The Influence of Anticodon–Codon Interactions and Modified Bases on Codon Usage Bias in Bacteria

Wenqi Ran and Paul G. Higgs*

Department of Physics and Astronomy, McMaster University, Hamilton, Ontario, Canada

***Corresponding author:** E-mail: higgsp@mcmaster.ca.

Associate editor: James McInerney

Abstract

Most transfer RNAs (tRNAs) can translate more than one synonymous codon, and most codons can be translated by more than one isoacceptor tRNA. The rates of translation of synonymous codons are dependent on the concentrations of the tRNAs and on the rates of pairing of each anticodon–codon combination. Translational selection causes a significant bias in codon frequencies in highly expressed genes in most bacteria. By comparing codon frequencies in high and low-expression genes, we determine which codons are preferred for each amino acid in a large sample of bacterial genomes. We relate this to the number of copies of each tRNA gene in each genome. In two-codon families, preferred codons have Watson–Crick pairs (GC and AU) between the third codon base and the wobble base of the anticodon rather than GU pairs. This suggests that these combinations are more rapidly recognized by the ribosome. In contrast, in four-codon families, preferred codons do not correspond to Watson–Crick rules. In some cases, a wobble-U tRNA can pair with all four codons. In these cases, A and U codons are preferred over G and C. This indicates that the nonstandard UU combination appears to be translated surprisingly well. Differences in modified bases at the wobble position of the anticodon appear to be responsible for the differences in behavior of tRNAs in two- and four-codon families. We discuss the way changes in the bases in the anticodon influence both the speed and the accuracy of translation. The number of tRNA gene copies and the strength of translational selection correlate with the growth rate of the organism, as we would expect if the primary cause of translational selection in bacteria is the requirement to optimize the speed of protein production.

Key words: codon usage, translational selection, tRNA, modified nucleotides, anticodon.

Research article

Introduction

It has long been known that, in many organisms, synonymous codons do not occur with equal frequency. In particular, selection acts to increase the frequency of preferred codons in high-expression genes, such as ribosomal proteins (Sharp and Li 1986). It is also well known that preferred codons tend to correspond to the transfer RNAs (tRNAs) that have the highest concentrations in cells (Ikemura 1981, 1985). It is presumed that high tRNA concentration leads to rapid translation and that this is an advantage to organisms that require a rapid rate of protein production. The total number of tRNA gene copies in bacterial genomes varies from fewer than 30 to more than 120 due to the presence of duplicate copies of some tRNA genes in some genomes. In cases where tRNA concentrations have been measured (Kanaya et al. 1999), it is found that concentrations are roughly proportional to gene copy numbers, although concentrations vary with growth conditions of the cell (Dong et al. 1996) and can depend on the position of the tRNA gene on the genome (Ardell and Kirsebom 2005). Nevertheless, it is presumed that if high tRNA concentrations are required, duplicate tRNA genes are required in order to provide a higher rate of tRNA transcription. Rocha (2004) showed that the total number of tRNA gene copies in bacterial genomes correlates with the doubling rate of the cells. Bacteria that require a high growth rate require duplicate tRNAs.

Translational selection also influences the relative frequencies of codons that are translated by the same tRNA. In particular, amino acids with two-codon families ending in U and C have a single type of tRNA that has a G at the wobble position. This G pairs with both bases at the third codon position (see fig. 1*a*). Sharp et al. (2005) showed that for a large majority of U+C codon families in bacterial genomes, the C codon is preferred in high-expression genes. This preference cannot be explained by variation of tRNA concentration and gene copy number. The effect must come from the relative efficiencies of pairing of the same tRNA anticodon with two different codons. As the direction of the preference for the C codon in U+C families is consistent across species, Sharp et al. (2005) used the strength of this preference as an indicator of the overall strength of translational selection, and was able to compare the selection strength in different genomes.

In general, many tRNAs are able to pair with more than one codon, and many codons may be translated by more than one kind of isoacceptor tRNA with different wobble bases in the anticodon. For example, in two-codon families ending in A and G, there is always a wobble-U tRNA that pairs with both codons, and there is sometimes a wobble-C tRNA that is presumed to pair with only the G codon (fig. 1*b*). In four-codon families, there may be three different kinds of tRNAs with wobble bases U, G, and C. Grosjean et al. (2010) have reviewed which tRNA combinations occur in which organisms in detail. Where all three of these

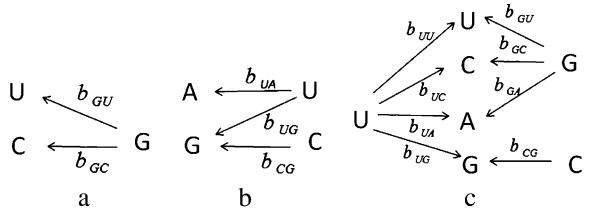

Fig. 1. Pairing possibilities for three types of codon family: (*a*) U+C families, (*b*) A+G families, and (*c*) four-codon families. An arrow labeled b_{ji} indicates that a tRNA with wobble base *j* can pair with a codon with third position base *i*. The parameter b_{ji} denotes the relative rate of translating this codon by this type of tRNA. Only those combinations of anticodons and codons that are labeled by arrows are assumed to occur.

tRNAs occur, Grosjean et al. (2010) call this "sparing strategy 1." There are also cases where a combination of wobble-U and wobble-G tRNAs occurs ("sparing strategy 2"), and where only the wobble-U tRNA is present ("sparing strategy 3"). As far as we know, the wobble-C tRNA only pairs with the G codon. It might be expected that the wobble-G pairs principally with U and C and the wobble-U pairs principally with A and G (see fig. 3 of Grosjean et al. (2010)). However, we know that U can pair with all four codons when it is the only tRNA, and we also know that GA pairing occurs in some mitochondria. Therefore, our figure 1*c* includes all these cases. This means that more than one tRNA can translate most codons, and the total rate of translation of each codon is dependent on the sum of the rates at which all tRNA types translate the codon. This is an important part of our model of selection on codon usage. In addition to the above, there are a few cases where tRNAs with an A at the wobble position occur in four-codon families, and the A is modified to inosine (I). These tRNAs are not included in our standard model (fig. 1*c*) but will be discussed as special cases.

The relative translation rates of codons, and hence the direction of selection on codon bias, depend on the numbers of tRNA gene copies of each type and the relative rates of pairing of each anticodon with each codon. Now that we have large numbers of complete bacterial genomes, we are in a position to observe which codons are preferred when different combinations of tRNA gene copies are present and to use these observations to deduce the relative rates of translation of each codon by each tRNA.

In our previous paper (Higgs and Ran 2008), we used a simple model of translation kinetics to calculate mean translation times of each codon and proposed that the selection coefficient for one synonymous codon over another was proportional to the difference in these times. This theory predicts which codon is preferred for each combination of tRNA gene copy numbers. Codon frequencies and tRNA copy numbers evolve toward coadapted stable states in which neither can change without the other. We showed that, in many cases, there is more than one possible stable state. This explains why the direction of selection sometimes differs between codon families of the same type in the same organism (e.g., for A+G families, the A codon might be preferred for one amino acid but the G codon for another). We also included the possibility of tRNA gene duplication and deletion and showed that it will pay to duplicate genes for organisms in which the intrinsic strength of selection is high. Our theory therefore explains why high tRNA gene copy numbers are found in species with strong codon bias and why both high copy numbers and strongly biased codon usage are found in fast multiplying organisms.

The key unknown parameters in our theory are the relative rates of pairing of the different anticodon–codon combinations, that is, the *b* parameters in figure 1. The main objective in this paper is to deduce as much as possible about these rate parameters by comparing the observed codon usage with the theory. We show that the presence of modified bases in the anticodon influences the rate parameters and leads to observable effects on codon usage. We then discuss our results in the context of the debate over the relative importance of speed and accuracy in translational selection.

