellularization. Gly, Cys and Pro cannot easily be placed into any of the five categories shown.

Griffiths, 2007, Nature Rev. Mol. Cell Biol.

Dispensability of Amino Acids in Today's Proteins

- A number of proteins in prokaryotes are entirely devoid of basic residues, Lys and Arg (McDonald and Storrie-Lombardi 2010).

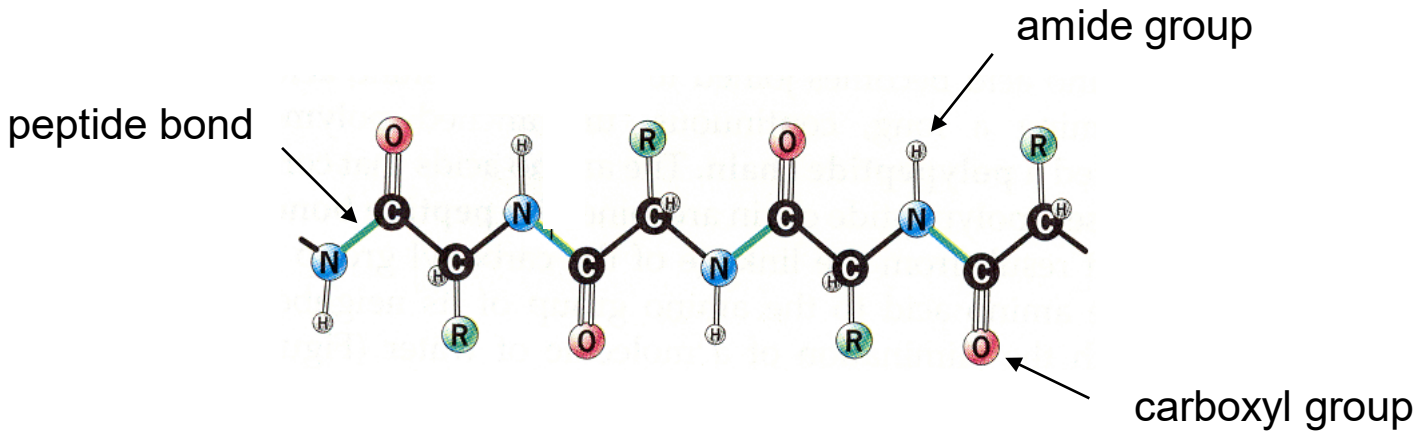
- A 213-residue protein involved in pyrimidine biosynthesis has been modified to function in the absence of seven AAs, with 188 positions being occupied by just nine Aas (Akanuma et al. 2002).

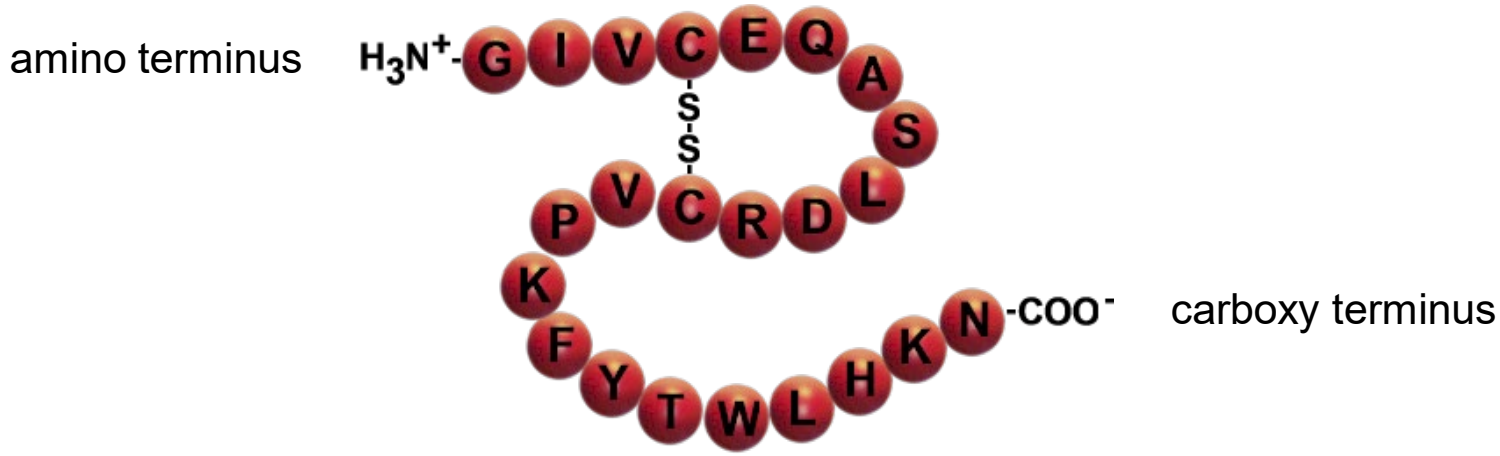
- A simplified version of archaeal chorismate mutase has been engineered to contain just nine AAs (MacBeath et al. 1998; Walter et al. 2005).

- Antifreeze protein in a flounder fish with only seven different residues (Sicheri and Yang 1995).

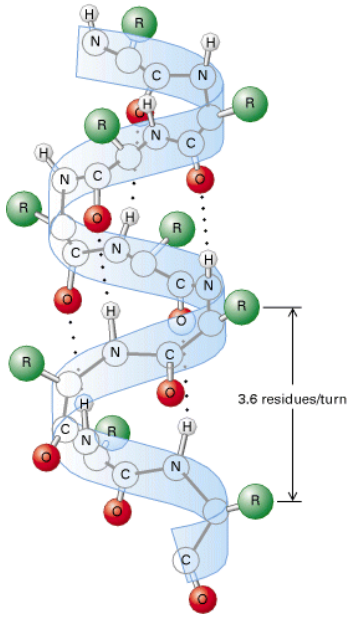
- Random-sequence proteins constructed from a 12 AA alphabet are more soluble on average than natural 20-AA proteins (Tanaka et al. 2010).

Proteins consist of primary chains of amino acids, which generally fold into secondary subunits, such as helices and sheets, which further arrange into tertiary globular structures essential for function.



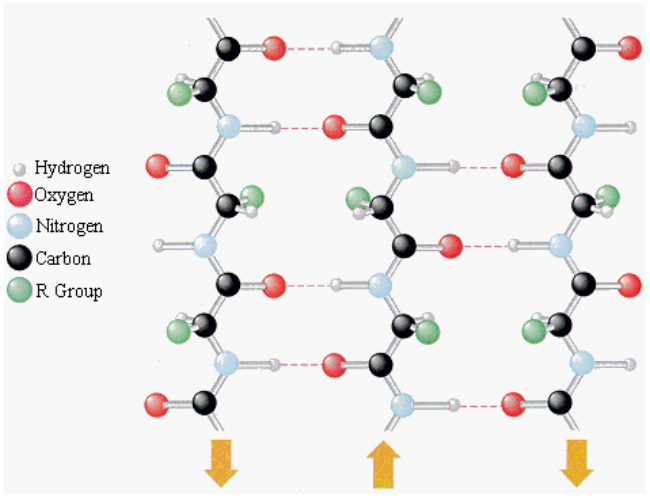


Alpha helix: N-H group of every AA donates a hydrogen bond to the backbone C=O group of the AA four residues earlier.

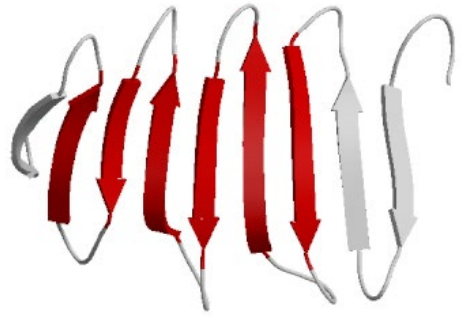


- Total chain length is typically 10 to 15 residues.
- Met, ala, leu, glu, and lys have high helix-forming potential.
- Gly has poor helix-forming propensities.
- Pro kinks a helix because it cannot donate an N-H bond.

Beta sheet: multiple strands are connected laterally by a network of hydrogen bonds.



Anti-parallel Sheet



- Strands are typically 3 to 10 residues long.

Protein Lengths in Eukaryotes Are Substantially Expanded

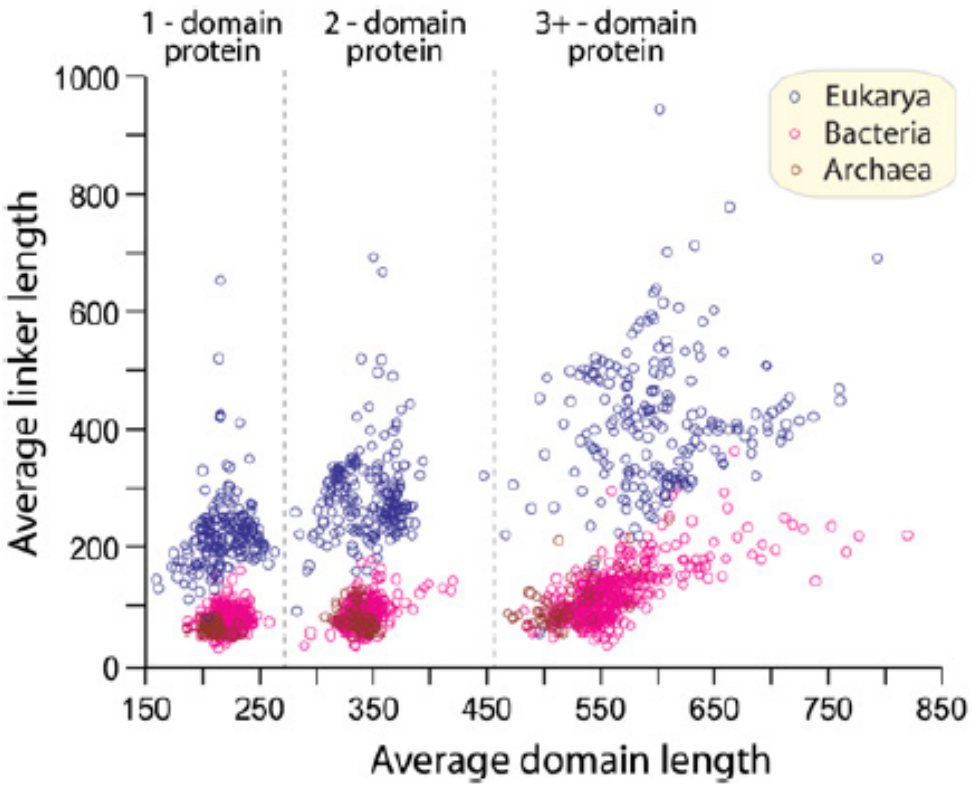


Fig. 1. Plot shows average lengths (amino acid numbers) of domains and linkers in 745 genomes, including 215 Eukarya (blue circles), 478 Bacteria (pink circles), and 52 Archaea (brown circles). Mean values of proteins with different domain numbers within the same genome could be separated well (dash lines) because of the increasing aggregate lengths.

