æ

12. THE PROTEIN WORLD

17 May 2022

The vast majority of cellular functions involve the use of one or more proteins. Although these biomolecules have a myriad of specialized functions, many of which are covered in subsequent chapters, we focus here on general issues. Proteins are composed of linear strings of amino-acid residues, parts of which generally assemble into simple secondary structures, e.g., helices and sheets held together by hydrogen bonds. These, in turn, arrange into tertiary (three-dimensional and frequently globular) architectures. Quaternary structures, the subject of Chapter 13, arise when separate proteins associate into higher-order assemblages via binding interfaces. Whereas most genomes encode for several thousand proteins, only a few hundred protein-coding genes are shared across all species (Harris et al. 2003; Koonin 2003), implying that lineage-specific gains and losses of genes are common.

Three general topics will be explored in this chapter. First, we consider the fundamental biochemical and biophysical properties of the twenty major amino acids that serve as the building blocks of virtually all proteins. It is highly unlikely that all twenty amino acids entered the biological world at the same moment of time, so it is of interest to consider the potential order of evolutionary entry, as well as the consequences of a presumably simpler early amino-acid alphabet.

Second, one of the central problems of protein science concerns the stable folding of proteins into their so-called native states. Levinthal (1968) famously pointed out that proteins longer than a few dozen residues cannot possibly examine all feasible configurations en route to final assembly, concluding that folding pathways must be guided by information in the primary amino-acid sequence. The underlying guidelines must operate on time scales short enough to enable rapid responses to gene-expression demands, and they must be accurate enough to ensure the proper assembly of catalytic sites and to avoid the energetic wastage associated with the management and disposal of improper assemblies. Poorly folded proteins impose the additional risk of initiating inappropriate aggregations with self and nonself proteins. Closely related to the problem of protein folding is the matter of stability once folded.

Third, in light of these features, we will review the evolutionary constraints on the amino-acid sequences found in different proteins, in different regions of proteins, and in different phylogenetic lineages. The central questions here concern the degree to which various pairs of amino acids are substitutable for each other, the extent to which evolution at one particular site is independent of that of others, and the overall capacity of natural selection to counter the relentless input of amino-acid altering mutations.

The Essential Features of Proteins

Proteins are composed of variable amino-acid chain lengths, parts of which typically fold into more compact localized domains. Average domain sizes are roughly constant across prokaryotes and eukaryotes, but the linkers between domains tend to average several-fold longer in eukaryotes, leading to $\sim 50\%$ longer total chain lengths in the latter (an average of ~ 530 residues in eukaryotes, and ~ 350 in prokaryotes; Wang et al. 2011; Rebeaud et al. 2021). Given that each amino acid is chemically unique with respect to molecular weight, charge, hydrophobicity, polarity, etc. (Table 12.1), when primary sequences are further combined into three-dimensional forms, the resultant combinatorial diversity endows protein repertoires with an essentially boundless array of structures and functions.

Each amino-acid consists of a central carbon atom attached to one hydrogen atom, one NH_2 (amino) group, one CO_2 (carbonyl) group, and a unique cognate side chain (Figure 12.1). Peptide chains are assembled by mRNA-translating ribosomes, with the amide group of each consecutive amino acid reacting with the carboxyl group of the adjacent member of the growing chain (Figure 12.2). Glycine has the simplest side chain, just a single hydrogen atom, and is therefore symmetrical and quite flexible. Two residues contain sulfur (cysteine and methionine), whereas several have side chains containing nitrogen, and serine and threonine side chains uniquely carry an OH group. Proline is exceptional in that the side chain is covalently bonded to the nitrogen atom of the peptide backbone, and as a consequence is the only amino acid lacking an amide hydrogen atom for use in hydrogen bonding. Alanine, valine, leucine, and isoleucine have simple side chains ending in CH_3, and like glycine and proline are highly hydrophobic.

Amino-acid composition. To a large extent, the different properties of amino acids dictate where they are deployed within proteins and define the biochemical and structural consequences of mutations. To acquire functionality, proteins need to achieve proper folds, which are strongly dependent upon backbone hydrogen bonds between residues (often located distantly on the polypeptide chain). In addition, hydrophobic residues, which avoid water molecules, tend to be buried within the cores of proteins. Exposure of hydrogen bonds and hydrophobic residues of protein cores leads to folding instability, which increases stickiness and the potential to engage in inappropriate protein-protein interactions. Thus, the surfaces of proteins are typically well wrapped with hydrophilic residues in ways that minimize the intrusion of water molecules into the core.

It is unlikely that the amino-acid alphabet had reached the current twenty-residue state when the first protein-based cells emerged some four billion years ago. This raises the question as to how much functional protein diversity might have been achieved in a setting involving a smaller number of amino acids. The potential seems large, given that many proteins in today's world do not contain the full set of twenty amino acids. An extreme case is an antifreeze protein in a flounder fish that contains only seven different residues (Sicheri and Yang 1995), and a number of proteins in prokaryotes are entirely devoid of basic residues (McDonald and Storrie-Lombardi 2010). Moreover, gene-sequence manipulations of modern-day proteins

show that, provided the catalytic site is not compromised, substantial reductions in the number of distinct amino acids used in primary sequences can be achieved without loss of function. For example, Akanuma et al. (2002) were able to modify a 213-residue protein involved in pyrimidine biosynthesis to function in the absence of seven amino acids, with 188 positions being occupied by just nine amino acids. A bovine pancreatic trypsin inhibitor sequence modified to contain $> 33\%$ alanine residues retained its native fold and functions (Islam et al. 2008). In addition, a simplified version of an archaeal chorismate mutase has been engineered to contain just nine amino acids (MacBeath et al. 1998; Walter et al. 2005). Several other such studies are reviewed in Longo and Blaber (2012) and Longo et al. (2013).

Although a diverse protein repertoire can be derived from a restricted set of amino acids, laboratory evolution experiments also suggest that enhanced enzyme efficiency would have been promoted by expansion of the amino-acid alphabet (Müller et al. 2013). Moreover, in experiments where the twenty canonical amino acids are supplemented with noncanonical forms, enzymes can be engineered to have still higher catalytic rates than found in natural populations (Windle et al. 2017; Zhao et al. 2020; de la Torre and Chin 2021), indicating that the canonical set of twenty amino acids upon which all life depends constitutes a constraint on natural selection's ability to promote proteins with optimal features.

Origin of amino acids. Given that the substantial differences among amino-acid features (e.g., positive vs. negative charge, hydrophilic vs. hydrophobic) define their potential contributions to various cellular transactions, an understanding of the temporal order of evolutionary incorporation of the amino acids into the early proteome might help clarify the origin of cellular features. All of the numerous attempts devoted to such inference rely on assumptions with tenuous validity, and the initial functions of some amino acids may have been totally unrelated to their use in today's proteins (e.g., charged amino acids might have been deployed to cell surfaces to improve adhesion to counter-charged surfaces). With these caveats in mind, the following is a brief survey of the conclusions reached by various approaches.

Davis (1999) postulated that the earliest arriving amino acids would be those with the simplest production mechanisms, i.e., with the fewest steps in today's biosynthetic pathways. Most amino-acid biosynthesis initiates at hubs of central metabolism – the citric-acid cycle, the pentose phosphate cycle, or the central trunk that connects the two, allowing the derivation of proximity measures for all twenty amino acids (Figure 12.3). For example, alanine, aspartic acid, asparagine, glutamic acid, and glutamine are just one to two steps removed from their metabolic-byproduct precursors, whereas biosynthesis of histidine, lysine, phenylalanine, tryptophan, and tyrosine requires 10 to 13 additional steps. Under Davis' hypothesis, the earliest amino acids were aspartic acid, glutamic acid, asparagine, and glutamine (for a variety of reasons, he viewed alanine as a later addition). These four building blocks are often referred to as the "nitrogen-fixing" amino acids as the first two are, respectively, produced by the addition of an amino group to oxaloacetate and α-ketoglutarate (both components of the citric acid cycle), with secondary amino additions then leading to asparagine and glutamine.

One concern with this type of reasoning is that variation exists in the pathways used in amino-acid biosynthesis by different species (Chapter 19), leaving the

generality of statements about the number of steps required in the production of various amino acids uncertain in ancestral pre-LUCA species. An additional concern is the observation that a number of prokaryotes are capable of producing certain amino acids "on demand" by converting one to another after loading onto a tRNA synthetase (the molecules that relay specific amino acids to cognate tRNAs). The conversion of glutamic acid to glutamine by addition of an NH_3 group is one such example.

An alternative approach to inferring the temporal ordering of amino-acid appearance relies on phylogenetic analysis. For example, if one is willing to assume that the amino-acid content of the most strongly conserved protein sequences across the Tree of Life reflects the availability of amino acids at the times of protein origin, one is led to conclude that alanine, glycine, aspartic acid, and valine were early arrivals, with cysteine, tryptophan, tyrosine, phenylalanine, glutamine, and glutamic acid being among the late arrivals (Brooks and Fresco 2002; Brooks et al. 2002). Using a rather different approach, involving simple pairwise comparisons of sequences in sister taxa with an outgroup to infer the directionality of amino-acid substitutions, Jordan et al. (2005) suggested a universal trend across the Tree of Life toward an increase in cysteine, methionine, histidine, serine, and phenylalanine, and a decrease in proline, alanine, glycine, and glutamic acid. The underlying assumption here is that amino acids that are declining in frequency represent the pool of early arrivals. A clear concern with these approaches is the assumption that there has been insufficient time in the history of life for the complete erasure of information on amino-acid compositions at the pre-LUCA stage. It is difficult to reconcile this view with the vast stretch of post-LUCA time and known rates of mutation (Chapter 4).

Despite the uncertainties in our ability to project backwards to the primordial amino-acid pool, by integrating the above sorts of analyses with empirical observations on the ease of synthesizing amino acids in potential settings for the origin of life, a loose argument has been made for a limited set of ten prebiotic amino acids: alanine, aspartic acid, glutamic acid, glycine, isoleucine, leucine, proline, serine, threonine, and valine (Higgs and Pudritz 2009; Longo and Blaber 2014). Notably absent from this list are two of the earliest arrivers under Davis' hypothesis, asparagine and glutamine. If roughly correct, this list of early amino acids has implications for the temporal ordering of the emergence of cellular biochemistry and the features of early proteins. For example, an absence of the basic, positively charged amino acids (arginine and lysine) likely would have limited the potential for intimate relationships between proteins and acidic nucleic acids (with negatively charged backbones).

Table 12.1. Properties of the twenty major amino acids (with conventional abbreviations). MW denotes the molecular weight in grams/mol. Hydropathy is recorded as the log of a coefficient that measures the propensity of a molecule to dissociate from water into a nonpolar solvent (Wolfenden et al. 2015). Interface denotes the log of the ratio of the incidence of an amino acid on interfaces to that on exposed surfaces of proteins; these numbers are taken from *E. coli,* although similar results are obtained in other species (Levy et al. 2012). GC is the average fractional G/C content within codons in the primary genetic code (Figure 12.1). Cost is the biosynthetic cost of a single amino-acid in units of ATP hydrolyses, which assumes a starting point of glucose, and includes both the loss of ATP generation due to the diversion of precursors (with ATP-generating power) and the direct use of ATP

in biosynthesis (derived in Chapter 17).

Amino acid	Polarity	Charge	MW	Hydropathy	Interface	GC	Cost
Alanine (Ala, A)	nonpolar	0	89	2.11	0.01	0.83	16
Arginine (Arg, R)	polar	+	174	-4.32	-0.09	0.83	31
Asparagine (Asn, N)	polar	0	132	-4.88	-0.27	0.17	16
Aspartic acid (Asp, D)	polar	−	133	-3.29	-0.75	0.50	14
Cysteine (Cys, C)	nonpolar	0	121	1.53	1.04	0.50	17
Glutamic acid (Glu, E)	polar	−	147	-2.26	-0.79	0.50	20
Glutamine (Gln, Q)	polar	0	146	-4.07	-0.41	0.50	21
Glycine (Gly, G)	nonpolar	0	75	0.20	-0.18	0.83	12
Histidine (His, H)	polar	+	155	-3.49	0.12	0.50	32
Isoleucine (Ile, I)	nonpolar	0	131	4.24	1.11	0.11	39
Leucine (Leu, L)	nonpolar	0	131	4.24	0.91	0.38	44
Lysine (Lys, K)	polar	+	146	-0.27	-1.18	0.17	36
Methionine (Met, M)	nonpolar	0	149	1.91	1.01	0.33	25
Phenylalanine (Phe, F)	nonpolar	0	165	2.64	1.27	0.17	62
Proline (Pro, P)	nonpolar	0	115	3.75	-0.18	0.83	26
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	20
Tryptophan (Trp, W)	nonpolar	0	204	1.83	0.79	0.68	71
Tyrosine (Tyr, Y)	polar	0	181	-0.31	0.88	0.17	57
Valine (Val, V)	nonpolar	0	117	4.09	0.76	0.50	31

Protein Folding and Stability

To acquire their enzymatic or structural features, individual polypeptide chains generally must undergo a developmental stage of folding into specific three-dimensional configurations. The overall architecture of an entire amino-acid chain is referred to as its tertiary structure, and the most appropriate functional configuration is referred to as its native state.

In the process of complete folding, numerous substructures are initially formed, the most common of which are α helices and β sheets. In α helices, the amide group of every amino-acid donates a hydrogen bond to the backbone carboxyl group of the amino acid four residues earlier in the polypeptide chain. Total helix-chain lengths are typically on the order of 10 to 15 residues (Figure 12.4). Methionine, alanine, leucine, glutamic acid, and lysine have high helix-forming propensities, whereas glycine is poor in this regard, and a proline residue will break or kink a helix because it cannot donate an amide hydrogen bond. In contrast, β sheets consist of sets of chains (each chain typically 3 to 10 residues long) held together by backbone hydrogen bonds between residues in adjacent chains. Such sheets can consist of parallel or anti-parallel chains, usually four to five but as many as ten, with the physical distance between hydrogen-bonding residues in the primary sequence depending on the length of the strands within the sheet.

Higher-order structures are commonly assembled from α helices and β sheets. For example, coiled coils result from the interlacing of two or three adjacent α helices, with appropriate spacing of hydrophobic residues. Helix-loop-helix repeats can yield a variety of different higher-order geometric forms, depending on the angular features

of the loop. The $(\beta\alpha)_8$ barrel, one of the most common enzyme folds throughout the Tree of Life, consists of eight alternating units of β strands and α helices, which fold to become an internal curved β sheet surrounded by α helices.

The reliance of almost all proteins on a moderate number of fold types is unlikely to simply be an evolutionary fossil of common ancestry. Rather, commonly observed folds appear to be natural outcomes of the fundamental features of peptide chains, including the intrinsic ability to hydrogen-bond and form hydrophobic associations with each other. Indeed, random sequences of amino acids (even those involving reduced sets of amino acids, including homopolymers) commonly generate stably folded proteins (Doi et al. 2005; Zhang et al. 2006; López de la Osa et al. 2007; Go et al. 2008; Labean et al. 2011). This suggests that the compact globular nature of proteins is an expectation based on physical properties, and hence need not be entirely a product of the guiding hand of natural selection (Alva et al. 2015). Thus, the majority of common folds in today's proteins were likely present even before the establishment of the full genetic code.

The rate of protein folding. Given the large number of fold types in the protein world, the specific folding pathways utilized by different proteins must be extraordinarily diverse. Nonetheless, considerable effort has gone towards identifying general solutions to the "protein-folding" problem that transcend the details of secondary structure. This is a highly technical field, far from fulfilling its ultimate goals, but enough quantitative information now exists to yield insight into the typical time scales and energetic forces involved in productive protein folding. We start by focusing on folding unassisted by outside factors, deferring until Chapter 14 consideration of the cellular mechanisms that have evolved to assist with the process.

One conceptual solution to the Levinthal paradox invokes the metaphor of a folding funnel, with an energetically favorable bias in the landscape of possible folds acting to progressively channel a protein towards the relatively stable native state (Dill and McCallum 2012; Englander and Mayne 2014, 2017; Wolynes 2015; Neupane et al. 2016). The hallmark of a stable protein is a well-packed hydrophobic core, nearly universally viewed as resulting from the favorable association of nonpolar surfaces in water, although the actual underlying molecular mechanisms driving such clustering remain unclear (Ball 2008; Snyder et al. 2011). Additional factors involved in protein folding and stability include hydrogen bonds in α helices and β sheets, electrostatic interactions between residues with different charges, and disulfide bonds between cysteine residues. Consistent with there being a multifactorial basis, the folding times of most proteins are quite resilient to sequence changes, with random mutagenesis (sometimes involving multiple residues) generally causing no more than a ten-fold increase in the mean folding time, and as many as half of residue changes causing reduced folding times (Kim et al. 1998; Plaxco et al. 2000).