Population Genetics and Translational Kinetics Model

The population genetics model for two-codon families has been described in several previous papers (Li 1987; Shields 1990; Bulmer 1991; Higgs and Ran 2008). Here, we will summarize it in a way that is valid for both two- and four-codon families. For each codon *i* in a given family, we set the fitness to be $1 + s_i$. We choose a reference codon and define $s_i = 0$ for the reference codon. For the other codons, s_i is the selective advantage or disadvantage of this codon with respect to the reference. Let π_i be the stationary frequency of base *i* under the mutation process. The mutation rate from the reference codon to codon *i* is $u\pi_i$ and the reverse mutation rate is $u\pi_{\mathrm{ref}}$. The value of *u* determines the absolute values of the mutation rates, but the expected frequencies of the codons do not depend on *u*. Let φ_i and φ_{ref} be the expected frequencies of the two codons under mutation–selection–drift balance. Using the population genetics theory described in the papers cited above, we find

$$\frac{\phi_i}{\phi_{\mathrm{ref}}} = \frac{\pi_i \exp(S_i)}{\pi_{\mathrm{ref}}}. \quad (1)$$

$$\phi_i = \frac{\pi_i \exp(S_i)}{\sum_j \pi_j \exp(S_j)}, \quad (2)$$

where $S_i = 2N_e s_i$, and N_e is the effective population size. If a different codon were chosen as reference, this would shift all the S_i values up or down by a constant, but this would make no difference to φ_i. Therefore, it makes no difference which codon is used as reference.

As in Higgs and Ran (2008) and Sharp et al. (2005), a set of highly expressed genes has been identified in which selection on codon usage is presumed to be significant and where codon frequencies should be given by equation (2). The majority of genes in the genome are assumed to be low-expression genes where translational selection is negligible and where codon frequencies are equal to the

frequencies under mutation, π_i. Let $n_i^{\text{low}}$and n_i^{high} be the number of occurrences of codon i in a codon family for one particular amino acid in one particular genome for the high and low-expression genes, respectively. If the low-expression genes are only influenced by mutation, then the π parameters can be estimated from the observed codon counts in the low-expression genes, that is

$$\pi_i = \frac{n_i^{\text{low}}}{\sum_j n_j^{\text{low}}}. \quad (3)$$

Similarly, the frequencies under selection can be estimated from the observed codon counts in the high-expression genes,

$$\phi_i = \frac{n_i^{\text{high}}}{\sum_j n_j^{\text{high}}}. \quad (4)$$

Now, rearranging (1) and using (3) and (4) gives

$$S_i = \ln\left(\frac{\phi_i}{\phi_{\text{ref}}}\frac{\pi_{\text{ref}}}{\pi_i}\right) = \ln\left(\frac{n_i^{\text{high}} n_{\text{ref}}^{\text{low}}}{n_{\text{ref}}^{\text{high}} n_i^{\text{low}}}\right). \quad (5)$$

Thus, the selection parameters S_i can be estimated directly from the codon counts.

For a given codon family, let N_j be the number of tRNA gene copies with wobble base j. We suppose that the concentration of a tRNA is proportional to the number of gene copies, that is, the concentration is $c_0 N_j$, where c_0 is the concentration resulting from a single tRNA gene. As discussed in the introduction, this is a reasonable approximation in species where concentrations have been measured. It has been shown that some tRNAs may be charged less efficiently than others with their appropriate amino acid (Elf et al. 2003). This factor is ignored in the present model because we do not want to introduce extra recharging rates into the model and because this may only be important in starvation conditions. Also, incompletely charged tRNAs cannot explain selection between codons translated by the same tRNA, which is one of the major effects that we see in the data.

The rate constant for translation of a codon ending in base i by a tRNA with wobble base j is taken to be $k_0 b_{ji}$ where k_0 is a typical rate and b_{ji} is a dimensionless number of order 1, which describes the relative rate of translation by this anticodon–codon combination. In general, the rate of translation of codon i is the sum of the rates at which it is translated by all the tRNAs that interact with that codon:

$$r_i = c_0 k_0 \sum_j N_j b_{ji}. \quad (6)$$

The mean time to translate the codon is $1/r_i$. The selective advantage or disadvantage of codon i relative to the reference codon is assumed to be proportional to the difference in the times, Δt, between the two codons

$$s_i = s_0 \Delta t = s_0\left(\frac{1}{r_{\text{ref}}} - \frac{1}{r_i}\right) \quad (7)$$

where s_o is a genome-specific constant that determines the strength of selection in that genome. Note that s_i is positive if $r_i > r_{\text{ref}}$. We suppose that s_0 varies among genomes because the extent to which translational speed is important to an organism depends on its lifestyle. Organisms that need to multiply rapidly should be under significant selection to increase translational speed and should have a high s_0.

It will be useful to define relative rates ρ_i as

$$\rho_i = \frac{r_i}{c_0 k_0} = \sum_j N_j b_{ji}. \quad (8)$$

From this, the selection parameters S_i, which can be compared with the data, can be written as

$$S_i = 2N_e s_i = K\left(\frac{1}{\rho_{\text{ref}}} - \frac{1}{\rho_i}\right). \quad (9)$$

For convenience, we combined several parameters into a single parameter, $K = \frac{2N_e s_o}{c_o k_o}$. In order to make the role of the effective population size more explicit, we changed the notation slightly from Higgs and Ran (2008). In that paper, a symbol σ was used, which can be written as $\sigma = 2N_e s_o$. in the current notation.

We will consider the three types of codon families described in the introduction and figure 1. Only those combinations shown by an arrow in figure 1 are assumed to occur at significant rate. We can now write down the relative rates specifically for the three types of codon family.

U+C families—

$$\rho_U = N_G b_{GU}, \quad \rho_C = N_G b_{GC}. \quad (10)$$

A+G families—

$$\rho_A = N_U b_{UA}, \quad \rho_G = N_U b_{UG} + N_C b_{CG}. \quad (11)$$

Four-codon families—

$$\begin{aligned} \rho_U &= N_U b_{UU} + N_G b_{GU}, \quad \rho_C = N_U b_{UC} + N_G b_{GC}, \\ \rho_A &= N_U b_{UA} + N_G b_{GA}, \quad \rho_G = N_U b_{UG} + N_C b_{CG}. \end{aligned} \quad (12)$$

It should be remembered that the U and G bases at the wobble position are modified in many tRNAs and that this is likely to have a significant effect on the rates of interaction with the codons. We will consider these effects in detail later, but for simplicity of notation, we will use the unmodified form of the base in the subscripts. Moreover, when analyzing the data, we wish to group cases with the same anticodons from a large number of different organisms, and the nature of the base modification is not known in every case. For this reason, we are obliged to consider only the unmodified form of the base at this stage.

The relative rates should be properties of the tRNA structures, the anticodon–codon interaction, and the

way that recognition of the correct tRNA occurs in the ribosome. However, if the mechanism of translation is similar in different organisms, the relative rates should not depend on the organism. If a particular anticodon–codon combination leads to rapid translation in one organism, then the same combination should be rapid in another organism. Thus, we expect to see a consistent relationship between preferred codons and tRNA genes that applies across many organisms. Equation (9) makes it clear that the direction of selection depends on the relative rates, which are tRNA-dependent properties, whereas the magnitude of selection depends on K, which is organism dependent.

Analysis of Codon Usage in U+C Codon Families

Here, we summarize the results on U+C codon families from our previous paper (Higgs and Ran 2008). A set of 80 genomes was used that spans the full range of complete bacterial genomes, as previously selected by Sharp et al. (2005). We obtained S_C for the C codon in each family, using the U codon as reference (as in eq. 5). We found that there is consistent preference for the C codon in most species. Using equations (9, 10), we have

$$S_C = K\left(\frac{1}{\rho_U} - \frac{1}{\rho_C}\right) = \frac{K}{N_G}\left(\frac{1}{b_{GU}} - \frac{1}{b_{GC}}\right). \qquad (13)$$

Preference for C indicates that $b_{GC} > b_{GU}$. Our interpretation is that the strongly interacting GC pair is processed more rapidly by the ribosome than the weakly interacting GU pair.

