# Codon Preferences in Free-Living Microorganisms

SIV G. E. ANDERSSON AND C. G. KURLAND*

Department of Molecular Biology, Uppsala University Biomedical Center, Box 590, Uppsala, Sweden 751 24

---

## INTRODUCTION

The standard version of the genetic code includes 61 sense codons and three stop codons. Although almost all organisms have made the same codon assignments for each amino acid, the preferred use of individual codons varies greatly among taxonomic groups. For example, whereas the related bacteria *Escherichia coli* and *Salmonella typhimurium* have very similar codon preferences (51), the taxonomically unrelated *Bacillus subtilis* has a quite different preferential codon usage pattern (71, 89). In addition, a considerable heterogeneity exists within species; in this case, individual genes tend to favor characteristic codon distributions (5, 18, 35, 86, 89). In vertebrates, for example, codon choice depends mainly on the GC bias of the particular DNA segment harboring the gene (3, 6–8). In unicellular organisms there is a strong connection between protein expressivity and the degree of codon bias, which, in the extreme case, leads to the so-called major codon bias (5, 35, 37, 42, 86).

One of our principal aims here will be to account for the major codon bias of unicellular organisms such as *E. coli* and *Saccharomyces cerevisiae*. To do this, we discuss at some length the connections between codon usage and a variety of functional parameters such as the rates of translation, protein expression levels, and cellular growth rates.

Our starting point is the notion that the degeneracy of the genetic code may be used in ways that solve a number of different problems for the translation apparatus of a cell. In effect, the programming of amino acid sequences of proteins may be only the beginning of the list of functions for individual codons. For example, the degeneracy of the genetic code may be exploited to regulate gene expression or to modulate the performance of the translation system.

Furthermore, we do not assume that codon assignments for amino acids are permanent. Rather, we emphasize that codon assignments as well as codon preferences can replace one another systematically in response to the interplay of mutational pressure and the selective consequences of alternative assignments (18, 20, 53, 73, 74, 83–85; S. G. E. Andersson and C. G. Kurland, unpublished data). Here we wish to indicate how some of the selective pressures for the assignments of synonymous codons may be developed. To stress the dynamics of codon assignments and to illustrate the idea that the genetic code has a structure that can be related to its history, we will begin with a brief summary of current views of the origins of the genetic code.

---

## EVOLUTION OF THE CODE

No obvious clues to the composition of the primeval code are apparent in modern genomes. Nor do we know for sure which amino acids were first encoded by the original genetic systems. However, we may assume for heuristic purposes that the very earliest peptides were composed of the amino acids most abundant at that time. These in turn can be guessed from data about amino acid accumulation in atmospheres assumed to be prebiotic. One of the surprising observations here is that the products synthesized under such putative prebiotic conditions contain in high yield a relatively small number of amino acids (67). Likewise, studies of meteorites and their content of organic molecules reveal a small, related group of relevant compounds (67) (Table 1).

It is worth emphasizing that the most abundant amino acids in this group differ from the other amino acids abundant in modern proteins in that they are not synthesized from other amino acids (99). Thus, glycine, valine, aspartic acid, glutamic acid, and alanine are apparently both chemically and metabolically the simplest amino acids to accumulate. Accordingly, they may very well have been the most abundant amino acids in the primitive biosphere (67).

A basic requirement for the function of a primitive translation system lacking ribosomes is that the bond between the equivalents of the adapter and the messenger molecules should be strong enough not to come apart until the polypeptide chain is transferred to the amino acid attached to the next adapter. It is therefore generally believed that the primitive RNA species involved in the early translation mechanism favored G+C-rich structures. According to Crick et al. (28), maximal binding stability would be promoted by a codon of the form RRY, assuming that the bases surrounding the anticodon are similar to those of present-day...
tRNAs. Eigen and Winkler-Oswatitsch have favored a less extreme coding sequence consisting of repeating RNY triplets (31).

If the RNY code is combined with a preferred use of G’s and C’s, we might guess that the primitive code preferentially used GGC and GCC followed by GAC and GUC. It is suggestive that these codons now code for the amino acids most frequently synthesized in putative primitive atmospheres (Table 1). Apparently, the naive arguments concerning the selective virtues of G-C pairs in primitive translation systems seem to be relevant. Furthermore, it has been observed that modern codon assignments tend to be enriched for RNN codons and in particular for GNN codons (87, 88, 92). Likewise, the amino acids derived directly from prebiotic synthesis have a higher frequency in today’s proteins than those believed to have been derived from inventive biosynthesis (100). Finally, codon degeneracy is higher for the first group of amino acids than for the remainder (100).

These features of the code have been identified by some as the fossil remains of the premeval code (87, 88). In contrast, others have argued that this pattern of codon and amino acid preferences is a selected pattern (92, 100). Indeed, it must be said that even if the primitive genetic apparatus had preferred GNN codons, it is hard to believe that such a bias would have survived if it were not associated with some functional advantages to modern organisms; one such advantage will be discussed below.

### CODON ASSIGNMENT

An intriguing feature of the genetic code is its virtual universality: with a few exceptions the same codon assignments are used in all organisms. Different proposals have been forwarded to explain this characteristic. At one extreme, Crick has suggested that modern codon assignments have been preserved as a frozen accident in which an original set of coincidental codon assignments have been conserved because the occurrence of amino acid replacements in proteins during a phase of shifting codon assignments would be lethal (25). In this hypothesis there is no obligatory relationship between the structure of the amino acids and their codon assignments. The other extreme is, of course, the conjecture that the modern code is universal because it is a functionally optimal code based on structural characteristics of the amino acids and their codon assignments (98). As discussed below, these are not exclusive accounts of the evolution of the genetic code.