Under this general model of folding, the approach to the native state can be viewed as a series of stochastic samplings of alternative states, with the initial establishment of local fold structures causing a progressive reduction in the multiplicity of routes to the final native state. Lin and Zewail (2012) go so far as to suggest that the force resulting from the mere presence of random hydrophobic residues is generally sufficient to induce a polypeptide chain of < 200 residues to collapse to a relatively compact form within an appropriate biological time frame. Indeed,

despite the apparent complexity of the process, as a first-order approximation, the known folding rates of proteins can be explained by knowledge of just the total chain length (i.e., the number of amino-acid residues, L). At least in the range of $L = 20$ to 300, there is a dramatic reduction in the spontaneous folding rate (here, given in units of sec^{-1}) with increasing L, with the function

$$k_f \simeq (1.1 \times 10^8)e^{-1.3\sqrt{L}} \tag{12.1}$$

explaining $\sim 78\%$ of the variance in the folding rate (k_f) among proteins (Dill et al. 2011). Over an order of magnitude increase in chain length, k_f declines approximately seven orders of magnitude from 3×10^4 to 2×10^{-3} sec^{-1} (Figure 12.5). Notably, the protein-folding rates that this formula derives from have been almost universally estimated with *in vitro* methods. This raises concerns because the macromolecular crowding within cells (Dhar et al. 2010) and the attachment of nascent chains to the ribosome during translation (Kaiser et al. 2011) can modulate folding by reducing the formation of inappropriate folds. Unfortunately, the technical difficulties of quantifying protein assembly *in vivo* remains formidable.

Numerous attempts have been made to improve the accuracy of prediction of folding rates by incorporating additional information. For example, Ivankov and Finkelstein (2004) suggest a refinement that subtracts the subsets of residues incorporated into α helices from L. However, the resultant regression yields only a marginal improvement over the preceding expression with the added expense of requiring a detailed understanding of the protein's secondary structure. Incorporation of further information on secondary structure, amino-acid composition, and/or the number of chain contacts has little added effect on the accuracy of prediction (Grantcharova et al. 2001; Ivankov and Finkelstein 2004; Prabhu and Bhuyan 2006; Galzitskaya 2008; Huang et al. 2012), and given that k_f itself is subject to measurement error, there may be little room for improvement beyond the pattern summarized in Equation 12.1.

Finally, it bears emphasizing that although chain length alone is a fairly good predictor of the folding rate, this need not exclude the importance of other factors, but simply means that any additional factors must be either tightly correlated with chain length or of minor significance. In other words, chain length may provide an overall summary measure with good predictive ability but possibly with little mechanistic relevance. In addition, not too much should be read into the significance of the $L^{0.5}$ scaling in Equation 12.1, as exponents in the range of 0.1 to 0.7 yield fits that are nearly equally as good.

How many contortions might a protein go through en route to achieving a proper fold? Because the mean time for a chain to switch from one configuration to another is estimated to be $\simeq 10^{-9}$ sec (Zana 1975), taking the reciprocal of Equation 12.1 as the approximate mean time to complete a search, $\simeq 10^{-8}e^{\sqrt{L}}$, the average number of configurations sampled prior to finding the proper fold solution is $\simeq 10e^{\sqrt{L}}$. For $L = 200$, this implies an average of 13×10^6 configurations searched in a time span of ~ 0.013 sec. For $L = 300$, this jumps to $\sim 33 \times 10^7$ configurations searched over 0.33 sec, and with $L = 500$ to $\sim 51 \times 10^9$ searches over 51 sec. This implies that beyond chain lengths of 200 to 300 residues, unassisted folding times rapidly approach biologically unrealistic levels, a point to which we will return to in Chapter 14. Thus, it may not be a coincidence that protein domains exceeding 300 residues

in length are rare (Wheelan et al. 2000; Wang et al. 2011; Lin and Zewall 2012), and that average lengths of entire proteins are commonly on the order of 300 residues in most species.

To what extent do observed folding rates approach the maximum rates possible from the standpoint of biophysical limitations? Following the suggestion that an upper bound to the rate of folding of a single-domain protein $\simeq (10^8/L)$ sec^{-1} (Kubelka et al. 2004), by comparison with Equation 12.1, it can be seen that empirically observed folding rates are typically far below the maximum. For example, for a 100-residue protein, the maximum folding rate is predicted to be $\simeq 10^6$/sec, whereas Equation 12.1 implies an average observed rate of only 249/sec. Because even the most rapidly folding proteins currently known do so one to two orders of magnitude more slowly than this proposed protein-folding speed limit (Kubelka et al. 2004), it appears that natural selection has generally been unable to achieve perfection in folding rates.

Does the level of refinement of folding rates vary among species? Although an ideal comparison of orthologous proteins in different phylogenetic lineages has not been performed, it has been argued that proteins of equivalent length fold at least ten times more rapidly in bacteria than in eukaryotes (Galzitskaya et al. 2011). Proteins also tend to be longer in eukaryotes than in prokaryotes, which will further exacerbate the protein-folding challenges in the former group. In one of the only comparative studies of protein-folding pathways, Lim et al. (2018; Lim and Marqusee 2018) found for the protein ribonuclease H that although different bacterial lineages use the same type of folding intermediate, the pathways to get there differ; and another study of a protease revealed dramatic differences in the folding mechanism (Nixon et al. 2021). Thus, even within bacteria, there are apparently evolutionary paths open to divergence of folding mechanisms without compromising folding rate.

Stability of folding. As with protein folding rates, folding stability (the tendency to remain folded after achieving the native state) is largely a function of the total chain length, with additional information on sequence and secondary structure again not greatly improving predictability (Robertson and Murphy 1997; Ghosh and Dill 2009; Khan and Vihinen 2010; Dill et al. 2011; Jarzab et al. 2020). The mechanisms by which stability is achieved are diverse, and include packing effects of hydrophobic residues, backbone hydrogen bonds, and favorable electrostatic interactions (Miller et al. 2010). Thus, not surprisingly, protein folding rates and stability are not independent attributes (Plaxco et al. 2000; Sato et al. 2001). Proteins that fold rapidly are often also quite stable, but conflicts can also exist. For example, whereas proteins are positively selected to fold into their proper native states, negative selection may operate to avoid folding too rapidly and/or too stably into misfolded states. Random mutagenesis with a model protein demonstrated that although a substantial fraction of mutations result in faster folding times, nearly all of these have the side effect of reducing stability, suggesting that natural selection places a premium on the latter (Kim et al. 1998).

Indirect evidence supports the idea that selection on folding stability does indeed play a central role in amino-acid sequence evolution. For example, there is a strong correlation between the thermostability of individual proteins and the optimal growth temperature of bacterial species, and random amino-acid substitutions

in protein cores are more deleterious than those for surface residues (Dehouck et al. 2008; Jarzab et al. 2020), and random amino-acid substitutions in the structural cores of proteins are more deleterious than those for surface residues (Tripathi et al. 2016). In particular, the total usage of seven amino acids – four hydrophobic (Ile, Val, Trp, and Leu), one polar (Tyr), and two charged (Arg and Glu) – is highly correlated with optimal growth temperature (Zeldovich et al. 2007). The usage of this particular mix of residues has been proposed to represent a compromise between the conflicting challenges of folding rapidly and avoiding stable misfolded configurations (Berezovsky et al. 2007). Under this hypothesis, the reduced incidence of these seven residues at lower temperatures is viewed as a by-product of the relaxed intensity of selection on folding mechanisms in less extreme thermal backgrounds.

Protein stability is deemed to be positively associated with fitness in the sense that destabilized proteins are prone to loss of function, aggregation, and/or direct toxicity. Nonetheless, most proteins sit on the "margin of stability" in the sense that only one or two mutations are often sufficient to induce complete loss of stability. Although it is commonly argued that marginal stability is required for proper protein function, with excess stability somehow reducing protein performance, this has not held up to close scrutiny. It is relatively easy to create more stable proteins by mutagenesis (Matsuura et al. 1999; Bershtein et al. 2013; Sullivan et al. 2012), and the individual residues contributing to stability typically interact in an additive fashion (Wells 1990; Serrano et al. 1993; Zhang et al. 1995). Numerous proteins have been engineered to have increased stability with few, if any, consequences for enzyme efficiency (e.g., Giver et al. 1998; van den Berg et al. 1998; Taverna and Goldstein 2002; Borgo and Havranek 2012; Moon et al. 2014). Indeed, when wide phylogenetic comparisons are used to generate consensus sequences of proteins, the resultant synthesized peptides are more often than not more stable than the extant proteins, each of which has a subset of the overall stabilizing consensus residues (Sternke et al. 2019).

An alternative explanation for all of these observations is that marginal stability evolves as a simple consequence of the diminishing benefits of increased stability. This would be the case, for example, if fitness is a hyperbolic function of the energy associated with the forces holding a protein together (Govindarajan and Goldstein 1997; Taverna and Goldstein 2002; Bloom et al. 2005; Wylie and Shakhnovich 2011; Serohijos and Shakhnovich 2014). Under this model, proteins are expected to be pushed by natural selection to more stable configurations until reaching the point where any further fitness improvement is small enough to be offset by the vagaries of random genetic drift and/or mutation pressure towards less stable states (Chapter 5). In essence, under any particular population genetic-environment, a quasi-steady-state distribution of stability is expected to evolve to the point at which the rates of fixation of beneficial and deleterious mutations are equal (Figure 12.6). The overall prediction of this drift-barrier hypothesis is that the mean folding stability of proteins will evolve to higher values in populations with larger effective population sizes. This same hypothesis may explain the higher folding rates in prokaryotes than in eukaryotes noted above.

A more mechanistic view of these issues can be acquired by considering the typical features of evolved proteins. The folding stability of proteins is often on the order of $\Delta G = -3$ to -20 kcal/mol (Plaxco et al. 2000; Dill et al. 2011). With the

expected fraction of folded proteins being $\simeq e^{-\Delta G/RT}/(1+e^{-\Delta G/RT})$ at thermodynamic equilibrium, where $RT \simeq 0.6$ kcal/mol, a protein with $\Delta G = -3$ is expected to be folded $> 99\%$ of the time. This diminishes to 96.5 and 84% for $\Delta G = -2$ and -1, respectively. A survey of experimental assays of mutational effects suggests an average $\Delta\Delta G \simeq 0.6$ kcal/mol (SD = 1.1) associated with individual surface residues, and higher destabilizing effects (1.4 kcal/mol; SD = 1.7) for core residues; additions of these values to ΔG yields the post-mutational stability. The distributions of both kinds of effects are roughly normal (Figure 12.7), so the overall distribution of site-specific effects for an entire protein is essentially a mixture of normals. Because smaller proteins have a higher fraction of surface residues, the average $\Delta\Delta G$ is expected to be smaller. To put this in perspective, the average energy associated with single hydrogen bonds in peptides is thought to be on the order of $\Delta G = -2$ kcal/mol (Sheu et al. 2003; Wendler et al. 2010).

If the drift-barrier hypothesis does indeed provide an explanation for the evolution of marginal stability, the distribution of $\Delta\Delta G$ values seen in such surveys must reflect the natural outcome of the joint forces of mutation, selection, and drift in positioning a population on the fitness-stability function (Wylie and Shakhnovich 2011). Unfortunately, as in most areas of cell biology, there are few comparative studies bearing on this issue. However, an *in vitro* evaluation of the folding stability of the dihydrofolate reductase enzyme from 36 species of mesophilic bacteria revealed a substantial range of variation among species, with the standard deviation being roughly 10% of the mean (Figure 12.8).

Determinants of Protein-sequence Evolution

Within a given protein, there can be substantial variation in the evolutionary rates of substitution among amino-acid sites. Not surprisingly, positions involved in catalytic sites are generally under strong purifying selection, and as a consequence, proteins often retain the ability to function appropriately in foreign cellular backgrounds after very long periods of evolutionary divergence. For example, a survey of the performance of over 400 different human proteins in yeast (lineages that separated over a billion years ago) revealed that nearly half were able to complement the absence of the native yeast gene (Kachroo et al. 2015). Remarkably, however, although between 60 and 90% of genes involved in various aspects of metabolism were able to complement, $\sim 50\%$ of genes involved in transcription, $\sim 65\%$ involved in DNA replication and repair, and nearly all involved in cell growth were unable to complement. Thus, proteins whose functions are most closely related to fitness need not remain highly conserved at the protein-sequence level. Here, we explore a wide range of issues bearing on the mechanisms responsible for the substantial evolutionary-rate variation that exists among proteins and among sites within them.

Lessons from phylogenetic comparisons and experimental mutagenesis. Comparisons of the sequences of orthologous protein-coding genes over a vast array of species have left little doubt that amino-acid sequences undergo slow but relentless change over evolutionary time. Not all amino-acid substitutions are acceptable in all contexts, and there is substantial variation in evolutionary rates among different

proteins and different phylogenetic lineages, but only a tiny fraction of amino-acid sites are invariant across the entire Tree of Life.

The most common approach to estimating protein evolutionary rates starts at the level of DNA-sequence analysis, and compares the rates of nucleotide substitution at amino-acid replacement and silent sites where, respectively, nucleotide substitutions do or do not elicit a change at the amino-acid level. Owing to the nature of the genetic code, $\sim 25\%$ of nucleotide sites in a protein-coding gene are typically silent. For example, third-positions in codons for the eight amino acids for which A, C, G, or T lead to the same residue (Figure 12.1) are referred to as four-fold redundant sites. The usual assumption is that such sites evolve in a neutral fashion, owing to their invisibility at the amino-acid level. If this is the case, then the rate of nucleotide substitution (meaning the rate at which one nucleotide type is displaced by another at the population level) at silent sites is expected to equal the mutation rate per site per generation (u) in accordance with the neutral theory (Chapter 4). The total expected silent-site divergence between two lineages separated by t time units (in this case, generations) would then be $2tu$ mutational changes per site, the 2 appearing because mutations accumulate independently down each lineage.

Having such a benchmark of neutral divergence is informative, as it provides a means for interpreting rates of amino-acid substitution, in particular factoring out the contribution of mutation pressure from that associated with selection. If substitutions at replacement sites are selected against, which is generally the case (Kimura 1983; Nei and Kumar 2000; Yang 2014), their rate of divergence should be lower than the expected neutral benchmark. Letting ϕ_f be the probability of fixation of a newly arisen replacement mutation, the rate of divergence at such sites has expectation $2t \cdot (2Nu) \cdot \phi_f$, where N is the absolute population size, and $2Nu$ is the rate at which mutations arise within each population (assumed to be diploid) per nucleotide site. If mutations at replacement sites are neutral, the fixation probability is simply equal to the initial frequency of a mutation, $1/(2N)$, and the overall amount of divergence is equal to the neutral expectation given above, $2tu$.

Generally, we do not have an accurate measurement of the divergence time t between two species, nor of the mutation rate u. However, if one simply takes the ratio of the observed divergences at replacement and silent sites (denoted d_N and d_S, respectively, with the N referring to nonsynonymous or amino-acid replacement sites), the resultant ratio has expectation $[2t \cdot (2Nu) \cdot \phi_f]/(2tu) = 2N \cdot \phi_f$, assuming silent sites do indeed evolve in a neutral fashion. When rewritten as $\phi_f/[1/(2N)]$, this ratio is seen to be equivalent to the fixation probability at replacement sites relative to the neutral expectation. Thus, under appropriate conditions, d_N/d_S provides a biologically interpretable measure of the degree of selective constraint on a protein-coding gene – assuming that the majority of mutations are either neutral or deleterious, d_N/d_S is equivalent to the fraction of amino-acid altering mutations that evade the eyes of natural selection, and for that reason is sometimes referred to as the width of the selective sieve.

There are many caveats with respect to this sort of analysis. First, it is assumed that silent sites are neutral, whereas we know that these can experience some selection at various levels from, for example, preferential tRNA recognition of certain nucleotides in third positions, mRNA structural constraints, and influence of translation speed on folding efficiency (Sharp et al. 2005; Zhou et al. 2010; Lawrie et

al. 2013; Long et al. 2018; Walsh et al. 2020). Second, d_N is generally measured as an average over multiple sites within a gene, obscuring the fact that although many substitutions can be strongly selected against, a minority may nonetheless be advanced by positive selection. Third, there is the difficult matter of accurately estimating d_N and d_S from highly divergent sequences, as multiple substitutions at individual sites will lead to an undercounting of the actual numbers of changes that have accrued, especially at more rapidly evolving silent sites.