In some cases, the U codon is found to be more frequent in the high-expression genes; hence, S_C is found to be negative in equation (5). This could be explained if b_{GC} were less than b_{GU} for some tRNAs. However, the cases with negative S_C occur in species with slow growth rates, few tRNAs, and generally weak translational selection. Thus, we interpret the increase of U in the high-expression genes in these cases as being a mutational effect that is not accounted for in the model, for example, base frequencies may differ between leading and lagging strands or according to position of a gene relative to the origin of replication, and high-expression genes may be nonrandomly positioned on the genome. These effects might be present to some extent in all species, but they would show up as anomalies in cases where translational selection is weak. Fourteen of the original 80 species were excluded from the subsequent analysis because they did not show a consistent preference for C in U+C families, and it appeared that translational selection was weak in these species. The remaining 66 species showed a consistent preference for C, and it was assumed that translational selection was a significant effect.

Analysis of Codon Usage in A+G Codon Families

The following analysis was carried out using the 66 species selected in the previous section. We included the A+G codon families for Leu (UUR), Gln (CAR), Lys (AAR), Glu (GAR), and Arg (AGR). For each amino acid in each genome, the observed codon usages π_i and φ_i in the low and high-expression genes were obtained from sequence data (using eqs. 3, 4). We used A as the reference codon and determined the selection strength of G relative to A using equation (5). We define the preferred codon as the one whose frequency increases in the high-expression genes. From equations (9, 11), the selection strength of the G codon relative to A is

$$S_G = K\left(\frac{1}{\rho_A} - \frac{1}{\rho_G}\right) = K\left(\frac{1}{N_U b_{UA}} - \frac{1}{N_U b_{UG} + N_C b_{CG}}\right). \qquad (14)$$

This may be positive or negative depending on the number of tRNA gene copies of the two types and on the values of the b parameters.

Cases were grouped according to the combination N_U:N_C of tRNA genes that is present for that amino acid. table 1 shows the number of cases for which A and G codons are preferred for each combination of tRNAs. The mean values of the codon frequencies in high and low-expression genes are also shown for each tRNA combination. Frequencies that increase in the high-expression genes are underlined. In table 1, we see that when $N_U > N_C$, the A codon is preferred in the majority of cases, and the mean frequency of A is higher in the high-expression genes than the low-expression genes. In contrast, when $N_U \leq N_C$, the G codon is preferred in the majority of cases, and the mean frequency of G is higher in the high than the low-expression genes.

These results give information on the b parameters. As G is preferred in the 1:1 case, we obtain $b_{UG} + b_{CG} > b_{UA}$, and as A is preferred in the 2:1 case, we obtain $2b_{UG} + b_{CG} < 2b_{UA}$. These inequalities can be rewritten as $(1-b_{UG}/b_{UA}) < b_{CG}/b_{UA} < 2(1-b_{UG}/b_{UA})$, which defines the region

Table 1. Codon Usage and tRNA Content in A+G Families.

tRNA Gene Copies		Number of Cases	Number of Cases Where Each Codon Is Preferred		Low Exp Genes		High Exp Genes	
N_U	N_C	Total	#A	#G	π_A	π_G	φ_A	φ_G
1	0	61	45	16	0.736	0.264	0.787	0.213
2	0	21	20	1	0.715	0.285	0.833	0.167
3	0	16	12	4	0.750	0.250	0.814	0.186
4	0	15	14	1	0.684	0.316	0.788	0.212
5	0	8	6	2	0.675	0.325	0.748	0.252
6	0	5	4	1	0.685	0.315	0.729	0.271
7	0	2	1	1	0.703	0.300	0.686	0.314
4	1	1	1	0	0.943	0.057	1.000	0.000
7	2	1	1	0	0.696	0.304	0.735	0.265
3	1	12	9	3	0.695	0.305	0.746	0.254
2	1	31	20	11	0.642	0.358	0.664	0.336
4	2	1	1	0	0.673	0.327	0.721	0.279
3	2	1	1	0	0.592	0.408	1.000	0.000
1	1	136	52	84	0.459	0.541	0.448	0.552
2	2	4	0	4	0.489	0.511	0.372	0.628
1	2	6	1	5	0.248	0.752	0.083	0.917
1	3	6	0	6	0.209	0.791	0.099	0.901
0	1	2	0	2	0.460	0.540	0.360	0.640
0	2	1	0	1	0.049	0.951	0.000	1.000

NOTE.—Underlined figures indicate codons whose frequency in the high expression genes is greater than in the low expression genes.

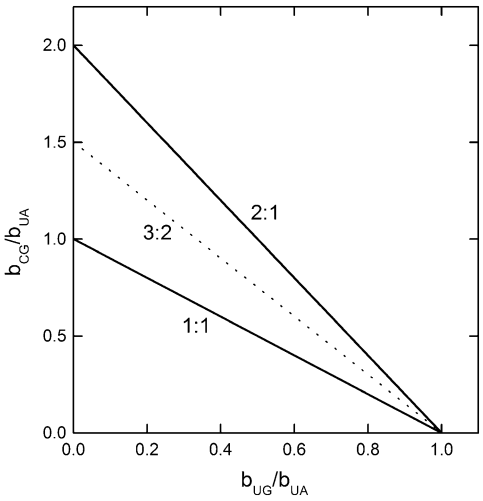

FIG. 2. Regions of parameter space for the relative rates of translation in A+G codon families.

between the two solid lines in figure 2. Most of the other combinations (such as 3:1 and 1:2) give inequalities that are less restrictive than the two above. Therefore, any set of parameters that falls between these lines will also correctly explain the direction of selection for the other tRNA combinations. Note that the parameters that we chose as examples previously (Higgs and Ran 2008) fall in the middle of this range—$b_{CG}/b_{UA} = 1$ and $b_{UG}/b_{UA} = 0.4$. The only case that is more restrictive than these two is 3:2, for which A is preferred. This gives $b_{CG}/b_{UA} < 3/2\ (1 - b_{UG}/b_{UA})$, which restricts parameters to the region below the dotted line in figure 2. However, there is a single example of the 3:2 case, so this condition is not as well substantiated as the other two.

Three cases at the bottom of table 1 have $N_U = 0$. This is impossible according to our model if only the U tRNA can translate the A codon. The two cases with $N_U{:}N_C = 0{:}1$ do not have particularly strong codon bias, so it may be that there is a misannotation of the tRNA genes in these cases, and there is really a wobble-U tRNA present. The single case with $N_U{:}N_C = 0{:}2$ stands out because the A codon is completely absent from the high-expression genes and is very rare in the low-expression genes. It is therefore possible that this case is real and that it illustrates extremely strong selection against the A codon.

Analysis of Codon Usage in Four-Codon Families

We now consider four-codon families in the same set of 66 bacterial genomes. Eight-codon families are considered: Leu (CUN), Val (GUN), Ser (UCN), Pro (CCN), Thr (ACN), Ala (GCN), Arg (CGN), and Gly (GGN). Although Leu, Ser, and Arg have six codons, these can be divided into a group of two and a group of four that are translated by separate tRNAs. Therefore, we consider the group of four codons for these amino acids in this section. We choose U as a reference codon because U often has a high frequency in the high-expression genes. We define the preferred codon to be the one with the highest S. The reference codon has $S_{ref} = 0$. It may be that the reference codon is preferred, in which case, the other three codons all have $S < 0$.

In tables 2–5, each line corresponds to one combination of tRNA gene copy numbers. Table 2 includes cases where only N_U is nonzero. The fact that there are many such cases that clearly shows that wobble-U tRNAs can pair with all four codons. In this table, the A and the U codons are most frequently preferred. Based on Watson–Crick pairing, we would expect A to be preferred, but the large number of cases where U is preferred is surprising. In a four-codon family, the number of codons that increase in frequency in the high-expression genes can be one, two, or three. Table 2 shows that, on average, both U and A codons increase (as shown by the underlined figures). We also find that the S value for the C codon is usually the lowest, that is, C usually decreases the most in the high-expression genes. In terms of the rate parameters, these results show that b_{UA} and b_{UU} are both high and are roughly equal to one another. Both of these are higher than b_{UG}, and all three are higher than b_{UC}. These predictions on the relative rates could be tested in future experiments on translational kinetics.