A traditional way to explore the structural origins of the code is to try to find chemical correlations between anti-codons or codons and their cognate amino acids. One such correlation would be to relate the hydropobicity of an amino acid to that of its cognate anticodon on the grounds that this might indicate a physical link between amino acids and their cognate adapters (59, 70). Unfortunately, previous searches for direct nucleic acid-amino acid interactions found little if any specificity in the weak complexes formed with mono- and trinucleotides (70). In contrast, a very recent study has revealed sequence-specific binding of L-arginine to an intron within a self-splicing ribosomal precursor RNA from *Tetrahymena.* Nevertheless, although amino acids, such as arginine, with a bulky, polar side chain may be able to interact stereospecifically with a nucleotide sequence, there is no evidence suggesting that the other amino acids in modern proteins can form site-specific complexes with short nucleotide sequences.

An alternative way to view the orderly development of the genetic code is to relate its evolution to that of amino acid metabolism (99). Thus, it can be argued that amino acids that were rare under prebiotic conditions, but eventually be synthesized from other amino acids, would have entered the biosphere at a relatively advanced stage of its evolution. Such a biosynthetic elaboration could occur with the simultaneous reassignment of codons within the groups of biosynthetically related amino acids (99). The data, such as they are, support this scenario.

At present, roughly 80 different enzymes are required for amino acid biosynthesis. In several of these pathways, one amino acid is required as a precursor in the production of another amino acid. However, seven amino acids do not depend on other amino acids for their biosynthesis. These seven primary amino acids and their products are shown in Table 2. Five of these seven amino acids are those expected to have been present at a high concentration in the prebiotic soup (Table 1). Furthermore, the codons of these groups of precursor and product amino acids are internally related by single-base substitutions. Accordingly, it has been proposed that the addition of a new member to an amino acid group occurred concomitantly with the assignment of new codons (99).

Additional support for this view is found in the observation that some of the families of amino acids seem to be internally related via their respective activation systems. For example, a sequence comparison of tRNA^Pro^ and tRNA^Thr^ reveals expected homologies (39). Likewise, sequence homology is found in comparisons of the Met and the Ile synthetases as well as between their tRNAs (15). Clearly, much more must be done to explore these homologies. Nevertheless, the available data suggest that at least part of the structure of the genetic code is a reflection of its coevolution with the metabolic pathways for the amino acids (99).

### TABLE 1. Relative abundance of amino acids in the Murchison meteorite and in an electric discharge synthesis

<table>
<thead>
<tr>
<th>Amino acid</th>
<th>Murchison meteorite</th>
<th>Electric discharge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly</td>
<td>++++</td>
<td>++++</td>
</tr>
<tr>
<td>Ala</td>
<td>++++</td>
<td>++++</td>
</tr>
<tr>
<td>Val</td>
<td>++++</td>
<td>++</td>
</tr>
<tr>
<td>Asp</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Glu</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Pro</td>
<td>+++</td>
<td>+</td>
</tr>
</tbody>
</table>

---

<table>
<thead>
<tr>
<th>Primary amino acid</th>
<th>Product amino acid(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gly</td>
<td>Glu, Lys, Asn, Thr</td>
</tr>
<tr>
<td>Ala</td>
<td>Lys</td>
</tr>
<tr>
<td>Val</td>
<td>Lys, Asn, Thr</td>
</tr>
<tr>
<td>Asp</td>
<td>Pro, Arg, Gln</td>
</tr>
<tr>
<td>Glu</td>
<td>Ser</td>
</tr>
<tr>
<td>Ser</td>
<td>Trp, Cys</td>
</tr>
<tr>
<td>Phe</td>
<td>Tyr</td>
</tr>
</tbody>
</table>

---

* Data are from reference 99 (with permission).
The accidental discovery of the translation of UGA as tryptophan (69, 73). According to our interpretation, the signal UGA is not ambiguous translation in which alternative interpretations of the reassigned codon will be expressed simultaneously. Instead, the transition will be accompanied by a period of ambiguity of codon assignment, tolerated in the transition from one assignment to the other? First, we would expect codon assignments to occur most often at the least frequently used codons because such a bias would minimize their disruptive effects on protein structure. Further, animal-mitochondrial genomes are rapidly evolving (16), which generates the conditions for rare codons to disappear and reappear in a reassigned form (74–76). And, mutatis mutandis, termination codons, which are normally used at low frequencies compared with sense codons, are highly favored for reassignment, but in this case there are additional virtues to be exploited.

When a termination codon is recruited by an aminoacyl-tRNA, the addition of the new amino acids would be expected to have a minimal effect on the performance of the modified protein unless the C-terminal sequence is critical for function (60). Furthermore, it is often observed that genes are punctuated with multiple termination codons. Here, too, the recruitment of one or another of the termination codons by an aminoacyl-tRNA would be expected to have only minimal effects on the functions of the resulting protein.