These and many other matters have been taken up in detail in the technical field of DNA-sequence analysis, but justified or not, the d_N/d_S ratio remains a central parameter determined in almost all molecular-evolution studies. With few exceptions, proteome-wide studies of d_N/d_S in comparisons of closely to moderately related species, yield average ratios on the order of 0.05 to 0.25 (Kuo et al. 2009; Stanley and Kulathinal 2016; Lynch et al. 2017). This implies that on *average* only 5 to 25% of amino-acid alterations of proteins are typically acceptable in nature. There is, however, a wide range of variation among proteins and among species. Given the possibility of selection on silent sites, especially in large-N_e species, such differences also need to be cautiously interpreted as implying lineage-specific differences in the efficiency of natural selection. Moreover, d_N/d_S analyses leave unresolved the degree to which neutral vs. beneficial nonsynonymous mutations contribute to the pool of fixed amino-acid replacement substitutions.

Comparative analyses have led to a number of other general observations that leave little doubt that the majority of amino-acid altering mutations are removed by purifying selection in nature. First, most amino-acid substitutions involve exchanges of amino acids with similar chemical properties, with radical exchanges being more common in low-N_e species (Bergman and Eyre-Walker 2019; Weber and Whelan 2019). Second, there is a premium on the use of amino acids with relatively low biosynthetic costs, conditional on maintaining a level of residue diversity necessary for maintaining stable and functional proteins (Krick et al. 2014; Venev and Zeldovich 2018). Third, substitution rates are higher for residues on protein surfaces than for those in hydrophobic cores (Suckow et al. 1996; Goldman et al. 1998; Bustamente et al. 2000; Ramsey et al. 2011; Roscoe et al. 2013; Firnberg et al. 2014; Sarkisyan et al. 2016; Moutinho et al. 2019). Fourth, membrane proteins exposed to the external environments evolve more rapidly, perhaps in response to adaptive challenges, than do cytosolic proteins confined to the more homeostatic internal cellular environment (Sojo et al. 2016).

As an alternative to deriving indirect inferences from long diverged sequences, the degree of constraint on protein sequences can be directly evaluated with the performance of randomly mutagenized sequences. However, although such an approach has the advantage of avoiding problems of sequence saturation, silent-site selection, etc., it has the strong limitation of only being able to identify residue changes with major effects. The central issue here is that selection in nature is capable of eradicating deleterious mutations with selective disadvantages down to order $1/N_e$, where N_e (the effective population size) is typically in the range of 10^4 to 10^9 (Chapter 7), whereas lab experiments are generally unable to detect deleterious mutations with fitness effects smaller than 10^{-3}. Thus, random mutagenesis experiments certainly underestimate the fraction of amino-acid substitutions that are eliminated by purifying selection in nature.

This being said, such experiments have been illuminating in a number of ways. For example, Yampolsky and Stoltzfus (2005) summarized the relative exchangeabilities of amino-acid pairs observed in such studies. Hydrophobic residues tend to be most substitutable with other hydrophobic residues, and hydrophilic residues with each other, whereas exchanges between these two extreme groups tend to be unacceptable.

The protein most extensively studied in this way is β-lactamase, a bacterial protein that hydrolyzes antibiotics such as penicillin. The functional consequences of every possible amino-acid substitution at every position in β-lactamase has been characterized (Deng et al. 2012; Jacquier et al. 2013; Firnberg et al. 2014). In agreement with evolutionary divergence data, this work reveals that the number of acceptable amino acids at individual sites is frequently below 15 (and often much lower), with surface residues being generally being more receptive to change (Figure 12.9). However, a number of surface positions far from the active site are highly sensitive to mutations, ruling out the generality that all surface residues are under relatively relaxed selection. Several residues can be altered in ways that increase molecular stability, and the overall distribution of effects is bimodal, with most acceptable variants having functionality just slightly below the norm and a small fraction being nonfunctional (Figure 12.10).

Surprisingly, a number of sites that are known to vary among β-lactamase sequences from natural isolates are intolerant to amino-acid substitutions, an observation that has been seen in other proteins (Mishra et al. 2016). This raises questions about the common assumption that sites with high natural levels of variability experience low functional constraints. It also suggests the importance of context dependence (Chapter 5), with certain sites being more or less accepting of alterations depending on the state of other sites within the protein (Bershtein et al. 2006; Salverda et al. 2011), a point to which we will return to below. A strong role for contingency is derived from the observation that when mutations at two sites that are acceptably exchangeable on their own are combined in this protein, they commonly lead to nonfunctional molecules (Axe 2000). Moreover, chimeric molecules obtained by splicing together halves of different natural variants are completely nonfunctional (Axe 2000).

Assays from numerous random mutagenesis experiments with other proteins are in general agreement with the preceding results. For example, Guo et al. (2004) examined the performance of $\sim 10^5$ single amino-acid substitutions in 3-methyladenine DNA glycosylase, a DNA repair enzyme in humans, and found that 34% of exchanges led to enzyme inactivation; substitutions in α-helices were about twice as exchangeable as those in β-strands (as seen in other studies; Silverman et al. 2001; Firnberg et al. 2014), and those in turns and loops were still more acceptable. Similar analyses with different proteins have yielded estimates of 30 to 80% for fractions of nonfunctional mutations (Guo et al. 2004; Axe et al. 1996; Materon and Palzkill 2001). Again, although the definition of nonfunctional varies among studies (with most incapacitated enzymes retaining at least a small amount of functionality), owing to measurement limitations, all such studies must greatly underestimate the total fraction of mutations that would be removed by purifying selection in nature.

From an evolutionary standpoint, it is more desirable to know the net consequences of mutations for organismal fitness not simply for molecular function, and

to have such measurements on a continuous scale rather than a yes/no scale with an arbitrary cutoff. Although the data are more limited here, they generally point in the same direction. In the case of β-lactamase, the distribution of fitness effects for single amino-acid substitutions has a mode near zero, with a small fraction being favorable, and a long tail to the left (denoting deleterious effects), with $\sim 40\%$ of these having selection coefficients $0 < s < 0.1$, and only $\sim 6\%$ completely obliterating enzyme function (Figure 12.10).

Direct fitness assays of random mutations in other genes are generally consistent with the observations for β-lactamase (Figure 12.10; see also Roscoe et al. 2013; Lind et al. 2016; Sarkisyan et al. 2016; Lundin et al. 2018). Typically, the main peak in the distribution of fitness effects is near zero, with only a small fraction of mutations improving fitness, the majority of mutations reducing fitness by no more than 10%, and on average just 1%, and with a secondary peak associated with mutations lacking entirely in activity. Notably, in a number of cases, a few silent-site substitutions have discernible negative effects, implying effects on transcription, translation, and/or folding efficiency. Ribosomal protein genes are particularly pronounced in this regard, having similar distributions of fitness effects for both silent and replacement substitutions (Figure 12.10).

Expression level and the propensity for sequence change. It has long been thought that the evolutionary rate of a protein is inversely related to its functional significance – the higher the relevance to fitness, the lower the acceptability of amino-acid changes. However, there is no formal way to rank functional significance, and simply invoking low d_N/d_S introduces a circularity. To avoid falling into this seductive trap, an exploration of alternative explanations is warranted.

As noted above, residues buried within a protein generally evolve at substantially lower rates than those exposed on protein surfaces. Buried polar residues involved in hydrogen bonding are especially conserved, although the constraint on molecular evolution declines with increasingly large internal cores of proteins, presumably because stability is distributed across more residues (Franzosa and Xia 2009; Worth and Blundell 2009, 2010). Proteins with especially low rates of substitution for surface residues have even more exceptionally low rates for the core residues, leading to the suggestion that alterations in surface residues facilitate the acceptance of mutations in the core (Toth-Petrósky and Tawfik 2011). However, it remains unclear whether this is a causal relationship or simply a consequence of some proteins being under greater overall constraint at all positions.

Although a plausible argument for reduced rates of evolution in core positions is the intimate involvement of backbone hydrogen bonds and hydrophobic effects in the maintenance of folding stability, the actual mechanisms may be more complicated, as other features are correlated with interior vs. exterior residues. For example, the tendency to engage in unproductive aggregations with other proteins is a function of surface residues, and amino-acid substitutions in such regions might sometimes even be driven by positive selection to avoid aggregation (Wright et al. 2005). Moreover, the relative packing density of residues is strongly correlated with solvent accessibility, and when these two are jointly accounted for in a multiple regression, the former accounts for more of the variance in evolutionary rate than the latter (Toft and Fares 2010; Yeh et al. 2014).

These observations on the consequences of mutations for protein stability and adhesivity help explain a general observation on relative rates of protein evolution. As noted above, it was long thought that variation in evolutionary rates would be dictated by the functional significance of a protein, but as single-gene knockout studies raised questions about this interpretation, it became clear that the best evolutionary-rate predictor is the expression level of a protein (Pal et al. 2001; Zhang and Yang 2015).

One interpretation of this pattern invokes the idea that natural selection operates to minimize the likelihood of improper folding and of instability once properly folded (Serohijos et al. 2012). Misfolded proteins might commonly arise as a consequence of erroneous protein sequences resulting from transcriptional or translational errors (Drummond and Wilke 2008; Yang et al. 2010). That the latter is a significant challenge is made plausible by the fact that proteins with low amino-acid substitution rates also have low substitution rates at silent sites, which might reflect selection to avoid amino-acid misloading by noncognate transfer RNAs at nonoptimal codons (Chapter 20). The fact that most mutations influencing protein performance do so by eliciting changes in protein folding and stability rather than by directly compromising the catalytic core provides further motivation for the misfolding hypothesis (Bloom et al. 2007; Shi et al. 2012). Under this view, the consequences of misfolding are proposed to be more significant in an abundant protein simply because the absolute number of problematical molecules is greater, although this would only follow if the number rather than the fraction of such proteins is of over-riding importance.

There are, however, alternative (and not necessarily mutually exclusive) explanations for low rates of evolution in highly expressed protein-coding genes. For example, the misinteraction hypothesis postulates purifying selection for surface residues that avoid promiscuous interactions / aggregations with inappropriate proteins (Levy et al. 2012; Yang et al. 2012). With a focus on surface residues, this explanation differs from the emphasis of the misfolding hypothesis on the importance of residues in protein cores to folding and stability.

That inappropriate protein-protein interactions are a nontrivial selective challenge is highlighted by the fact that $\sim 20\%$ of protein molecules are typically bound with nonspecific partners in yeast and metazoan cells (Zhang et al. 2008). Under the misinteraction hypothesis, the efficiency of selection against amino-acid substitutions in surface residues is again expected to be especially elevated in more highly expressed proteins. Thus, it is of interest that in *E. coli,* the more abundant proteins have a lower tendency to aggregate (de Groot and Ventura 2010), apparently because of their reduced surface hydrophobicity (Ishihama et al. 2008). The same is true for human proteins (Tartaglia et al. 2007).

To investigate this idea further, Levy et al. (2012) ordered the full set of amino acids with respect to their tendency to adhere to other molecules, using information on their degree of participation in natural interfaces. They found a negative correlation between a protein's cellular abundance and the predicted adhesiveness of its surface. This effect diminishes from *E. coli* to yeast to human, consistent with an expected reduction in the efficiency of selection against mildly deleterious mutations in species with reduced effective population sizes. Notably, in bacteria, the disparity in evolutionary rates between highly and lowly expressed genes is greatest in species with rapid cell-division rates (Vieira-Silva et al. 2011), which might reflect the latter

species having larger effective population sizes and hence a higher level of efficiency of natural selection.

Mutation pressure and biased amino-acid usage. The particular amino acids deployed within a protein need not simply be outcomes of selection. As discussed in Chapter 5, the likelihood of occupancy of a particular residue at any position within a protein is a joint function of the mutation biases towards and away from individual allelic variants and the ratio of the power of selection to drift. Thus, to understand the relative roles of these two determinants, it is necessary to consider why genome-wide G+C nucleotide compositions range from ~ 0.25 to ~ 0.80 among different species (Lynch 2007), and whether such biases have cascading effects on encoded amino-acid composition.

If genomic G+C composition reflects the prevailing pressure of mutation, it ought to be correlated with the expectations based on known mutational spectra, as recorded in mutation-accumulation experiments (Long et al. 2018). Letting u be the mutation pressure of A+T nucleotides to C+G, and v be the reciprocal rate, the expected equilibrium frequency of G+C under mutation pressure alone is simply $u/(u + v)$. Across the Tree of Life, average genome-wide G+C compositions are indeed strongly correlated with this neutral expectation (Figure 12.11). Nonetheless, despite the positive correlation, almost all genomes also have an excess G+C content relative to the neutral expectation. Notably, the deviations of G+C composition from the neutral expectations are particularly large at silent sites, supporting the idea that contrary to popular belief (noted above), such sites do not generally evolve in a neutral fashion. The general interpretation of these results is that whereas there is biased mutation pressure towards A+T content in most species, as indicated by most neutral G+C expectations being < 0.5, there is near universal selection for G+C.

For the organisms in Figure 12.11, the ratio of mutation rates from G+C $\rightarrow$ A+T to the reverse ranges from 16 (very low G/C-content genomes) to 0.4 (moderately high G/C-content genomes). As most biologists assume all variation to be a product of selection, it is often suggested that nucleotide-composition bias is a product of lineage-specific selection pressures, e.g., to generate base compositions conducive to producing the amino-acid compositions of proteins most compatible with the challenges of specific environments (Mendez et al. 2010). However, as noted in Chapter 4, explaining phylogenetic variation in the mutation rate itself with optimization arguments has not been easy, and explanations for a fine-tuning of the molecular spectrum are even more challenging. There is no direct evidence that mutation spectra are driven by selection, and the possibility that the substantial level of divergence may have been governed largely by effectively neutral processes cannot be ruled out (Haywood-Farmer and Otto 2003). Selection operates on the genome-wide mutation rate, driving this down to some level beyond which further advantages are offset by the power of random genetic drift, but conditional on any particular overall rate, the mutational spectrum may be free to wander over evolutionary time.

Why is there near-universal selection for G+C composition, regardless of the magnitude of mutation pressure towards A+T? The bioenergetic costs of all four nucleotides are very similar (Chapter 17), so selective discrimination on this basis

is unlikely. It has been argued that high-temperature environments impose selection for higher G+C composition because G:C pairs involve three hydrogen bonds (as opposed to two for A:T), rendering a higher degree of DNA (and RNA) stability (Musto et al. 2004; Basak and Ghosh 2005). However, although there are correlations between G+C content and optimal growth temperatures within narrow phylogenetic groups of bacteria, this is not true on a broader phylogenetic scale. Moreover, the G+C composition of silent sites does not exhibit such correlations, contrary to expectations if there is genome-wide selection for duplex stability (Hurst and Merchant 2001). Adenine and guanine (purine) nucleotides contain three more nitrogen atoms than do pyrimidines, so long-term residence in nitrogen limiting environments might select for genomes enriched with Cs and Ts (Rocha and Danchin 2002; Luo et al. 2015), but as DNA consists of A:T and G:C bonds, such selection would have to occur at the RNA level and be efficient enough to discriminate a difference of two nitrogen atoms against a total-cell backdrop of billions of these. Finally, gene conversion (a result from the repair of heteroduplex DNA arising from recombination between two nonidentical sequences) is thought to be weakly biased towards Cs and Gs (when mismatches with As and Ts arise) across the Tree of Life (Lassalle et al. 2015), providing still another pressure on nucleotide composition, depending on the level of recombination.

Regardless of the mechanisms driving genome-wide nucleotide composition, the central question here is whether such biases have repercussions at the level of amino-acid composition across phylogenetic lineages. Owing to the structure of the genetic code, the codons for some amino acids are much richer in GC content than others (Figure 12.1); e.g., 83% for alanine, arginine, glycine, and proline, but $\leq 17\%$ for asparagine, isoleucine, lysine, phenylalanine, and tyrosine. Moreover, the biochemical features of amino acids are not independent of the GC composition of their codons – amino acids encoded by GC-rich codons tend to be less hydrophobic but also less energetically expensive to synthesize (Table 12.1). Thus, we wish to know whether certain population-genetic environments promote the use of particular amino acids independent of their immediate functional significance.

From Figure 12.11, this can be seen to be the case – species with very strong mutation pressure towards A+T also gravitate to codons with low G+C composition. Among species, genome-wide G+C-content at first and second positions of codons (which mostly consist of amino-acid replacement sites) is correlated with that at third positions (which are largely silent sites) (Gu et al. 1998; D'Onofrio et al. 1999; Bastolla et al. 2004; Chen et al. 2004). The proteome-wide usage of amino acids with GC-rich codons in different species more than doubles across the range of genome-wide GC composition at silent sites, whereas that of the AT-rich group declines by more than 50% (Knight et al. 2001; Li et al. 2015). Thus, although there is strong selection for amino-acid composition at key sites within most proteins, mutation pressure is frequently sufficient to overcome the weak selection for amino-acid usage in a substantial fraction of sites.

