Roles for Synonymous Codon Usage in Protein Biogenesis

Julie L. Chaney[1] and Patricia L. Clark[1,2]

[1]Department of Chemistry and Biochemistry, [2]Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, Indiana 46556; email: jchaney@nd.edu, pclark1@nd.edu

Annu. Rev. Biophys. 2015. 44:143–66

First published online as a Review in Advance on February 26, 2015

The *Annual Review of Biophysics* is online at biophys.annualreviews.org

This article's doi: 10.1146/annurev-biophys-060414-034333

Keywords

translation, ribosome, cotranslational folding, protein aggregation, protein degradation

Abstract

Owing to the degeneracy of the genetic code, a protein sequence can be encoded by many different synonymous mRNA coding sequences. Synonymous codon usage was once thought to be functionally neutral, but evidence now indicates it is shaped by evolutionary selection and affects other aspects of protein biogenesis beyond specifying the amino acid sequence of the protein. Synonymous rare codons, once thought to have only negative impacts on the speed and accuracy of translation, are now known to play an important role in diverse functions, including regulation of cotranslational folding, covalent modifications, secretion, and expression level. Mutations altering synonymous codon usage are linked to human diseases. However, much remains unknown about the molecular mechanisms connecting synonymous codon usage to efficient protein biogenesis and proper cell physiology. Here we review recent literature on the functional effects of codon usage, including bioinformatics approaches aimed at identifying general roles for synonymous codon usage.

Contents

INTRODUCTION

The genetic code is degenerate: 61 trinucleotide codons are used to encode the 20 standard amino acids. Whereas both methionine and tryptophan are encoded by only one codon, each of the other 18 amino acids is encoded by two to six synonymous codons (**Figure 1**). Because synonymous codons do not alter the encoded amino acid sequence, they were historically considered equivalent and interchangeable. Accordingly, DNA mutations that alter synonymous codon usage are referred to as "silent" mutations and are often assumed to be neutral in models of mutation and selection (29, 62, 72, 78, 117). Indeed, commonly used models of molecular evolution are based on the assumption that synonymous codon changes are merely genomic background noise (62, 72). However, a growing body of evidence has revealed that synonymous codon choice is nonrandom and can affect multiple aspects of protein biogenesis in diverse organisms, including humans. Here we review the current status of our understanding of synonymous codon usage and its impacts on diverse aspects of protein biogenesis, including transcription, translation, cotranslational folding, secretion, and posttranslational modifications. We also provide an overview of algorithms used to identify and quantify patterns in synonymous codon usage, as well as our current understanding of the effects of these patterns on protein biogenesis.

Codon Bias

Usage frequencies of synonymous codons within the same genome often differ, in some cases by more than an order of magnitude (56, 57). Although almost all organisms use the same genetic code (**Figure 1*a***), the specific synonymous codons that are rare versus common vary between organisms and correlate with genome %GC bias (19, 40). For many sets of synonymous codons, the base composition at the first and second codon positions is fixed and varies only at the third

a

	NTN		NCN		N		NGN	
TNN	TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
	TTC		TCC		TAC		TGC	
	TTA	Leu	TCA		TAA	STOP	TGA	STOP
	TTG		TCG		TAG		TGG	Trp
CNN	CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
	CTC		CCC		CAC		CGC	
	CTA		CCA		CAA	Gln	CGA	
	CTG		CCG		CAG		CGG	
ANN	ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
	ATC		ACC		AAC		AGC	
	ATA		ACA		AAA	Lys	AGA	Arg
	ATG	Met	ACG		AAG		AGG	
GNN	GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
	GTC		GCC		GAC		GGC	
	GTA		GCA		GAA	Glu	GGA	
	GTG		GCG		GAG		GGG	

b

```
Met-Leu-Asp-Gly-Pro
ATG TTA GAT GGT CCT
    TTG GAC GGC CCC
    CTT     GGA CCA
    CTC     GGG CCG
    CTA
    CTG
```

Figure 1

(*a*) The standard genetic code. (*b*) A hypothetical peptide sequence and its possible coding sequences. Most amino acids are encoded by a set of synonymous codons that differ only at the wobble position. There are five exceptions: Leucine, serine, and arginine are each encoded by six codons and therefore also include different nucleotides at the first (all three) and second (serine only) codon positions. Methionine and tryptophan are each encoded by only one codon.

(wobble) position (**Figure 1*b***). For example, proline is encoded by CCN, where N can be any nucleotide. Thus, for protein coding sequences, the average %GC content can be adjusted by altering the %GC content at the third codon position (%GC3). In genomes with a high average %GC, common synonymous codons typically end in G or C, whereas in low average %GC genomes, common codons are more likely to end in A or T (91). Historically, it was thought that average %GC was driven primarily by an organism's growth temperature (90), although it now appears that other effects can also be important drivers (87, 101, 102, 134).

Direct Effects of Synonymous Codon Usage

%GC3: percentage of guanine (G) and cytosine (C) nucleotides at the third nucleotide position in codons

Compared with more common synonymous codons, rare codons generally correlate with lower levels of cognate tRNA or weaker interactions at the wobble position with the same tRNA (27, 39, 59, 63, 64, 116). For these reasons, rare codons are generally thought to be translated more slowly than common codons, and experimental evidence supports this hypothesis (80, 98, 114, 115). Indeed, although average translation rates are ~20 amino acids (aa)/s in prokaryotes and ~5 aa/s in

eukaryotes, local translation rates can vary by more than an order of magnitude (20, 114). However, codon rarity is not the only determinant of translation rate, and a detailed understanding of all factors that affect translation rate has been limited by the technical challenges of measuring absolute translation rates of individual codons in vivo. Translation rates measured from in vitro translation systems are typically much slower and less variable than rates measured in vivo (18, 103, 123). Ribosome profiling is a new method that measures relative ribosome occupancy of mRNA sites on a transcriptome-wide level (65, 81), and in one case it has been possible to convert these occupancies to absolute translation rates (66). Most ribosomal profiling studies have not identified a broad connection between codon usage and ribosome occupancy except in a few specific cases caused by the depletion of charged tRNA or the presence of Shine-Dalgarno (SD) motifs (81). However, most ribosome profiling studies have focused on the most extreme translational pauses across an entire proteome. In contrast, a new analysis of yeast ribosome profiling data, designed to quantify comparatively subtle differences in translation rate associated with synonymous codon usage, discovered that rare codons are on average translated more slowly than synonymous common codons (50).

Shine-Dalgarno (SD) motif: an mRNA nucleotide motif (consensus AGGAGG) in bacteria that interacts with the 16S rRNA and may introduce translational pauses

Two systems were recently developed to measure the relative effects of individual or small segments of synonymous codons on translation rates in vivo. The first is a fluorescent biosensor called YKB. YKB consists of three fluorescent half-domains connected by flexible linkers (104) (**Figure 2*b***). The N-terminal Y half-domain and the C-terminal B half-domain compete with each other to interact with the central K half-domain, and the outcome of this competition determines whether a yellow or cyan fluorescent protein is formed. Translational pauses in a switch region at the 5′ end of the B half-domain allow additional time for a Y-K interaction to form cotranslationally, thus increasing the proportion of Y-K (yellow) fluorescence observed. Introducing synonymous rare codons into the switch region slowed translation rates, resulting in increased yellow fluorescence and decreased cyan fluorescence. This yellow/cyan fluorescence ratio changed predictably with the rareness of the synonymous codons used, and showed directly for the first time that synonymous codons, presumably because of their effects on translation rate, can alter not just a protein folding pathway but also the final, folded structure of the encoded protein (104).

While the YKB biosensor represents an engineered approach to measuring codon-specific effects on translation rate, a second translation rate sensor system is based on the naturally occurring His operon (20). Fast translation of the His leader peptide results in transcriptional termination, reducing expression of the downstream genes, whereas slow translation enhances expression. Replacing the downstream genes with β-galactosidase enabled researchers to relatively compare the translation rates for codons inserted into the leader peptide coding sequence and demonstrate that translation rate varies considerably for different codons, including synonymous codons decoded by the same tRNA (see also the sidebar, What Determines the Translation Rate of a Codon?). Codon context (the identity of the 5′ and 3′ flanking codons) also affected translation rates (20).

Although beyond the scope of this review, the amino acid sequence of the nascent chain can also perturb the local translation rate. Positively charged amino acids can reduce the translation rate because of their interaction with negative charges in the ribosome exit tunnel (16, 82). Rates of peptide bond formation vary among amino acids (137) and are especially slow for proline, leading to translational pauses at polyproline motifs (31, 132, 139). Specific peptide motifs, including the well-characterized *Escherichia coli* SecM (secretion monitor) stalling sequence (23, 41, 67, 139), can also stall translation. Nascent peptide effects add an extra layer of complexity to the interpretation of translation rate changes, as codon usage effects may be hidden among these and other translation rate regulatory mechanisms.