Finally, we would not expect that the reassignment of a codon would occur as a single mutational event, but, rather, we suggest that it would take place gradually in a series of mutational transformations of the relevant tRNA species. This means that the evolving translation system will not go instantly from one to another assignment of the codon. Instead, the transition will be accompanied by a period of ambiguous translation in which alternative interpretations of the reassigned codon will be expressed simultaneously. Recent work with bacterial mutants and antibiotics that raise the translation error frequencies has indicated that bacteria grow surprisingly well with dramatically enhanced error frequencies (32, 66). Indeed, there is mounting evidence that limited amino acid replacements most often have a minimal effect on the structure and performance of proteins (29, 58, 63).

In summary, we believe that there are acceptable, if vague, scenarios for the evolution of codon assignments even in modern organisms. It is against this background that we wish to view the evolution of extreme preferences for one or another group of synonymous codons. In particular, we favor the view that the evolution of codon reassignments in
mitochondria has been driven by the selection of a translation system that tends toward a minimal number of different tRNA species. Likewise, we will suggest that the evolution of gene-specific codon preferences is dependent on mechanisms to reduce the abundance of certain tRNA species under favorable growth conditions.

**MAJOR CODON PREFERENCES**

In general, the codon composition of unicellular organisms follows the base composition of the genome (68). A direct relationship was found between the G+C content of several species and the number of CNN anticodons (73). Accordingly, it has been suggested that for many genes the mutation bias of the DNA polymerases has been the main determinant for codon and anticodon composition (68, 73).

However, although the codon usage for most genes seems to reflect the average nucleotide composition of the genome as a whole, there is a subset of genes for which the codon choice is strongly biased toward a group of "major" codons (18, 86, 89). In the very highly biased group, containing, for example, genes for the ribosomal proteins, the elongation factors, and outer membrane proteins, there may be as much as a 100-fold variation in the usage frequency for preferred as opposed to avoided codons. The codon usage of such typically highly biased genes, as well as that of genes with a low codon bias, is shown in Tables 3 and 4. Furthermore, the major codon usage bias can be correlated with the expression level of the protein molecules (5, 35, 37, 42, 86). It is observed that the higher the protein production level, the higher the tendency to use only a subset of codons in the gene. In other words, for the highly expressed genes, codon usage seems not to be determined solely by the mutation rate, but seems to be under strong selection pressure (18, 83, 89).

The selective origin of the major codon preference is also apparent in comparisons of the mutation rates for different genes in a group of related enterobacteria (50, 85). Here, pairwise comparisons of homologous genes in, for example, E. coli and S. typhimurium reveal characteristic synonymous codon substitution rates: these are relatively low for genes with marked major codon preferences and more pronounced for genes with less highly biased codon preferences. In contrast, the nonsynonymous codon substitution rates are not strongly related to codon bias. Such results suggest that the synonymous codon preferences of the major proteins are constrained by selective forces operating most probably at the level of translation (83, 85).

A more direct indication that the constraints on synonymous codon preferences arise at the level of translation is found in the strong correlation observed between the major codons and the concentrations of their cognate tRNA species in the bacteria (34, 47-51). These data are illustrated in Fig. 1. In effect, a bacterium growing under normal laboratory conditions is synthesizing a protein population that is highly biased toward the products of a relatively small number of genes. Furthermore, the mRNA population coding for these proteins is constituted from a very biased subset of codons that are translated by a matching tRNA population dominated by a selected subset of isoacceptor species.

One extreme interpretation of the codon usage-tRNA isoacceptor correlation suggests that the codon composition of the mRNA pool has been "adjusted" to the isoacceptor distribution of the tRNA pool to provide a balanced flow of amino acids (46). The tacit assumption here is that although codon composition can be selected, tRNA concentrations are, for some undisclosed reason, immutable. In fact, tRNA isoacceptor concentrations are regulatable and respond, for example, to changes in growth conditions (V. Emilson and C. G. Kurland, unpublished data; C. X. Fournier and A. D. McKay, personal communication). Thus, it seems that both tRNA isoacceptor levels and codon composition can evolve together (17), and the question remains; why do they do so in such a biased way?

**OPTIMAL CODONS**

It seems natural to identify the major codon preference with a subset of particularly "good" codons. One theory to account for the superiority of these codons focuses on the notion that the stability of the codon-anticodon interaction must be optimized (42, 43). According to this view, a tRNA with the capacity to interact with two or more isocodons will be optimally matched when the G+C content of the codon-anticodon interaction is intermediate between the extremes of weak A·U and strong G·C pairings.

This physical model is contradicted by several observations. First, direct physical measurements of the stabilities of interactions between the complementary anticodons of suitable tRNA pairs reveal that the differences between G+C-rich and A+U-rich pairings are reduced by the loop
TABLE 4. Codon usage in very highly biased and very lowly biased genes in *S. cerevisiae*  

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Codon</th>
<th>Usage in:</th>
<th>Usage in:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Highly biased genes</td>
<td>Lowly biased genes</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>UUU</td>
<td>6 126</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUC</td>
<td>64 64</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>UUA</td>
<td>15 130</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UUG</td>
<td>152 142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUU</td>
<td>2 64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUC</td>
<td>0 46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUA</td>
<td>6 74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CUG</td>
<td>0 59</td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>AUU</td>
<td>61 171</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUC</td>
<td>61 75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AUA</td>
<td>0 116</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>AUG</td>
<td>39 103</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>GGU</td>
<td>103 102</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUC</td>
<td>87 70</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUA</td>
<td>0 75</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GUG</td>
<td>1 66</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>UAU</td>
<td>4 98</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAC</td>
<td>55 84</td>
<td></td>
</tr>
<tr>
<td></td>
<td>UAG</td>
<td>0 5</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>CAU</td>
<td>9 53</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td>39 34</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CAA</td>
<td>63 165</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAA</td>
<td>20 90</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>AAU</td>
<td>3 255</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>88 155</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG</td>
<td>0 42</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>AAA</td>
<td>33 253</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AAG</td>
<td>191 169</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>GAU</td>
<td>32 204</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td>91 107</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>GAA</td>
<td>131 218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAG</td>
<td>2 106</td>
<td></td>
</tr>
</tbody>
</table>