These observations are of relevance to the question as to whether isolated lineages are likely to evolve completely independently at the molecular level. When two separate lineages independently acquire the same novel phenotype from the same starting state, the change is said to be parallel, whereas independent acquisition of the same state from different initial conditions represents convergent evolution

(Zhang and Kumar 1997; Storz 2016).

Evolutionary substitutions to certain types of amino acids at particular sites within proteins occur more frequently than expected by chance (e.g., Bazykin et al. 2007; Rokas and Carroll 2008; Elias and Tawfik 2012; Ayuso-Fernández et al. 2018; Cano et al. 2022), and there is little question that lineages do occasionally respond to the same selective challenge in parallel manners. A dramatic example was revealed in replicated *E. coli* populations exposed to an increasing gradient of the antibiotic trimethoprim, which exhibited a similar temporal ordering of similar mutations conferring resistance in the dihydrofolate reductase gene (Toprak et al. 2011). However, demonstrating that convergent/parallel evolution is an outcome of shared selective pressures is difficult without rigorous statistical and/or empirical analysis, and numerous examples exist in which parallel evolution has inspired arguments about the channeling of molecular adaptations, only to be overturned by subsequent evaluation (Storz 2016).

This being said, it has been consistently observed that, relative to the neutral expectation, the incidence of amino-acid convergence events becomes progressively less common with more distantly related lineages (Goldstein et al. 2015; Shah et al. 2015; Zou and Zhang 2015). A compelling explanation for such behavior is that as a protein accepts amino-acid changes at a variety of sites in different lineages, the local selective environment at other sites is altered, thereby diminishing the likelihood of effectively neutral mutations being channeled to the same set of residues. Such a model implies a predominance of both effectively neutral substitutions (allowing change to occur at individual sites) and of epistasis (interaction effects between individual sites).

Epistasis and compensatory mutation. The preceding sections provided numerous examples in which the effects of mutations in protein-coding sequences are epistatic with respect to fitness. That is, the fitness effects of individual mutations often depend on local context. Direct evidence for such interactions derives from experimental mutagenesis experiments, such as that of Bank et al. (2015), who in an analysis of > 1000 double mutants in the binding domain of a yeast heat-shock protein found a preponderance of pairs with negative combined effects on fitness (beyond the additive expectations based on single-mutational effects). In this study, very few pairs exhibited positive epistatic effects.

In a somewhat different study, Lunzer et al. (2010) substituted (one at a time) 168 amino acids in the isopropymalate dehydrogenase protein in *E. coli* to match the differences in the orthologous protein in *Pseudomonas aeruginosa.* On the *E. coli* background, 63 of these single substitutions were functionally compromised, whereas only one had improved performance. In another comparative study, Starr et al. (2018) reconstructed estimated ancestral states in a yeast heat shock protein (Hsp70) and then laboriously substituted amino acids from the modern-day sequence into the ancestral form, and vice versa. Although Hsp70 has retained a highly conserved function over a billion years of evolution, $> 75\%$ of these single-residue exchanges were deleterious, even though they must have been acceptable over the course of evolution. All of these observations are consistent with stochastic lineage-specific additions of mutations conditional upon earlier changes progressively altering the permissive environment for substitution.

Further indirect evidence for the long-term evolutionary significance of epistasis derives from a number of different comparative analyses. For example, in a study of 16 eukaryotic proteins, each with > 1000 sequences available from a wide variety of phylogenetic lineages, Breen et al. (2012) found that the average amino-acid site is occupied by just 8 different amino-acids, even though ample evolutionary time has elapsed for all mutation types to have appeared at each site, i.e., on average each site can shift to seven alternative amino acids. The authors reasoned that d_N/d_S ought to be $7/19 = 0.36$ if amino-acid altering mutations accumulate in a noninteractive way, i.e., 74% of amino-acid replacements would be expected to be unacceptable. However, the average observed d_N/d_S ratio (measured from sequence divergence between species) for these proteins averages about $7\times$ lower than this expectation, leading to the conclusion that negative epistatic fitness effects must be pervasive among mutations. The implication is that if a particular amino-acid fixes at one particular site, it apparently creates a local environment that prevents the fixation of the majority of amino-acid altering mutations at other sites, leaving an average of only ~ 1 permissible change per site at any point in evolutionary time.

A second compelling line of evidence for the role of epistasis in protein evolution derives from the observation that many amino-acid changes that cause human pathologies (and are therefore rare in the human population) are nonetheless well-established (with no pathogenic effects) in other mammalian species (Kondrashov et al. 2002; Gao and Zhang 2003). Very similar observations have been made with mutations known to be pathogenic in *Drosophila,* but established in other insect species (Kulathinal et al. 2004). As the frequency of such compensated deviations does not increase with the evolutionary distance of a lineage, this suggests that they accrue relatively rapidly, rather than awaiting long-term protein remodeling. The effects of such mutations must be context dependent.

Finally, it has been noted that amino-acid changes in proteins tend to be clustered within a sequence, generally on a chain-length scale of < 10 residues, and also tend to preserve the local charge of the protein (Callahan et al. 2011). Notably, the average physical distance between central carbon atoms of amino acids in folded proteins plateaus at chain distances than > 10 residues, implying that on average residues separated by < 10 positions have a high likelihood of physical interaction. The fact that silent-site substitutions are not clustered argues against the pattern being a result of regional mutational hot spots. Additional work shows that long-range epistatic interactions are not uncommon (Sharir-Ivry and Xia 2018).

A General Model for Protein Evolution

A key point emerging from the previous discussion is that, more often than not, many cumulative amino-acid changes have little impact on the immediate functionality of a gene. Rather, much of protein evolution appears to reflect little more than a restricted random walk down nearly-neutral pathways (Figure 12.12). Some of these pathways may involve the fixation of effectively neutral but slightly deleterious mutations, which then allow the fixation of a compensatory mutation that was insignificantly favorable (or even deleterious) on the prior ancestral background but now more favorable in its new context. Such compensatory changes are not

necessarily epistatic with respect to the long-term enhancement of total fitness, although they are epistatic with respect to the physical structure of the protein.

Thus, an emerging view of protein-sequence evolution is that at any point in time the number of degrees of freedom for change at individual amino-acid sites is small, with the identities of exchangeable amino acids shifting with fortuitous prior fixations elsewhere in the molecule (Goldstein and Pollock 2017; Xie et al. 2021). In part, restricted sequence walks are governed by the nature of the genetic code, by which single mutations at each replacement nucleotide site can generate at most three alternative amino acids. More generally, however, the structural environment of the protein itself will dictate the subset of permissible (effectively neutral) amino-acid exchanges. Over time, slight shifts in the amino-acid constitution of the protein, each nearly neutral incrementally, alter the local protein-structural environment, further restricting the degrees of freedom for future changes, but in a progressively divergent way, allowing the long-term degrees of freedom for change at a large fraction of sites to wander to levels as high as 19.

Such cycles of modification of the background environment, and subsequent channeling of permissible mutations allows for an expansive set of paths open to evolutionary change across the Tree of Life, while rendering individual lineages victims of historical contingency. Under this model, because slightly deleterious mutations can sometimes fix, such events also pave the way for the subsequent fixation of compensatory beneficial mutations without significant consequences for long-term adaptation in terms of protein function. Moreover, by this process, amino-acid changes that were originally effectively neutral may sometimes become entrenched to the point of being essential to protein functionality and hence nearly irreversible evolutionarily.

Summary

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

- At the dawn of the protein world, only a fraction of the twenty amino acids used in today's organisms would have been in play, and other noncanonical amino acids might have been used. Nonetheless, enormous functional diversity of proteins can still be generated by a reduced amino-acid alphabet, although an expanded vocabulary allows for further refinement in catalytic activity and efficiency.

- One of the major challenges of proteins is their initial need to fold into three-dimensional structures essential for functionality. Although there are a number of important substructural influences, folding rates are largely determined by the amino-acid chain length, and those in excess of ~ 250 residues are generally incapable of folding on their own on reasonable time scales.

- Despite the 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 non-canonical amino acids. Folding rates and stability are also less than their maximum possible values, and potentially more so in eukaryotes than prokaryotes. These observations suggest that the efficiency of natural selection is stalled by either a drift barrier and/or constraints imposed by the restricted set of canonical amino acids.

- Based on phylogenetic comparisons of sequence data, only 5 to 25% of amino-acid altering mutations are acceptable in nature, although experimental substitutions of random amino acids consistently indicate that much larger fractions do not entirely eliminate protein function. The distribution of fitness effects associated with amino-acid exchanges generally has a mode not significantly different from zero, a long tail towards deleterious effects, and only a small tail containing favorable changes. The overall conclusion is that the majority of mutations at the protein level are mildly deleterious. However, the details of the distribution in the range of very small effects, which is most critical to evolutionary theory, remains uncertain.

- One of the primary determinants of the rate of evolution of a protein is its level of expression. This is thought to be a consequence of strong purifying selection for the maintenance of folding stability to avoid the production of wasted or harmful by-products and/or selection for surface residues that minimize misinteractions with other key proteins.

- The magnitude of mutation bias varies widely among phylogenetic lineages, but is usually in the direction of A and T nucleotides Via the structure of the genetic code, this can sometimes drive the biased deployment of particular amino acids in the proteome, leading to parallel evolution in different lineages with little involvement of selection.

- Amino-acid altering mutations frequently have context-dependent fitness effects, whereby the incorporation of earlier mutations can dictate whether specific subsequent substitutions are deleterious, beneficial, or effectively 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 amino-acid substitutions to the point of becoming near essential to protein functionality. In retrospect, such progressive changes may appear to involve adaptive fixations, the entire process may unfold with only minor consequences for overall fitness. This view of protein evolution is entirely compatible with long-term wandering of amino-acid sequences along the drift barrier.

Literature Cited

Akanuma, S., T. Kigawa, and S. Yokoyama. 2002. Combinatorial mutagenesis to restrict amino acid usage in an enzyme to a reduced set. Proc. Natl. Acad. Sci. USA 99: 13549-13553.

Akashi, H., and T. Gojobori. 2002. Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc. Natl. Acad. Sci. USA 99: 3695-3700.

Alva, V., J. Söding, and A. N. Lupas. 2015. A vocabulary of ancient peptides at the origin of folded proteins. eLife 4: e09410.

Axe, D. D. 2000. Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. J. Mol. Biol. 301: 585-595.

Axe, D. D., N. W. Foster, and A. R. Fersht. 1996. Active barnase variants with completely random hydrophobic cores. Proc. Natl. Acad. Sci. USA 93: 5590-5594.

Ayuso-Fernández, I., F. J. Ruiz-Dueñas, and A. T. Martínez. 2018. Evolutionary convergence in lignin-degrading enzymes. Proc. Natl. Acad. Sci. USA 115: 6428-6433.

Ball, P. 2008. Water as an active constituent in cell biology. Chem. Rev. 108: 74-108.

Bank, C., R. T. Hietpas, J. D. Jensen, and D. N. Bolon. 2015. A systematic survey of an intragenic epistatic landscape. Mol. Biol. Evol. 32: 229-238.

Basak, S., and T. C. Ghosh. 2005. On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biochem. Biophys. Res. Commun. 330: 629-632.

Bastolla, U., A. Moya, E. Viguera, and R. C. van Ham. 2004. Genomic determinants of protein folding thermodynamics in prokaryotic organisms. J. Mol. Biol. 343: 1451-1466.

Bazykin, G. A., F. A. Kondrashov, M. Brudno, A. Poliakov, I. Dubchak, and A. S. Kondrashov. 2007. Extensive parallelism in protein evolution. Biol. Direct 2: 20.

Berezovsky, I. N., K. B. Zeldovich, and E. I. Shakhnovich. 2007. Positive and negative design in stability and thermal adaptation of natural proteins. PLoS Comput. Biol. 3: e52.

Bergman, J., and A. Eyre-Walker. 2019. Does adaptive protein evolution proceed by large or small steps at the amino acid level? Mol. Biol. Evol. 36: 990-998.

Bershtein, S., W. Mu, A. W. Serohijos, J. Zhou, and E. I. Shakhnovich. 2013. Protein quality control acts on folding intermediates to shape the effects of mutations on organismal fitness. Mol. Cell 49: 133-144.

Bershtein, S., M. Segal, R. Bekerman, N. Tokuriki, and D. S. Tawfik. 2006. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. Nature 444: 929-932.

Bershtein, S., A. W. Serohijos, S. Bhattacharyya, M. Manhart, J. M. Choi, W. Mu, J. Zhou, and E. I. Shakhnovich. 2015. Protein homeostasis imposes a barrier on functional integration of horizontally transferred genes in bacteria. PLoS Genet. 11: e1005612.

Bloom, J. D., A. Raval, and C. O. Wilke. 2007. Thermodynamics of neutral protein evolution. Genetics 175: 255-266.

Bloom, J. D., M. M. Meyer, P. Meinhold, C. R. Otey, D. MacMillan, and F. H. Arnold. 2005. Evolving strategies for enzyme engineering. Curr. Opin. Struct. Biol. 15: 447-452.

Borgo, B., and J. J. Havranek. 2012. Automated selection of stabilizing mutations in designed and natural proteins. Proc. Natl. Acad. Sci. USA 109: 1494-1499.

Breen, M. S., C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov. 2012. Epistasis as the primary factor in molecular evolution. Nature 490: 535-538.

Brooks, D. J., and J. R. Fresco. 2002. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. Mol. Cell. Proteomics 1: 125-131.

Brooks, D. J., J. R. Fresco, A. M. Lesk, and M. Singh. 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. 19: 1645-1655.

Bustamante, C. D., J. P. Townsend, and D. L. Hartl. 2000. Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica.* Mol. Biol. Evol. 17: 301-308.

Callahan, B., R. A. Neher, D. Bachtrog, P. Andolfatto, and B. I. Shraiman. 2011. Correlated evolution of nearby residues in drosophilid proteins. PLoS Genet. 7: e1001315.

Cano, A. V., H. Rozhoňová, A. Stoltzfus, D. M. McCandlish, and J. L. Payne. 2022. Mutation bias shapes the spectrum of adaptive substitutions. Proc. Natl. Acad. Sci. USA 119: e2119720119.

Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. USA 101: 3480-3485.

Davis, B. K. 1999. Evolution of the genetic code. Prog. Biophys. Mol. Biol. 72: 157-243.

de la Torre, D., and J. W. Chin. 2021. Reprogramming the genetic code. Nat. Rev. Genet. 22: 169-184.

de Groot, N. S., and S. Ventura. 2010. Protein aggregation profile of the bacterial cytosol. PLoS One 5: e9383.

Dehouck, Y., B. Folch, and M. Rooman. 2008. Revisiting the correlation between proteins' thermoresistance and organisms' thermophilicity. Protein Eng. Des. Sel. 21: 275-278.

Deng, Z., W. Huang, E. Bakkalbasi, N. G. Brown, C. J. Adamski, K. Rice, D. Muzny, R. A. Gibbs, and T. Palzkill T. 2012. Deep sequencing of systematic combinatorial libraries reveals β-lactamase sequence constraints at high resolution. J. Mol. Biol. 424: 150-167.

Dhar, A., A. Samiotakis, S. Ebbinghaus, L. Nienhaus, D. Homouz, M. Gruebele, and M. S. Cheung. 2010. Structure, function, and folding of phosphoglycerate kinase are strongly perturbed by macromolecular crowding. Proc. Natl. Acad. Sci. USA 107: 17586-17591.

Dill, K. A., K. Ghosh, and J. D. Schmit. 2011. Physical limits of cells and proteomes. Proc. Natl. Acad. Sci. USA 108: 17876-17882.

Dill, K. A., and J. L. MacCallum. 2012. The protein-folding problem, 50 years on. Science 338: 1042-1046.

Doi, N., K. Kakukawa, Y. Oishi, and H. Yanagawa. 2005. High solubility of random-sequence proteins consisting of five kinds of primitive amino acids. Protein Eng. Des. Sel. 18: 279-284.

D'Onofrio, G., K. Jabbari, H. Musto, and G. Bernardi. 1999. The correlation of protein hydropathy with the base composition of coding sequences. Gene 238: 3-14.

Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 134: 341-352.

Elias, M., and D. S. Tawfik. 2012. Divergence and convergence in enzyme evolution: parallel evolution of paraoxonases from quorum-quenching lactonases. J. Biol. Chem. 287: 11-20.

Englander, S. W., and L. Mayne. 2014. The nature of protein folding pathways. Proc. Natl. Acad. Sci. USA 111: 15873-15880.