- Functional domain sizes are roughly constant across prokaryotes and eukaryotes.

- Linkers, any non-domain sequence (which may be used for binding other proteins or nucleic acids), are substantially expanded in eukaryotes.

Wang et al. (PNAS, 2011)

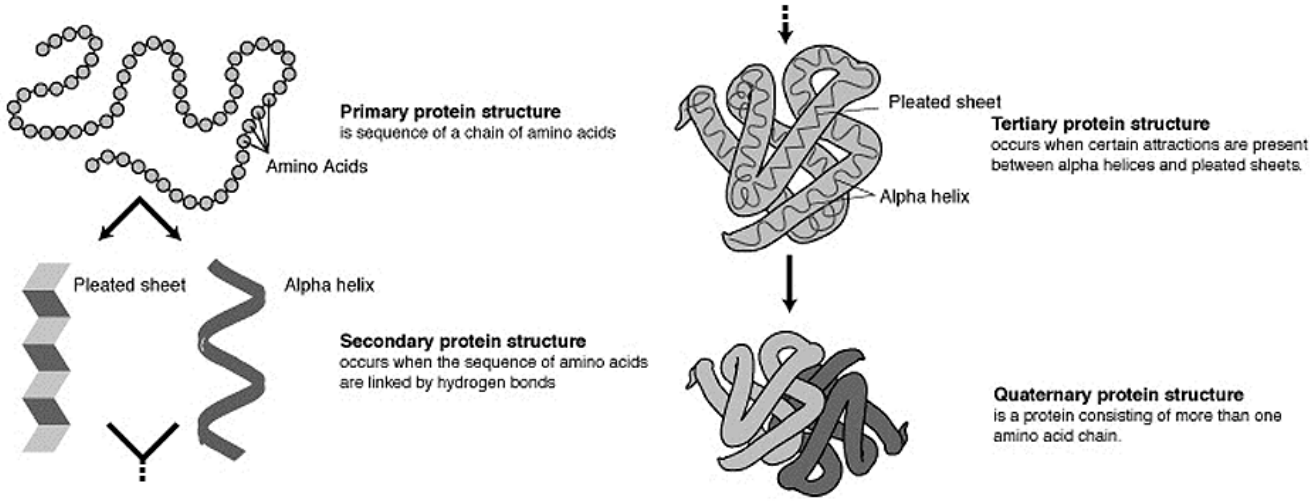
Central Challenges for Assembling and Maintaining Productive Proteins

- **Structural stability and proper folding** – the idea that the information for folding a protein is entirely contained within its amino-acid sequence is sometimes referred to as the “second half of the genetic code”.

- **Avoidance of aggregation** – proteins typically tend to aggregate only with themselves or very similar chains.

The Protein-folding Problem

- Protein folding involves a balance between intramolecular hydrogen bonding between polar residues (e.g., alpha helices and beta sheets), and the hydrophobic effect between nonpolar core residues, which expels water.
- Backbone hydrogen bonds are stabilized by the removal of nearby water molecules, which is accomplished by wrapping such structures with hydrophobic residues.

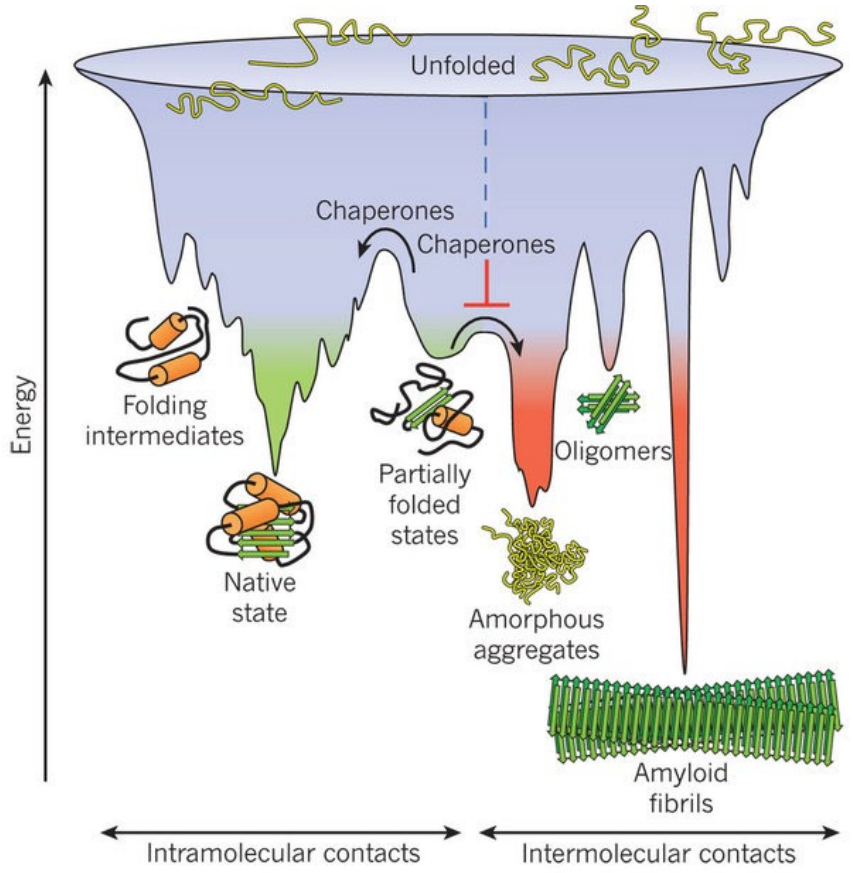


Degeneracy of the Protein-folding Code: many sequences can fold to the same native conformation.

- The “Levinthal Paradox” – a protein cannot possibly attain its native state by an exhaustive random search; for a protein of just 150 residues, this would take longer than the age of the universe.

- Good folders have large energy gaps between the lowest energy state and the next lowest energy misfold.

- Have similar structural features in different proteins arisen by divergence of a single ancestral structure, or by convergence from dissimilar sequences/structures?



Protein Folding Rates Vary by Orders of Magnitude, to a Large Extent Based on Chain Length

- The folding rate declines approximately exponentially with the square root of the chain length, nearly six orders of magnitude with a 10-fold increase in chain length.
- Beyond a chain length of ~300 residues, folding cannot be achieved in a reasonable time without some form of assistance (chaperones).

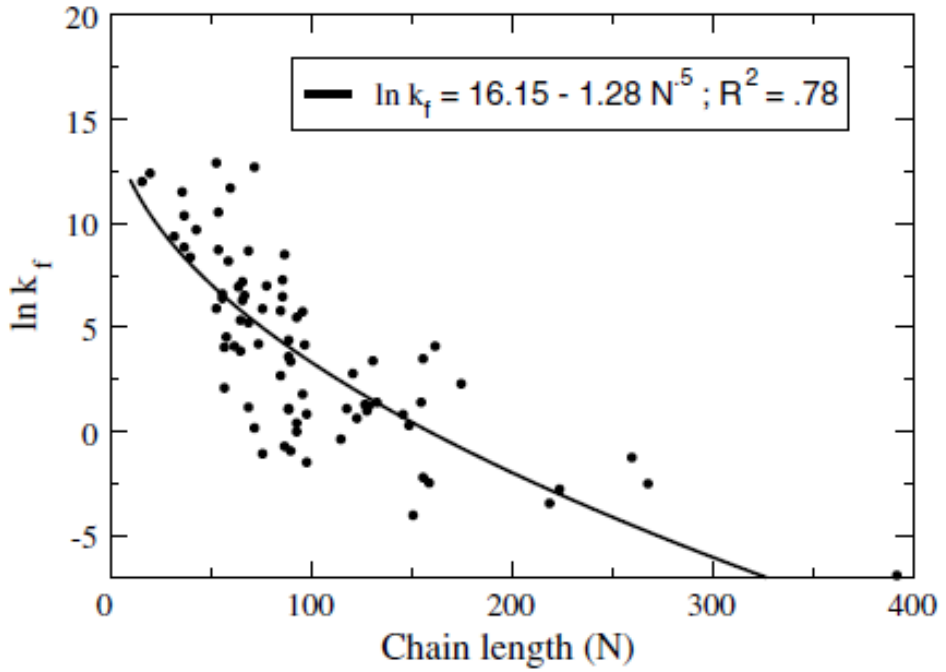


Fig. 7. Chain-length dependent fit (Eq. 16 using result from ref. 61) to the folding rates of 80 proteins including two-state and multistate folders (62). Data in circles and fit in solid line.