* Data are from reference 18 (with permission).

constraint of the anticodon (41). Second, the predicted difference in translational efficiency between G+C-rich and A+U-rich synthetic mRNA species is not observed in vitro (2). Third, the postulated bias toward codon-anticodon pairs of intermediate stability is absent in bacteria as different as *E. coli* (18), *B. subtilis* (89), *M. capricolum* (69), and *Micrococcus luteus* (72). Finally, different organisms use different subsets of major codons (36–38), which argues strongly against there being any intrinsic characteristic of the codons that can be the basis of their preferred character. In fact, virtually every single codon is used as a major codon by some system and as a minor codon by another.

It seems reasonable to assume that codon composition in the highly expressed genes is under the same mutational pressure at the DNA level as it is in the weakly expressed genes. In fact, the G+C content of the third codon position of ribosomal protein genes varies by as much as a factor of 10 in different bacteria. However, in each individual case that has been studied, this variation is biased in the same direction as the G+C content of the whole genome (68). Furthermore, highly expressed and poorly expressed genes in *E. coli* have very similar G+C contents in the third codon position, even though they have different codon preferences. This strongly suggests that a pronounced mutational pressure has been exerted uniformly on all of the genome. In fact, in some of these species, such as *M. capricolum* and *M. luteus*, it appears that mutational biases completely dominate codon usage.

None of this is meant to imply that there are no translational differences among codons. Rather, we suggest that an optimization of translational efficiency is possible for any arbitrarily chosen codon, for example by adjusting tRNA concentrations. Therefore, there is no reason to suppose that the properties of the codon-anticodon interaction provide the sole grounds for the selection of codon preferences.

**REGULATORY CODONS**

A favorite class of theoretical codon preference strategies postulates a role in the regulation of gene expression for a subset of codons. One particularly persistent rumor concerns the regulatory functions of the most infrequently used codons, the so-called rare codons. This conjecture has at least two forms. One is that rare codons modulate the expression levels of proteins present in low copy numbers (42, 44, 93). The other is that regulatory proteins themselves are encoded preferentially by rare codons to keep their expression levels low (55) and, similarly, that rare codons have been selected in signal sequences to reduce the rate of translations (21). Both forms of this conjecture are insupportable.

Thus, a recent compilation of codon usage data shows that regulatory genes and signal peptides do not have a significantly higher frequency of rare codons that a great number of other genes, expressed at moderate to low levels (84). In fact, even in these latter genes there is a slight tendency to avoid rare codons, although this bias is much less pronounced than in highly expressed genes. Most important, no gene has been sequenced yet that shows a preference for rare codons. Accordingly, although rare codons are avoided by major proteins, they are not preferred by minor ones.

Similarly, an inverse relationship between the occurrence of rare codons and the yield of protein is assumed in these...
models. This expectation is based on the notion that poorly translated codons will lower the expression levels. However, as we shall see below, there is no necessary relationship between expression level and translation rate. In fact, high expression levels are routinely achieved for genes with a relatively high content of rare codons. For example, even though the chloramphenicol acetyltransferase gene has a codon usage very different from that of an optimal E. coli gene, it can be expressed to high levels when cloned under a strong promoter (82). Similarly, bacteriophage lambda can be highly expressed when cloned under a strong promoter, even though its codon usage is similar to that of the weakly expressed host genes (46). The same phenomenon has been observed in S. cerevisiae, in which hepatitis B virus core antigen was expressed to a level of 40% of soluble proteins even though it has a very low index of codon bias (54). In this case, it was shown that the high expression level was very dependent on 5'- and 3'-flanking sequences derived from yeast sequences.

Finally, if rare codons were selected for encoding regulatory proteins, they would be expected to have relatively low mutation rates in the corresponding genes. However, two such regulatory genes, dnaG and araC, accumulate synonymous substitutions at a rate matching that of genes with little codon bias, which contradicts the expectation of the regulatory model (85). On the other hand, rare codons may have an effect on attenuation by being clustered in the leader sequences (97).

We can turn the problem on its head by anticipating a different sort of codon strategy, in which a subset of genes are very highly expressed under certain growth states. We find that these genes are selected with a strong preference for a subset of codons (the major codon bias). Next, we consider the fate of the codons that are avoided in this subset of genes, assuming that there are no other codon preferences at work. First, it seems clear that all codons must be translatable in the system so that random mutations do not lead to lethal events with significant probability. However, this constraint need not be an extreme one. Instead, we would expect that the selection pressure to retain rare codons is slight in weakly expressed genes. Supporting this view is the finding of an inverse relationship between codon usage bias and the rate of synonymous nucleotide substitutions (85). Thus, it seems likely that rare codons in regulatory and weakly expressed genes result from the absence of strong negative selection rather than from the presence of strong positive selection.