Englander, S. W., and L. Mayne. 2017. The case for defined protein folding pathways. Proc. Natl. Acad. Sci. USA 114: 8253-8258.

Firnberg, E., J. W. Labonte, J. J. Gray, and M. Ostermeier. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. Mol. Biol. Evol. 31: 1581-1592.

Franzosa, E. A., and Y. Xia. 2009. Structural determinants of protein evolution are context-sensitive at the residue level. Mol. Biol. Evol. 26: 2387-2395.

Galzitskaya, O. V., N. S. Bogatyreva, and A. V. Glyakina. 2011. Bacterial proteins fold faster than eukaryotic proteins with simple folding kinetics. Biochemistry (Mosc.) 76: 225-235.

Galzitskaya, O. V., D. C. Reifsnyder, N. S. Bogatyreva, D. N. Ivankov, and S. O. Garbuzynskiy. 2008. More compact protein globules exhibit slower folding rates. Proteins 70: 329-332.

Gao, L., and J. Zhang. 2003. Why are some human disease-associated mutations fixed in mice? Trends Genet. 19: 678-681.

Ghosh, K., and K. A. Dill. 2009. Computing protein stabilities from their chain lengths. Proc. Natl. Acad. Sci. USA 106: 10649-10654.

Giver, L., A. Gershenson, P. O. Freskgard, and F. H. Arnold. 1998. Directed evolution of a thermostable esterase. Proc. Natl. Acad. Sci. USA 95: 12809-12813.

Go, A., S. Kim, J. Baum, and M. H. Hecht. 2008. Structure and dynamics of *de novo* proteins from a designed superfamily of 4-helix bundles. Protein Sci. 17: 821-832.

Goldman, N., J. L. Thorne, and D. T. Jones. 1998. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics 149: 445-458.

Goldstein, R. A., S. T. Pollard, S. D. Shah, and D. D. Pollock. 2015. Nonadaptive amino acid convergence rates decrease over time. Mol. Biol. Evol. 32: 1373-1381.

Goldstein, R. A., and D. D. Pollock. 2017. Sequence entropy of folding and the absolute rate of amino acid substitutions. Nat. Ecol. Evol. 1: 1923-1930.

Govindarajan, S., and R. A. Goldstein. 1997. Evolution of model proteins on a foldability landscape. Proteins 29: 461-466.

Grantcharova, V., E. J. Alm, D. Baker, and A. L. Horwich. 2001. Mechanisms of protein folding. Curr. Opin. Struct. Biol. 11: 70-82.

Gu, X., D. Hewett-Emmett, and W.-H. Li. 1998. Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. Genetica 102/103: 383-931.

Guo, H. H., J. Choe, and L. A. Loeb. 2004. Protein tolerance to random amino acid change. Proc. Natl. Acad. Sci. USA 101: 9205-9210.

Harris, J. K., S. T. Kelley, G. B. Spiegelman, and N. R. Pace. 2003. The genetic core of the universal ancestor. Genome Res. 13: 407-412.

Haywood-Farmer, E., and S. P. Otto. 2003. The evolution of genomic base composition in bacteria. Evolution 57: 1783-1792.

Higgs, P. G., and R. E. Pudritz. 2009. A thermodynamic basis for prebiotic amino acid synthesis and the nature of the first genetic code. Astrobiol. 9: 483-490.

Huang, J. T., D. J. Xing, and W. Huang. 2012. Relationship between protein folding kinetics and amino acid properties. Amino Acids 43: 567-572.

Hurst, L. D., and A. R. Merchant. 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis amongst prokaryotes. Proc. Biol. Sci. 268: 493-497.

Ishihama, Y., T. Schmidt, J. Rappsilber, M. Mann, F. U. Hartl, M. J. Kerner, and D. Frishman D. 2008. Protein abundance profiling of the *Escherichia coli* cytosol. BMC Genomics 9: 102.

Islam, M. M., S. Sohya, K. Noguchi, M. Yohda, and Y. Kuroda. 2008. Crystal structure of an extensively simplified variant of bovine pancreatic trypsin inhibitor in which over one-third of the residues are alanines. Proc. Natl. Acad. Sci. USA 105: 15334-15339.

Ivankov, D. N., and A. V. Finkelstein. 2004. Prediction of protein folding rates from the amino acid sequence-predicted secondary structure. Proc. Natl. Acad. Sci. USA 101: 8942-8944.

Jacquier, H., A. Birgy, H. Le Nagard, Y. Mechulam, E. Schmitt, J. Glodt, B. Bercot, E. Petit, J. Poulain, G. Barnaud, et al. 2013. Capturing the mutational landscape of the β-lactamase TEM-1. Proc. Natl. Acad. Sci. USA 110: 13067-13072.

Jarzab, A., N. Kurzawa, T. Hopf, M. Moerch, J. Zecha, N. Leijten, Y. Bian, E. Musiol, M. Maschberger, G. Stoehr, et al. 2020. Meltome atlas – thermal proteome stability across the Tree of Life. Nat. Methods 17: 495-503.

Jordan, I. K., F. A. Kondrashov, I. A. Adzhubei, Y. I. Wolf, E. V. Koonin, A. S. Kondrashov, and S. Sunyaev. 2005. A universal trend of amino acid gain and loss in protein evolution. Nature 433: 633-638.

Kachroo, A. H., J. M. Laurent, C. M. Yellman, A. G. Meyer, C. O. Wilke, and E. M. Marcotte. 2015. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. Science 348: 921-925.

Kaiser, C. M., D. H. Goldman, J. D. Chodera, I. Tinoco, Jr., and C. Bustamante. 2011. The ribosome modulates nascent protein folding. Science 334: 1723-1727.

Khan, S., and M. Vihinen. 2010. Performance of protein stability predictors. Hum. Mutat. 31: 675-684.

Kim, D. E., H. Gu, and D. Baker. 1998. The sequences of small proteins are not extensively optimized for rapid folding by natural selection. Proc. Natl. Acad. Sci. USA 95: 4982-4986.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge, UK.

Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol. 2: RESEARCH0010.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc. Natl. Acad. Sci. USA 99: 14878-14883.

Koonin, E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nature Rev. Microbiol. 1: 127-136.

Krick, T., N. Verstraete, L. G. Alonso, D. A. Shub, D. U. Ferreiro, M. Shub, and I. E. Sánchez. 2014. Amino acid metabolism conflicts with protein diversity. Mol. Biol. Evol. 31: 2905-2912.

Kubelka, J., J. Hofrichter, and W. A. Eaton. 2004. The protein folding 'speed limit'. Curr. Opin. Struct. Biol. 14: 76-88.

Kulathinal, R. J., B. R. Bettencourt, and D. L. Hartl. 2004. Compensated deleterious mutations in insect genomes. Science 306: 1553-1554.

Kuo, C. H., N. A. Moran, and H. Ochman. 2009. The consequences of genetic drift for bacterial genome complexity. Genome Res. 19: 1450-1454.

Labean, T. H., T. R. Butt, S. A. Kauffman, and E. A. Schultes. 2011. Protein folding absent selection. Genes (Basel) 2: 608-626.

Lassalle, F., S. Périan, T. Bataillon, X. Nesme, L. Duret, and V. Daubin. 2015. GC-Content evolution in bacterial genomes: the biased gene conversion hypothesis expands. PLoS Genet. 11: e1004941.

Lawrie, D. S., P. W. Messer, R. Hershberg, and D. A. Petrov. 2013. Strong purifying selection at synonymous sites in *D. melanogaster.* PLoS Genet. 9: e1003527.

Levinthal, C. 1968. Are there pathways for protein folding? J. Chim. Physique 65: 4445.

Levy, E. D., S. De, and S. A. Teichmann. 2012. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc. Natl. Acad. Sci. USA 109: 20461-20466.

Lim, S. A., E. R. Bolin, and S. Marqusee. 2018. Tracing a protein's folding pathway over evolutionary time using ancestral sequence reconstruction and hydrogen exchange. eLife 7: e38369.

Lim, S. A., and S. Marqusee. 2018. The burst-phase folding intermediate of ribonuclease H changes conformation over evolutionary history. Biopolymers 109: e23086.

Li, J., J. Zhou, Y. Wu, S. Yang, and D. Tian. 2015. GC-Content of synonymous codons profoundly influences amino acid usage. G3 (Bethesda) 5: 2027-2036.

Lin, M. M., and A. H. Zewail. 2012. Hydrophobic forces and the length limit of foldable protein domains. Proc. Natl. Acad. Sci. USA 109: 9851-9856.

Lind, P. A., L. Arvidsson, O. G. Berg, and D. I. Andersson. 2016. Variation in mutational robustness between different proteins and the predictability of fitness effects. Mol. Biol. Evol. 34: 408-418.

Lind, P. A., O. G. Berg, and D. I. Andersson. 2010. Mutational robustness of ribosomal protein genes. Science 330: 825-827.

Long, H., W. Sung, S. Kucukyildirim, E. Williams, S. F. Miller, W. Guo, C. Patterson, C. Gregory, C. Strauss, C. Stone, et al. 2018. Evolutionary determinants of genome-wide nucleotide composition. Nat. Ecol. Evol. 2: 237-240.

Longo, L. M., and M. Blaber. 2012. Protein design at the interface of the pre-biotic and biotic worlds. Arch. Biochem. Biophys. 526: 16-21.

Longo, L. M., and M. Blaber. 2014. Prebiotic protein design supports a halophile origin of foldable proteins. Front. Microbiol. 4: 418.

Longo, L. M., J. Lee, and M. Blaber. 2013. Simplified protein design biased for prebiotic amino acids yields a foldable, halophilic protein. Proc. Natl. Acad. Sci. USA 110: 2135-2139.

López de la Osa, J., D. A. Bateman, S. Ho, C. González, A. Chakrabartty, and D. V. Laurents. 2007. Getting specificity from simplicity in putative proteins from the prebiotic earth. Proc. Natl. Acad. Sci. USA 104: 14941-14946.

Lundin, E., P. C. Tang, L. Guy, J. Näsvall, and D. I. Andersson. 2018. Experimental determination and prediction of the fitness effects of random point mutations in the biosynthetic enzyme HisA. Mol. Biol. Evol. 35: 704-718.

Lunzer, M., G. B. Golding, and A. M. Dean. 2010. Pervasive cryptic epistasis in molecular evolution. PLoS Genet. 6: e1001162.

Luo, H., L. R. Thompson, U. Stingl, and A. L. Hughes. 2015. Selection maintains low genomic GC content in marine SAR11 lineages. Mol. Biol. Evol. 32: 2738-2748.

Lynch, M. 2007. The Origins of Genome Architecture. Sinauer Assocs., Inc., Sunderland, MA.

Lynch, M. 2018. Phylogenetic diversification of cell biological features. eLife 7: e34820.

Lynch, M., M. Ackerman, K. Spitze, Z. Ye, and T. Maruki. 2017. Population genomics of *Daphnia pulex*. Genetics 206: 315-332.

MacBeath, G., P. Kast, and D. Hilvert. 1998. A small, thermostable, and monofunctional chorismate mutase from the archaeon *Methanococcus jannaschii*. Biochemistry 37: 10062-10073.

Materon, I. C., and T. Palzkill. 2001. Identification of residues critical for metallo-β-lactamase function by codon randomization and selection. Protein Sci. 10: 2556-2565.

Matsuura, T., K. Miyai, S. Trakulnaleamsai, T. Yomo, Y. Shima, S. Miki, K. Yamamoto, and I. Urabe. 1999. Evolutionary molecular engineering by random elongation mutagenesis. Nat. Biotechnol. 17: 58-61.

McDonald, G. D., and M. C. Storrie-Lombardi. 2010. Biochemical constraints in a protobiotic earth devoid of basic amino acids: the "BAA(-) world". Astrobiol. 10: 989-1000.

Mendez, R., M. Fritsche, M. Porto, and U. Bastolla. 2010. Mutation bias favors protein folding stability in the evolution of small populations. PLoS Comput. Biol. 6: e1000767.

Miller, C., M. Davlieva, C. Wilson, K. I. White, R. Couñago, G. Wu, J. C. Myers, P. Wittung-Stafshede, and Y. Shamoo. 2010. Experimental evolution of adenylate kinase reveals contrasting strategies toward protein thermostability. Biophys. J. 99: 887-896.

Mishra, P., J. M. Flynn, T. N. Starr, and D. N. Bolon. 2016. Systematic mutant analyses elucidate general and client-specific aspects of Hsp90 function. Cell Rep. 15: 588-598.

Moon, S., D. K. Jung, G. N. Phillips, Jr., and E. Bae. 2014. An integrated approach for thermal stabilization of a mesophilic adenylate kinase. Proteins 82: 1947-1959.

Moutinho, A. F., F. F. Trancoso, and J. Y. Dutheil. 2019. The impact of protein architecture on adaptive evolution. Mol. Biol. Evol. 36: 2013-2028.

Müller, M. M., J. R. Allison, N. Hongdilokkul, L. Gaillon, P. Kast, W. F. van Gunsteren, P. Marliére, and D. Hilvert. 2013. Directed evolution of a model primordial enzyme provides insights into the development of the genetic code. PLoS Genet. 9: e1003187.

Musto, H., H. Naya, A. Zavala, H. Romero, F. Alvarez-Valín, and G. Bernardi. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 573: 73-77.

Nei, M. and S. Kumar. 2000. Molecular Evolution and Phylogenetics. Oxford Univ. Press, Oxford, UK.

Neupane, K., D. A. Foster, D. R. Dee, H. Yu, F. Wang, and M. T. Woodside. 2016. Direct observation of transition paths during the folding of proteins and nucleic acids. Science 352: 239-242.

Nixon, C. F., S. A. Lim, Z. R. Sailer, I. N. Zheludev, C. L. Gee, B. A. Kelch, M. J. Harms, and S. Marqusee. 2021. Exploring the evolutionary history of kinetic stability in the α-lytic protease family. Biochemistry 60: 170-181.

Pál, C., B. Papp, and L. D. Hurst. 2001. Highly expressed genes in yeast evolve slowly. Genetics 158: 927-931.

Plaxco, K. W., K. T. Simons, I. Ruczinski, and D. Baker. 2000. Topology, stability, sequence, and length: defining the determinants of two-state protein folding kinetics. Biochemistry 39: 11177-11183.

Prabhu, N. P., and A. K. Bhuyan. 2006. Prediction of folding rates of small proteins: empirical relations based on length, secondary structure content, residue type, and stability. Biochemistry 45: 3805-3812.

Ramsey, D. C., M. P. Scherrer, T. Zhou, and C. O. Wilke. 2011. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. Genetics 188: 479-488.

Rebeaud, M. E., S. Mallik, P. Goloubinoff, and D. S. Tawfik. 2021. On the evolution of chaperones and cochaperones and the expansion of proteomes across the Tree of Life. Proc. Natl. Acad. Sci. USA 118: e2020885118.

Robertson, A. D., and K. P. Murphy. 1997. Protein structure and the energetics of protein stability. Chem. Rev. 97: 1251-1268.

Rocha, E. P., and A. Danchin. 2002. Base composition bias might result from competition for metabolic resources. Trends Genet. 18: 291-294.

Rokas, A., and S. B. Carroll. 2008. Frequent and widespread parallel evolution of protein sequences. Mol. Biol. Evol. 25: 1943-1953.

Roscoe, B. P., K. M. Thayer, K. B. Zeldovich, D. Fushman, and D. N. Bolon. 2013. Analyses of the effects of all ubiquitin point mutants on yeast growth rate. J. Mol. Biol. 425: 1363-1377.

Salverda, M. L., E. Dellus, F. A. Gorter, A. J. Debets, J. van der Oost, R. F. Hoekstra, D. S. Tawfik, and J. A. de Visser. 2011. Initial mutations direct alternative pathways of protein evolution. PLoS Genet. 7: e1001321.

Sarkisyan, K. S., D. A. Bolotin, M. V. Meer, D. R. Usmanova, A. S. Mishin, G. V. Sharonov, D. N. Ivankov, N. G. Bozhanova, M. S. Baranov, O. Soylemez, et al. 2016. Local fitness landscape of the green fluorescent protein. Nature 533: 397-401.

Sato, S., S. Xiang, and D. P. Raleigh. 2001. On the relationship between protein stability and folding kinetics: a comparative study of the N-terminal domains of RNase HI, *E. coli* and *Bacillus stearothermophilus* L9. J. Mol. Biol. 312: 569-577.

Serohijos, A. W., Z. Rimas, and E. I. Shakhnovich. 2012. Protein biophysics explains why highly abundant proteins evolve slowly. Cell Rep. 2: 249-256.