Dill et al. (PNAS, 2011)

Environmental Modulation of Folding-Stability Evolution: Enrichment of a Subset of AAs with Increasing Temperature

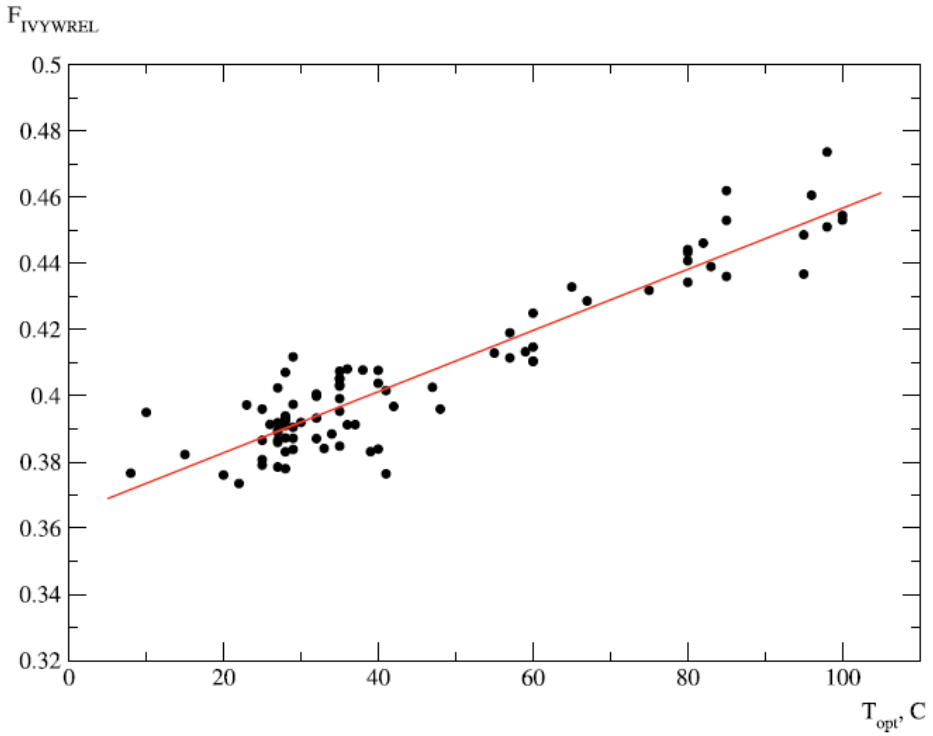


Figure 1. Correlation between the Sum *F* of Fractions of Ile, Val, Tyr, Trp, Arg, Glu, and Leu (IVYWREL) Amino Acids in 86 Proteomes and the OGT of Organisms T_{opt}

The linear regression (red line) corresponds to the correlation coefficient $R = 0.93$. The OGT T_{opt} (in degrees Celsius) can be calculated from the total fraction *F* of IVYWREL in the proteome according to $T_{opt} = 937F - 335$. By construction, the IVYWREL set is the most precise predictor of OGT among all possible combinations of amino acids; other combinations statistically yield a larger error of prediction of OGT.

doi:10.1371/journal.pcbi.0030005.g001

- All amino acids in this set are serviced by Type-I tRNA synthetases. Could these have been the first set of AAs?
- Mixture of features of AA side chains: Ile, Val, Trp, and Leu are hydrophobic; Tyr is polar; Arg and Glu are charged.
- The use of both hydrophobic (IVLW) and charged (RE) AAs is thought to jointly stabilize the native state and destabilize misfolded states.

Zeldovich et al. (2007, PLoS Comp. Biol.)
Berezovsky et al. (2007, PLoS Comp. Biol.)

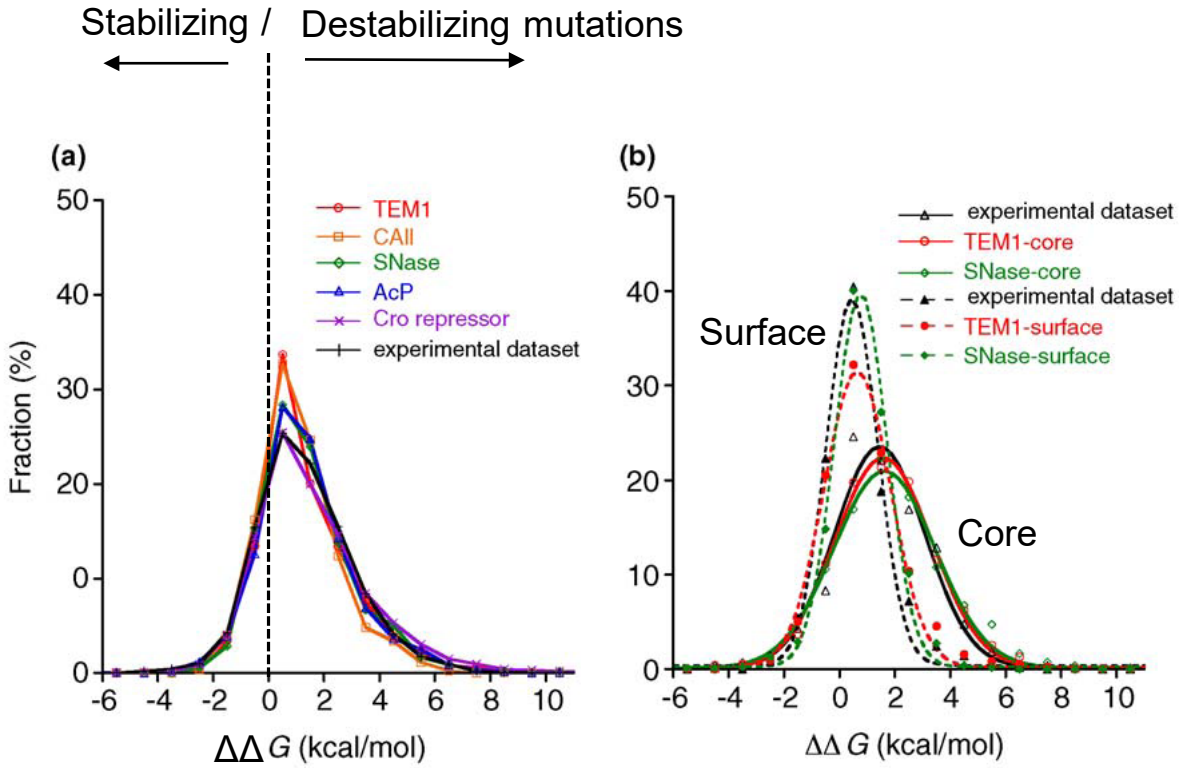
Proteins Are Typically on the Margin of Stability

- The overall stability of most proteins is on the order of ΔG = 10 kcal / mol.

$$P_{\text{nat}} = e^{-\beta\Delta G_i}/(1 + e^{-\beta\Delta G_i})$$

- ΔΔG for amino-acid substitution mutations is often > 2 kcal/mol, which is near the point at which protein stability is compromised.

- Mutations to surface residues are less destabilizing than those to the core.



The universal distribution of stability effects of mutations [31]. ΔΔG values are presented in histograms using 1 kcal/mol bins. **(a)** The predicted ΔΔG values by FoldX for all possible mutations in many proteins (shown are few characteristic examples), and the experimentally measured ΔΔG values for 1285 mutations, all give similar asymmetric distributions with larger destabilizing shoulders (ΔΔG > 0). **(b)** Separated ΔΔG distributions of core and surface residues. Residues were divided according to their accessible surface area (ASA) values, and the ΔΔG values for all possible mutations were arranged in histograms and fitted to a single Gaussian.

Tokuriki et al., 2007, J. Mol. Biol.; Tokuriki and Tawfik, 2009, Curr. Opin. Struct. Biol.

Substantial Interspecific Variation in Folding Stability (resistance to unfolding once correctly folded)

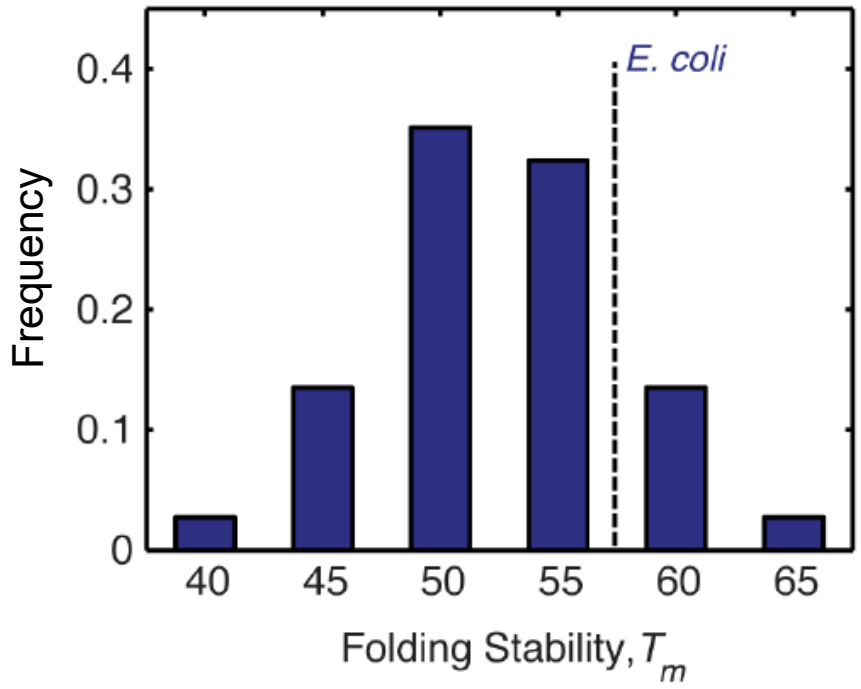
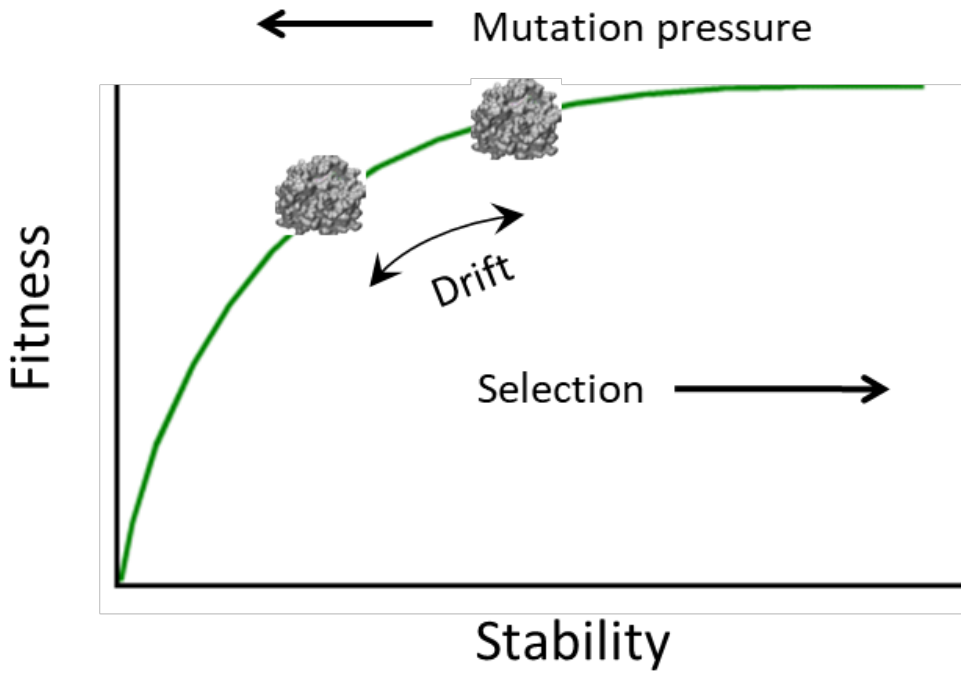