**Expression level is not translation rate**

All of this is not to say that certain codons do not influence translation rates or that it is, in principle, not possible to regulate protein expression levels by the judicious use of certain codons. However, even if an mRNA is constituted by a large number of "slow" codons, it will not necessarily be expressed at a low level, even if its translation rate is significantly depressed by its codon complement. The point is that the overall rate of which an mRNA is translated can directly influence the expression level of its polypeptide product if and only if the mRNA species can capture a major fraction of the ribosomes in the cell. To understand this constraint, we shall consider the relation of translation rate and expression levels in a little more detail.

In the steady state that corresponds to exponential cell growth, the expression level for a particular gene will be determined by the rate of synthesis of its mRNA and the number of ribosomes that complete the translation of each of these mRNA species. Under normal conditions, even the most highly expressed genes produce an mRNA pool that is only a minor fraction of the total mRNA population. In other words, if the mRNA pool of a cell consisted of only one mRNA species, an increase in the speed of its translation would lead (all other things being equal) to a higher expression level for the corresponding gene product, because the number of ribosomes that translate the mRNA per unit time is increased. However, when we consider the influence of an increase in the speed of translation of an mRNA that represents a very small fraction of the total mRNA pool, this effect is reduced to a corresponding degree. The reason is that a ribosome which has completed translation of an mRNA that is a minor fraction of the total mRNA pool will, with overwhelmingly high probability, be captured afterwards by a different mRNA species. This means that the average number of ribosomes that translate a particular mRNA per unit time will not be influenced by a variation of its rate of translation. Accordingly, the rate of translation of mRNA species will normally not influence the expression level of the corresponding protein.

On the other hand, it is in principle possible, through a judicious selection of codons at the beginning of the mRNA, to influence the queuing of ribosomes on the mRNA; this could lead to a modulation of the expression level for the corresponding proteins (19, 61). In this case we imagine that some codons are translated quickly and some are translated slowly. Accumulating the slower codons at the beginning of the gene sequence would tend to lower the ribosomal loading frequency, and accumulating the fast codons at the beginning would tend to raise the loading frequency. The finding that rare, presumably slow codons are accumulated preferentially at the beginning of a few genes that are normally expressed at a low level is consistent with such a queuing model for the regulation of gene expression by a biased codon sequence (61). However, a general review of a larger number of genes in E. coli as well as in the yeast S. cerevisiae has revealed that in general codon bias is less extreme at the beginning of genes than further on in the sequences; indeed, there is a remarkable similarity in the low codon bias of the initial sequences in weakly, moderately, and highly expressed genes (19). It therefore seems highly unlikely that these organisms regulate gene expression by codon-modulated ribosomal queuing.

**Codon-Specific Rates**

We have completed some of the preliminaries, and now we shall focus our attention on some relevant molecular details of the translation machinery as well as their relationship to cellular growth rates. The reason for this combination of concerns is that organisms such as enterobacteria and certain yeasts are the ones that are characterized by a well-developed major codon preference. That they also represent organisms for which competitive cellular growth rates provide a significant selective advantage is an important clue to understanding the major codon strategy.

From the very first observations of a direct relationship between codon bias and relative concentration of cognate tRNA species, it was tempting to guess that the rates of translation at particular codons were proportional to the amounts of cognate tRNA present in cells (22, 34). In contrast, microbiologists were taught by the Copenhagen school that biosynthetic efficiency always required ribosomes to work at maximum rates (62). Since maximum rates
are obtained only in the presence of excess substrate, it seemed, from this point of view, that translation rates could not be limited by tRNA availability. However, it might be that intrinsic differences exist for the maximum rates of translation at different codons (42, 43, 50). Accordingly, the questions that naturally arise are as follows. Do translation rates vary from codon to codon? Do ribosomes normally translate at their maximum rates for different codons? How different are these rates? To what extent are codon-dependent rate differences due to limiting tRNA concentrations?

The first clear indication that codons are translated at nonuniform rates came from studies of so-called pause sites for proteins that are destined for export through bacterial membranes (81, 94). Thus, for some of these proteins the translation process is markedly uneven, and incomplete polypeptide intermediates accumulate at sites where the ribosomes are retarded in their transit of the mRNA. It could be shown that such pause sites are correlated with short strings of seldom-used or rare codons (94, 95). It was concluded that these particular codons are translated slowly because their tRNA species are in short supply and that the selection of aminoacyl-tRNA to match these codons is rate limiting for peptide bond formation (48, 94). In this case the alternative interpretation that ribosomes translate in vivo at rates independent of the tRNA concentration but dependent on the identity of the codons was not favored.

A number of other studies with modified gene sequences have shown that the introduction of strings of rare codons can lower the expression level of the corresponding protein products (62). Similarly, codon substitutions in the leader peptide of pyrE affected the frequency of transcription past the attenuator (11–13). The interpretation of these results was that rare codons regulate gene expression level by lowering translation rates. However, in none of these studies has the rate of translation been measured. Since we have seen that there is no automatic connection between expression level and translation rate, the meaning of the results of these experiments is unclear. For example, none of these studies could distinguish the effects of the codon insertions on translation from those on transcription or mRNA half-life. In fact, replacing an increasing number of major codons with synonymous minor ones in the phosphoglycerate kinase gene expressed in S. cerevisiae results in a drastic decrease in the steady-state level of the corresponding mRNA (45).