Table 3 includes the cases where N_G and N_U are nonzero, listed in order of increasing ratio of $N_G{:}N_U$. In the upper half of the table, $N_G < N_U$. Here, both U and A codons increase in frequency on average in the high-expression genes, as with table 2. U and A are most frequently the preferred codons. In the lower half of table 3, $N_G \geq N_U$. Here, only the U codon increases in the high-expression genes, and there is a large majority of cases where U is the preferred codon. We expect that wobble-G tRNAs pair principally with U and C codons. It would appear that U is preferred in these cases because both the wobble-G and the wobble-U

Table 2. Codon Usage in Four-Codon Families Where Only Wobble-U tRNAs Are Present.

tRNA Gene Copies	Number of Cases	Number of Cases Where Each Codon Is Preferred				Low Exp Genes				High Exp Genes			
N_U	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	33	13	1	14	5	0.461	0.065	0.408	0.066	0.473	0.039	0.437	0.052
2	18	5	0	12	1	0.385	0.064	0.445	0.100	0.387	0.005	0.563	0.045
3	16	8	0	7	1	0.370	0.113	0.345	0.173	0.487	0.033	0.417	0.063
4	6	5	0	1	0	0.406	0.107	0.348	0.139	0.570	0.030	0.342	0.058
5	2	2	0	0	0	0.403	0.100	0.341	0.156	0.618	0.024	0.295	0.062
6	3	2	0	1	0	0.474	0.135	0.320	0.071	0.572	0.036	0.353	0.040
7	1	0	0	1	0	0.466	0.112	0.338	0.083	0.521	0.034	0.389	0.056

NOTE.—Underlined figures indicate codons whose frequency in the high expression genes is greater than in the low expression genes.

Table 3. Codon Usage in Four-Codon Families With Both Wobble-U and Wobble-G tRNAs.

tRNA Gene Copies		Number of Cases	Number of Cases Where Each Codon Is Preferred				Low Exp Genes				High Exp Genes			
N_G	N_U	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	5	5	5	0	0	0	0.270	0.200	0.302	0.228	0.494	0.042	0.334	0.130
2	10	1	0	0	1	0	0.295	0.157	0.251	0.298	0.278	0.005	0.624	0.093
1	4	15	14	0	1	0	0.276	0.169	0.300	0.255	0.523	0.051	0.321	0.105
2	7	1	0	0	1	0	0.254	0.224	0.172	0.351	0.302	0.005	0.545	0.148
1	3	20	10	0	10	0	0.344	0.129	0.376	0.152	0.443	0.031	0.466	0.061
2	5	4	4	0	0	0	0.221	0.272	0.175	0.332	0.555	0.107	0.232	0.105
1	2	19	13	0	5	1	0.364	0.151	0.321	0.163	0.494	0.035	0.388	0.083
2	4	5	5	0	0	0	0.252	0.279	0.168	0.302	0.624	0.064	0.239	0.073
4	7	1	1	0	0	0	0.281	0.052	0.586	0.081	0.403	0.052	0.541	0.004
2	3	10	10	0	0	0	0.284	0.249	0.207	0.259	0.540	0.122	0.223	0.114
1	1	64	37	9	9	9	0.405	0.175	0.269	0.151	0.475	0.149	0.262	0.115
2	2	9	9	0	0	0	0.289	0.224	0.271	0.217	0.522	0.150	0.258	0.070
3	3	2	2	0	0	0	0.231	0.314	0.261	0.194	0.463	0.248	0.220	0.068
4	4	1	1	0	0	0	0.371	0.125	0.363	0.141	0.638	0.155	0.194	0.012
4	3	2	2	0	0	0	0.215	0.268	0.313	0.203	0.500	0.165	0.284	0.051
3	2	1	1	0	0	0	0.358	0.217	0.291	0.135	0.553	0.176	0.241	0.030
2	1	3	3	0	0	0	0.316	0.280	0.241	0.163	0.558	0.206	0.192	0.044
3	1	1	1	0	0	0	0.433	0.294	0.164	0.108	0.784	0.170	0.038	0.007
6	2	1	1	0	0	0	0.388	0.363	0.118	0.131	0.659	0.327	0.014	0.000
7	2	1	1	0	0	0	0.391	0.380	0.103	0.127	0.709	0.280	0.010	0.000
4	1	1	1	0	0	0	0.436	0.247	0.151	0.166	0.800	0.142	0.039	0.018
11	2	1	1	0	0	0	0.447	0.358	0.102	0.092	0.683	0.309	0.009	0.000
6	1	1	1	0	0	0	0.380	0.389	0.103	0.128	0.654	0.316	0.020	0.011

NOTE.—Cases are listed in order of increasing ratio of N_G:N_U. In the upper half, $N_G < N_U$. In the lower half $N_G \geq N_U$.
Underlined figures indicate codons whose frequency in the high expression genes is greater than in the low expression genes.

tRNAs pair well with the U codon. Wobble-G tRNAs can also pair with A codons in some circumstances (see the section on modified bases). However, there are no observed cases where $N_G > 0$, $N_C > 0$, and $N_U = 0$. The probable reason for this is that the wobble-G tRNA on its own is not sufficient to translate the A codon. This shows b_{GA} must be small and could probably be eliminated from figure 1*c*. Table 2 shows that the principal effect of increasing N_G is to reduce the preference for A and to increase the preference for U. This is what we would expect if b_{GA} is small.

Table 4 includes cases where N_C is nonzero. The cases where $N_C = 1$ are rather similar to tables 2 and 3. The preferred codon is usually U. If $N_G = 0$ or $N_G < N_U$, then both U and A codons increase in frequency in the high-expression genes, whereas when $N_G > N_U$, usually only the U codon increases in frequency in the high-expression

Table 4. Codon Uses in Four-Codon Families With Combinations of Wobble-C tRNAs With Wobble-U and Wobble-G tRNAs.

tRNA Gene Copies			Number of Cases	Number of Cases Where Each Codon Is Preferred				Low Exp Genes				High Exp Genes			
N_G	N_U	N_C	Total	#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
0	3	1	1	0	0	1	0	0.449	0.046	0.440	0.064	0.363	0.004	0.617	0.016
0	2	1	2	2	0	0	0	0.218	0.203	0.323	0.256	0.467	0.135	0.359	0.041
1	4	1	1	1	0	0	0	0.359	0.221	0.280	0.141	0.398	0.184	0.276	0.142
1	3	1	4	3	0	1	0	0.218	0.291	0.254	0.237	0.370	0.223	0.297	0.110
1	2	1	12	10	0	2	0	0.288	0.201	0.261	0.251	0.442	0.124	0.297	0.136
2	3	1	3	3	0	0	0	0.166	0.387	0.118	0.330	0.336	0.380	0.099	0.186
1	1	1	137	60	41	7	29	0.188	0.334	0.146	0.332	0.208	0.259	0.117	0.317
2	2	1	1	1	0	0	0	0.053	0.440	0.126	0.382	0.164	0.447	0.131	0.257
2	1	1	24	21	2	0	1	0.140	0.471	0.083	0.307	0.268	0.520	0.034	0.177
4	2	1	1	1	0	0	0	0.308	0.194	0.398	0.100	0.652	0.113	0.215	0.020
3	1	1	9	7	2	0	0	0.154	0.525	0.078	0.243	0.265	0.554	0.028	0.152
4	1	1	3	3	0	0	0	0.258	0.502	0.086	0.153	0.576	0.402	0.009	0.014
1	2	2	1	0	0	0	1	0.249	0.236	0.171	0.343	0.181	0.043	0.162	0.614
1	1	2	8	3	1	0	4	0.126	0.225	0.079	0.570	0.141	0.219	0.048	0.592
2	1	2	2	2	0	0	0	0.129	0.378	0.104	0.390	0.334	0.449	0.027	0.190
3	1	2	1	1	0	0	0	0.016	0.545	0.031	0.409	0.059	0.650	0.003	0.287
1	5	3	1	0	0	1	0	0.196	0.228	0.135	0.441	0.180	0.024	0.302	0.494
3	1	3	1	1	0	0	0	0.022	0.527	0.044	0.407	0.088	0.650	0.011	0.252
1	1	4	2	0	0	0	2	0.144	0.135	0.056	0.665	0.055	0.040	0.002	0.902

NOTE.—Underlined figures indicate codons whose frequency in the high expression genes is greater than in the low expression genes.