Figure 9.8. The distribution of folding stabilities of the enzyme dihydrofolate reductase for 36 species of mesophilic bacteria, measured as the midpoint temperature (°C) required for 50% unfolding *in vitro*. From Behrstein et al. (2015).

Selection-Mutation Balance and the Margin of Protein Stability

- Protein stability is deemed to be positively associated with fitness, as destabilized proteins are prone to loss of function, aggregation, and/or direct toxicity.

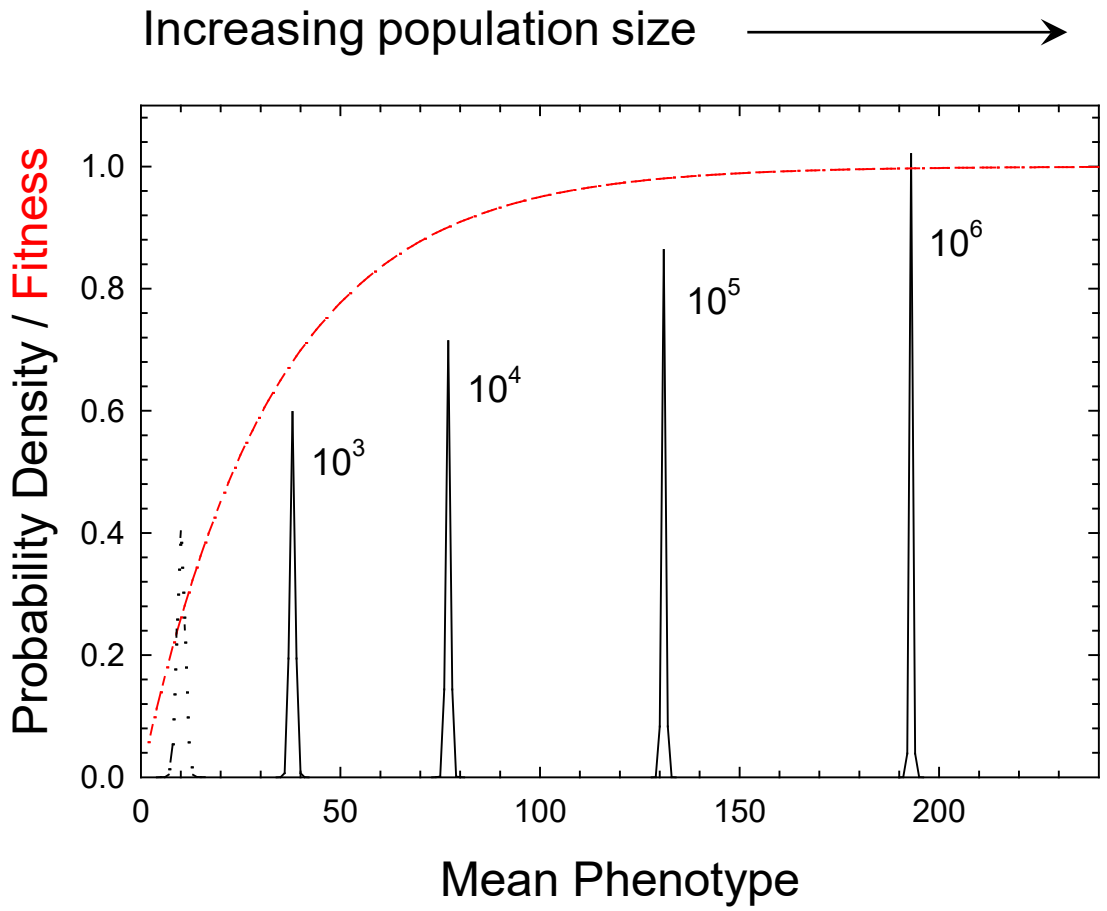
- One potential explanation for marginal stability is that overly rigid proteins compromise protein function, but this argument is inconsistent with observations indicating that proteins engineered to have higher stability often have normal enzyme function.



- Folding rates and stability are potentially lower in eukaryotes than prokaryotes.

- Despite their high level of refinement, the functionality of proteins has not reached the limits set by biophysics: catalytic rates can be improved by the use of noncanonical amino acids.

Long-term Evolutionary Distributions of Folding Stability in Populations with Different Sizes



- Biased mutation toward destabilizing residues will shift the steady-state distributions further to the left.

Primary Determinants of Amino-Acid Sequence Evolution

- Surface vs. core residues.
- Expression level: avoidance of misfolding and aggregation.
- Functional significance.
- Mutation bias.

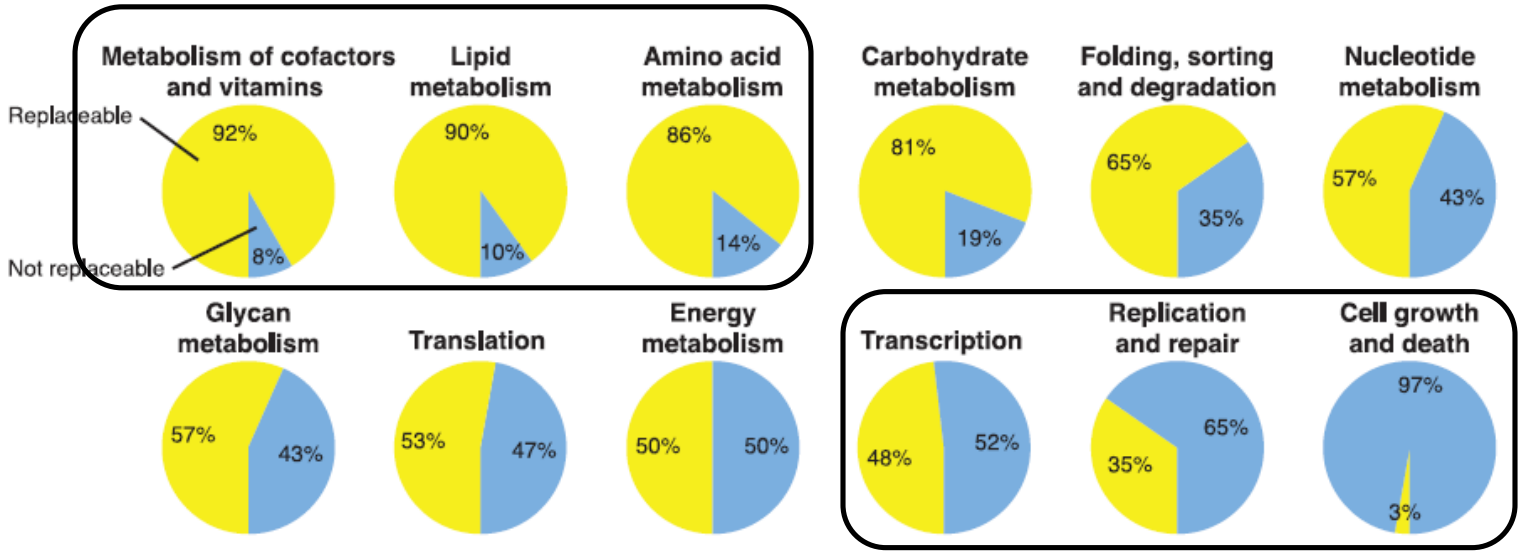
Proteins Often Have Conserved Functions After Millions of Years and Substantial Sequence Divergence:

"Humanized" Proteins in Yeast (Kachroo et al. 2015)

- Of over 400 different human proteins expressed in yeast (lineages that separated over a billion years ago), nearly half were able to complement the absence of the native gene.

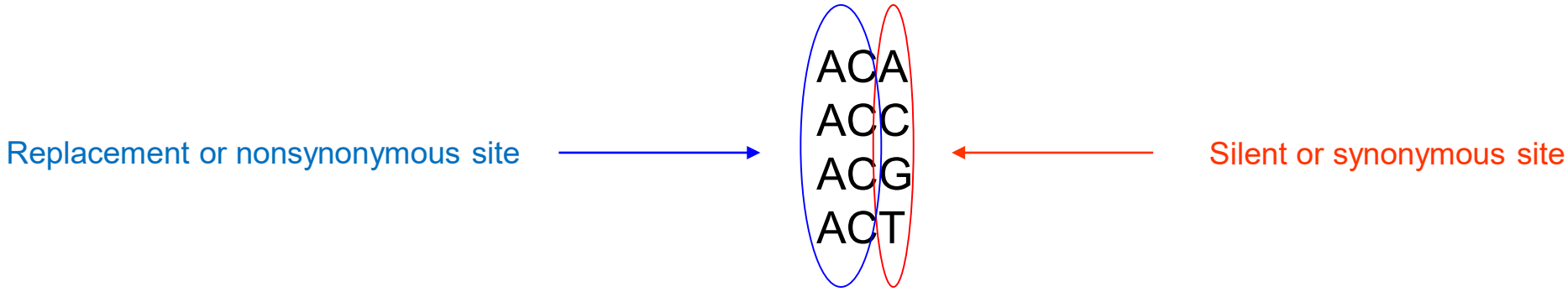
- ~50% of genes involved in transcription, ~65% involved in DNA replication and repair, and nearly all involved in cell growth and death are unable to complement, suggesting substantial changes in the proteins most closely related to fitness.