Serohijos, A. W., and E. I. Shakhnovich. 2014. Contribution of selection for protein folding stability

in shaping the patterns of polymorphisms in coding regions. Mol. Biol. Evol. 31: 165-176.

Serrano, L., A. G. Day, and A. R. Fersht. 1993. Step-wise mutation of barnase to binase. A procedure for engineering increased stability of proteins and an experimental analysis of the evolution of protein stability. J. Mol. Biol. 233: 305-312.

Shah, P., D. M. McCandlish, and J. B. Plotkin. 2015. Contingency and entrenchment in protein evolution under purifying selection. Proc. Natl. Acad. Sci. USA 112: E3226-E3235.

Sharir-Ivry, A., and Y. Xia. 2018. Nature of long-range evolutionary constraint in enzymes: insights from comparison to pseudoenzymes with similar structures. Mol. Biol. Evol. 35: 2597-2606.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden, and R. E. Sockett. 2005. Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res. 33: 1141-1153.

Sheu, S. Y., D. Y. Yang, H. L. Selzle, and E. W. Schlag. 2003. Energetics of hydrogen bonds in peptides. Proc. Natl. Acad. Sci. USA 100: 12683-12687.

Shi, Z., J. Sellers, and J. Moult. 2012. Protein stability and *in vivo* concentration of missense mutations in phenylalanine hydroxylase. Proteins 80: 61-70.

Sicheri, F., and D. S. Yang. 1995. Ice-binding structure and mechanism of an antifreeze protein from winter flounder. Nature 375: 427-431.

Silverman, J. A., R. Balakrishnan, and P. B. Harbury. 2001. Reverse engineering the $(\beta/\alpha)_8$ barrel fold. Proc. Natl. Acad. Sci. USA 98: 3092-3097.

Snyder, P. W., J. Mecinovic, D. T. Moustakas, S. W. Thomas, 3rd, M. Harder, E. T. Mack, M. R. Lockett, A. Héroux, W. Sherman, and G. M. Whitesides. 2011. Mechanism of the hydrophobic effect in the biomolecular recognition of arylsulfonamides by carbonic anhydrase. Proc. Natl. Acad. Sci. USA 108: 17889-17894.

Sojo, V, C. Dessimoz, A. Pomiankowski, and N. Lane. 2016. Membrane proteins are dramatically less conserved than water-soluble proteins across the Tree of Life. Mol. Biol. Evol. 33: 2874-2884.

Stanley, C. E., Jr., and R. J. Kulathinal. 2016. flyDIVaS: a comparative genomics resource for *Drosophila* divergence and selection. G3 (Bethesda) 6: 2355-2363.

Starr, T. N., J. M. Flynn, P. Mishra, D. N. A. Bolon, and J. W. Thornton. 2018. Pervasive contingency and entrenchment in a billion years of Hsp90 evolution. Proc. Natl. Acad. Sci. USA 115: 4453-4458.

Sternke, M., K. W. Tripp, and D. Barrick. 2019. Consensus sequence design as a general strategy to create hyperstable, biologically active proteins. Proc. Natl. Acad. Sci. USA 116: 11275-11284.

Storz, J. F. 2016. Causes of molecular convergence and parallelism in protein evolution. Nat. Rev. Genet. 17: 239-250.

Suckow, J., P. Markiewicz, L. G. Kleina, J. Miller, B. Kisters-Woike, and B. Müller-Hill. 1996. Genetic studies of the Lac repressor. XV: 4000 single amino acid substitutions and analysis of the resulting phenotypes on the basis of the protein structure. J. Mol. Biol. 261: 509-523.

Sullivan, B. J., T. Nguyen, V. Durani, D. Mathur, S. Rojas, M. Thomas, T. Syu, and T. J. Magliery. 2012. Stabilizing proteins from sequence statistics: the interplay of conservation and correlation in triosephosphate isomerase stability. J. Mol. Biol. 420: 384-399.

Tartaglia, G. G., S. Pechmann, C. M. Dobson, and M. Vendruscolo. 2007. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. Trends Biochem. Sci. 32: 204-206.

Taverna, D. M., and R. A. Goldstein. 2002. Why are proteins marginally stable? Proteins 46: 105-109.

Toft, C., and M. A. Fares. 2010. Structural calibration of the rates of amino acid evolution in a search for Darwin in drifting biological systems. Mol. Biol. Evol. 27: 2375-2385.

Tokuriki, N., F. Stricher, J. Schymkowitz, L. Serrano, and D. S. Tawfik. 2007. The stability effects of protein mutations appear to be universally distributed. J. Mol. Biol. 369: 1318-1332.

Tokuriki, N., and D. S. Tawfik. 2009. Chaperonin overexpression promotes genetic variation and enzyme evolution. Nature 459: 668-673.

Toprak, E., A. Veres, J. B. Michel, R. Chait, D. L. Hartl, and R. Kishony. 2011. Evolutionary paths to antibiotic resistance under dynamically sustained drug selection. Nat. Genet. 44: 101-105.

Tóth-Petróczy, A., and D. S. Tawfik. 2011. Slow protein evolutionary rates are dictated by surface-core association. Proc. Natl. Acad. Sci. USA 108: 11151-11156.

Tripathi, A., K. Gupta, S. Khare, P. C. Jain, S. Patel, P. Kumar, A. J. Pulianmackal, N. Aghera, and R. Varadarajan. 2016. Molecular determinants of mutant phenotypes, inferred from saturation mutagenesis data. Mol. Biol. Evol. 33: 2960-2975.

van den Berg, P. A., A. van Hoek, C. D. Walentas, R. N. Perham, and A. J. Visser. 1998. Flavin fluorescence dynamics and photoinduced electron transfer in *Escherichia coli* glutathione reductase. Biophys. J. 74: 2046-2058.

Venev, S. V., and K. B. Zeldovich. 2018. Thermophilic adaptation in prokaryotes is constrained by metabolic costs of proteostasis. Mol. Biol. Evol. 35: 211-224.

Vieira-Silva, S., M. Touchon, S. S. Abby, and E. P. Rocha. 2011. Investment in rapid growth shapes the evolutionary rates of essential proteins. Proc. Natl. Acad. Sci. USA 108: 20030-20035.

Walsh, I. M., M. A. Bowman, I. F. Soto Santarriaga, A. Rodriguez, and P. L. Clark. 2020. Synonymous codon substitutions perturb cotranslational protein folding in vivo and impair cell fitness. Proc. Natl. Acad. Sci. USA 2020 117: 3528-3534.

Walter, K. U., K. Vamvaca, and D. Hilvert. 2005. An active enzyme constructed from a 9-amino acid alphabet. J. Biol. Chem. 280: 37742-37746.

Wang, M., C. G. Kurland, and G. Caetano-Anollés. 2011. Reductive evolution of proteomes and protein structures. Proc. Natl. Acad. Sci. USA 108: 11954-11958.

Weber, C. C., and S. Whelan. 2019. Physicochemical amino acid properties better describe substitution rates in large populations. Mol. Biol. Evol. 36: 679-690.

Wells, J. A. 1990. Additivity of mutational effects in proteins. Biochemistry 29: 8509-8517.

Wendler, K., J. Thar, S. Zahn, and B. Kirchner. 2010. Estimating the hydrogen bond energy. J. Phys. Chem. A. 114: 9529-9536.

Wheelan, S. J., A. Marchler-Bauer, and S. H. Bryant. 2000. Domain size distributions can predict domain boundaries. Bioinformatics 16: 613-618.

Windle, C. L., K. J. Simmons, J. R. Ault, C. H. Trinh, A. Nelson, A. R. Pearson, and A. Berry.

2017. Extending enzyme molecular recognition with an expanded amino acid alphabet. Proc. Natl. Acad. Sci. USA 114: 2610-2615.

Wolfenden, R., C. A. Lewis, Jr., Y. Yuan, and C. W. Carter, Jr. 2015. Temperature dependence of amino acid hydrophobicities. Proc. Natl. Acad. Sci. USA 112: 7484-7488.

Wolynes, P. G. 2015. Evolution, energy landscapes and the paradoxes of protein folding. Biochimie 119: 218-230.

Worth, C. L., and T. L. Blundell. 2009. Satisfaction of hydrogen-bonding potential influences the conservation of polar side chains. Proteins 75: 413-429.

Worth, C. L., and T. L. Blundell. 2010. On the evolutionary conservation of hydrogen bonds made by buried polar amino acids: the hidden joists, braces and trusses of protein architecture. BMC Evol. Biol. 10: 161.

Wright, C. F., S. A. Teichmann, J. Clarke, and C. M. Dobson. 2005. The importance of sequence diversity in the aggregation and evolution of proteins. Nature 438: 878-881.

Wylie, C. S., and E. I. Shakhnovich. 2011. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc. Natl. Acad. Sci. USA 108: 9916-9921.

Xie, V. C , J. Pu, B. P. Metzger, J. W. Thornton, and B. C. Dickinson. 2021. Contingency and chance erase necessity in the experimental evolution of ancestral proteins. eLife 10: e67336.

Yampolsky, L. Y., and A. Stoltzfus. 2005. The exchangeability of amino acids in proteins. Genetics 170: 1459-1472.

Yang, J. R., B. Y. Liao, S. M. Zhuang, and J. Zhang. 2012. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc. Natl. Acad. Sci. USA 109: E831-E840.

Yang, J. R., S. M. Zhuang, and J. Zhang. 2010. Impact of translational error-induced and error-free misfolding on the rate of protein evolution. Mol. Syst. Biol. 6: 421.

Yang, Z. 2014. Molecular Evolution: a Statistical Approach. Oxford Univ. Press, Oxford, UK.

Yeh, S. W., J. W. Liu, S. H. Yu, C. H. Shih, J. K. Hwang, and J. Echave. 2014. Site-specific structural constraints on protein sequence evolutionary divergence: local packing density versus solvent exposure. Mol. Biol. Evol. 31: 135-139.

Zana, R. 1975. On the rate-determining step for helix propagation in the helix-coil transition of polypeptides in solution. Biopolymers 14: 2425-2428.

Zeldovich, K. B., P. Chen, and E. I. Shakhnovich. 2007. Protein stability imposes limits on organism complexity and speed of molecular evolution. Proc. Natl. Acad. Sci. USA 104: 16152-16157.

Zhang, J., and S. Kumar. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. Mol. Biol. Evol. 14: 527-536.

Zhang, J., S. Maslov, and E. I. Shakhnovich. 2008. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. Mol. Syst. Biol. 4: 210.

Zhang, J., and J. R. Yang. 2015. Determinants of the rate of protein sequence evolution. Nat. Rev. Genet. 16: 409-420.

Zhang, X. J., W. A. Baase, B. K. Shoichet, K. P. Wilson, and B. W. Matthews. 1995. Enhancement of protein stability by the combination of point mutations in T4 lysozyme is additive. Protein Eng. 8: 1017-1022.

Zhang, Y., I. A. Hubner, A. K. Arakaki, E. Shakhnovich, and J. Skolnick. 2006. On the origin and highly likely completeness of single-domain protein structures. Proc. Natl. Acad. Sci. USA 103: 2605-2610.

Zhao, J., A. J. Burke, and A. P. Green. 2020. Enzymes with noncanonical amino acids. Curr. Opin. Chem. Biol. 55: 136-144.

Zhou, T., W. Gu, and C. O. Wilke. 2010. Detecting positive and purifying selection at synonymous sites in yeast and worm. Mol. Biol. Evol. 27: 1912-1922.

Zou, Z., and J. Zhang. 2015. Are convergent and parallel amino acid substitutions in protein evolution more prevalent than neutral expectations? Mol. Biol. Evol. 32: 2085-2096.

Figure 12.1. A pictorial guide to the structures of the twenty amino acids generally used in proteins. The side-chains unique to each amino acid are shown to the left of the amino (NH_2) group; all unlabeled nodes are carbon atoms, each of which has four bonds (all of which are to hydrogens, unless labeled otherwise). The genetic (DNA) codes for each amino acid are those for the so-called "universal" genetic code; slight variants of this code exist in the nuclear genomes of some unicellular eukaryotes, as well as in the mitochondrial genomes of many eukaryotes. As denoted by the index at the bottom, the color-coded circles denote some of the unique biochemical properties of individual residues. Aromatic residues contain closed carbon rings. Acidic residues are negatively charged, whereas basic residues are positively charged.

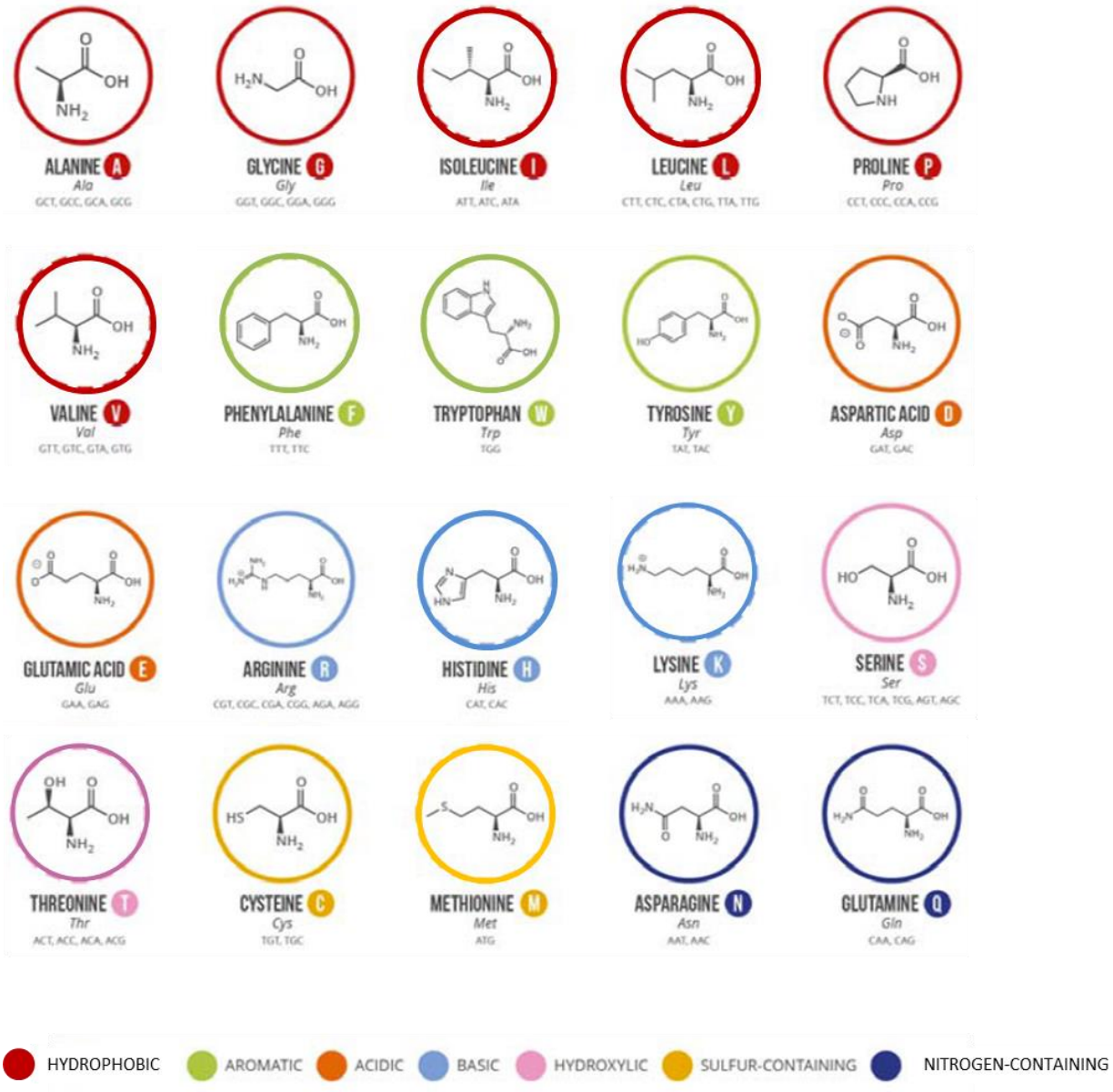
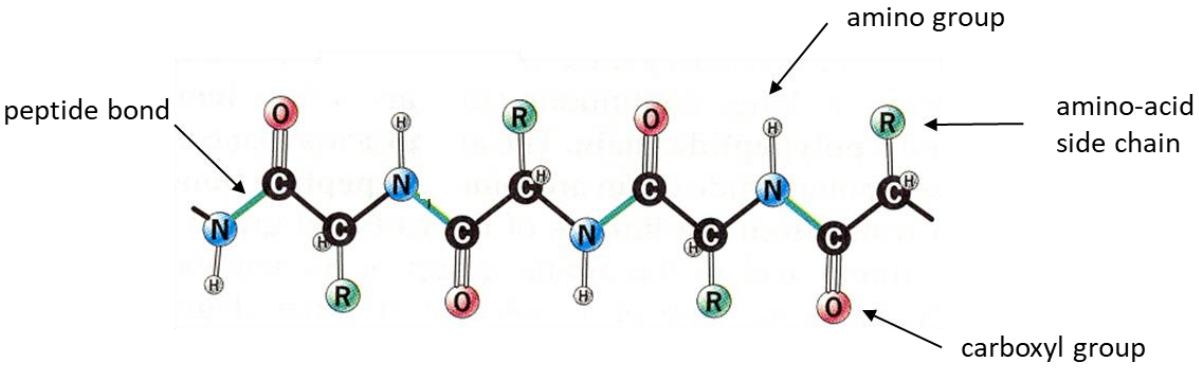


Figure 12.2. Second- and third-order structure of proteins. **Above**) Polypeptide chains are built by covalent bonding between the amide and carboxyl groups of adjacent amino acids (blue lines). The specific features of each amino acid then being represented as side chains (R) off the overall backbone. **Below**) The emerging chain typically goes through a folding process, which in some cases allows for further covalent bonding between the sulfide residues of two cysteines. Unbound amide and carboxyl groups reside at the first and final amino acid.