Translational accuracy is slightly lower for rare codons than for common codons (69, 75, 110), although as for translation rates, experimental data for the translational accuracy of specific codons

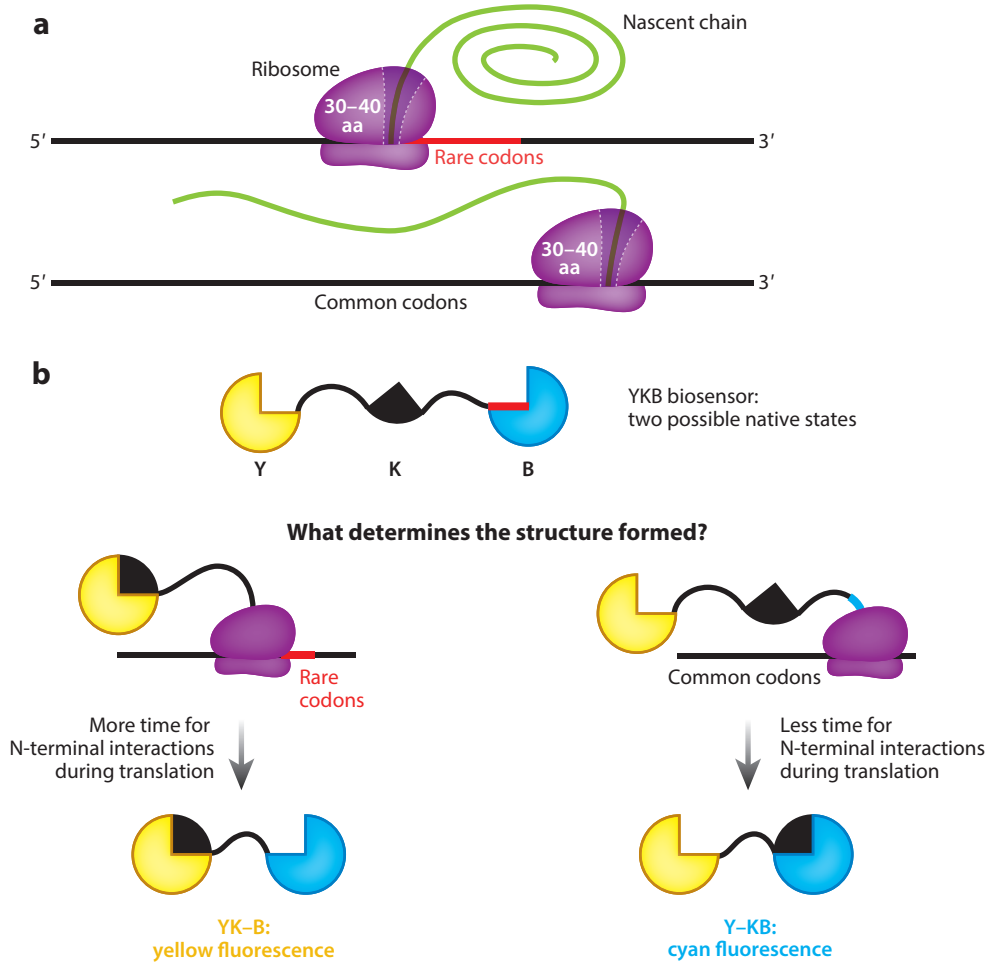

Figure 2

(*a*) Rare codons are hypothesized to introduce translational pauses that modulate cotranslational protein folding. A translational pause could allow additional time for the nascent chain to form a 3D structure. The most C-terminal 30–40 amino acids of the nascent protein are constrained in the ribosomal exit tunnel and unable to form a 3D structure. Hence a rare codon cluster that promotes the folding of a protein structural unit is expected to be offset 30–40 codons from the region encoding that structural unit. (*b*) The YKB biosensor is a split fluorescence system designed to detect changes in translation rates due to synonymous codon usage (104). YKB consists of three fluorescent half-domains: The N-terminal Y half-domain and the C-terminal B half-domain compete to interact with the central K half-domain, and the resulting native structure determines the color of fluorescence observed. YK/KB fluorescence ratios show a stronger correlation with relative codon usage than do tRNA levels, a tRNA wobble metric, %GC, or mRNA stability (24, 104).

is limited. Certain codons may also be rare in specific genomes because they form deleterious nucleotide motifs and are thus subject to negative selection. For example, CG dinucleotide pairs have an increased mutational susceptibility in methylated eukaryotic genomes (99). As a result, in the human genome NCG is rarely used to encode amino acids encoded by NCN codons (149).

Synonymous codon usage within coding sequences can also alter the stability of mRNA structure, affecting cell physiology. An analysis of mutations in a bacterial long-term evolution experiment indicated that mutations that alter mRNA stability were selected against (21).

WHAT DETERMINES THE TRANSLATION RATE OF A CODON?

Codon usage frequencies generally correlate with levels of the cognate tRNAs. However, due to wobble pairing rules, some tRNAs recognize multiple synonymous codons, and the strength of the wobble pairing interaction can also influence the translation rate. If the translation rate is determined by the tRNA level alone, codons decoded by the same tRNA should be translated at the same rate.

For example, *Escherichia coli* has only one Glu tRNA, which decodes both Glu codons, GAA and GAG (see **Supplemental Figure 1**; follow the **Supplemental Material link** from the Annual Reviews home page at **http://www.annualreviews.org**). Sørensen & Pedersen (115) compared the translation rates of the GAA and GAG codons. The rarer GAG codon was translated three times more slowly than the common codon GAA, indicating that tRNA levels are not the sole determinant of translation rates. In addition to wobble pairing, other effects, including codon context, can modulate translation rates (see main text).

Supplemental Material

Stable mRNA structures can cause translational pauses (129), so mRNA structure provides an additional mechanism to regulate translation speed and, by extension, the temporal coordination of cotranslational folding and/or interactions of the nascent chain with other cellular components (see Cotranslational Protein Folding and Cotranslational Interactions with Other Cellular Components, below). For example, in the polytopic membrane protein CFTR, a synonymous mutation (Ile507 ATC→ATT) exacerbates the protein folding defect caused by the ΔF508 mutation linked to cystic fibrosis. This synonymous mutation is hypothesized to alter mRNA structure by introducing large single-stranded loops that lead to slower translation (7, 79).

Selection for mRNA structure is an example of synonymous codon selection that can affect protein biogenesis, and hence cell physiology and fitness, for reasons not directly related to the decoding of those codons at the ribosome. Functionally significant RNA motifs are another example. For example, synonymous single nucleotide polymorphisms (SNPs) may either weaken or strengthen splice sites (105, 120). Likewise, ribosomal profiling experiments have shown that bacterial ribosomes accumulate at internal SD motifs (consensus sequence AGGAGGT), which bind to the complementary anti-SD motif found within small subunit rRNA (81). SD motifs may occur in any reading frame and are underrepresented within coding sequences, suggesting that they are generally under negative selection. The selection against SD motifs appears to influence synonymous codon usage in *E. coli*: Glycine-glycine pairs are unlikely to be encoded by GGA-GGT and GGA is the rarest glycine codon in *E. coli*. In support of this interpretation, translation of internal SD motifs by ribosomes that lack an anti-SD motif did not increase ribosome occupancy at these motifs (81).

Single nucleotide polymorphism (SNP): a single nucleotide variant common in a population, sometimes associated with a disease

Selective Pressures on Codon Usage: Why Are Rare Codons Used at All?

Rare codons were historically viewed as mildly deleterious because of their lower rates of translation speed and accuracy. Scientists assumed that evolution favored common codons in general, but because the negative selective pressure was not strong enough to completely eliminate rare codons, some continued to be incorporated as a result of mutational drift (14, 108, 113). However, synonymous rare codons are not distributed randomly across coding sequences; instead, they tend to appear in clusters within the protein coding sequences of most species (24, 25, 96). This discovery led to the hypothesis that rare codons are under positive selection in certain locations because of some functional role (25). In support of the selection hypothesis, Pechmann & Frydman

(97) discovered that rare codons from homologous coding sequences of 10 closely related yeast species tend to occur in similar locations. Analogous to the conservation of functionally important amino acid residues, their observation suggested that synonymous rare codons could be functionally important within these protein families. However, other selection forces could also cause the appearance of rare codon clustering. For example, if selection favors common codons, but this selection is strong in some parts of coding sequences and weak or absent in others, mutational drift will result in a higher concentration of rare codons in the regions of weak selection. If these regions of weak codon selection occur in equivalent regions of homologous coding sequences, this can create the appearance of conservation.

Because of their lower translational accuracy, rare codons are hypothesized to be under stronger negative selection in coding sequence regions encoding functionally important amino acids (35, 151). Highly expressed sequences tend to be encoded by more common codons (8, 51, 52, 63), and this finding has led to the hypothesis that common codons are under stronger positive selection in highly expressed genes because of their faster translation rate, leading to more efficient protein production (55, 109). Thus, rare codons would be expected to cluster in more lowly expressed genes with weaker selection for translational efficiency. However, in many organisms numerous highly expressed genes, such as those encoding ribosomal proteins, contain rare codon clusters (24). The ambiguity of these results underscores the importance of examining experimental data for the functional importance of codon usage.

EFFECTS OF SYNONYMOUS CODON USAGE ON PROTEIN BIOGENESIS

Cotranslational Protein Folding

Protein synthesis occurs at the peptidyltransferase active site of the ribosome, which is located within the core of its roughly spherical, ~25-nm-diameter structure. As the ribosome decodes the mRNA sequence, the newly synthesized nascent chain initially passes through a narrow, ~10-nm-long exit tunnel (6, 143) (**Figure 2*a***). This tunnel constrains the conformations of the most recently synthesized 30–40 aa of the nascent protein chain, as it is too narrow to permit the nascent chain to adopt significant tertiary structure (although it is wide enough to permit, and even stabilize, an α-helical conformation) (138). After a nascent chain is 30–40 aa long, its N-terminus emerges at the surface of the ribosome, and these conformational restrictions are significantly reduced.

The vectorial appearance of the nascent chain during its synthesis means that in vivo the N-terminus of a polypeptide chain has an opportunity to begin folding before its C-terminus appears. To what extent this folding vector can alter folding mechanisms or the final folded structure of an encoded protein remains controversial. To date, researchers have studied protein folding primarily in vitro, by diluting full-length protein chains from a chemical denaturant, using small, simple model proteins that fold robustly, meaning they can assume their native structure from a wide variety of unfolded conformations (22). Results from this approach have led to the current dominant thermodynamic paradigm, developed by Anfinsen (1), that the amino acid sequence of a protein provides all the information necessary to specify its native, 3D structure. Under this paradigm, the structure that a protein adopts should be unaffected by the starting conformation, and synonymous variations within the mRNA coding sequence that alter local translation rate should not affect the folded structure of the protein.