Table 5. Codon Usage in Four-Codon Families Involving Wobble-A (or I) tRNAs.

tRNA Gene Copies				Number of Cases Total	Number of Cases Where Each Codon Is Preferred				Low Exp Genes				High Exp Genes			
N_A	N_G	N_U	N_C		#U	#C	#A	#G	π_U	π_C	π_A	π_G	φ_U	φ_C	φ_A	φ_G
1	0	0	0	1	1	0	0	0	0.122	0.548	0.068	0.262	0.262	0.685	0.008	0.044
1	0	0	1	17	14	1	0	2	0.223	0.427	0.113	0.237	0.368	0.473	0.045	0.115
2	0	0	2	1	1	0	0	0	0.263	0.520	0.074	0.142	0.634	0.247	0.004	0.016
2	0	0	1	16	16	0	0	0	0.427	0.321	0.140	0.112	0.687	0.289	0.011	0.013
3	0	0	1	4	4	0	0	0	0.509	0.197	0.199	0.096	0.813	0.172	0.014	0.001
4	0	0	1	5	5	0	0	0	0.310	0.400	0.124	0.167	0.672	0.314	0.006	0.007
6	0	0	1	3	3	0	0	0	0.424	0.384	0.126	0.065	0.762	0.231	0.005	0.002
8	0	0	1	1	1	0	0	0	0.469	0.343	0.163	0.026	0.779	0.215	0.006	0.000
1	0	1	0	10	8	0	1	1	0.570	0.147	0.213	0.071	0.737	0.089	0.107	0.066
1	0	2	0	2	2	0	0	0	0.418	0.197	0.230	0.155	0.922	0.033	0.038	0.007
1	0	3	0	1	1	0	0	0	0.476	0.138	0.285	0.101	0.935	0.027	0.031	0.007
1	0	1	1	2	2	0	0	0	0.158	0.407	0.170	0.264	0.480	0.415	0.051	0.053
2	0	2	1	1	1	0	0	0	0.561	0.163	0.163	0.114	0.792	0.205	0.000	0.003
1	1	1	1	1	1	0	0	0	0.246	0.223	0.278	0.353	0.537	0.151	0.141	0.171

NOTE.—Underlined figures indicate codons whose frequency in the high expression genes is greater than in the low expression genes.

genes. Cases where $N_C \geq 2$ (at the bottom of table 4) are fairly rare. When there is a high proportion of wobble-C tRNAs, such as the cases 1:1:2, 1:1:4, and 1:2:2, the G codon is often preferred. This suggests that b_{CG} is reasonably large. However, in cases where both N_G and N_C are in high proportion, such as 2:1:2, 3:1:3, and 3:1:2, the U codon is preferred over G. So this tells us that b_{CG} cannot be particularly large in comparison with b_{GU} and b_{UU}.

Table 5 shows the cases of four-codon families in which there are tRNAs with an A at the wobble position in the tRNA gene sequence. The A base is known to be modified to inosine (I) in many cases in the tRNA molecule, but we will use the notation N_A for the number of copies of this gene. Almost all these cases occur for the Arg AGN codon family, but there are also a small number of examples for Leu CUN and one example for Thr ACN. There is a single example where N_A:N_G:N_U:N_C = 1:0:0:0, which shows that the I can pair with all four codons. More typically, this tRNA occurs with a wobble-C tRNA, which suggests that I pairs efficiently with U, C, and A codons but not with G codons and that the wobble-C tRNA is usually required to translate the G codon. In almost all the examples in table 5, the preferred codon is U. Thus, b_{IU} must be the largest of the rates involving the wobble-I base. It is also seen that U is preferred relative to G, even though there is usually a wobble-C tRNA present to translate the G codon. Thus, b_{IU} must be high with respect to b_{CG} as well.

Does Selection Follow Preexisting Mutational Bias?

In our analysis, we were careful to distinguish codon bias arising from biased mutation rates from that caused by translational selection. We defined the preferred codon as the one with the highest S because this is the codon whose frequency increases the most due to selection in the high-expression genes. The preferred codon is not always the most frequent codon, as we now discuss.

Table 1 shows that the tRNA combinations where A is preferred correspond to cases where $\pi_A > \pi_G$ in the low-expression genes, whereas the reverse is true for combinations where G is preferred. We suppose that the frequencies in the low-expression genes are determined by mutational bias. The results show that the selection arising from the tRNAs is, on average, acting in the same direction as the mutational bias. Hence, the mean frequencies in the high-expression genes are more strongly biased than the low-expression genes and the bias is in the same direction. However, there are a significant number of cases that do not follow the average trend. Of the 330 cases in this table, the preferred codon is less frequent in the low-expression genes in 92 cases (28%). We have previously shown (Higgs and Ran 2008) that multiple stable states of codon usage and tRNA copy number can exist because the two quantities are coadapted. In a stable state, neither the tRNAs nor the codon frequencies can change without decreasing the translational efficiency. In some of these states, the direction of selection is opposite to that of the mutation bias. The observation of these states in the data is an important confirmation of the coevolution theory in our previous paper.

The following argument suggests why selection should follow the pre-existing mutational bias in the majority of cases. We have shown (Higgs and Ran 2008) that organisms with weak translational selection are likely to have only one tRNA for an A+G family. This must be a wobble-U tRNA whatever the GC content of the organism. If translational selection is stronger, it pays to have duplicate tRNAs because the benefit from increased translation rate outweighs the cost of the additional tRNAs. The additional tRNAs could be simple duplications of the wobble-U or duplications followed by an anticodon mutation to give wobble-C tRNAs. If there is a preexisting mutational bias in the codon frequencies of the low-expression genes, we would expect the additional tRNAs to follow this existing bias because the greatest speedup in translation will occur if the additional tRNA matches the codons that are already most frequent. Thus, when $\pi_A > \pi_G$, we expect to see tRNA combinations with increased N_U, and when $\pi_G > \pi_A$,

we expect to see increased N_C, as is actually the case in table 1.The most likely way that a genome could reach a state in which mutation and selection are opposed is that it was initially in a state where the selection followed the mutation bias, but the direction of mutation bias subsequently changed (e.g., because of changes in DNA replication enzymes). The codon frequencies in the low-expression genes would then change to follow the new mutation bias, but those in the high-expression genes would remain biased in the original direction because of selection from the existing tRNAs. Only if the mutational bias in the opposite direction became very strong would the existing coevolved state of tRNA copy numbers and codon usage become unstable. Stability conditions are discussed by Higgs and Ran (2008). Coevolution of tRNAs and codon usage can therefore exhibit hysteresis.

The results on four-codon families also illustrate the relevance of preexisting mutational bias. In table 2 is that the situation where only wobble-U tRNAs are present occurs when π_U and π_A are high in the low-expression genes. U and A are the preferred codons in this case; therefore, the direction of translational selection is in the same direction as the mutational bias for most of these cases. In table 3, the G and C frequencies in the low-expression genes are slightly higher than those in table 2. This shows that if the mutational process is such that the number of C codons is moderately large, then it pays to have a wobble-G tRNA to translate them. Cases where all three tRNAs exist (table 4) usually correspond to GC-rich codons in the low-expression genes because wobble-G and wobble-C tRNAs are more useful for translation in high GC species. However, because the U codon is preferred in many of these cases, these are examples where selection is acting in the opposite direction to the mutation bias.

There are several cases where a strong GC skew is apparent in the low-expression genes (i.e., the G and C frequencies are widely different). For example, π_C is much greater than π_G in cases where N_G is in high proportion (such as 6:1, 6:2, 7:2, and 11:2 in table 3), whereas π_G is much greater than π_C in cases where N_C is in high proportion (such as 1:1:2 and 1:1:4 in table 4). This shows that tRNA duplications are influenced by existing mutational bias. We presume that the reason the GC skew exists in the low-expression genes is because of context-dependent mutation. This means that the π frequencies can be different in different four-codon families. We will not pursue this point here, although we have discussed it in detail in the case of codon usage in mitochondrial genes (Jia and Higgs 2008). Context-dependent effects in codon frequencies in prokaryotic and eukaryotic genomes have also been found by Moura et al. (2007).