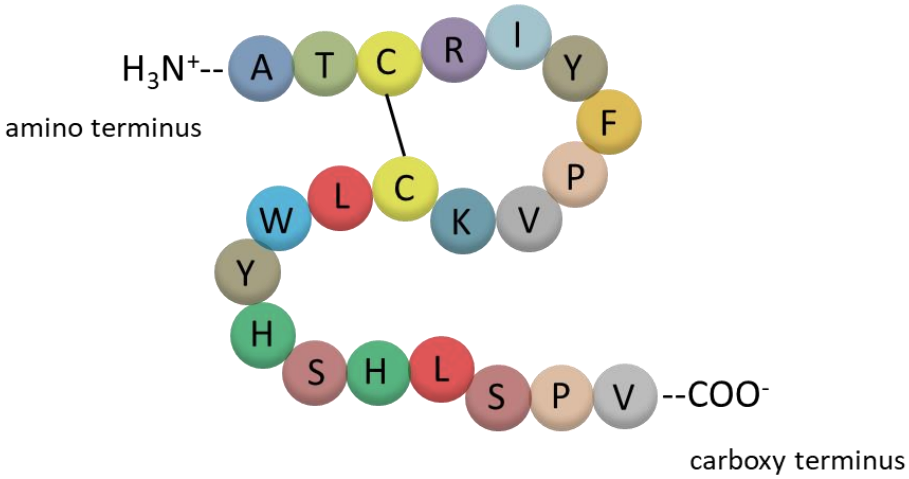


Figure 12.3. Pathways for synthesis of amino acids from precursors generated in central metabolism.

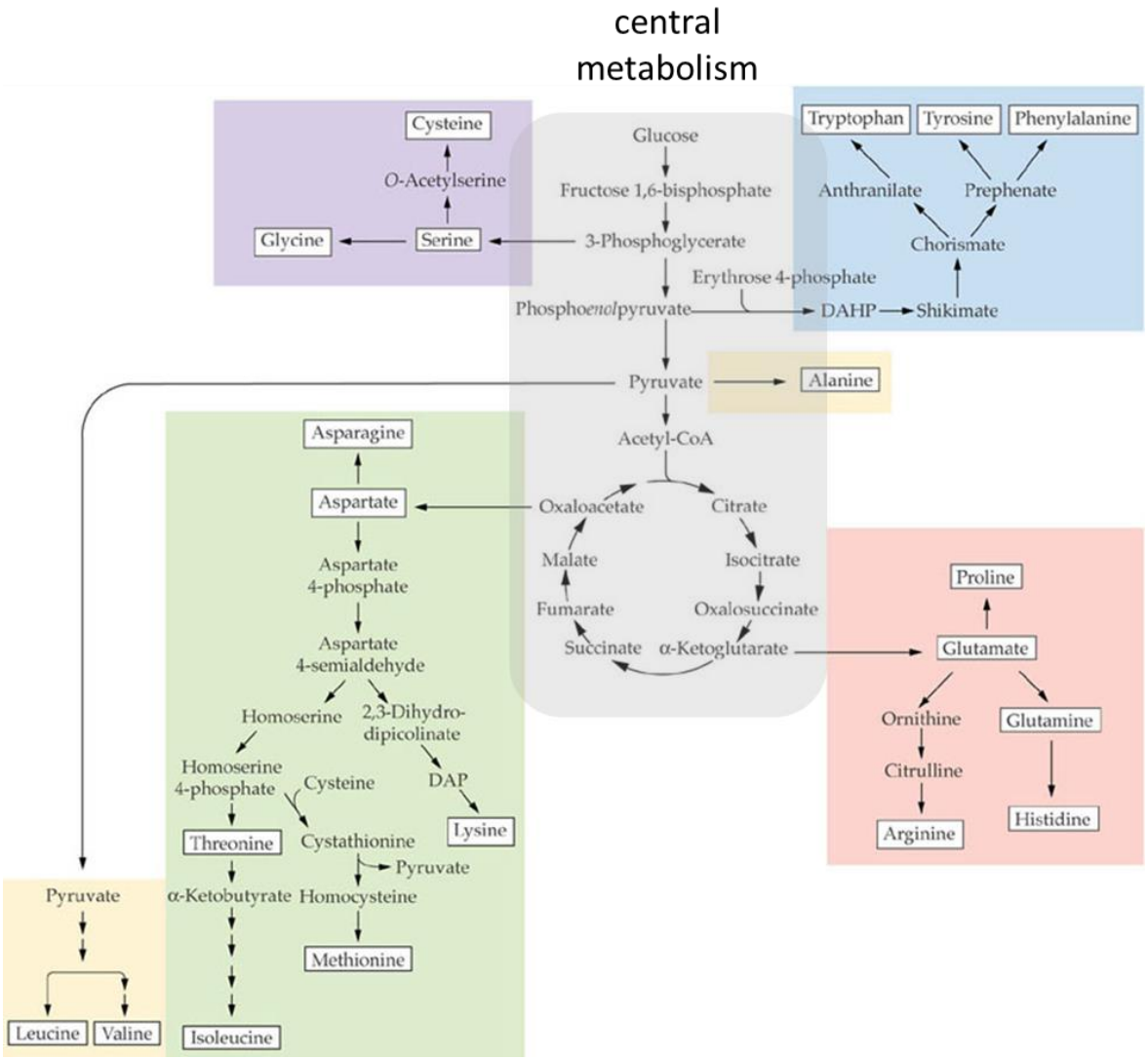
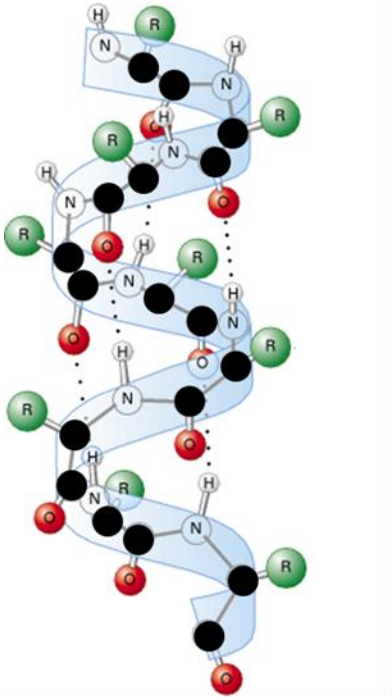
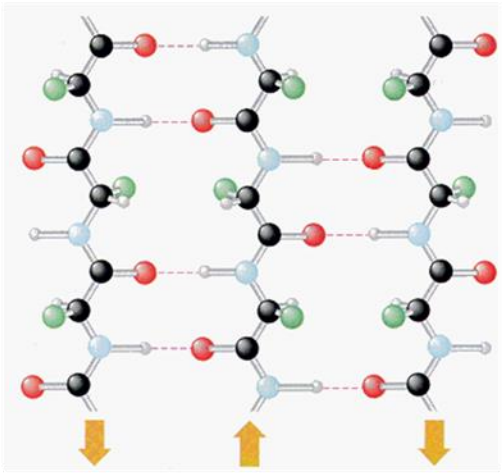


Figure 12.4. Two common types of protein substructures. **Left**) Hydrogen bonds between amino acids separated by three steps, shown as dotted lines, hold alpha helices together. **Right**) Here the bonds are between more distant amino acids on adjacent chains, creating beta sheets. The ribbon diagram to the lower right depicts an anti-parallel beta sheet.



Alpha helix



Beta sheet

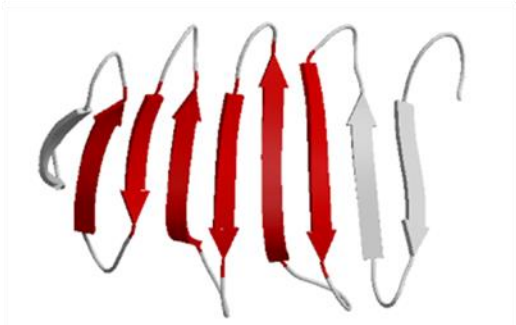


Figure 12.5. Reduction in the folding rate (k_f) with increasing chain length (L), as described by Equation 12.1. The fit is based on 80 proteins known to engage in two- and multi-state folding. From Dill et al. (2011).

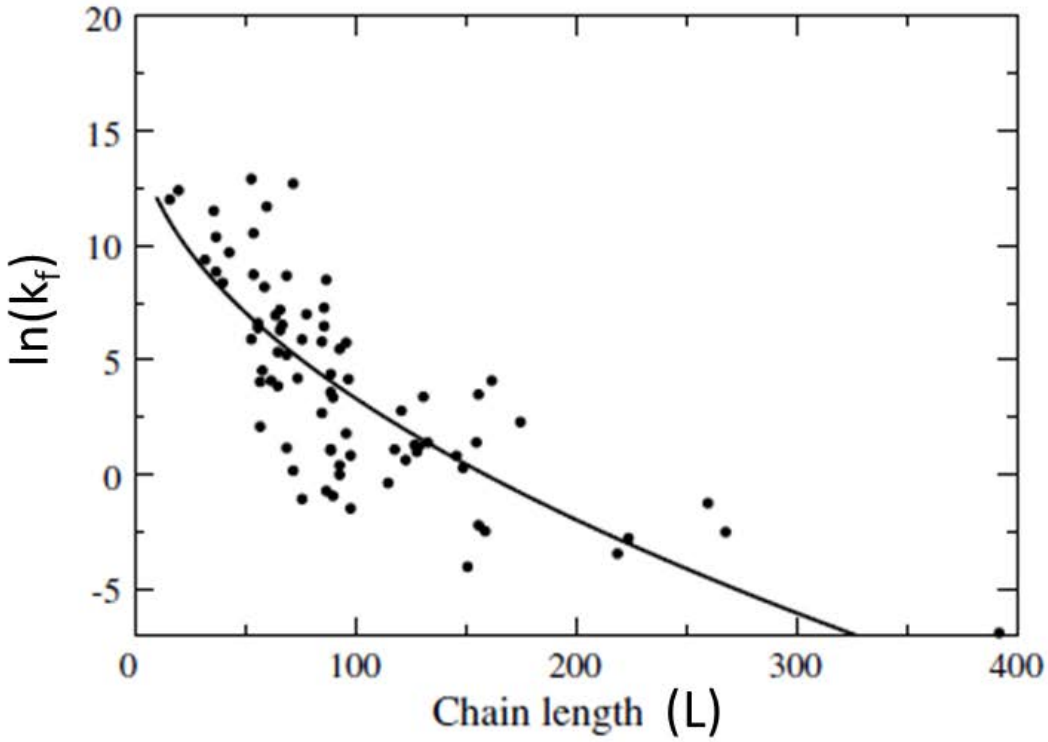


Figure 12.6. Evolution of population mean phenotypes on a hyperbolic fitness function (given by the dashed red line). Shown are steady-state distributions of mean phenotypes under a balance between the forces of mutation, selection, and random genetic drift. It is assumed a large number of factors contribute to the trait, in this case folding stability, with each factor having two alternative states (positive or negative with respect to stability). The dashed black line gives the distribution under effective neutrality (i.e., when the effective population size is small enough that the power of selection is overcome by random genetic drift), the position and width of the distribution being a function of the relative rates of mutations to stabilizing vs. destabilizing mutations. The distribution shifts to the right with larger effective population sizes (given as the four labeled peaks, $N_e = 10^3$ to 10^6), as the efficiency of natural selection becomes greater. Details on the methods used to generate these results are given in Chapter 5 and in Lynch (2018).

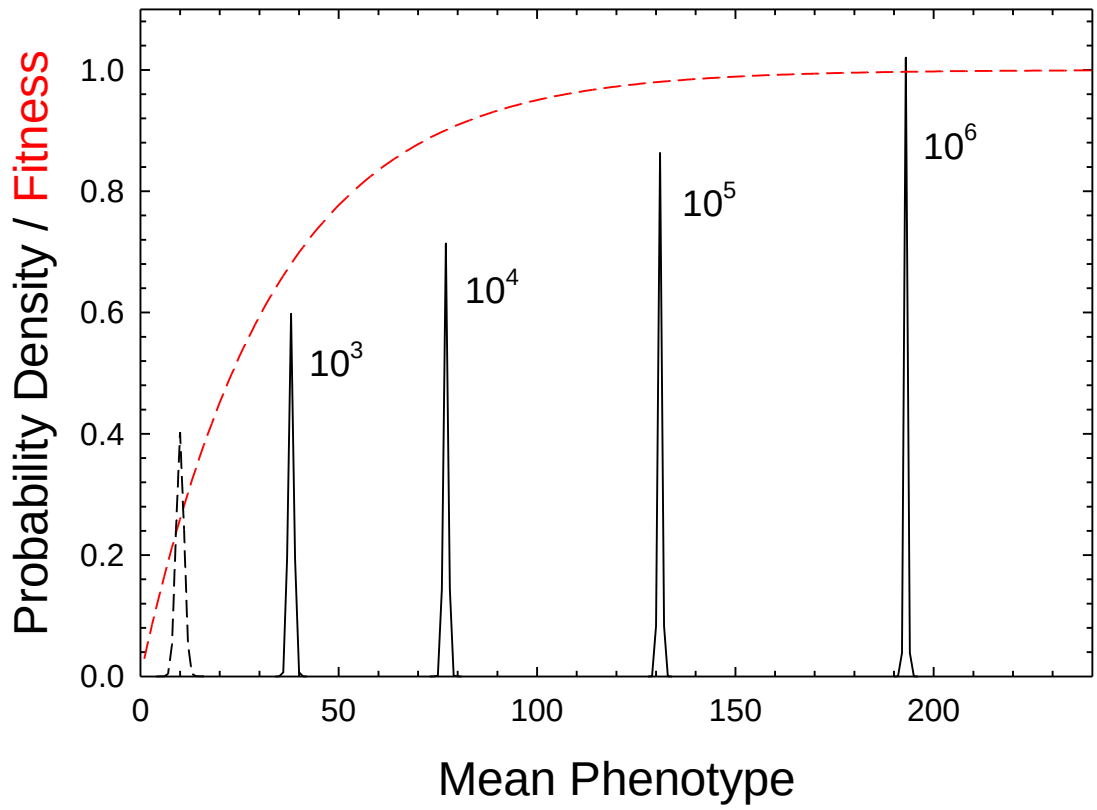


Figure 12.7. Frequency distributions of the destabilizing effects of mutations for protein folding. ΔG is a measure of folding stability, with negative values denoting greater stability; whereas $\Delta\Delta G$ is the difference in ΔG between the mutant and native protein. Positive values of $\Delta\Delta G$ denote destabilizing mutations; negative values denote stabilizing mutations. Plots on the left map the distribution of $\Delta\Delta G$ for random amino-acid exchanges at sites across the entire protein; results are given for five specific proteins and a larger set of experimental data. In the right panel mutations are subdivided into those affecting surfaces and internal cores of proteins. From Tokuriki et al. (2007) and Tokuriki and Tawfik (2009).