However, the model proteins commonly used to study folding do not wholly capture the complexities of protein structural and folding properties across the proteome (12), nor do in

vitro folding experiments accurately mimic in vivo folding conditions. For example, the native structure of a protein need not represent a unique thermodynamic energy minimum: For some proteins, the native structure represents a kinetically trapped state (95, 140), which can lead to the population of more than one native structure (5, 15, 83, 112, 130). Compared with current folding models, larger proteins with more complex structural properties might be more dependent on cotranslational folding to avoid protein misfolding and aggregation. Indeed, multidomain proteins can undergo misfolding owing to interactions between adjacent structural domains, and the amino acid sequences of these proteins have evolved to avoid such interactions (11). Vectorial folding during translation could enable an N-terminal domain to fold before the appearance of a potentially interfering C-terminal domain, and modulating the translation rate via synonymous codon substitutions could provide still more time for the N-terminus to fold. In this fashion, synonymous codon usage would add an additional layer of information to the coding sequence by encoding the translation rate as well as the amino acid sequence (73, 93, 100, 104) (**Figure 2**).

Supporting this hypothesis, several studies have shown that synonymous rare codon substitutions inhibit the function of specific proteins, presumably by disfavoring the formation of the functional native structure. Two broad types of synonymous substitution folding effects have been observed: (*a*) those that alter the structure and/or function of the encoded protein, and (*b*) those that produce even more dramatic effects, altering the partitioning between proper folding and aggregation and thereby reducing the yield of correctly folded proteins.

A classic example of the first type is chloramphenicol acetyltransferase (CAT). The wild-type CAT coding sequence contains two large rare codon clusters. Replacing one of these clusters with more common synonymous codons resulted in faster translation but reduced CAT specific activity by ~20%, suggesting that some portion of CAT translated with more common codons, although soluble, adopts a structure distinct from the functional native structure (74). Sun et al. (119) identified in a library of galactose oxidase mutants with altered substrate specificity similar synonymous mutations that reduce enzymatic activity. A more recent study showed that synonymous codon substitutions can also impact organism physiology. Zhou et al. (150) examined the effects of codon usage in FRQ, a fungal circadian clock regulator protein encoded largely by rare codons. Replacing the rare codons with more common synonymous codons impaired FRQ function and the circadian properties of the fungus. FRQ protein translated from the common codon sequence showed different protease susceptibility than FRQ translated from the wild-type coding sequence, suggesting that the functional effects were caused by a difference in FRQ protein structure.

Alternatively, slowly translated, rare synonymous codons can reduce protein misfolding and aggregation (Type *b* from above). Cortazzo et al. (26) examined the in vivo folding of synonymous variants of a fatty-acid-binding protein. Replacing five rare codons with common synonymous substitutions lowered the yield of soluble proteins and induced a cellular stress response, suggesting misfolding occurred (26). Increasing cellular tRNA levels is expected to raise the translation rate of codons with low levels of cognate tRNA, reducing translational pauses. Increasing tRNA levels in *E. coli* produced an overall decrease in soluble protein and increase in insoluble aggregates, suggesting translational pauses are important for the folding of many proteins (43).

Although these examples provide a compelling case for the impact of synonymous codon usage for proper cotranslational folding of some proteins, the extent to which synonymous codon usage affects cotranslational folding in general remains unknown. There is a pressing need to study codon usage in a systematic way in order to develop a predictive model for its effects on cotranslational protein folding, and to determine to what extent these effects are due to slow decoding versus alterations to mRNA structure or other effects. The development of such a predictive model is

further challenged because the factors that affect translation rate likely differ between organisms. For example, the prokaryotic ribosome has helicase activity (122, 127), and therefore prokaryotes are expected to be less susceptible to stalling induced by mRNA structures than eukaryotes are.

Cotranslational Interactions with Other Cellular Components

During translation, nascent chains may form cotranslational interactions with other cellular components and undergo covalent modifications (68, 76). Perhaps the most well-studied example is cotranslational interactions between N-terminal nascent chain signal sequences and the signal recognition particle, which initially arrests translation elongation to ensure subsequent cotranslational translocation of the nascent polypeptide chain once it arrives at the ER translocon (133, 135, 136). Because the 5′ ends of coding sequences of secreted proteins are more enriched in rare codons than those of cytosolic proteins, it has been hypothesized that they might contribute to increased membrane targeting and secretion efficiency (25). Rare-codon-induced translational pauses are hypothesized to allow extra time for other cotranslational interactions as well (24). For example, a rare-codon-induced reduction of translation rate in actin increases the likelihood of cotranslational arginylation (145), a covalent modification that regulates actin activity and prevents aggregation of actin filaments (70). In addition, many multimeric protein complexes can form cotranslationally (37). Although less well explored, codon usage and other mechanisms that control local translation rates might help regulate other well-characterized steps in nascent chain biogenesis including interactions with molecular chaperones (94) and enzymes that catalyze N-glycosylation (85).

Functional Regulation of Expression Level

In addition to affecting translation rates and cotranslational processes, codon usage can significantly alter protein expression levels. Indeed, because of the known negative effects of rare codons on expression yield, it is common practice when expressing a human gene in *E. coli* or another expression host to replace the arginine codons AGA and AGG, which are common in *Homo sapiens* but very rare in *E. coli*, with more common versions to maximize expression level (see Heterologous Protein Expression, below). In support of this approach, the coding sequences of many highly expressed genes are enriched in common codons. This observation is often interpreted as support for the hypothesis that selection favors common codons, allowing most rare codons to accumulate in lowly expressed sequences, which are thought to be under weaker selection (109).

However, recent results have challenged this longstanding view by demonstrating functional significance of expression level regulation by rare codons in certain sequences. Xu et al. (142) tested the effects of codon usage in cyanobacterium (*Synechococcus elongatus*) circadian clock proteins KaiB and KaiC, whose wild-type coding sequences contain many rare codons. Replacing these rare codons with common synonymous codons increased protein expression level, and this increase in expression disrupted the circadian growth rhythms of the organism (142). This result suggested that synonymous codons may be under selection to regulate protein expression level and are not merely a by-product of mutational drift in lowly expressed genes. Likewise, Subramaniam et al. (118) discovered that bacterial biofilm formation is inhibited by SinR, and the SinR coding sequence contains a number of serine codons susceptible to pausing under serine-starving conditions (81, 118). As a result, the production of SinR decreases when nutrients are limited, initiating biofilm formation. In support of the ideas that highly expressed transcripts are not the only coding sequences subject to synonymous codon selection and that synonymous

codon-related regulation such as this might be widespread, tRNA levels differ with cell cycle stage and growth conditions, as do the expression levels of genes with different codon usage (10, 124). Likewise, cell cycle–regulated genes show atypical codon usage (46).

Viral Codon Usage

Viruses provide another example of how genes adapt to expression conditions. Viral genes rely on host translational machinery for expression and, as a consequence, have evolved diverse mechanisms to exploit host codon usage biases. Viral coding sequences, particularly those with high expression, typically have codon usage biases similar to those of their host organisms (4). However, certain viral genes are enriched in codons that are rare in their hosts' genomes, and this rare codon usage is hypothesized to provide a mechanism to reduce viral protein expression and minimize host immune system responses (89, 128). Some viruses also encode their own tRNA (often a tRNA that is rare in the host) to allow more efficient, selective translation of viral transcripts (34).

Codon Usage in Horizontally Transferred Genes

In contrast to the examples of functionally significant rare codon usage provided above, some genes have unusual codon usage because they were horizontally transferred between different species or strains. Horizontal transfer often involves plasmids, and horizontally transferred genes in bacteria are often related to virulence or antibiotic resistance, but this process can also occur in eukaryotes and transferred genes can be incorporated into the genome (61, 121). Genes that have recently undergone horizontal transfer often have unusual GC content and codon usage compared with their host genome, and researchers have used this unusual codon usage to develop algorithms to identify putative horizontally transferred genes (49). However, not all horizontally transferred genes have unusual codon usage, and not all genes with unusual codon usage have been horizontally transferred (46). When evaluating a gene with unusual codon usage, researchers should consider the possibility of horizontal transfer before assuming that the unusual codon bias resulted from selection.

Heterologous Protein Expression

Analogous to naturally occurring horizontal gene transfer, genes are also frequently transferred between species in the laboratory. Because synonymous codon usage can be functionally significant, and the distribution of rare and common codons within the same coding sequence can be drastically different, differences in codon usage frequencies between organisms can have consequences for heterologous protein expression (**Figure 3**). This finding has led to a number of algorithms that adjust the codon usage of a sequence for expression in a particular organism (47, 58, 60). Because of the traditional focus on rare codons as deleterious, these algorithms typically seek to optimize a coding sequence by increasing the number of amino acids encoded by common codons (38). The emphasis on common codons is due partly to the association between codon commonness and high expression and to the desire to produce large amounts of protein from heterologous expression. However, this approach ignores potential contributions of rare codons to cotranslational folding or other processes, such as interactions with molecular chaperones. Hence, if the goal is to produce the highest possible yield of soluble, natively folded proteins, then designing a coding sequence for uniformly fast translation may not be the most effective approach. An alternative strategy, called codon harmonization, involves replicating the coding sequence's original distribution of rare and common codons in the context of the heterologous expression system (2). Harmonization was used to design mRNA sequences for *E. coli* expression of proteins