- Between 60 and 90% of genes involved in various aspects of metabolism do complement.



Evolutionary Analysis: Stringency of Selection on AA Replacement Mutations From Observations on Silent and Replacement Sites

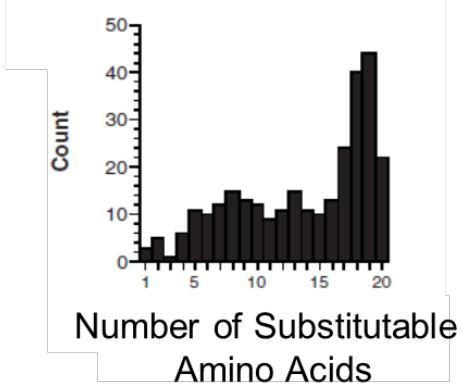
- Nucleotide sites in coding DNA can be subdivided into classes thought to be neutral vs. under selection.
- dN / dS is the ratio of substitution rates at Nonsynonymous and Synonymous sites.

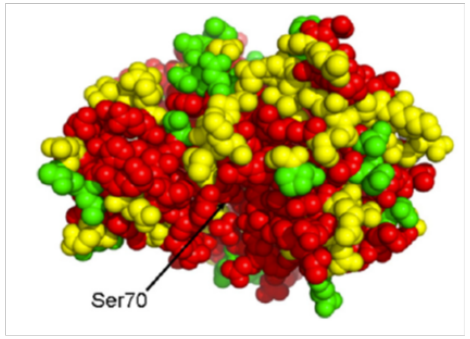


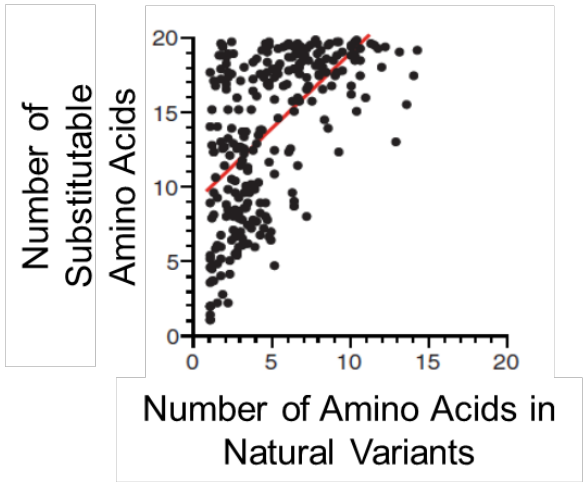
The Four Threonine Codons

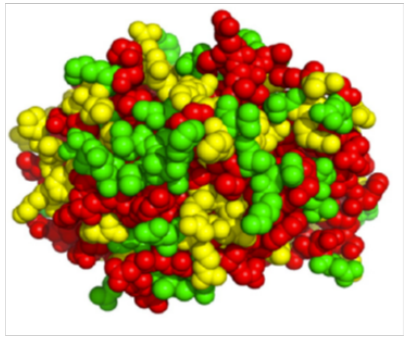
- Interpreted as the fraction of AA-altering mutations capable of fixing in nature, i.e., as the “effectively neutral” fraction. Typically, have genome-wide values of 0.1 to 0.3.
- Caveats: 1) selection on silent sites; 2) averages over all sites; 3) saturation of changes at silent sites.

Direct assay of effects of AA exchanges: β-lactamase – the effects of all 19 AA substitutions at all sites.





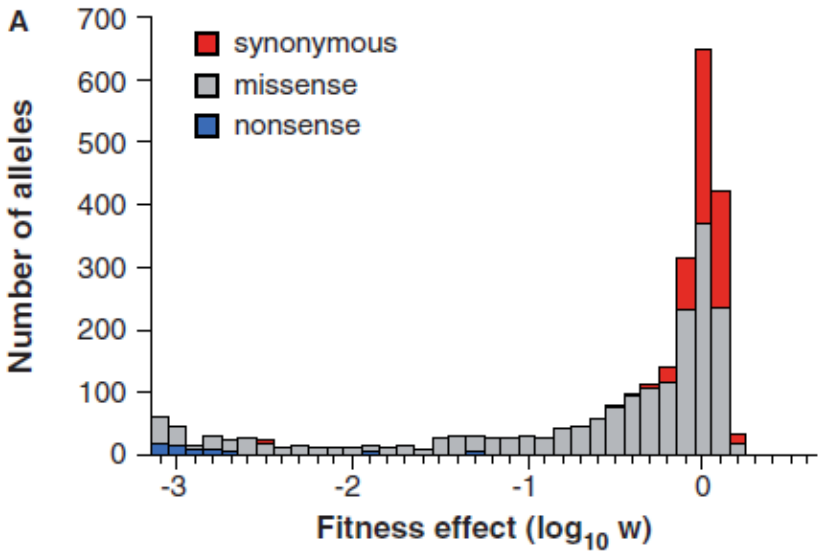




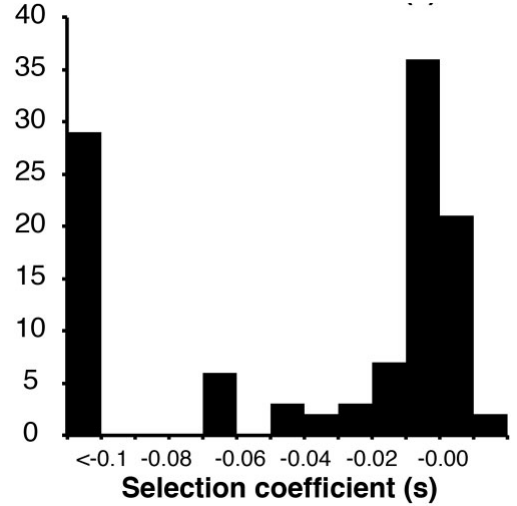
- For most sites, >5 of 19 changes have no *observable* effects on ampicillin resistance.
- Surface residues are more accepting of changes.
- Weak relationship between exchangeability at a site and variation observed in nature.
- Overall distribution of fitness effects is bimodal, and silent substitutions are not always neutral.
- Very few mutations improve performance.

Deng et al. (2012)

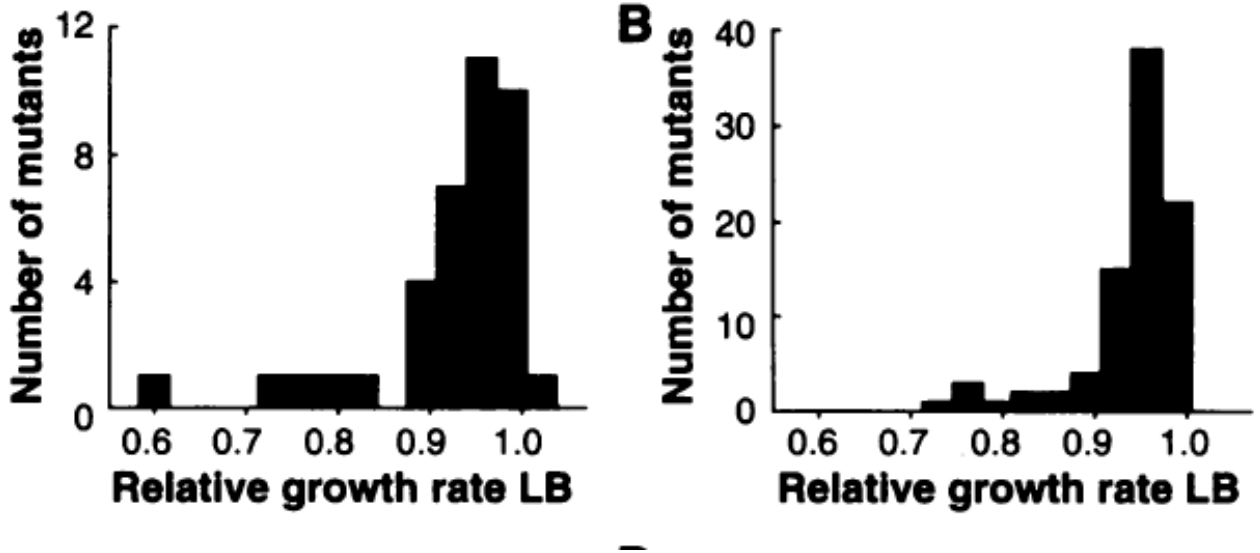
β-lactamase, Firnberg et al. (2014)



Proteins involved in arabinose metabolism, *Salmonella*, nonsynonymous, Lind et al. (2016)



Ribosomal proteins, *Salmonella*, synonymous (left) and nonsynonymous (right), Lind et al. (2014)



- Distribution of fitness effects for AA exchanges generally has mode near zero, and only a small fraction of favorable changes.
- However, the details of the distribution in the range of very small effects, which is most critical to evolutionary theory, is uncertain.