However, it has been recognized that strings of rare codons can in principle act synergistically to reduce their own translation rates (95). In this case the rare codons must be serviced by comparably rare tRNA species; the concentrations of these tRNA species must be rate limiting for translation, and the steady-state concentration of the rare codon strings being translated must be greater than those of their cognate tRNA species. If these unnatural conditions are met, as, for example, with a high-copy-number plasmid construction, the sequestering of the tRNA species at the codons in the string will amplify the translational delay that individual rare codons manifest. In addition to the sequestering effects, cis-acting of repeated rare codons may produce translational movement errors; indeed, a frameshift frequency of 50% was recently observed following the insertion of two neighboring AGG codons (91). Either or both of these effects may have played a role in the earlier experiments concerning the influence of short strings of rare codons on expression levels.

Genuine elongation rate measurements have been carried out for proteins encoded by mRNA species with different codon usage biases. These data reveal a clear positive correlation between the elongation rate and the prevalence of major codons (78). Another sort of experiment with genes containing inserted codon strings has provided additional evidence that translation rates vary from codon to codon (90). Since these data, as incomplete as they are, provide the most direct measurements to date of codon-specific translation rates, we shall inspect them more closely.

First, the inserts to be tested were versions of a ca. 20-codon sequence excised from a ribosomal gene; the virtue of this particular sequence is that when it is inserted into a gene in one reading frame it functions as a string rich in commonly used codons, whereas in another reading frame it provides a string rich in infrequently used codons. By measuring the delay in the transit time of the bacterial ribosomes caused by the inserts, it could be shown that common codons are translated at an average rate of 12 amino acids per s, whereas the uncommon ones are translated with an average rate of 2 amino acids per s (90).

The sixfold difference in translation rate observed for commonly used and seldom-used codons in these experiments is almost certainly an underestimate of the rate extremes for different codons, because it measures only averages for strings of codons. Thus, the biggest shortcoming of this sort of measurement is that it cannot yet be applied to individual codons. Nevertheless, these data show that translation rates at different codons can vary quite significantly. We next attempt to estimate the extent to which this variation is due to differences in tRNA concentrations or to the codon dependence of maximal translation rates.

In this context it is worth noting more recent experiments by Sørensen and Pedersen (Abstr. 13th Int. Transfer RNA Meet. 1989, abstr. no. mo-am-13). They have suggested that the same tRNA species may translate two isocodons at very different rates. They have measured the translation rate of strings of GAA, a major codon for Glu, compared with that of GAG, an uncommon codon for Glu. Both of these are thought to be translated by tRNA\textsubscript{Glu}, but the GAA string is translated at least three times faster in vivo than is the GAG string. Such observations provide the most convincing evidence that translation rates can be codon specific. Nevertheless, for our purposes it is the tRNA concentration dependence of the translation rate that is most relevant.

**HUNGRY RIBOSOMES**

The suggestion that ribosomes function optimally when they are driven at their maximum rates is based on the notion that the mass investment in the ribosome is much greater than that of any other component of the translation system (62). Indeed, if the mass investment in all of the other components of the system is negligible, the rate of translation per ribosome should be the only factor contributing to the rate optimization of the translation system. However, if the mass investment into the aminoacyl-tRNA ternary complexes with elongation factor Tu (TF-Tu) and GTP is assumed not to be negligible, the rate optimization of the translation system is more complex.

If the masses of ternary complex are included in the optimization, the Maalje maximization of the ribosome rate is found to be one theoretical limit of the optimization (30). It corresponds to the extreme of the highest conceivable growth rates in which the organism is doing little else than synthesizing proteins. At the other extreme of the lowest conceivable growth rates, the organism is doing very little
protein synthesis and is engaged mainly in building up amino acids, nucleotides, and other building blocks from a very simple medium. Here, the optimal mass investment is an equipartitioning of ternary complex and ribosome mass, and the kinetics of translation are limited by the concentrations of ternary complexes and ribosomes (30).

In reality, laboratory culture conditions support bacterial growth in states that are intermediate between these two theoretical extremes. The culture conditions supporting the fastest bacterial growth may begin to approach the extreme of translation rates limited by the maximum turnover rates of ribosomes. However, the expectation is that normal labora-

tion codon in the middle of the mRNA. Alternatively, ribosomes may shift reading frames and, in the new phase, complete the translation of the functional form of RF-2, which occurs at extraordinarily high frequencies in this particular sequence. In effect, there is a reading frame branch point in this mRNA which was exploited by Curran and Yarus (27, 28) to study the codon dependence of the initial selection kinetics in vivo. Thus, they have shown that when a sense codon replaces the nonsense codon in this sequence, its recognition by a tRNA species is competitive with the events leading to the frameshift. Accordingly, they could use this branch point to compare the initial kinetics for tRNA selection at different codons by measuring the relative frequencies of codon translation and frameshifting in a series of suitable mRNA constructions.

The results obtained in this assay for 29 different codons can be summarized as follows. First, the initial rates of codon recognition vary quite significantly, and within this group of codons the extremes vary as much as 25-fold. Of special interest is a reasonably good correlation between the more rapidly recognized codons and the commonly used codons. However, for the minor codons the situation is more complex. The frequencies for some minor codons can be correlated with the initial kinetics of the selection, but for the majority there is no correlation. Furthermore, the codon recognition rates for isocodons recognized by the same tRNA species were compared. These initial kinetics tend to be very similar; in no case did the relative frequencies for two isocodons differ by more than a factor of 2.