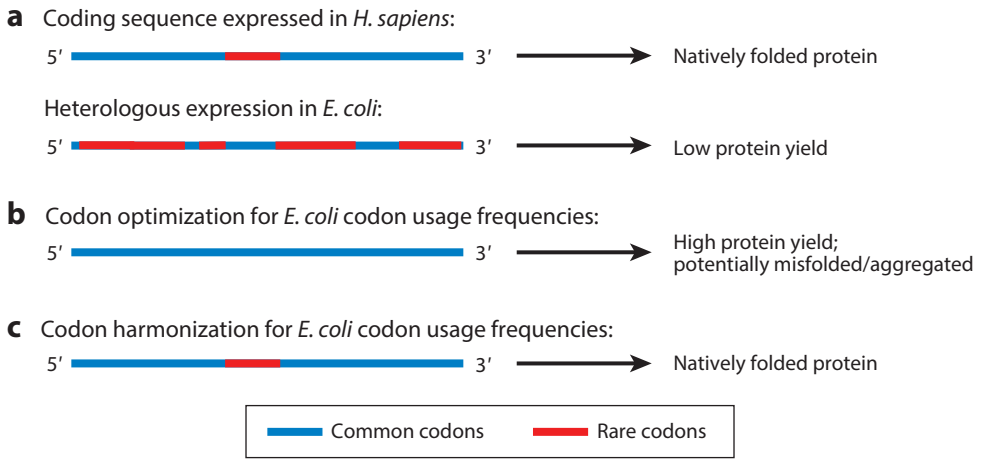

Figure 3

Hypothetical example of heterologous expression of a human coding sequence in *Escherichia coli*. (*a*) Because many codons that are common in *Homo sapiens* are rare in *E. coli*, a human coding sequence often contains more rare codons when expressed in *E. coli*. This can lower protein expression levels. (*b*) To solve this problem, coding sequences are frequently optimized for the codon usage of the heterologous expression host by selecting only synonymous codons that are common in the host. This approach often results in increased expression levels, but much of the resulting protein product may be misfolded. (*c*) Codon harmonization is based on the hypothesis that the original pattern of rare and common codon usage, under native expression conditions, promotes proper folding. Instead of choosing all common synonymous codons, codons that are rare in *H. sapiens* are replaced by codons that are rare in *E. coli*, and codons that are common in *H. sapiens* are replaced by codons that are common in *E. coli*.

from the eukaryotic pathogen *Plasmodium falciparum*. *E. coli* and *P. falciparum* have very different genome %GC bias and codon usage levels, and codon harmonization improved expression level and soluble protein yield relative to conventional codon optimization (2, 3). More generally, translating eukaryotic coding sequences in *E. coli* strains that translate more slowly, at a rate more similar to eukaryotic translation rates, increased folding yield for some eukaryotic proteins (111).

Rare Codons at Coding Sequence Termini

Rare codons are enriched at the 5′ termini of many coding sequences in many species (25, 131, 144); however, the functional significance of this enrichment is still a matter of debate (**Figure 4**). Ribosome profiling has identified higher ribosome occupancies at the 5′ termini of coding sequences, although it is not clear whether this was caused by slower translation or premature termination (65). A 5′ rare codon cluster is not expected to regulate cotranslational folding, because at the time of its translation the nascent chain has not yet exited the ribosome (**Figure 2*a***). According to one hypothesis, because rare codons reduce translation rate, they create a ramp at the 5′ end of the coding sequence, which optimizes ribosome spacing on the mRNA and prevents ribosome traffic jams (88, 131). In contrast, the 5′ enrichment of rare codons is proposed to be a by-product of selection for GC content. High %GC generally leads to more stable mRNA structures, and stable 5′ structures may interfere with translation initiation. Less stable 5′ mRNA structures are correlated with higher protein expression levels (77, 107, 148).

Evidence from recent large-scale screening studies suggests that the enrichment of rare codons at the 5′ termini does result primarily from selection for low %GC codons, at least in bacteria

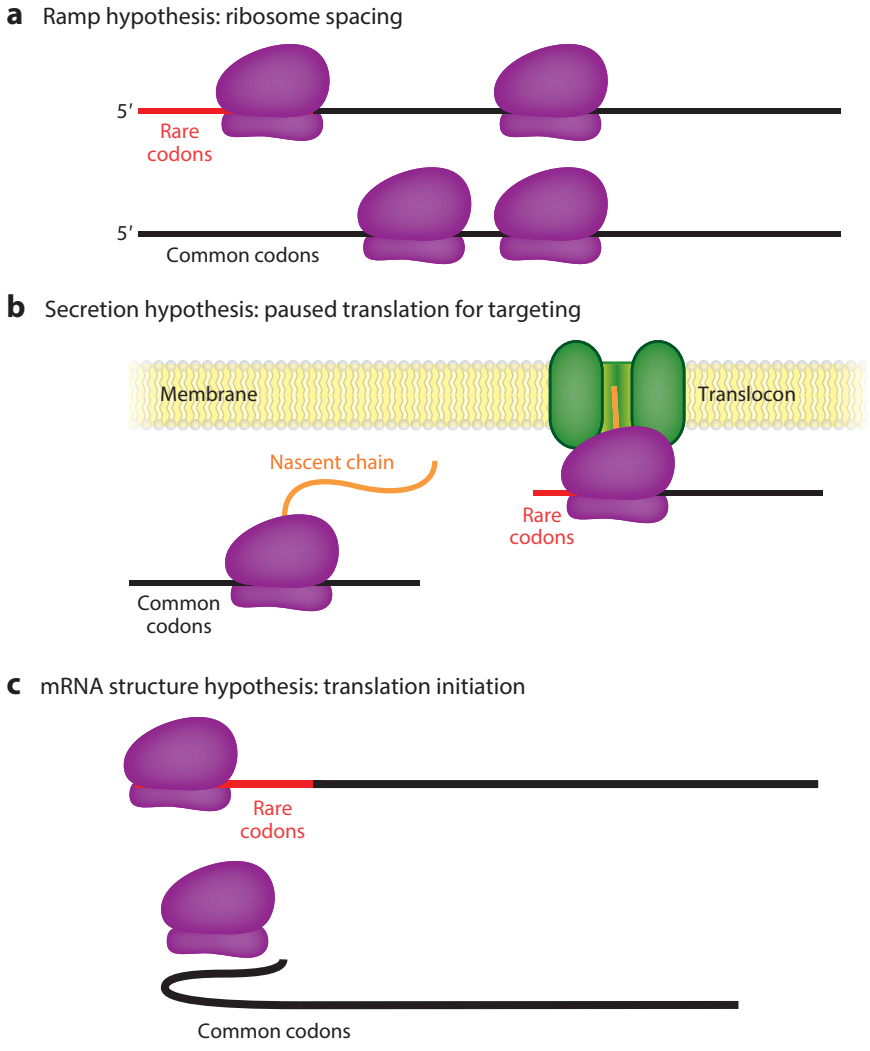

Figure 4

Proposed functions of 5′ rare codons. (*a*) The ramp hypothesis (131) suggests that 5′ rare codons introduce a region of slow translation that keeps ribosomes more widely spaced on the mRNA, as ribosomes will move more quickly after leaving the ramp. (*b*) The secretion hypothesis (25, 144) suggests that 5′ rare codons cause a translational pause to allow time for targeting before the nascent chain is translated. (*c*) The mRNA structure hypothesis (9, 54) suggests that 5′ codon usage is under selection for reduced mRNA structure so that translation may be initiated more efficiently.

(9, 44, 54). These studies indicated that low %GC codons are enriched at the 5′ termini. In organisms with high %GC content, most low %GC codons are rare, so selection for low %GC codons results in a corresponding coincidental enrichment of rare codons.

The rare codons at the 5′ termini in the transcripts of secreted proteins are hypothesized to slow translation and allow additional time for the ribosome to be targeted to the membrane (144). Consistent with this hypothesis, in bacteria secreted proteins are more likely to contain

synonymous rare codons near the 5′ end of their coding sequences (25), suggesting these rare codons might play a role in efficient translocation in addition to their role in enhancing translation initiation. A recent study of several eukaryotic species also showed a 5′ enrichment of rare codons in coding sequences of secreted proteins but not cytosolic proteins (84), supporting the hypothesis that a lowered translation rate could promote secretion.

In addition to the enrichment of rare codons at the 5′ termini, some species show an enrichment of rare codons at the 3′ termini of coding sequences (25). Compared with 5′-end rare codons, 3′-end rare codons have received far less attention in the literature, but they have been hypothesized to tether a synthesized protein to the ribosome to allow additional time for cotranslational folding before translation termination and release of the nascent chain (25).

Codon Usage and Human Health

Emerging connections between synonymous codon usage and human health further underscore the importance of understanding the mechanisms by which synonymous codon substitutions alter protein biogenesis. Although many disease-associated SNPs are nonsynonymous, some synonymous SNPs have been identified as disease associated (105). One example is the CFTR synonymous mutation Ile507 ATC→ATT, which alters mRNA stability and exacerbates the effects of the cystic fibrosis mutation ΔF508 (see Direct Effects of Synonymous Codon Usage, above). In *MDR1*, a gene encoding a transmembrane efflux pump involved in multidrug resistance, three disease-associated SNPs (two synonymous and one nonsynonymous) correlate with patient responses to chemotherapy (48, 71). *MDR1* expressed in a human cell line with these SNPs produced an efflux pump with an altered structure (assessed by protease susceptibility and antibody binding) and altered drug-pumping activity compared with wild-type *MDR1*. In addition to altering cotranslational folding, pathogenic synonymous SNPs can alter splicing motifs, leading to alternative splicing (28, 36, 42, 53, 86). Supek et al. (120) discovered that synonymous mutations are often driver mutations in human cancers, and speculated that some of these synonymous SNPs could affect protein folding, as they are enriched near the beginning of α-helical regions in oncogenes. This study also found that cancer-driving synonymous substitutions are both rare to common and common to rare, underscoring that increasing both codon rarity and commonness can be deleterious in specific circumstances.

MEASURING AND TRACKING SYNONYMOUS CODON USAGE

Quantifying Codon Usage and Predicting Translation Rates

Although the growing body of experimental results described above has demonstrated that synonymous codon selection can play a functional role in specific coding sequences, such results have not yet led to a general understanding of the roles for synonymous codon usage in protein biogenesis. Alongside functional studies, bioinformatics studies of codon usage patterns have provided a valuable strategy for developing and testing broad, general hypotheses about synonymous codon usage.

To study the functional importance of synonymous codon usage, it is first necessary to develop an algorithm to classify or quantify codon usage. A number of these algorithms have been developed, some of which merely quantify codon usage while others attempt to use codon usage frequencies and other information to predict translation rates (24, 33, 97, 109, 116, 147) (**Figure 5**). Whereas some algorithms quantify codon usage across an entire gene (returning a

Calculating synonymous codon usage

There are multiple algorithms to calculate synonymous codon usage. Some examples include the following:

Leu-Leu-Cys	Protein sequence
CTG-CTA-TGC	Codons

a Absolute rareness

Uses codon usage frequencies from the genome. This method does not control for the influence of the amino acid sequence (in this example, Cys is a rare amino acid).