The Role of Modified Bases at the Wobble Position

There are many cases where the wobble base of the tRNA is modified in a way that influences anticodon–codon pairing (Curran 1998; Agris 2004, 2008, Grosjean et al. 2010). In this section, we consider the way that modified bases might influence the interpretation of the codon preference data above.

In U+C families, the G base at the wobble position is often modified to queuosine, Q, in tRNAs for Tyr, His, Asn, and Asp (Romier et al. 1998; Morris and Elliott 2001). We presume that this modification occurs in the majority of the bacteria in our sample. However, a detailed study of *Mycoplasma capricolum* (Andachi et al. 1989) shows that the G remains unmodified, and other cases outside bacteria are also known where the G in these tRNAs is unmodified (Morris and Elliott 2001; Jühling et al. 2009). Furthermore, the G is unmodified in tRNAs for other U+C codon families (Phe, Cys, and SerAGY) and for the three-codon Ile family. In *Escherichia coli*, Urbonavicius et al. (2001) found that both Q and G interact more efficiently with the C codon than the U codon. However, Meier et al. (1985) found that $tRNA^{His}$ in *Drosophila* showed a definite preference for the C codon when the G was unmodified but a slight preference for U when the Q modification was present. Also Morris et al. (1999) showed by molecular modeling that the Q modification promotes the ability of this type of tRNA to pair with the U codon. It seems likely that the reason for the presence of the Q modification is to increase the efficiency of pairing with the U codon. However, our observation is that the C codon is still preferred, even when Q is present; therefore, we conclude that the Q modification does not change the direction of the codon preference. Another possible function of Q is that it may reduce pairing with the A codon, which might occur if the wobble-G were unmodified. Loss of the Q modification and subsequent GA pairing is relevant with respect to the origin of several variant genetic codes in mitochondrial genomes (Yokobori et al. 2001; Sengupta et al. 2007).

One of the most surprising results from the codon usage analysis is that the U codon is often preferred in four-codon families when wobble-U tRNAs are the only ones (table 2) or the most frequent ones (table 3). This shows that there is an unexpectedly high efficiency of pairing of the UU combination, with a rate b_{UU} that is similar to b_{UA} and higher than b_{UG}. Nevertheless, when wobble-U tRNAs occur in two-codon A+G families, the UU pairing rate cannot be high because this would lead to high rates of mistranslation. Thus, the codon usage data show that wobble-U tRNAs behave differently in two- and four-codon families. Base modifications seem to be an important part of the explanation for this. In tRNAs for four-codon families, the U base at the wobble position is modified to 5-methoxyuridine (mo^5U) or uridine-5-oxyacetic acid (cmo^5U) in most bacteria (Curran 1998; Agris 2004, 2008; Jühling et al. 2009). We will denote this class of mutations as xo^5U. In tRNAs for two-codon families, the U base usually has a 5-methylaminomethyl modification and often a 2-thio modification as well. Examples occurring in bacteria are 5-methylaminomethyl uridine (mnm^5U), 5-carboxymethylaminomethyl uridine ($cmnm^5U$), 5-methylaminomethyl 2-thiouridine (mnm^5s^2U), and 5-carboxymethylaminomethyl 2-thiouridine ($cmnm^5s^2U$). We denote this class as xm^5U.

Several cases of xo^5U modifications have been studied experimentally (Lustig et al. 1993; Kothe and Rodnina

2007; Näsvall et al. 2007; Weixlbaumer et al. 2007; Vendeix et al. 2008). The main conclusion from these examples is that the xo^{5}U modification enhances the ability of the tRNA to pair with all four codons. Nevertheless, in *Mycoplasma*, the wobble-U base is unmodified in tRNAs for four-codon families (Andachi et al. 1989), and the same is also true in animal mitochondria (Yokobori et al. 2001; Jia and Higgs 2008). This shows that an unmodified U can pair with all four codons, at least to some extent.

In contrast, experiments show that the xm^{5}U modification restricts pairing to only A and G codons, as is necessary to prevent mistranslation in two-codon families. Experiments on tRNA^{Lys} in *E. coli* (Hagervall et al. 1998) showed that the xm^{5}U modification reduces the rate at which this tRNA misreads Asn codons. However, Ashraf et al. (1999) and Yarian et al. (2002) found that the 5-methylaminomethyl and 2-thio modifications were essential to allow binding of the tRNA^{Lys} to its own A and G codons, so these modifications seem to help with correct codon pairing as well as in elimination of mispairing. For tRNA^{Glu}, these modifications also alter the relative affinities of pairing to own A and G codons (Krüger et al. 1998). We are not aware of any examples of wobble-U tRNAs for A+G families in bacteria that lack the xm^{5}U modification. It appears to be significant that the xm^{5}U modifications occur even in *Mycoplasma* (Andachi et al. 1989), whereas the xo^{5}U and Q modifications discussed above have both been lost.

One reason why the two types of modification function in different ways is because they have different effects on the 3D configurations of the ribose. Yokoyama et al. (1985) showed that the xo^{5}U modification increases the stability of the C2′-endo form relative to the C3′-endo form; hence, it makes UU pairing easier. On the other hand, the xm^{5}U modification makes the C2′-endo form less stable; hence, it prevents UU pairing. This is consistent with our observation from the codon usage data that UU pairing is fast, and U codons are often preferred when the xo^{5}U modification is present. Yokoyama et al. (1985) did not give a structure for the UC pair, but we know that this must occur, as there are many cases where this is the only tRNA (table 2). Experiments also demonstrate formation of the UC pair (Näsvall et al. 2007; Kothe and Rodnina 2007). Table 2 shows that the C codon is the least preferred in cases where only wobble-U tRNAs are present; therefore, we conclude that UC pairing is possible but is still weak, even in the presence of the xo^{5}U modification.

It will be seen from the results tables that wobble-C tRNAs are much less frequent than wobble-U and G tRNAs. It is presumed that C only pairs with G codons. This is essential for Met and Trp, which have only one codon, but in two- and four-codon families, wobble-C tRNAs are usually an optional extra because the wobble-U tRNA can pair fairly well with the G codon. In table 5, N_A denotes the number of genes with A at the wobble position, but the A is almost always modified to I in the mature tRNA. I is usually thought to pair with U and C and to a lesser extent with A but not with G (Curran 1998). For this reason, the wobble-I tRNAs usually occur in combination with wobble-C (or sometimes wobble-U) tRNAs that translate the G codon (table 5). The U codon is preferred in almost all these cases. Also, the A codon decreases in frequency in high-expression genes, which is consistent with there being weak interaction between I and A. It is not clear why an unmodified A is rare at the wobble position. Boren et al. (1993) showed that a tRNA^{Gly} with a wobble position that was changed to A was able to read all four codons. They speculated that wobble-A bases are generally avoided because A would be indiscriminate. This argument makes sense in split codon boxes but does not apply in four-codon families. Also, if A pairs well with four codons, there must be a reason why it is necessary to modify it to I. Presumably, the I modification must speedup translation of at least the U and C codons relative to the unmodified A, but the reason for avoidance of the unmodified A still seems rather unclear.

We have focused on modifications at the wobble position (tRNA position 34) because these have a direct interaction with the third codon position. However, base modifications in other positions are also significant in terms of translation and possibly codon usage. In particular, modifications often occur at position 37 (the base that follows the anticodon). Removal of these modifications has been shown to have a detrimental effect on either speed or accuracy of translation in some cases (Yarian et al. 2002; Agris 2004, 2008). Base modifications at positions 34 and 37 have also been found to be important for proper translocation of a tRNA from the A site to the P site in the ribosome (Phelps et al. 2004) and in maintenance of the correct reading frame (Urbonavicius et al. 2001).