Undeniably, the most disconcerting aspect of the experiment is that it relates information about two undefined events which are nested within the complete peptide elongation cycle. One of these events is part of the initial discrimination step for aminoacyl-tRNA acquisition. The other is the event that initiates the unusual high-frequency frameshift event characteristic of this sequence. This is almost certainly not a normal part of the peptide elongation cycle, because it requires the participation of a Shine-Dalgarno sequence; accordingly, it is not clear what the frameshift event is reporting. In summary, we are not getting an unambiguous signal that reports information about the whole peptide elongation cycle from these experiments. Nevertheless, the experiment is state of the art.

When the ribosome is translating at its maximum rate, the EF-Tu-dependent steps leading to peptide bond formation and the translocation steps mediated by EF-G take up roughly equal parts of the peptide cycle (10). However, when the cognate ternary complex is below saturating concentrations, the time required to make the peptide bond will lengthen because the acquisition time for the tRNA has lengthened. As a consequence, the measurements made by Curran and Yarus (28) for rare codons translated by tRNA species that are equally rare more accurately reflect the time required to complete a peptide bond than do the measurements made on the major codons translated by tRNA species that are at the highest concentrations. Accordingly, we can be reasonably confident from the data on the slowest codons that there really is a large variation of codon-dependent translation times. This supports the conclusions drawn by Pedersen and colleagues (78, 90) concerning the codon dependence of elongation. Likewise, for the subset of relatively slow codons whose translation times are correlated with the relative abundance of their cognate tRNAs, it is not unlikely that the translation times are determined primarily by the availability of a matching ternary complex.

In summary, all the available data show that there is a great range of translation times for different codons.
analysis of the ribosome mutants with altered kinetic characteristics provides strong support of the interpretation that translation rates for most codons are limited by the availability of the cognate tRNA species. In addition, there seem to be intrinsic kinetic differences in the rates of recognition and translation for some of these codons. However, the fact that two isocodons have different translation rates does not mean that they are not rate limited as well by the availability of their shared tRNA species. For another group of codons, the translation rate clearly seems to be determined by the degree to which individual codons are starved for their cognate ternary complexes. In particular, the major codons belong to the group that is read by the most abundant tRNA species. According to Pedersen and colleagues (78, 90), these are translated at the higher rates, and according to Curran and Yarus (28), they are recognized at the higher rates. Our next concern will be to relate the speed of translation to a selective parameter that will account for the major codon preference.

GROWTH MAXIMIZATION STRATEGIES

We noted above that if there were only one mRNA species to be translated by a fixed number of ribosomes, the number of proteins produced per unit time in the steady state would be proportional to the average rate of translation per codon. This is so because the availability of ribosomes to start a new polypeptide is influenced by the speed with which they complete the transit of the mRNA. Therefore, in this case, the use of fast rather than slow codons would make a difference in the translational efficiency. Similarly, a heterogeneous collection of mRNA species that used the same subset of codons would be translated faster or slower depending on whether the codon bias is for fast or slow codons. We wish to suggest that this sort of coupling between codon bias and translation rate is the selective virtue of the major codon preference in organisms such as *E. coli*. To see how this would work, we must introduce a connection between translational efficiency and growth rates.

The biosynthetic machinery of a cell reproduces itself as well as the other working components of the cell. If the flow of amino acids into proteins under particular growth conditions is limited, the time required to produce the proteins of the cell will be minimized if the rate of translation is maximized and the mass of the translation machinery is minimized. In other words, the growth rate will be influenced not simply by the rate of translation but by this rate normalized to the mass of the translational machinery (30).

In addition, the extent to which translational efficiency influences growth rate will depend on the fraction of the total biomass that is invested in translation (30). This means that the significance of translational efficiency for the growth rate depends on the growth conditions. At high growth rates in rich media, translation is the dominant cellular function of a bacterium, and, accordingly, translation efficiency should be a definitive parameter. At very low growth rates in poor media, the translation machinery is a small fraction of the cell mass, and the translational efficiency should have a correspondingly small influence on the growth rates. This view of the variable influence of translational efficiency on growth rate has been directly verified with a series of ribosome mutants. It has been shown that the growth rate impairment due to defective ribosomes is maximal at the highest growth rates and decreases systematically as the quality of the growth medium is lowered (66).

A related parameter that is growth rate dependent is the heterogeneity of the protein population (62). Under poor growth conditions a broad spectrum of enzymes as well as core components such as translational, transcriptional, and membrane proteins are being produced. In contrast, at the highest growth rates the same core components completely dominate the protein population. This means that the mRNA population will become progressively less heterogeneous as the growth rates increase, and at the highest growth rates the species that code for core components will represent most of the mRNA species. If we suppose that only a subset of codons are used preferentially to code for the core proteins, a situation is created that could markedly influence the efficiency of the translation system at the highest growth rates (57).

Maximization of the rates of translation would require increasing the aminoacyl-tRNA ternary complex concentrations sufficiently to drive the ribosomes at near-maximum rates. However, if this mass increase were uniform for all ternary complexes, it would tend to lower significantly the efficiency of translation. On the other hand, if the core proteins were encoded preferentially by a subset of codons, the mass investment in ternary complexes corresponding to these codons could be compensated, at least in part, by decreasing the relative concentration of the ternary complexes corresponding to the codons that are avoided in the core protein genes (57).