CTG	CTA	TGC
51.1	3.9	6.4

b Codon optimality: Codon Adaptation Index (CAI)

This method compares the codon usage of a sequence to the codon usage of a set of highly expressed genes (in this case, *E. coli* ribosomal proteins). Highly expressed genes are hypothesized to use more translationally optimal codons.

CTG	CTA	TGC	
379	1	17	Codon frequency
379	379	17	Most common codon frequency

The frequency of each codon in the codon usage set is divided by the frequency of the most common codon in that synonymous set. The CAI for the full sequence is the geometric mean:

$$[(379/379) \times (1/379) \times (17/17)]/3 = 0.00088$$

c Relative codon usage: %MinMax

Relative measures of codon usage control for amino acid frequencies: The usage frequency of a codon is compared to others in its synonymous set. The %MinMax algorithm was designed to highlight clusters of rare codons. Usage frequencies of the codons in a coding sequence are compared with the maximum, minimum, and average possible for a given amino acid sequence.

CTG	CTA	TGC	
51.1	3.9	6.4	Genome-wide usage frequencies
51.1	51.1	6.4	Max. for synonymous set
3.9	3.9	5.2	Min. for synonymous set
17.3	17.3	5.8	Avg. for synonymous set

%MinMax scores are typically calculated for a sequence window:

$$\%\text{MinMax} = 100x(\text{Actual} - \text{Avg.})/(\text{Max} - \text{Avg.}) =$$

$$100x\frac{((51.1 + 3.9 + 6.4) - (17.3 + 17.3 + 5.8))}{((51.1 + 51.1 + 6.4) - (17.3 + 17.3 + 5.8))} = 30.7$$

Figure 5

Metrics commonly used to calculate and compare the codon usage of mRNA sequences. Examples of calculations using (*a*) absolute rareness, (*b*) Codon Adaptation Index (CAI) (109), and (*c*) %MinMax (24). For absolute rareness and %MinMax calculations, using data from the Codon Usage Database (92), the codon usage frequencies are shown in units of frequency per 1,000 codons in genome open reading frames.

single codon usage value per coding sequence), position-specific codon usage algorithms that analyze individual codons or short sliding windows along the coding sequence are more useful for identifying local perturbations of codon usage that might affect cotranslational folding and other aspects of protein biogenesis.

While some algorithms focus on codon usage, others classify codons according to levels of cognate tRNA. Unlike codon usage frequencies, which are available for all fully sequenced genomes, absolute tRNA concentrations are available for only a few species. Although tRNA levels in general correlate with tRNA gene copy number, and in many cases gene copy numbers have been used as

a proxy for tRNA concentration, cellular levels of tRNA and charged tRNA vary with cell cycle and growth conditions (32, 81) and between different cell types in multicellular organisms (30), limiting these approximations. Some algorithms incorporate additional factors predicted to affect the connection between codon usage and translation rate, for example, by including an additional term representing the strength of wobble pairing between a given codon and anticodon (116). There is some experimental evidence that codon usage metrics in addition to codon rarity do correlate with translation rates, although the difficulty of measuring the absolute translation rate of individual codons or local sequence regions has hindered comparisons between these metrics and actual translation rates. For example, Spencer et al. (116) used pulse-chase analysis to compare the translation rates of synonymous luciferase variants with predicted fast or slow translation based on a wobble velocity metric. Experimental results were globally consistent with predictions, but the experimental methods were not sufficiently sensitive to test predictions of translation rate in local sequence regions. Zhang et al. (146) identified partially translated products, showing that clusters of slow-translating codons (as predicted by their tRNA level algorithm) caused translational pauses. Similar results were observed for the %MinMax algorithm, a measure of relative codon rareness (24) (**Figure 6**). Translation-rate-dependent cotranslational folding of the YKB biosensor (**Figure 2**; described in Direct Effects of Synonymous Codon Usage, above) was more

%MinMax: an algorithm based on relative codon usage used to quantify the codon usage within a gene

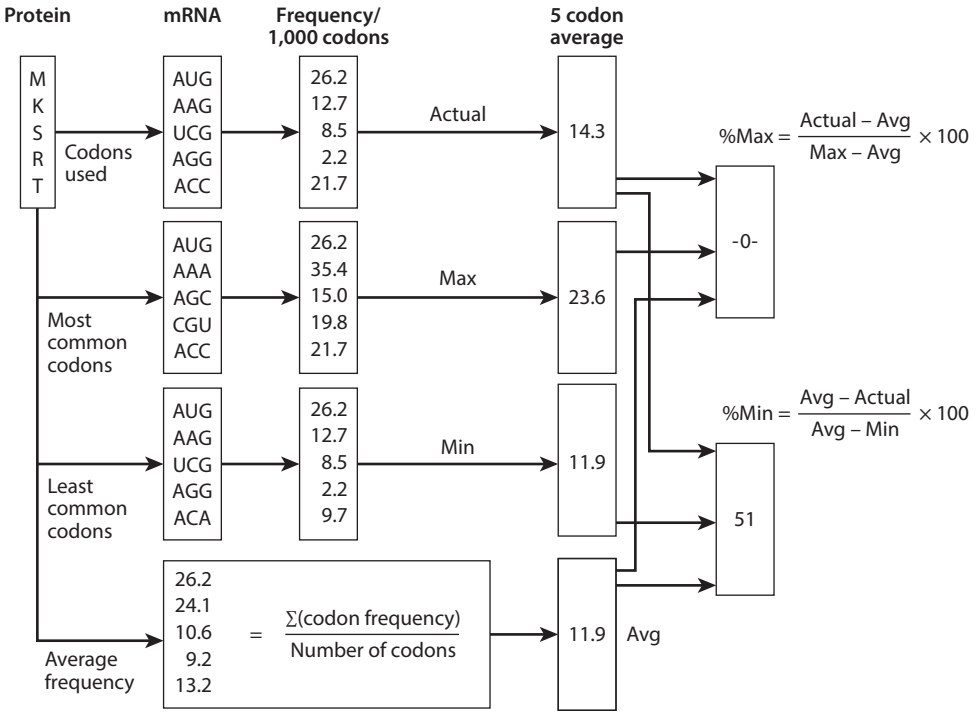

Figure 6

Flowchart for %MinMax analysis. The minimum, maximum, and average usage frequencies are calculated for each set of synonymous codons. The usage frequency of the actual codon used is compared with these values. Common codons result in positive %MinMax scores and rare codons result in negative %MinMax scores. In this simple example results are averaged over a window of 5 codons; in practice, a sliding window of 17 codons is typically used. Figure reproduced from Reference 24 with permission.

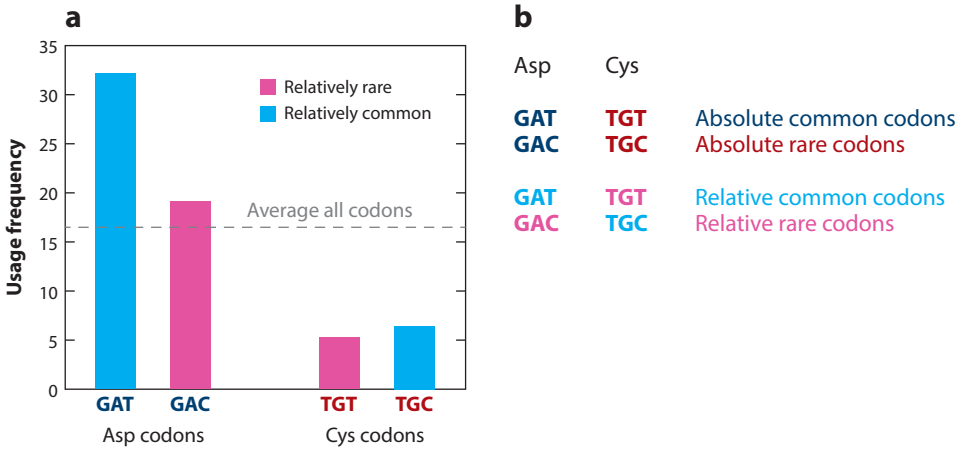

Figure 7

Absolute versus relative codon rareness. (*a*) Absolute codon usage frequencies of aspartate and cysteine codons in the *Escherichia coli* genome (frequency is defined as codons per 1,000 genome codons). The gray dashed line indicates the average usage frequency of all 61 sense codons. Both aspartate codons have high absolute usage frequencies, and both cysteine codons have low absolute usage frequencies, because aspartate is a common amino acid and cysteine is a rare one. (*b*) If codons are classified according to absolute rareness, all cysteine codons are considered rare and all aspartate codons are considered common. However, measures of relative rareness control for amino acid frequencies and classify a codon as rare or common on the basis of how it compares to other codons encoding the same amino acid.

accurately predicted by an algorithm that considers only codon usage frequency than by several other algorithms, including those based on tRNA concentration and wobble velocity (104).

Even among algorithms that evaluate only codon usage frequency, comparisons can be based on either absolute or relative codon usage frequencies. Because amino acid usage frequencies differ widely, these results are not equivalent (**Figure 7**). A codon is rare in absolute terms if its usage frequency is lower than the average of all codons that encode amino acids, but it is rare in relative terms if its usage frequency is lower than the usage frequency of other codons encoding the same amino acid. Relative rareness is often more useful for studies of codon usage because it allows for correction based on amino acid conservation. For example, cysteine is a rare amino acid, so even the most common cysteine codon may be rare in absolute terms, but its presence in a codon sequence indicates selection for cysteine, not for a rare codon. Some algorithms, including the widely used Codon Adaptation Index (CAI) (**Figure 5**), are based on codon usage frequencies in highly expressed genes rather than whole genomes, assuming that highly expressed genes are under stronger selection and contain more optimal codons (109).