Within-gene Variation: Surface Residues Evolve More Rapidly Than Core Residues

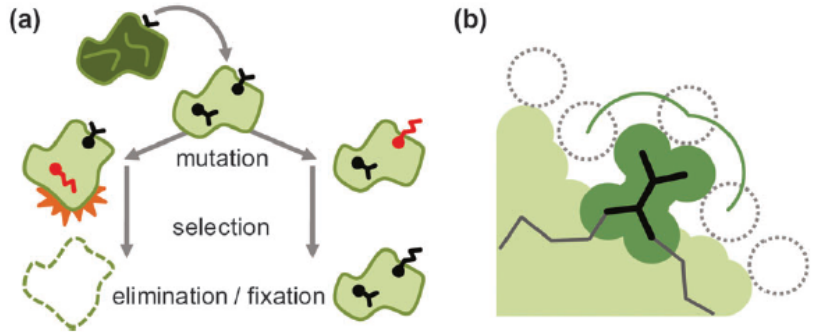
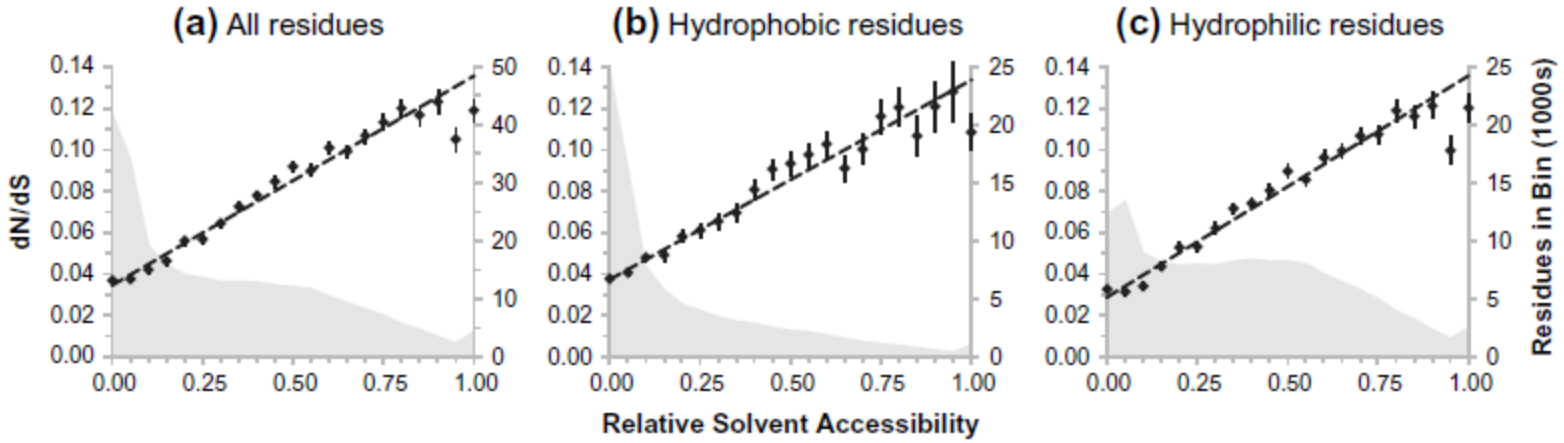


FIG. 1.—Evolutionary implications of the residue microenvironment. (*a*) A cartoon diagram of a protein is shown in cross section, highlighting two residues: one completely buried in the core and another partially exposed to solvent. Mutations occur at both sites, but whether or not they go to fixation depends on the properties of the residue microenvironments and their effects on the overall stability and function of the protein. (*b*) One quantitative property of the residue microenvironment is shown in detail. Here, a solvent molecule (dotted circle) traces the solvent accessible surface of a particular residue, shown in heavy wireframe.



Franzosa and Xia (2009)

One of the strongest predictors of evolutionary rate is the expression level of a gene.

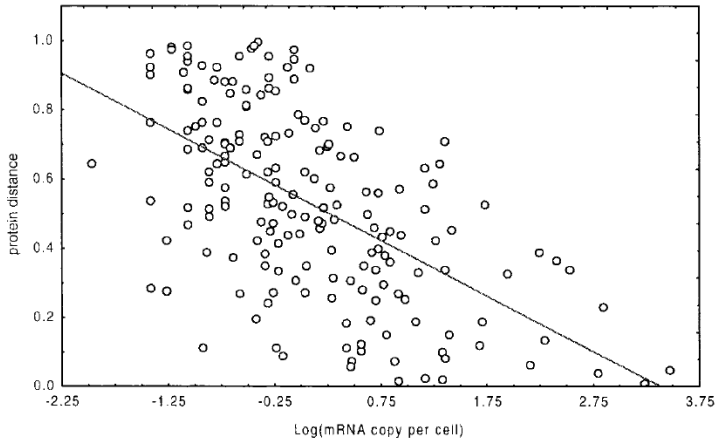
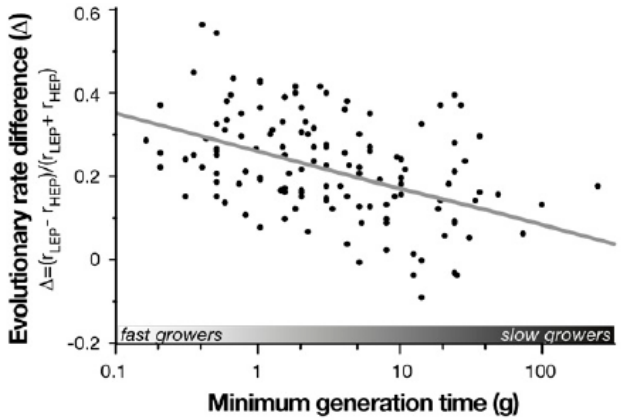


FIGURE 1.—Correlation between the mRNA level and protein distance in duplicated genes of yeast. $r = -0.584$, $P < 10^{-6}$.

Pal et al. (2001, Genetics)

Microbes with rapid generation times have higher disparities in rates of evolution of highly vs. lowly expressed genes.

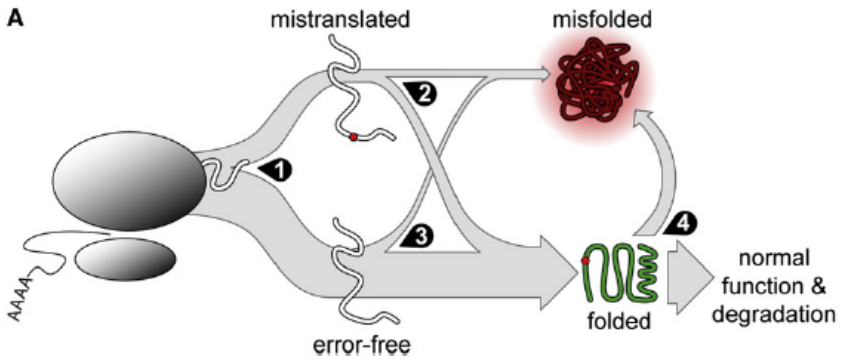


Scaled difference in evolutionary rate between classes of highly vs. lowly expressed genes.

Vieira-Silva et al. (2011, PNAS)

Mistranslation-induced protein misfolding as an explanation for the inverse relationship between expression level and evolutionary rate.

Stronger response in microbes than in vertebrates →



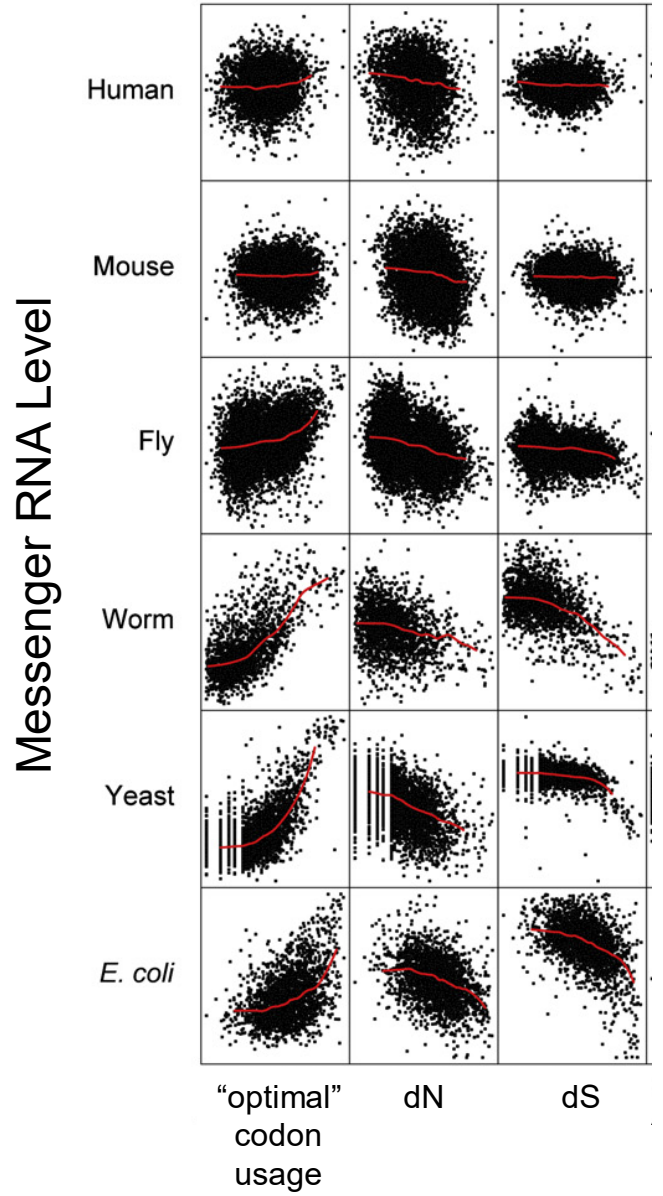


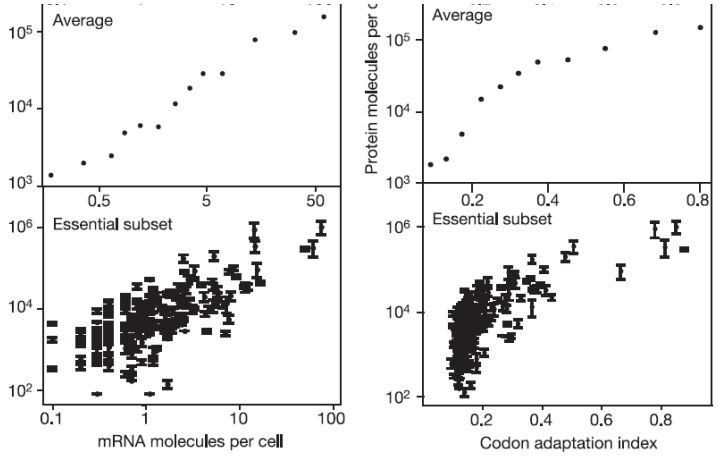
Figure 3. The Misfolding Hypothesis

(A) Outcomes of translation. Most proteins exit the ribosome (left) with no errors (bottom), but a substantial proportion contain at least one error (top). The probability of misfolding after correct translation is lower than after erroneous translation (center). Some proteins attain native state but then improperly unfold (right). Natural selection can act at four points: at (1), to reduce the frequency of translation errors in certain proteins; at (2), to reduce the proportion of error-containing proteins which misfold; at (3), to reduce the number of error-free proteins which misfold; and at (4), to reduce the number of proteins (with or without errors) that improperly unfold.