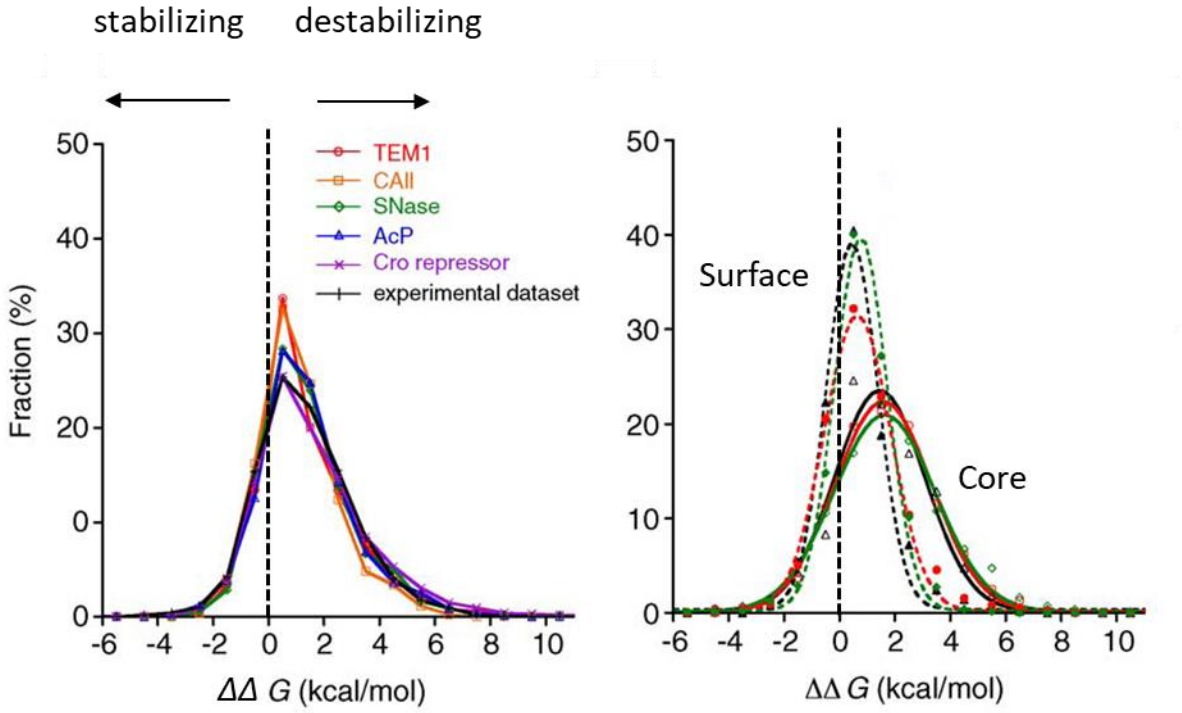


Figure 12.8. The distribution of folding stabilities of the enzyme dihydrofolate reductase for 36 species of mesophilic bacteria, measured as the temperature (°C) required for 50% unfolding *in vitro*. The position for *E. coli* is given as a reference point. From Bershtein et al. (2015).

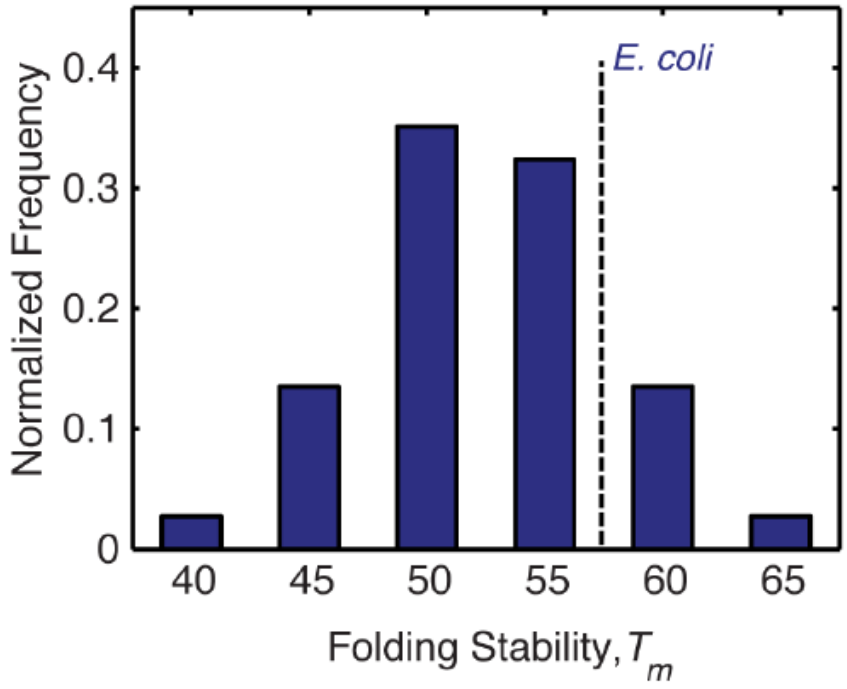
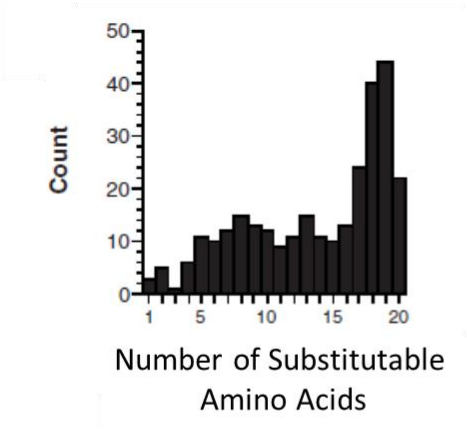
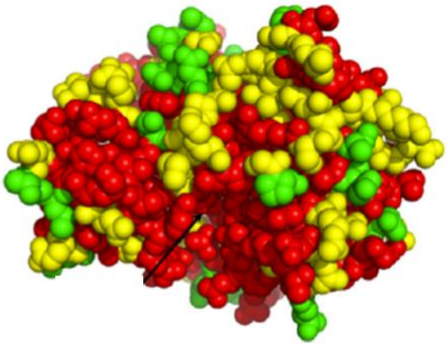
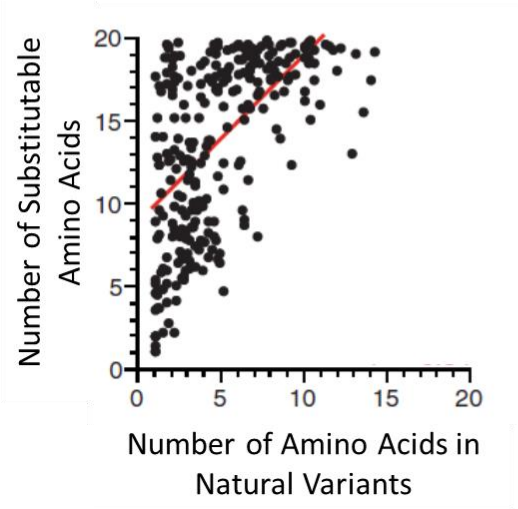


Figure 12.9. Exhangeability of amino acids in the TEM-1 variant of β-lactamase, based on observations of all 19 possible amino-acid alterations at each position in the protein. **Upper left**) The distribution of the number of possible amino acids residing at each site that retain enzyme function. **Lower left**) Weak relationship between the number of amino acids observed at particular positions in sequences of natural variants of β-lactamase and the number of the 19 possible amino-acid exchanges in laboratory manipulations that retain enzyme function. **Right**) Spatial pattern of amino-acid exhangeability on the surface residues, with red denoting four or fewer, yellow five to nine, and green ten or more amino acids conferring a functional enzyme. The upper panel illustrates the residues on the side of the molecule containing the active site, whereas the lower panel illustrates the opposite side of the molecule. Note that these are examinations of single amino-acid changes, and some exchanges become possible on different genetic backgrounds. From Deng et al. (2012) and Firnberg et al. (2014).







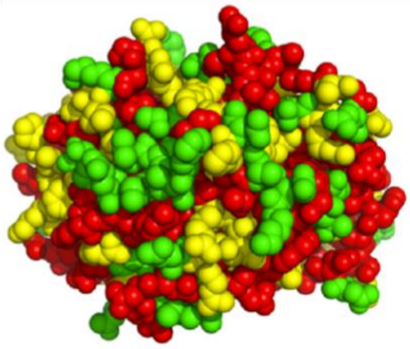


Figure 12.10. Examples of estimated distributions of fitness effects of random mutations in laboratory constructs.**Upper left**) β-lactamase, with fitness given on a log scale for three types of single amino-acid substitutions (from Firnberg et al. 2014). **Upper right**) Nonsynonymous substitutions in genes involved in arabinose metabolism in *Salmonella* (from Lind et al. 2016). **Lower left and right**) Respectively, synonymous and nonsynonymous substitutions in ribosomal protein genes in *Salmonella* (Lind et al. 2010).

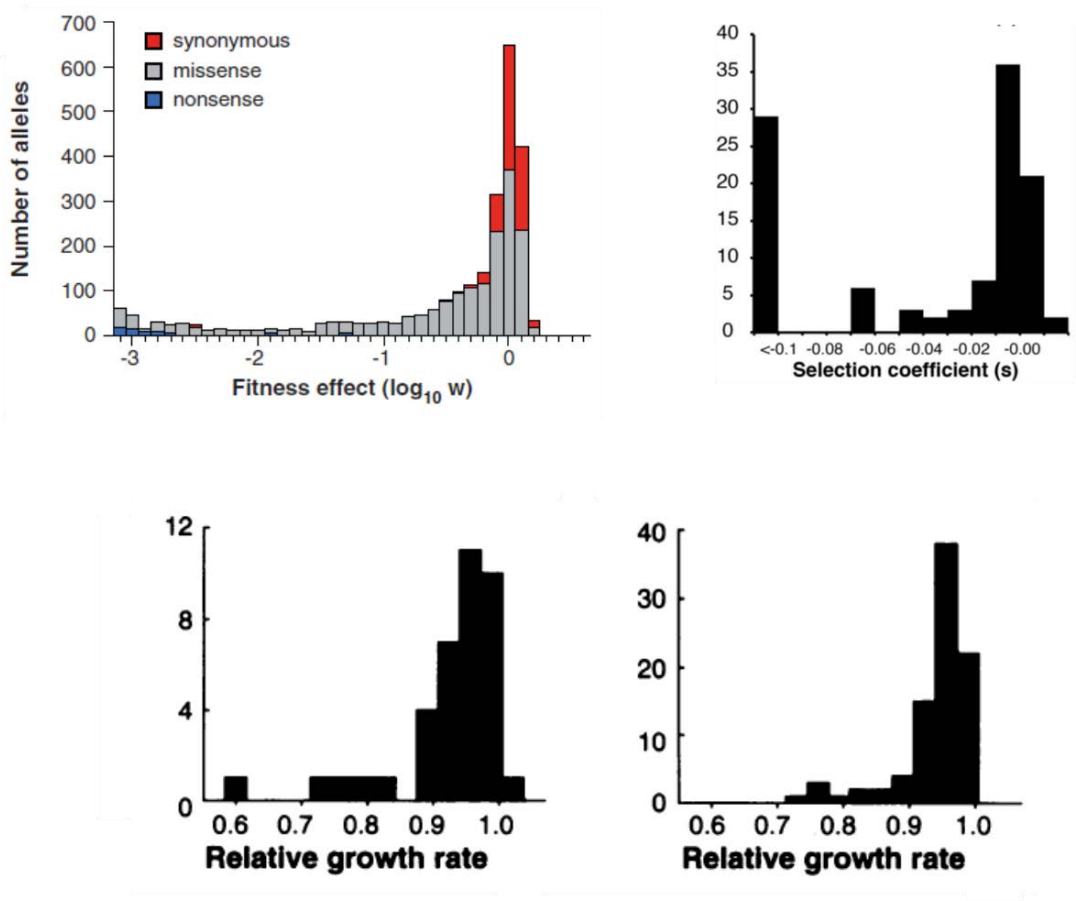


Figure 12.11. G+C compositions for a wide range of species relative to the expectations under mutation-pressure alone (the neutral equilibrium expectation), using directly estimated spectra of spontaneous mutations obtained from mutation-accumulation experiments (Long et al. 2018). The dashed diagonal lines give the null expectation for the situation in which G+C content is entirely driven by mutation. Replacement sites are positions within codons for which all mutations lead to an amino-acid substitution, whereas silent sites are those for which mutations have no effect on the encoded amino acid. As revealed by the distribution of points above the diagonal, nearly all species have excess G+C content, and the pattern is even more extreme for silent sites in protein-coding genes. The quantity $\phi N_e s$, where $\phi = 2$ for haploids and 4 for diploids, is a measure of the strength of selection relative to drift (Chapter 5), with $\phi N_e s \ll 1$ implying effective neutrality.

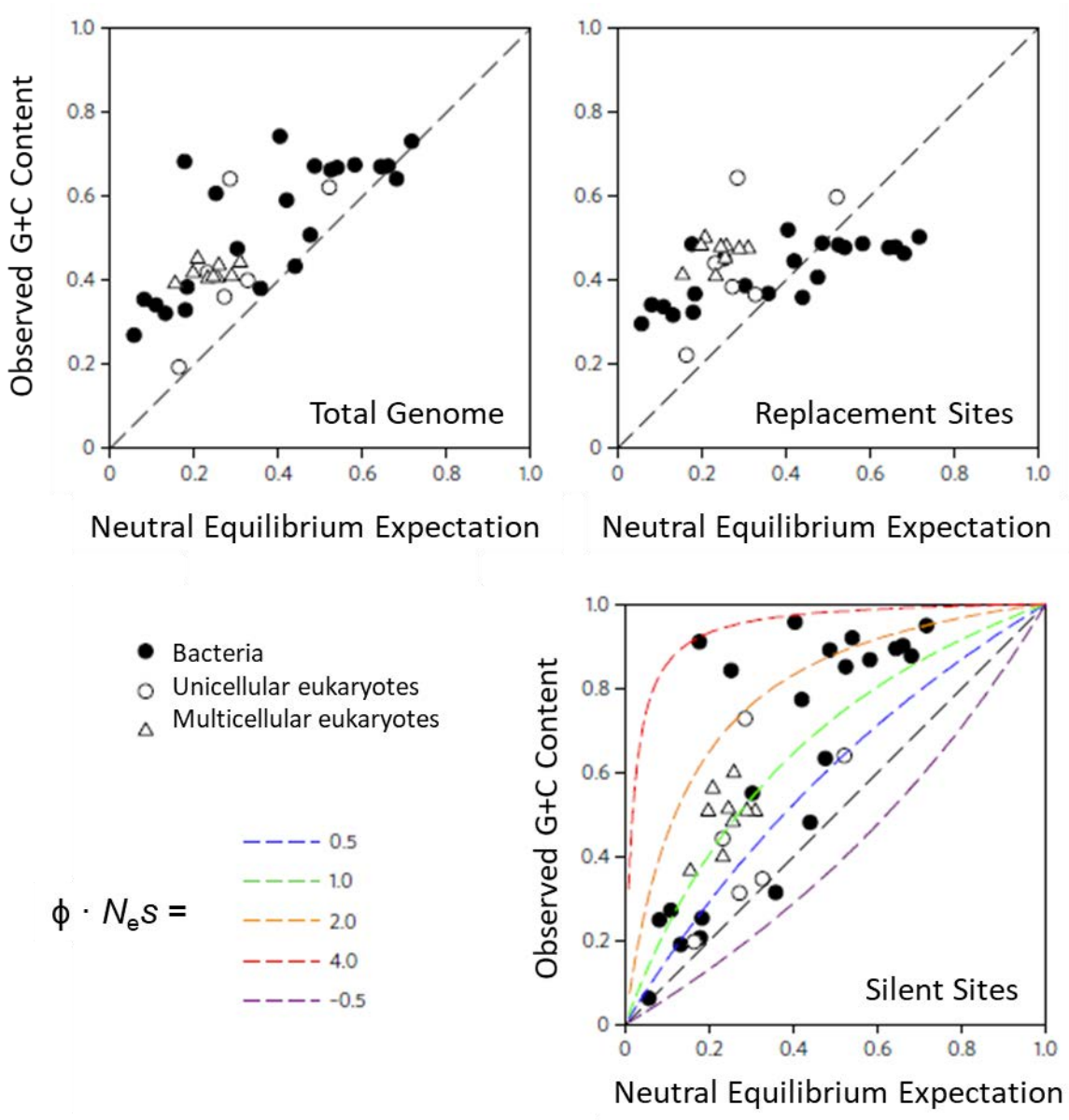


Figure 12.12. A cartoon view of the restricted paths to evolution for a three-residue protein with two alternative states at each site. Of the eight possible sequences, four (in black) are effectively inviable owing to their deleterious fitness effects being large enough for selection to overcome random genetic drift. The four states in red are either effectively neutral, or involve the substitution of a selectively beneficial change on the adjacent background. For example, A is permissible only on a bc, Bc, or BC background, and B is only permissible on a background containing A. Under this view, a lineage is free to wander back and forth along the red route. One of the two end states, abc and ABC, may be selectively superior to the other.