Codon Adaptation Index (CAI): an algorithm used to quantify the codon usage of an entire gene; based on comparisons to codon usage in highly expressed genes

As data for translation rates of specific codons become available, methods of quantifying codon usage can be refined. Until this becomes possible, bioinformatics analyses are limited to identifying coding sequence regions with statistically unusual codon usage. A crucial aspect for every bioinformatics study to consider is the development of an appropriate null model to control for the effects of amino acid sequence and other irrelevant variables on codon selection. A common choice is random reverse translations, in which control mRNA sequences are constructed by randomly selecting synonymous codons for a given protein sequence (24). Synonymous codons may be chosen according to probabilities determined from the species codon usage or by randomly shuffling the synonymous codons in a single protein.

Identifying Patterns in Synonymous Codon Usage

Many bioinformatics studies have sought to identify correlations between the locations of rare codon clusters and structural units in the encoded protein, on the basis of the hypothesis that rare codon clusters promote cotranslational folding of these units. However, the scientific community has not reached a clear consensus on the location and functional significance of rare codon clusters, primarily because different studies have yielded dramatically different results. This inconsistency is not surprising, given the differences in the algorithms and data sets used. In particular, studies that focus on proteins of known structure tend to rely on small data sets, which can introduce bias.

One attractive hypothesis is that rare codons might separate the cotranslational folding of individual domains in a multidomain protein. Several studies have tested whether rare codons are enriched at domain boundaries or, given the shielding by the ribosome exit tunnel of the nascent chain C-terminus, 30 to 40 codons downstream relative to domain boundaries (**Figure 2**). However, whereas Thanaraj & Argos (126) determined that rare codons were enriched at domain boundaries in a set of 37 multidomain *E. coli* proteins with solved structures, Saunders & Deane (106) observed the opposite trend and concluded that rare codons were underenriched at domain boundaries in human, yeast, and *E. coli* sequences. The sample size used by Saunders & Deane was larger than that used by Thanaraj & Argos, but still represented only a small fraction of the entire proteome: Their largest sample, from *E. coli*, consisted of only 121 proteins.

Different secondary structure types, which have different topologies and fold at different rates, are hypothesized to exhibit distinct codon usage patterns. An analysis of 54 *E. coli* proteins of known structure indicated α-helices are associated with common codon usage, whereas β-sheets and coil regions are associated with rare codon usage, which was hypothesized to be related to the slower folding rate of β-sheets (125). Similarly, Saunders & Deane (106) found that certain codons have different usage frequencies at regions of transition between secondary structure types. But in contrast to the study above, they found "fast" codons at the transition into irregular, random coil structures and "slow" codons at transitions into α-helices or β-sheets. Xie & Ding (141) examined associations between codon usage and secondary structure in 54 *E. coli* proteins and 107 mammalian (mostly human) proteins of known structure. They determined that two codons in *E. coli* and 17 codons in mammals had statistically significant different usage in different secondary structure types, but the authors did not draw any conclusions about preferences for rare and common codons in different secondary structure types. In a broader analysis, Brunak & Engelbrecht (13) analyzed 719 proteins of known structure and found no correlations between synonymous codon usage and secondary structure regions, leaving unclear the connections between codon selection and secondary structure of the encoded protein.

Others have hypothesized that the distribution of rare codons is influenced more by selection for translational accuracy than by selection for translation rate. According to this view, common codons are expected to be favored in regions where amino acid sequence is critical, whereas rare codons are clustered in regions (for example, unstructured linkers) where mistranslation is less harmful (151). However, the synonymous mutations studied by Cortazzo et al. (26) were located in a turn between two α-helices, so rare codons located between structural units (rather than in them) may still be functional.

Recent studies have sought to determine whether rare codons are conserved in homologous coding sequences during evolution, as this would imply a functional role for rare codons. Pechmann & Frydman (97) examined several yeast species and concluded that rare codons tend to align in homologs. Chartier et al. (17) reported that gene families within the Pfam database contain many conserved rare codon clusters. However, the Chartier et al. study relied on measures of absolute codon rareness and failed to control for amino acid content, so their analysis

may have detected conservation of rare amino acids rather than conservation of rare synonymous codons.

CONCLUSIONS

A growing body of experimental evidence has demonstrated that synonymous codon usage is under evolutionary selection and has functional significance for the biogenesis of some proteins. However, although experimental evidence for the functional significance of rare codons is accumulating, we still lack a general, predictive understanding of the mechanisms by which synonymous codon selection impacts protein biogenesis. An understanding of these mechanisms will enhance synthetic gene design and heterologous protein expression and further our understanding of the mechanisms underlying human diseases. Although it is not yet possible to accurately predict which synonymous mutations have functional impact, bioinformatics analyses based on wider, more representative data sets and more extensive statistical analyses, coupled with additional broadscale experimental testing, will be invaluable for clarifying the currently exciting but murky picture of the effects of codon usage on protein biogenesis.

SUMMARY POINTS

1. Synonymous codon usage can be functionally significant and impacts multiple steps of protein biogenesis.
2. Evidence suggests that synonymous codon usage is under evolutionary selection.
3. Synonymous codon changes can affect human health.
4. Although extensive evidence shows that synonymous codon usage affects biogenesis-specific proteins, we still lack a general, predictive understanding of these effects.

FUTURE ISSUES

1. Will broader bioinformatics studies reveal correlations between codon usage and protein structure?
2. Can we accurately predict which synonymous codon changes will most likely negatively impact protein biogenesis and lead to disease?
3. Will an improved understanding of synonymous codon usage lead to rules for codon selection that can be applied to improve folding yields for rational protein design?

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants R01 GM074807 and U54 GM105816.

LITERATURE CITED

1. Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science* 181:223–30
2. Angov E, Hillier CJ, Kincaid RL, Lyon JA. 2008. Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host. *PLOS ONE* 3:e2189
3. Angov E, Legler PM, Mease RM. 2011. Adjustment of codon usage frequencies by codon harmonization improves protein expression and folding. *Methods Mol. Biol.* 705:1–13
4. Bahir I, Fromer M, Prat Y, Linial M. 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. *Mol. Syst. Biol.* 5:311
5. Baker D, Agard DA. 1994. Influenza hemagglutinin: kinetic control of protein function. *Structure* 2:907–10
6. Ban N, Nissen P, Hansen J, Moore PB, Steitz TA. 2000. The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. *Science* 289:905–20
7. Bartoszewski RA, Jablonsky M, Bartoszewska S, Stevenson L, Dai Q, et al. 2010. A synonymous single nucleotide polymorphism in ΔF508 *CFTR* alters the secondary structure of the mRNA and the expression of the mutant protein. *J. Biol. Chem.* 285:28741–48
8. Bennetzen JL, Hall BD. 1982. Codon selection in yeast. *J. Biol. Chem.* 257:3026–31
9. Bentele K, Saffert P, Rauscher R, Ignatova Z, Blüthgen N. 2013. Efficient translation initiation dictates codon usage at gene start. *Mol. Syst. Biol.* 9:675
10. Berg OG, Kurland CG. 1997. Growth rate-optimised tRNA abundance and codon usage. *J. Mol. Biol.* 270:544–50
11. Borgia MB, Borgia A, Best RB, Steward A, Nettels D, et al. 2011. Single-molecule fluorescence reveals sequence-specific misfolding in multidomain proteins. *Nature* 474:662–65
12. Braselmann E, Chaney JL, Clark PL. 2013. Folding the proteome. *Trends Biochem. Sci.* 38:337–44
13. Brunak S, Engelbrecht J. 1996. Protein structure and the sequential structure of mRNA: alpha-helix and beta-sheet signals at the nucleotide level. *Proteins* 25:237–52
14. Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907
15. Burmann BM, Knauer SH, Sevostyanova A, Schweimer K, Mooney RA, et al. 2012. An α-helix to β-barrel domain switch transforms the transcription factor RfaH into a translation factor. *Cell* 150:291–303
16. Charneski CA, Hurst LD. 2013. Positively charged residues are the major determinants of ribosomal velocity. *PLOS Biol.* 11:e1001508
17. Chartier M, Gaudreault F, Najmanovich R. 2012. Large-scale analysis of conserved rare codon clusters suggests an involvement in co-translational molecular recognition events. *Bioinformatics* 28:1438–45
18. Chen C, Stevens B, Kaur J, Cabral D, Liu H, et al. 2011. Single-molecule fluorescence measurements of ribosomal translocation dynamics. *Mol. Cell* 42:367–77
19. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. *PNAS* 101:3480–85
20. Chevance FF, Le Guyon S, Hughes KT. 2014. The effects of codon context on in vivo translation speed. *PLOS Genet.* 10:e1004392
21. Chursov A, Frishman D, Shneider A. 2013. Conservation of mRNA secondary structures may filter out mutations in *Escherichia coli* evolution. *Nucleic Acids Res.* 41:7854–60
22. Clark PL. 2004. Protein folding in the cell: reshaping the folding funnel. *Trends Biochem. Sci.* 29:527–34
23. Clark PL, Ugrinov KG. 2009. Measuring cotranslational folding of nascent polypeptide chains on ribosomes. *Methods Enzymol.* 466:567–90
24. Clarke TF, Clark PL. 2008. Rare codons cluster. *PLOS ONE* 3:e3412
25. Clarke TF, Clark PL. 2010. Increased incidence of rare codon clusters at 5′ and 3′ gene termini: implications for function. *BMC Genomics* 11:118
26. Cortazzo P, Cervenansky C, Marin M, Reiss C, Ehrlich R, Deana A. 2002. Silent mutations affect in vivo protein folding in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 293:537–41
27. Curran JF. 1995. Decoding with the A-I wobble pair is inefficient. *Nucleic Acids Res.* 23:683–88
28. Daidone V, Gallinaro L, Grazia Cattini M, Pontara E, Bertomoro A, et al. 2011. An apparently silent nucleotide substitution (c.7056C>T) in the von Willebrand factor gene is responsible for type 1 von Willebrand disease. *Haematologica* 96:881–87

29. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, et al. 2012. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 22:1589–98
30. Dittmar KA, Goodenbour JM, Pan T. 2006. Tissue-specific differences in human transfer RNA expression. *PLOS Genet.* 2:e221
31. Doerfel LK, Wohlgemuth I, Kothe C, Peske F, Urlaub H, Rodnina MV. 2013. EF-P is essential for rapid synthesis of proteins containing consecutive proline residues. *Science* 339:85–88
32. Dong H, Nilsson L, Kurland CG. 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.* 260:649–63
33. dos Reis M, Savva R, Wernisch L. 2004. Solving the riddle of codon usage preferences: a test for translational selection. *Nucleic Acids Res.* 32:5036–44
34. Dreher TW. 2010. Viral tRNAs and tRNA-like structures. *Wiley Interdiscip. Rev. RNA* 1:402–14
35. Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. *Cell* 134:341–52
36. Du YZ, Dickerson C, Aylsworth AS, Schwartz CE. 1998. A silent mutation, C924T (G308G), in the L1CAM gene results in X linked hydrocephalus (HSAS). *J. Med. Genet.* 35:456–62
37. Duncan CD, Mata J. 2011. Widespread cotranslational formation of protein complexes. *PLOS Genet.* 7:e1002398
38. Elena C, Ravasi P, Castelli ME, Peirú S, Menzella HG. 2014. Expression of codon optimized genes in microbial systems: current industrial applications and perspectives. *Front. Microbiol.* 5:21
39. Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* 300:1718–22
40. Ermolaeva MD. 2001. Synonymous codon usage in bacteria. *Curr. Issues Mol. Biol.* 3:91–97
41. Evans MS, Ugrinov KG, Frese M-A, Clark PL. 2005. Homogeneous stalled ribosome nascent chain complexes produced in vivo or in vitro. *Nat. Methods* 2:757–62
42. Faa′ V, Coiana A, Incani F, Costantino L, Cao A, Rosatelli MC. 2010. A synonymous mutation in the *CFTR* gene causes aberrant splicing in an Italian patient affected by a mild form of cystic fibrosis. *J. Mol. Diagn.* 12:380–83
43. Fedyunin I, Lehnhardt L, Böhmer N, Kaufmann P, Zhang G, Ignatova Z. 2012. tRNA concentration fine tunes protein solubility. *FEBS Lett.* 586:3336–40
44. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. 2014. A comprehensive, high-resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* 31:1581–92
45. Frenkel-Morgenstern M, Danon T, Christian T, Igarashi T, Cohen L, et al. 2012. Genes adopt non-optimal codon usage to generate cell cycle-dependent oscillations in protein levels. *Mol. Syst. Biol.* 8:572
46. Friedman R, Ely B. 2012. Codon usage methods for horizontal gene transfer detection generate an abundance of false positive and false negative results. *Curr. Microbiol.* 65:639–42
47. Fuglsang A. 2003. Codon optimizer: a freeware tool for codon optimization. *Protein Expr. Purif.* 31:247–49
48. Fung KL, Pan J, Ohnuma S, Lund PE, Pixley JN, et al. 2014. *MDR1* synonymous polymorphisms alter transporter specificity and protein stability in a stable epithelial monolayer. *Cancer Res.* 74:598–608
49. Garcia-Vallve S, Guzman E, Montero MA, Romeu A. 2003. HGT-DB: a database of putative horizontally transferred genes in prokaryotic complete genomes. *Nucleic Acids Res.* 31:187–89
50. Gardin J, Yeasmin R, Yurovsky A, Cai Y, Skiena S, Futcher B. 2014. Measurement of average decoding rates of the 61 sense codons in vivo. *Elife* 3:e03735
51. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. 2003. Global analysis of protein expression in yeast. *Nature* 425:737–41
52. Goetz RM, Fuglsang A. 2005. Correlation of codon bias measures with mRNA levels: analysis of transcriptome data from *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 327:4–7
53. Gonzalez-Paredes FJ, Ramos-Trujillo E, Claverie-Martin F. 2014. Defective pre-mRNA splicing in *PKD1* due to presumed missense and synonymous mutations causing autosomal dominant polycystic disease. *Gene* 546:243–49
54. Goodman DB, Church GM, Kosuri S. 2013. Causes and effects of N-terminal codon bias in bacterial genes. *Science* 342:475–79

55. Gouy M, Gautier C. 1982. Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* 10:7055–74
56. Grantham R, Gautier C, Gouy M. 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Res.* 8:1893–912
57. Grantham R, Gautier C, Gouy M, Mercier R, Pave A. 1980. Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* 8:r49–r62
58. Grote A, Hiller K, Scheer M, Münch R, Nörtemann B, et al. 2005. JCat: a novel tool to adapt codon usage of a target gene to its potential expression host. *Nucleic Acids Res.* 33:W526–31
59. Guo FB, Ye YN, Zhao HL, Lin D, Wei W. 2012. Universal pattern and diverse strengths of successive synonymous codon bias in three domains of life, particularly among prokaryotic genomes. *DNA Res.* 19:477–85
60. Gustafsson C, Govindarajan S, Minshull J. 2004. Codon bias and heterologous protein expression. *Trends Biotechnol.* 22:346–53
61. Gyles C, Boerlin P. 2014. Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease. *Vet. Pathol.* 51:328–40
62. Hurst LD. 2002. The *Ka*/*Ks* ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18:486
63. Ikemura T. 1985. Codon usage and transfer-RNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.* 2:13–34
64. Ikemura T, Ozeki H. 1983. Codon usage and transfer RNA contents: organism-specific codon-choice patterns in reference to the isoacceptor contents. *Cold Spring Harb. Symp. Quant. Biol.* 47(Pt. 2):1087–97
65. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS. 2009. Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *Science* 324:218–23
66. Ingolia NT, Lareau LF, Weissman JS. 2011. Ribosome profiling of mouse embryonic stem cells reveals the complexity and dynamics of mammalian proteomes. *Cell* 147:789–802
67. Ito K, Chiba S, Pogliano K. 2010. Divergent stalling sequences sense and control cellular physiology. *Biochem. Biophys. Res. Commun.* 393:1–5
68. Jha S, Komar AA. 2011. Birth, life and death of nascent polypeptide chains. *Biotechnol. J.* 6:623–40
69. Kane JF. 1995. Effects of rare codon clusters on high-level expression of heterologous proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* 6:494–500
70. Karakozova M, Kozak M, Wong CC, Bailey AO, Yates JR 3rd, et al. 2006. Arginylation of β-actin regulates actin cytoskeleton and cell motility. *Science* 313:192–96
71. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, et al. 2007. A "silent" polymorphism in the *MDR1* gene changes substrate specificity. *Science* 315:525–28
72. Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16:111–20
73. Komar AA. 2009. A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* 34:16–24
74. Komar AA, Lesnik T, Reiss C. 1999. Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation. *FEBS Lett.* 462:387–91
75. Kramer EB, Farabaugh PJ. 2007. The frequency of translational misreading errors in *E. coli* is largely determined by tRNA competition. *RNA* 13:87–96
76. Kramer G, Boehringer D, Ban N, Bukau B. 2009. The ribosome as a platform for co-translational processing, folding and targeting of newly synthesized proteins. *Nat. Struct. Mol. Biol.* 16:589–97
77. Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–58
78. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 499:214–18
79. Lazrak A, Fu LW, Bali V, Bartoszewski R, Rab A, et al. 2013. The silent codon change I507-ATC → ATT contributes to the severity of the ΔF508 CFTR channel dysfunction. *FASEB J.* 27:4630–45
80. Letzring DP, Dean KM, Grayhack EJ. 2010. Control of translation efficiency in yeast by codon-anticodon interactions. *RNA* 16:2516–28
81. Li GW, Oh E, Weissman JS. 2012. The anti-Shine-Dalgarno sequence drives translational pausing and codon choice in bacteria. *Nature* 484:538–41