Finally in this section, we note that tRNAs with the same wobble position base may differ in their ability to pair with alternative codons because of structural differences that have nothing to do with base modifications. Lehmann and Libchaber (2008) argued that a single tRNA can pair with four codons when codon–anticodon interactions are strong but not when they are weak. For this reason, weakly interacting codon boxes can be split between two amino acids, whereas strongly interacting boxes must remain as four-codon families. This argument is interesting in the context of the evolution of the genetic code (Lehmann and Libchaber 2008; Higgs 2009), and in the present context, it provides another explanation of why the wobble-U tRNAs can translate all four codons in four-codon families but not in two-codon families.

Discussion and Conclusions

The model of translational selection that we have used to interpret the codon usage data above is based on the assumption that translational speed is the key factor. However, it is also possible that translational accuracy plays a role, that is, the preferred codons are those for which the error rate is smallest rather than those that are most rapidly translated. Here, we will discuss both possible causes of selection, and we will argue that there are important aspects of these data that can be most easily explained

in terms of selection for speed, although it is quite possible that selection for accuracy is operating at the same time.

First, we note that there is experimental evidence that there is significant difference in translation speeds between synonymous codons. Curran and Yarus (1989) and Sorensen and Pedersen (1991) found that codons that were preferred in *E. coli* according to codon usage data were indeed translated faster. It is also known that insertion of blocks of slow codons into a sequence has a significant effect on protein production rate (Mitarai et al. 2008) and that these effects can be well described by a model that considers the position of the fast and slow codons along the messenger RNA. More recent experiments have attempted to measure the rates of the different steps involved in the translation cycle for each codon (Rodnina and Wintermeyer 2001; Blanchard et al. 2004; Daviter et al. 2006; Gromadski et al. 2006; Pavlov et al. 2009), but it is not yet clear exactly which of the underlying steps leads to variation in the effective rate. Also, different groups use different kinetic schemes, as pointed out by Ninio (2006), so there is not yet complete agreement on what the underlying steps are. We wish to emphasize that selection on codon bias shows up in the simplest possible case where there is only one tRNA that pairs with two synonymous codons. Thus, if it is speed that is under selection, there must be a difference in the rate at which the ribosome recognizes and processes the tRNA that depends on the details of the codon–anticodon interaction.

Accuracy-based arguments assume that mistranslation has a cost because some mistranslated proteins are nonfunctional or that they misfold more often (Drummond and Wilke 2008). Selection for accuracy can explain observed differences in codon usage between conserved and variable sites (Akashi 1994; Stoletzki and Eyre-Walker 2006) seen in some species, which we would not expect from selection for speed alone. In general, the probability of mistranslating a codon is $m_i/(m_i + r_i)$, where r_i is the rate of correct translation and m_i is the rate of incorrect translation. If accuracy is the key factor, then in order to understand this fully, it will be necessary to measure the mispairing rates of each tRNA with all the non-cognate codons as well as the correct pairing rates with the cognate codons. The relative accuracy of synonymous codons has been measured in a few cases (Precup and Parker 1987; Kramer and Farabaugh 2007), although there is no systematic study of relative accuracy that covers all codons in a given organism. An interesting special case related to selection against inaccurate codons is the elimination of ambiguously translated codons during periods of codon reassignment, as has been observed with *Candida* (Butler et al. 2009).

If a tRNA is rare, then the rate of translation of its cognate codons will be slow, but it is also likely to be inaccurate because the ratio m_i/r_i will tend to be larger when r_i is smaller. For example, Kramer and Farabaugh (2007) showed that there is a relatively high rate of mistranslation of the AGR Arg codons in *E. coli*, for which the cognate tRNA is rare. This might explain selection between the CGN Arg codons and the AGR codons, but it is less clear that this argument can be used to explain selection between codons translated by the same tRNA, such as any of the U+C or A+G pairs considered above.

Despite these caveats regarding the possible relevance of accuracy as well as speed, there are two aspects of the data that point to the fundamental importance of translational speed. First, it has been shown by Rocha (2004) and in our previous paper (Higgs and Ran 2008) that codon bias is higher in bacteria with faster growth rates. This is a natural expectation if speed is important—bacteria living in a niche where rapid cell division is advantageous need to adapt to optimize their rate of protein synthesis. It is difficult to see why this correlation should occur if selection were solely due to accuracy. Second, it is found that there are more duplicate tRNA genes in bacteria that are rapidly multiplying and in species where the codon bias is strong (Rocha 2004; Higgs and Ran 2008). Duplication of a tRNA leads to higher tRNA concentration and hence increases translational speed. Our theory predicts in which circumstances duplications are favored by selection for translational speed (Higgs and Ran 2008). In contrast, it is not clear that selection for accuracy alone will favor gene duplications. Duplicating one tRNA should increase the rate of translation of its cognate codons and hence also increase their accuracy. However, it will also increase the rate at which this tRNA mispairs with non-cognate codons. It is not clear which of these is more important. Furthermore, if we make a general duplication of all genes, this will increase all the correct pairing rates and mispairing rates proportionately, so there should be no change in accuracy but a large increase in speed of all codons.

We emphasize that the above arguments apply only to bacteria, and it may well be that speed is less relevant in multicellular organisms than in bacteria. Several examples where arguments for translational accuracy have been made are multicellular eukaryotes (Drummond and Wilke 2008). However, eukaryotes tend to have large numbers of tRNA genes, and the number of copies of each type of tRNA is correlated with the codon frequencies, as has been shown, for example, in *Caenorhabditis elegans* (Duret 2000) and humans (Lavner and Kotlar 2005). Therefore, tRNA copy number and codon frequencies are also coevolving in eukaryotes, and this means that there is still a fundamental role for speed and efficiency even in multicellular organisms.

Our interpretation of the observed variation in the strength of codon bias among species is that the selection strength (s_0 in eq. 7) varies as a consequence of the difference in growth rates. However, the effective population size, N_e, also influences codon frequencies (eqs. 1, 2, 9). It is therefore possible that the species with high codon bias correspond to those with the highest N_e. We do not have estimates of N_e for all species in this data set but Lynch and Conery (2003) determined the product $N_e u$ for several species (where u is the mutation rate), including nine of the bacteria in our data set. We found no correlation between the strength of codon bias in our data

and $N_e u$ for these species, whereas codon bias was found to be correlated with growth rate in the same nine species (as in the full data set). Therefore, variation of N_e does not seem to be an important confounding factor in our results.

In conclusion, the idea of relating codon usage to tRNA concentrations dates back to some of the earliest papers that detected codon usage bias. However, very few studies have considered the nature of the anticodon–codon interaction and the modified bases in the anticodon. Previous papers have tended to assume a one-to-one relationship between codons and tRNAs. Our theory is unique in that it builds in the essential feature that each tRNA interacts with more than one codon and that often there is more than one tRNA that interacts with the same codon. This has allowed us to relate observations of codon usage more closely to tRNA structure and function and to experimental measurements on translation kinetics and the effects of modified bases on translation.

It is also a rather old idea that the codons that are used frequently in highly expressed genes such as ribosomal proteins are the ones that are preferred by translational selection. This is incorporated into measures of codon bias, such as codon adaptation index. However, our theory goes beyond this by making a careful distinction between codons that are frequent due to mutational bias and those that are frequent due to translational selection. We have shown that, in a majority of cases, coevolution of tRNAs and codon usage leads to states in which the direction of translational selection is the same as that of the mutational bias. However, there is a substantial minority of cases where selection prefers a different codon from the one that is most frequent under mutation. Our theory explains why these situations are sometimes stable. Therefore, we should not automatically assume that the highest frequency codons are the most preferred by translational selection.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada. We thank Paul Sharp for sharing his codon usage data and Joshua Plotkin, Dick Morton, and Herb Schellhorn for helpful advice and criticism.

References

Agris PF. 2004. Survey and summary decoding the genome: a modified view. *Nucleic Acids Res.* 32:223–238.

Agris PF. 2008. Bringing order to translation: the contributions of transfer RNA anticodon-domain modifications. *EMBO Rep.* 9:629–635.

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 164:1291–1303.