Drummond and Wilke (2008, Cell)

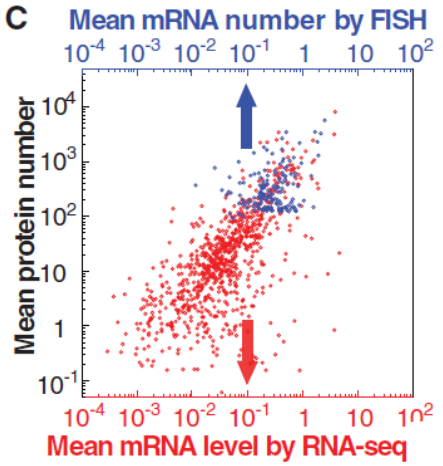
Standing levels of mRNAs are correlated with intracellular protein abundance, when averaged over populations of cells.

Yeast



Ghaemmaghami et al. (2003, Nature)

E. coli

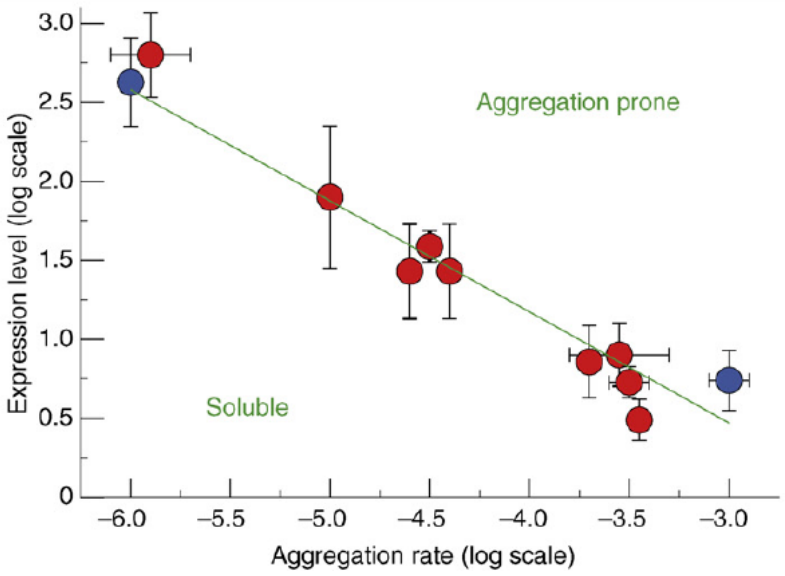


Taniguchi et al. (2010, Science)

- On average, ~5000 protein molecules / mRNA.

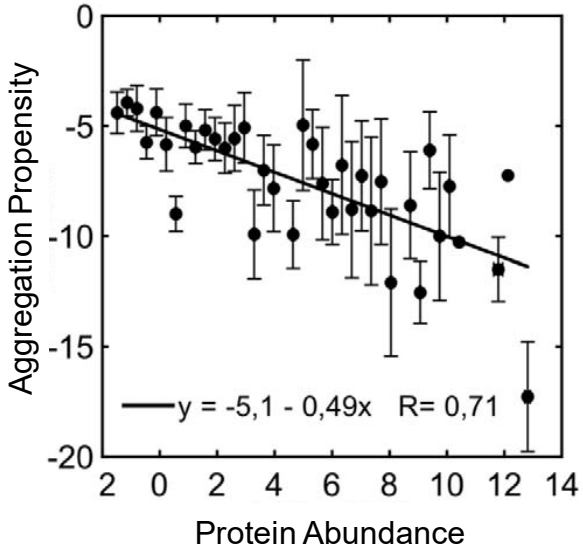
Proteins With High Expression Rates Are Structured to Avoid Aggregation

Human



Tartaglia et al. (2007, Trends Biochem. Sci.)

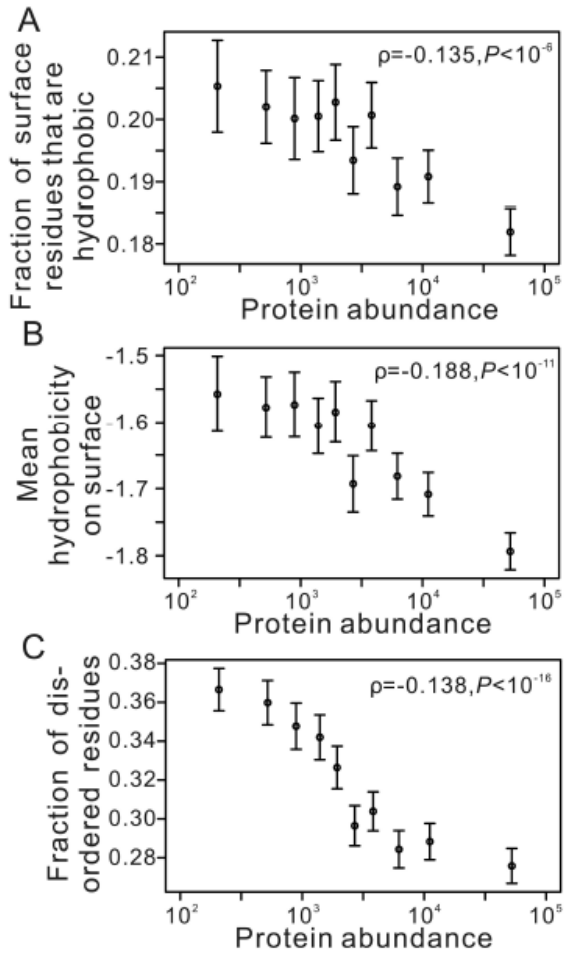
E. coli



De Groot and Ventura (2010, PLoS ONE)

Protein-misinteraction avoidance causes highly expressed proteins to evolve slowly

Jian-Rong Yang[1,2], Ben-Yang Liao[2,3], Shi-Mei Zhuang[1], and Jianzhi Zhang[2,*] PNAS (2012)



The misfolding hypothesis provides an inadequate explanation for the reduced rate of evolution of highly expressed genes.

- The predicted adhesiveness of yeast proteins decreases with expression level.

The Evolved Relationship Between Protein Abundance and Stickiness is Higher in Microbes