82. Lu JL, Deutsch C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J. Mol. Biol.* 384:73–86
83. Luo XL, Tang ZY, Xia GH, Wassmann K, Matsumoto T, et al. 2004. The Mad2 spindle checkpoint protein has two distinct natively folded states. *Nat. Struct. Mol. Biol.* 11:338–45
84. Mahlab S, Linial M. 2014. Speed controls in translating secretory proteins in eukaryotes–an evolutionary perspective. *PLOS Comput. Biol.* 10:e1003294
85. Malaby HLH, Kobertz WR. 2013. Molecular determinants of co- and post-translational N-glycosylation of type I transmembrane peptides. *Biochem. J.* 453:427–34
86. Meijer J, Nakajima Y, Zhang C, Meinsma R, Ito T, Van Kuilenburg AB. 2013. Identification of a novel synonymous mutation in the human β-ureidopropionase gene *UPB1* affecting pre-mRNA splicing. *Nucleosides Nucleotides Nucleic Acids* 32:639–45
87. Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* 21:984–90
88. Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translation efficiency: optimization by codon usage and mRNA destabilization. *J. Mol. Biol.* 382:236–45
89. Mueller S, Papamichail D, Coleman JR, Skiena S, Wimmer E. 2006. Reduction of the rate of poliovirus protein synthesis through large-scale codon deoptimization causes attenuation of viral virulence by lowering specific infectivity. *J. Virol.* 80:9687–96
90. Musto H, Naya H, Zavala A, Romero H, Alvarez-Valín F, Bernardi G. 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. *FEBS Lett.* 573:73–77
91. Muto A, Osawa S. 1987. The guanine and cytosine content of genomic DNA and bacterial evolution. *PNAS* 84:166–69
92. Nakamura Y, Gojobori T, Ikemura T. 2000. Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.* 28:292
93. Novoa EM, Ribas de Pouplana L. 2012. Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet.* 28:574–81
94. Oh E, Becker AH, Sandikci A, Huber D, Chaba R, et al. 2011. Selective ribosome profiling reveals the cotranslational chaperone action of trigger factor in vivo. *Cell* 147:1295–308
95. Park C, Zhou S, Gilmore J, Marqusee S. 2007. Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. *J. Mol. Biol.* 368:1426–37
96. Parmley JL, Huynen MA. 2009. Clustering of codons with rare cognate tRNAs in human genes suggests an extra level of expression regulation. *PLOS Genet.* 5:e1000548
97. Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nat. Struct. Mol. Biol.* 20:237–43
98. Pedersen S. 1984. *Escherichia coli* ribosomes translate in vivo with variable rate. *EMBO J.* 3:2895–98
99. Pfeifer GP. 2006. Mutagenesis at methylated CpG sequences. *Curr. Top. Microbiol. Immunol.* 301:259–81
100. Purvis IJ, Bettany AJ, Santiago TC, Coggins JR, Duncan K, et al. 1987. The efficiency of folding of some proteins is increased by controlled rates of translation in vivo. A hypothesis. *J. Mol. Biol.* 193:413–17
101. Raghavan R, Kelkar YD, Ochman H. 2012. A selective force favoring increased G+C content in bacterial genes. *PNAS* 109:14504–7
102. Rocha EPC, Danchin A. 2002. Base composition bias might result from competition for metabolic resources. *Trends Genet.* 18:291–94
103. Rosenblum G, Chen C, Kaur J, Cui X, Zhang H, et al. 2013. Quantifying elongation rhythm during full-length protein synthesis. *J. Am. Chem. Soc.* 135:11322–29
104. Sander IM, Chaney JL, Clark PL. 2014. Expanding Anfinsen's principle: contributions of synonymous codon selection to rational protein design. *J. Am. Chem. Soc.* 136:858–61
105. Sauna ZE, Kimchi-Sarfaty C. 2011. Understanding the contribution of synonymous mutations to human disease. *Nat. Rev. Genet.* 12:683–91
106. Saunders R, Deane CM. 2010. Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res.* 38:6719–28
107. Schauder B, McCarthy JEG. 1989. The role of bases upstream of the Shine-Dalgarno region and in the coding sequence in the control of gene expression in *Escherichia coli*: translation and stability of messenger RNAs in vivo. *Gene* 78:59–72

108. Sharp PM, Li WH. 1986. Codon usage in regulatory genes in *Escherichia coli* does not reflect selection for 'rare' codons. *Nucleic Acids Res.* 14:7737–49
109. Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.* 15:1281–95
110. Shpaer EG. 1986. Constraints on codon context in *Escherichia coli* genes. Their possible role in modulating the efficiency of translation. *J. Mol. Biol.* 188:555–64
111. Siller E, DeZwaan DC, Anderson JF, Freeman BC, Barral JM. 2010. Slowing bacterial translation speed enhances eukaryotic protein folding efficiency. *J. Mol. Biol.* 396:1310–18
112. Sinclair JF, Ziegler MM, Baldwin TO. 1994. Kinetic partitioning during protein-folding yields multiple native states. *Nat. Struct. Biol.* 1:320–26
113. Smith NGC, Eyre-Walker A. 2001. Why are translationally sub-optimal synonymous codons used in *Escherichia coli*? *J. Mol. Evol.* 53:225–36
114. Sørensen MA, Kurland CG, Pedersen S. 1989. Codon usage determines translation rate in *Escherichia coli*. *J. Mol. Biol.* 207:365–77
115. Sørensen MA, Pedersen S. 1991. Absolute in vivo translation rates of individual codons in *Escherichia coli*. The two glutamic acid codons GAA and GAG are translated with a threefold difference in rate. *J. Mol. Biol.* 222:265–80
116. Spencer PS, Siller E, Anderson JF, Barral JM. 2012. Silent substitutions predictably alter translation elongation rates and protein folding efficiencies. *J. Mol. Biol.* 422:328–35
117. Strauss BS. 1998. Hypermutability and silent mutations in human carcinogenesis. *Semin. Cancer Biol.* 8:431–38
118. Subramaniam AR, DeLoughery A, Bradshaw N, Chen Y, O'Shea E, et al. 2013. A serine sensor for multicellularity in a bacterium. *Elife* 2:e01501
119. Sun L, Petrounia IP, Yagasaki M, Bandara G, Arnold FH. 2001. Expression and stabilization of galactose oxidase in *Escherichia coli* by directed evolution. *Protein Eng.* 14:699–704
120. Supek F, Miñana B, Valcárcel J, Gabaldón T, Lehner B. 2014. Synonymous mutations frequently act as driver mutations in human cancers. *Cell* 156:1324–35
121. Syvanen M. 2012. Evolutionary implications of horizontal gene transfer. *Annu. Rev. Genet.* 46:341–58
122. Takyar S, Hickerson RP, Noller HF. 2005. mRNA helicase activity of the ribosome. *Cell* 120:49–58
123. Talkad V, Schneider E, Kennell D. 1976. Evidence for variable rates of ribosome movement in *Escherichia coli*. *J. Mol. Biol.* 104:299–303
124. Taylor RC, Webb Robertson B-J, Markillie LM, Serres MH, Linggi BE, et al. 2013. Changes in translational efficiency is a dominant regulatory mechanism in the environmental response of bacteria. *Integr. Biol.* 5:1393–406
125. Thanaraj TA, Argos P. 1996. Protein secondary structural types are differentially coded on messenger RNA. *Protein Sci.* 5:1973–83
126. Thanaraj TA, Argos P. 1996. Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* 5:1594–612
127. Thomas JO, Kolb A, Szer W. 1978. Structure of single-stranded nucleic acids in the presence of ribosomal protein S1. *J. Mol. Biol.* 123:163–76
128. Tindle RW. 2002. Immune evasion in human papillomavirus-associated cervical cancer. *Nat. Rev. Cancer* 2:59–65
129. Tu C, Tzeng TH, Bruenn JA. 1992. Ribosomal movement impeded at a pseudoknot required for frameshifting. *PNAS* 89:8636–40
130. Tuinstra RL, Peterson FC, Kutlesa S, Elgin ES, Kron MA, Volkman BF. 2008. Interconversion between two unrelated protein folds in the lymphotactin native state. *PNAS* 105:5057–62
131. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, et al. 2010. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. *Cell* 141:344–54
132. Ude S, Lassak J, Starosta AL, Kraxenberger T, Wilson DN, Jung K. 2013. Translation elongation factor EF-P alleviates ribosome stalling at polyproline stretches. *Science* 339:82–85
133. Walter P, Blobel G. 1981. Translocation of proteins across the endoplasmic reticulum III. Signal recognition protein (SRP) causes signal sequence-dependent and site-specific arrest of chain elongation that is released by microsomal membranes. *J. Cell Biol.* 91:557–61

134. Wang HC, Susko E, Roger AJ. 2006. On the correlation between genomic G+C content and optimal growth temperature in prokaryotes: data quality and confounding factors. *Biochem. Biophys. Res. Commun.* 342:681–84
135. Wickner W, Schekman R. 2005. Protein translocation across biological membranes. *Science* 310:1452–56
136. Wiedmann M, Huth A, Rapoport TA. 1986. A signal sequence is required for the functions of the signal recognition particle. *Biochem. Biophys. Res. Commun.* 134:790–96
137. Wohlgemuth I, Brenne S, Beringer M, Rodnina MV. 2008. Modulation of the rate of peptidyl transfer on the ribosome by the nature of substrates. *J. Biol. Chem.* 283:32229–35
138. Woolhead CA, McCormick PJ, Johnson AE. 2004. Nascent membrane and secretory proteins differ in FRET-detected folding far inside the ribosome and in their exposure to ribosomal proteins. *Cell* 116:725–36
139. Woolstenhulme CJ, Parajuli S, Healey DW, Valverde DP, Petersen EN, et al. 2013. Nascent peptides that block protein synthesis in bacteria. *PNAS* 110:E878–87
140. Xia K, Manning M, Hesham H, Lin Q, Bystroff C, Colón W. 2007. Identifying the subproteome of kinetically stable proteins via diagonal 2D SDS/PAGE. *PNAS* 104:17329–34
141. Xie T, Ding D. 1998. The relationship between synonymous codon usage and protein structure. *FEBS Lett.* 434:93–96
142. Xu Y, Ma PJ, Shah P, Rokas A, Liu Y, Johnson CH. 2013. Non-optimal codon usage is a mechanism to achieve circadian clock conditionality. *Nature* 495:116–20
143. Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, et al. 2001. Crystal structure of the ribosome at 5.5 Å resolution. *Science* 292:883–96
144. Zalucki YM, Beacham IR, Jennings MP. 2009. Biased codon usage in signal peptides: a role in protein export. *Trends Microbiol.* 17:146–50
145. Zhang FL, Saha S, Shabalina SA, Kashina A. 2010. Differential arginylation of actin isoforms is regulated by coding sequence-dependent degradation. *Science* 329:1534–37
146. Zhang G, Hubalewska M, Ignatova Z. 2009. Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.* 16:274–80
147. Zhang G, Ignatova Z. 2009. Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLOS ONE* 4:e5036
148. Zhang W, Xiao W, Wei H, Zhang J, Tian Z. 2006. mRNA secondary structure at start AUG codon is a key limiting factor for human protein expression in *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 349:69–78
149. Zhao Z, Jiang C. 2010. Features of recent codon evolution: a comparative polymorphism-fixation study. *J. Biomed. Biotechnol.* 2010:202918
150. Zhou M, Guo JH, Cha J, Chae M, Chen S, et al. 2013. Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* 495:111–15
151. Zhou T, Weems M, Wilke CO. 2009. Translationally optimal codons associate with structurally sensitive sites in proteins. *Mol. Biol. Evol.* 26:1571–80

RELATED RESOURCES

%MinMax calculator: **http://www.codons.org/**
CAIcal Server: **http://genomes.urv.cat/CAIcal/**
Codon Usage Database: **http://www.kazusa.or.jp/codon/**

Annual Review of Biophysics
Volume 44, 2015

Contents