The major proteins of bacteria grown under normal laboratory conditions are, in fact, those that we have identified as the core proteins, and their relative amounts decrease substantially in very slowly growing cells. If the major codon preference is, as we suggest, a strategy to maximize growth rates in relatively rich media, the high-concentration tRNA species that are matched with the major codons in rapidly growing (normal) cultures should diminish significantly at the very lowest growth rates. In other words, the growth optimization strategy requires that the steady-state concentrations of individual tRNA species are regulated in a growth rate-dependent manner.

The data concerning the isoacceptor levels at different growth rates are far from complete, but the fragmentary data available at this writing are encouraging. First, the relative amounts of the different tRNA species under different growth conditions are not constant, but vary in systematic ways (Emilsson and Kurland, unpublished; Fournier and McKay, personal communication). For example, analysis of the leucine family shows that the concentration of the major Leu isoacceptor progressively increases at higher growth rates. Likewise, the level of the one minor Leu species that also can translate the major Leu codon (tRNA_Leu; CUA, CUG) increases almost threefold, whereas the remaining three decrease by factors of 2 to 3 as the growth rate increases (Emilsson and Kurland, unpublished). In the same series of measurements it was found that the levels of all three members of the methionine acceptor family increase when the growth rates are increased. Since these read AUG, they are by definition translators of a major codon, and their increase with growth rate is the predicted behavior.

In summary, there are reasonable grounds to believe that the major codon preference is a strategy to adapt bacteria and yeasts to rapid growth. Although it is not a strategy to regulate protein expression levels, there is indeed a very important regulatory problem nested in the major codon strategy. This concerns the unknown mechanism that regulates in a growth-rate-dependent way the expression level for each individual tRNA isoacceptor species.
DIVERSE REMEDIES

Roughly speaking, there are three aspects of protein biosynthesis that can go wrong: (i) individual amino acid missense substitutions can occur; (ii) processivity can be interrupted by a reading frame error, a drop-off event, or an abortive termination event; and (iii) protein folding can go awry. There are suggestions that specific codon preference strategies are directed at minimizing these sorts of errors.

It should be said that although, as we have claimed above, cells are more tolerant of translational ambiguity than was initially expected, this does not mean that translational errors are not deleterious. Rather, we mean that tolerable mistakes are still deleterious in the sense that they represent a wastage of biosynthetic potential. Accordingly, strategies to minimize such wastage will be advantageous to cells.

Unfortunately, so few data are available that virtually nothing can be said about the codon dependence of the missense errors in vivo (77). However, it is clear enough that whenever a minor codon can be misread by a major tRNA species, there is the danger that unacceptably high missense error rates will be the result. This sort of consideration has led to the suggestion that the choice of the members of the major codon preference is determined by the codon’s out of phase (C) or in phase nucleotide motif. Similarly, a device is needed to explain why looping out of nucleotides in short mRNA sequence would not disturb this phasing mechanism by presenting alternative mRNA structures to the ribosome. In summary, there are good reasons to continue to explore the functions of the ubiquitous G-non-G-N motif in reading frame maintenance. Likewise, there are equally good reasons to be skeptical about whether it is functioning as has been suggested.

Finally, the correlation that we have discussed above that associates pause sites with short strings of rare codons naturally leads to a conjecture concerning the folding process for some proteins. Thus, a rare codon string that transiently slows down the transit of the ribosome over the mRNA might facilitate proper folding of a protein domain (56, 79, 80). In particular, when the incomplete polypeptides can be arranged in alternative folds, the order of the folding could be relevant. Here, a pause site that permits the one domain to be organized before an alternative possibility is elaborated might provide a smoother assembly pathway for certain proteins. There is no strong evidence either to support or to rule out this conjecture. However, the combination of site-directed mutagenesis and protein structure studies that have become feasible now certainly offers the opportunity for a direct test of this conjecture.

CONCLUSIONS

In this review, we have emphasized the notion that the genetic code is not a frozen code. In particular, we have presented a view of evolving codon assignments as well as codon usage patterns as the adaptive response of genomes to the solution of practical problems of gene expression. For example, we have associated the evolution of the sense codon reassignments of the vertebrate mitochondria with a strategy to reduce the entire genetic apparatus of these systems. Similarly, we have associated the tendency to use only a subset of codons in the highly expressed genes of microorganisms with a growth maximization strategy. These two strategies are related in the sense that both of them are dependent on the coevolution of codon patterns and tRNA abundances. More precisely, both strategies depend on a tendency to use codons in such a way that the complexity of...
tRNA populations is reduced. This similarity leads to the notion that codon reassignments are usefully viewed as extreme forms of codon preference strategies.

Our view of the evolution of codon strategies does not exclude a role for biased mutation pressure, a role that has been emphasized by others (73). However, we wish to draw attention to the selective role of a functional feedback between constraints on gene expression and the microstructure of genomes. In our view, codon usage patterns are evolving along with other characteristics of a genetic system as the result of an interplay between mutational and selective forces. The particular selective forces that we have emphasized here are directed toward a minimization of the biosynthetic apparatus of mitochondria and a maximization of microbial growth rates. There is no doubt in our minds that other such genomic strategies remain to be described.

LITERATURE CITED

44. Hinds, P. W., and R. D. Blake. 1985. Delineation of coding areas in DNA sequences through assignment of codon proba-
CODON PREFERENCES IN FREE-LIVING MICROORGANISMS

209