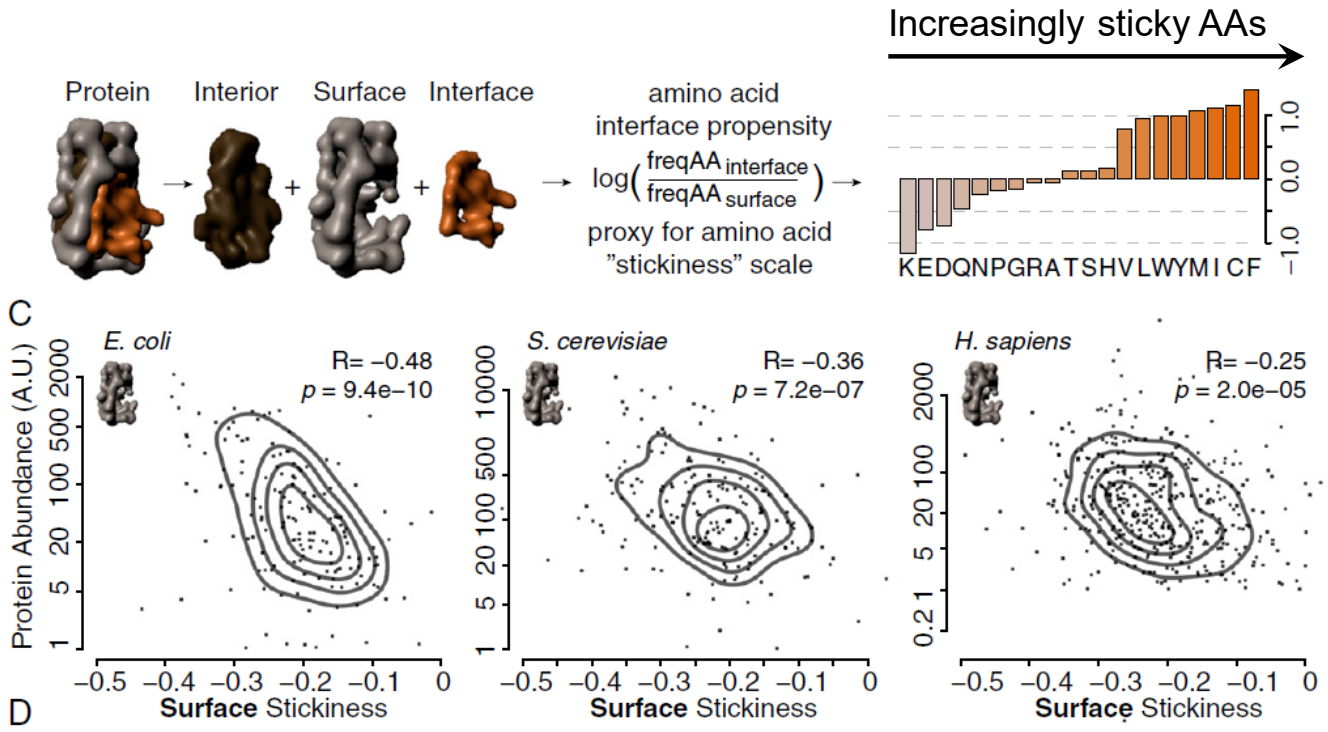
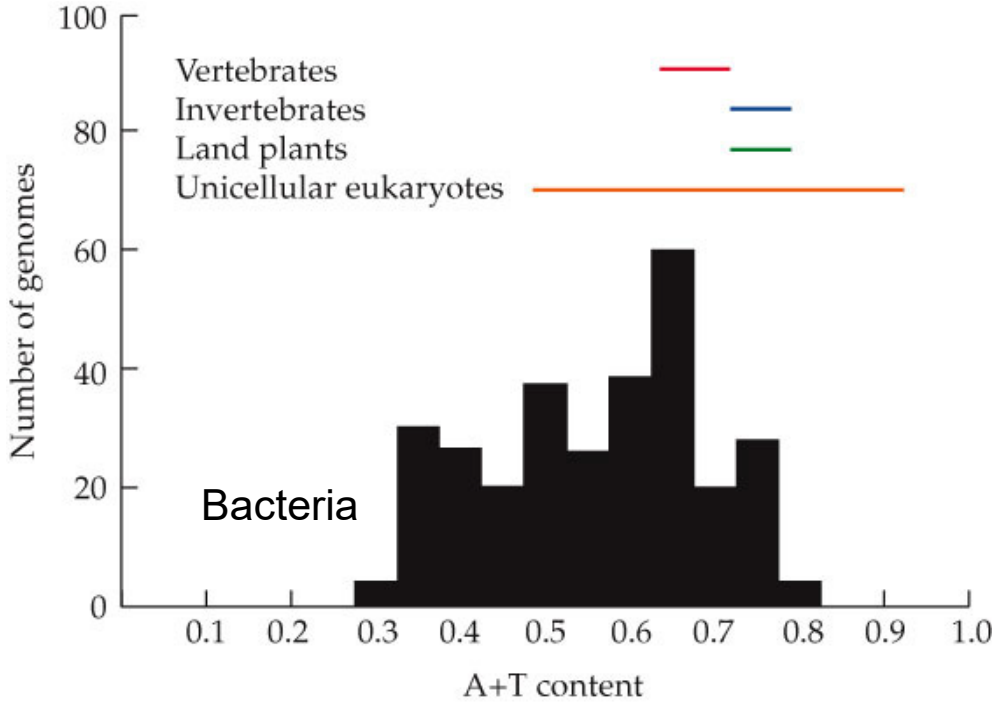


Fig. 1. The solvent-accessible surfaces of high-abundance proteins are enriched in nonsticky amino acids compared with low-abundance proteins. (*A*) Illustration of the approach taken in this study. (*B*) We first define a stickiness scale for each amino acid using its interface propensity. The propensity is defined by the log ratio of amino acid frequencies at interfaces versus surfaces. The definition of the structural regions used is explained in more detail in

Levy et al. (2012, PNAS)

What Drives Nucleotide Composition Bias?

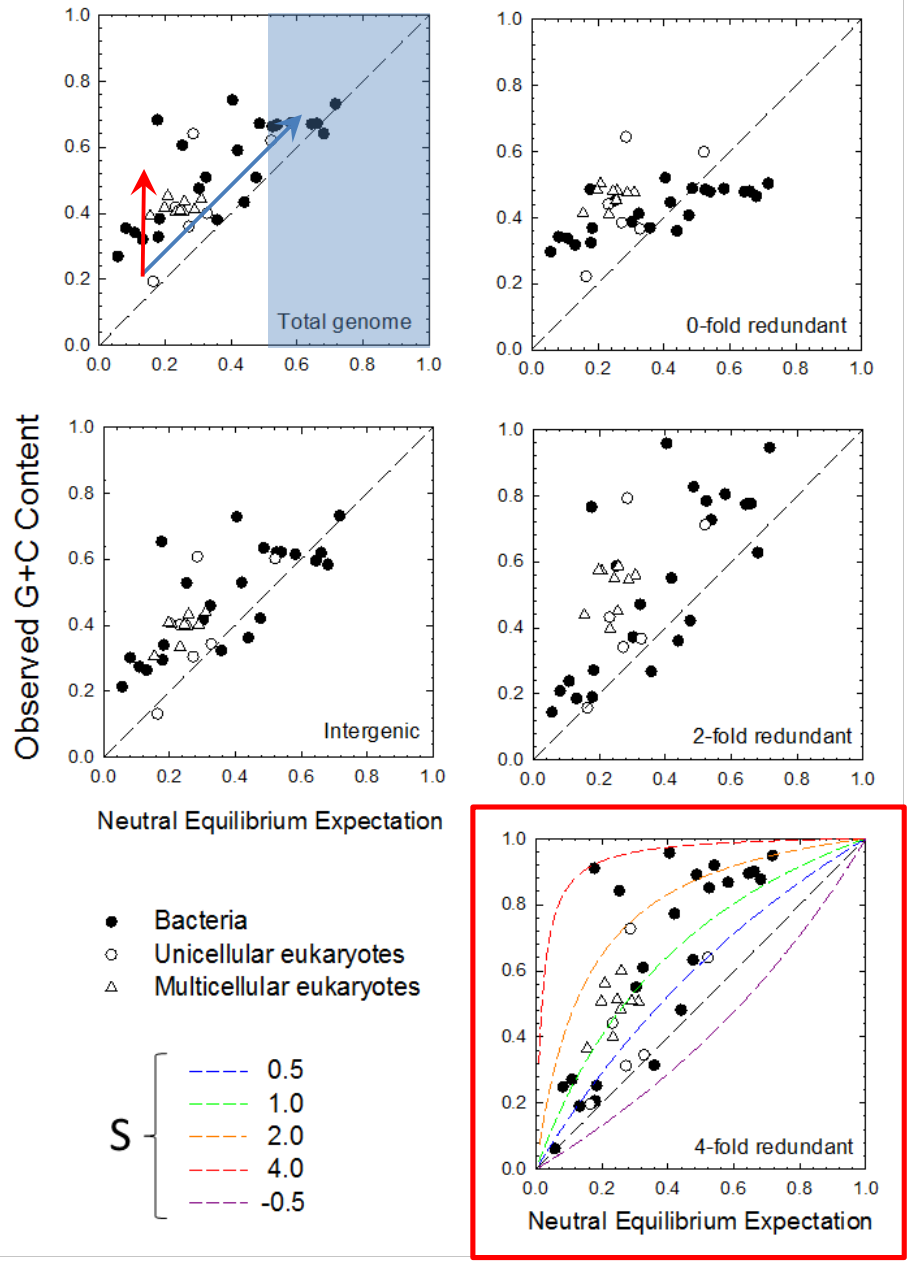


- Variation in mutation bias?
- Variation in selection pressures?
- Both?

- Neutral mutation expectation for AT composition = $u_{GC>AT} / (u_{GC>AT} + u_{AT>GC})$.
- Deviation from the neutral expectation is governed by the selection bias parameter, $S = \exp(2N_e s_{GC>AT})$.

What Drives Nucleotide Composition Bias?

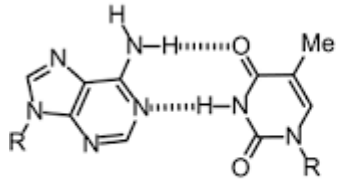
- Mutation pressure is usually in the AT direction.
- Genome-wide nucleotide composition is driven by both mutation and selection bias.
- Nearly universal selection pressure towards G + C.
- GC selection bias is stronger at 4-fold redundant (silent) sites than anywhere else in the genome.



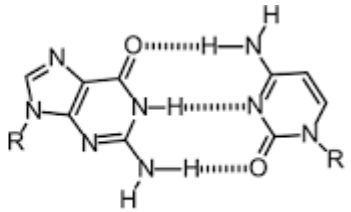
Long et al. (2017, Nature Ecol. Evol.)

Why Near Universal Selection for G:C?

- G:C bonds are more stable than A:T bonds.



A·T base pair



G·C base pair

- G:C pairs are slightly less expensive to synthesize A:T pairs.

- Amino acids with G/C rich codons tend to be energetically less expensive, and also less hydrophobic (less sticky).

Amino acid (abbrv.)	Cost
Alanine (Ala, A)	13
Arginine (Arg, R)	22
Asparagine (Asn, N)	12
Aspartic acid (Asp, D)	10
Cysteine (Cys, C)	25
Glutamic acid (Glu, E)	11
Glutamine (Gln, Q)	12
Glycine (Gly, G)	14
Histidine (His, H)	33
Isoleucine (Ile, I)	30
Leucine (Leu, L)	32
Lysine (Lys, K)	28
Methionine (Met, M)	30
Phenylalanine (Phe, F)	59
Proline (Pro, P)	16
Serine (Ser, S)	14
Threonine (Thr, T)	15
Tryptophan (Trp, W)	76
Tyrosine (Tyr, Y)	55
Valine (Val, V)	26

A General View of Protein-Sequence Evolution

- AA altering mutations frequently have context-dependent fitness effects, whereby the incorporation of earlier mutations can dictate whether specific subsequent substitutions are deleterious, beneficial, or neutral.
    - As a consequence, the fixation of effectively neutral (but mildly deleterious) mutations can pave the way for the future fixation of compensatory mutations that otherwise would not be beneficial.

- Over time, a series of such subtle remodeling events can lead to the entrenchment of previously neutral AA substitutions to the point of becoming near essential to protein functionality.

- Such progressive changes may appear to lead to major adaptive fixations, the entire process may unfold with only minor consequences for overall fitness.
    - This view of protein evolution is compatible with long-term wandering of AA sequences along the drift barrier.