Andachi Y, Yamao F, Muto A, Osawa S. 1989. Codon recognition patterns as deduced from sequences of the complete set of transfer RNAs species in *Mycoplasma capricolum*. *J Mol Biol.* 209:37–54.

Ardell DH, Kirsebom LA. 2005. The genomic pattern of tDNA operon expression in E. coli. *PLoS Comp. Biol.* 1(1):e12.

Ashraf SS, Sochacka E, Cain R, Guenther R, Malkiewicz A, Agris PF. 1999. Single atom modification (O→S) of tRNA confers ribosome binding. *RNA* 5:188–194.

Blanchard SC, Gonzalez RL Jr., Kim HD, Chu S, Puglisi JD. 2004. tRNA selection and kinetic proofreading in translation. *Nature Struct Mol Biol.* 11:1008–1014.

Boren T, Elias P, Samulelsson T, Claesson C, Barciszewska M, Gehrke CW, Kuo KC, Lustig F. 1993. Undescrimitating codon reading with adenosine in the wobble position. *J Mol Biol.* 230:739–749.

Bulmer M. 1991. The selection-mutation-drift theory of synonymous codon usage. *Genetics* 129:897–907.

Butler G, Rasmussen MD, Lin MF, Santos MAS, et al. (51 co-authors). 2009. Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes. *Nature* 459:657–662.

Curran JF. 1998. Modified nucleosides in translation. In: Grosjean H, Benne R, editors. Modification and editing of RNA. Washington (DC): ASM Press. p. 493–516.

Curran JF, Yarus M. 1989. Rates of aminoacyl-tRNA selection at 29 sense codons *in vivo*. *J Mol Biol.* 209:65–77.

Daviter T, Gromadski KB, Rodnina MV. 2006. The ribosome's response to codon-anticodon mismatches. *Biochimie* 88:1001–1011.

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.* 260:649–663.

Drummond DA, Wilke CO. 2008. Mistranslation-induced protein misfolding as a dominant constraint on codong-sequence evolution. *Cell* 134:341–342.

Duret L. 2000. tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes. *Trends Genet.* 16:287–289.

Elf J, Nilsson D, Tenson T, Ehrenberg M. 2003. Selective charging of tRNA isoacceptors explains patterns of codon usage. *Science* 300:1718–1722.

Gromadski KB, Daviter T, Rodnina MV. 2006. A uniform response to mismatches in codon-anticodon complexes ensures ribosomal fidelity. *Mol Cell.* 21:369–377.

Grosjean H, de Crécy-Lagard V, Marck C. 2010. Deciphering synonymous codons in the three domains of life: co-evolution with specific tRNA modification enzymes. *FEBS Lett.* 584:252–264.

Hagervall TG, Pomerantz SC, McCloskey JA. 1998. Reduced misreading of asparagine codons by *Escherichia coli* $tRNA^{Lys}$with hypomodified derivatives of 5-methylaminomethyl 2-thiouridine in the wobble position. *J Mol Biol.* 284:33–42.

Higgs PG. 2009. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. *Biol Direct.* 4:16.

Higgs PG, Ran W. 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol.* 25:2279–2291.

Ikemura T. 1981. Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes. *J Mol Biol.* 146:1–21.

Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. *Mol Biol Evol.* 2:13–34.

Jia WL, Higgs PG. 2008. Codon usage in mitochondrial genomes: distinguishing context-dependent mutation from translational selection. *Mol Biol Evol.* 25:339–351.

Jühling F, Mörl M, Hartmann RK, Sprinzl M, Stadler PF, Pütz J. 2009. tRNAdb2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.* 37:D159–D169.

Kanaya S, Yamada Y, Kudo Y, Ikemura T. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of Bacillus subtilis tRNAs. *Gene* 238:143–155.

Kothe U, Rodnina MV. 2007. Codon reading by tRNAAla with modified uridine in the wobble position. *Mol Cell* 25:167–174.

Kramer E, Farabaugh PJ. 2007. *The frequency of translational misreading errors in E. coli is largely determined by tRNA competition*. *RNA* 13:87–96.

Krüger MK, Pedersen S, Hagervall TG, Sorensen MA. 1998. The modification of the wobble base of tRNAGlu modulates the translation rate of glutamic acid codons in vivo,. *J Mol Biol*. 284:621–631.

Lavner Y, Kotlar D. 2005. Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene* 345:127–138.

Lehmann J, Libchaber A. 2008. Degeneracy of the genetic code and stability of the base pair at the second position of the anticodon. *RNA* 14:1264–1269.

Li WH. 1987. Models of nearly neutral mutation with particular implications for nonrandom usage of synonymous codons. *J Mol Evol*. 24:337–345.

Lustig F, Borén T, Claesson C, Simonsson C, Barciszewska M, Lagerkvist U. 1993. The nucleotide in position 32 of the tRNA anticodon loop determines ability of anticodon UCC to discriminate among glycine codons,. *Proc Natl Acad Sci U S A*. 90:3343–3347.

Lynch M, Conery JS. 2003. The origins of genome complexity. *Science* 302:1401–1404.

Meier F, Suter B, Grosjean H, Keith G, Kubli E. 1985. Queuosine modification of the wobble base in $tRNA^{His}$ influences *in vivo* decoding properties. *EMBO J*. 4:823–827.

Mitarai N, Sneppen K, Pedersen S. 2008. Ribosome collisions and translational efficiency: optimization by codon usage and mRNA destabilization. *J Mol Biol*. 382:236–245.

Morris RC, Brown KG, Elliott MS. 1999. The effect of queuosine on tRNA structure and function. *J Biomol Struct Dyn*. 16:757–774.

Morris RC, Elliott MS. 2001. Queuosine modification of tRNA: a case for convergent evolution. *Mol Genet Metab*. 74:147–159.

Moura G, Pinheiro M, Arrais J, Gomes AC, Carreto L, Freitas A, Oliveira JL, Santos MAS. 2007. Large scale comparative codon-pair context analysis unveils general rules that fine-tune evolution of mRNA primary structure. *PLoS One*. 2(9):e847.

Näsvall SJ, Chen P, Björk R. 2007. The wobble hypothesis revisited: uridine-5-oxyacetic acid is critical for reading of G-ending codons. *RNA* 13:2151–2164.

Ninio J. 2006. Multiple stages in codon-anticodon recognition: double-trigger mechanisms and geometric constraints. *Biochimie* 88:963–992.

Pavlov MY, Watts RE, Tan Z, Cornish VW, Ehrenberg M, Forster AC. 2009. Slow peptide bond formation by proline and other N-alkylamino acids in translation. *Proc Natl Acad Sci U S A*. 106:50–54.

Phelps SS, Malkiewicz A, Agris PF, Joseph S. 2004. Modified nucleotides in $tRNA^{Lys}$ and $tRBA^{Val}$ are important for translocation. *J Mol Biol*. 338:439–444.

Precup J, Parker J. 1987. Missense misreading of asparagine codons as a function of codon identity and context. *J Biol Chem*. 23:11351–11355.

Rocha EPC. 2004. Codon usage bias from the tRNA's point of view: redundancy, specialization, and efficient decoding for translational optimization. *Genome Res*. 14:2279–2286.

Rodnina MV, Wintermeyer W. 2001. Fidelity of aminoacyl-tRNA selection on the ribosome: kinetic and structural mechanisms. *Annu Rev Biochem*. 70:415–435.

Romier C, Ficner R, Suck D. 1998. Structural basis of base exchange by tRNA-guanine transglycosylases. In: Grosjean H, Benne R, editors. Modification and editing of RNA. Washington (DC): ASM Press. p. 493–516.

Sengupta S, Yang X, Higgs PG. 2007. The mechanisms of codon reassignments in mitochondrial genetic codes. *J Mol Evol*. 64: 662–688.

Sharp PM, Li WH. 1986. Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. *Nucleic Acids Res*. 14:7737–7749.

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE. 2005. Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res*. 33:1141–1153.

Shields DC. 1990. Switches in species-species codon preferences: the influence of mutation bias. *J Mol Evol*. 31:71–80.

Sorensen MA, Pedersen S. 1991. Absolute *in vivo* translation rates of individual codons in *Escherichia coli*. *J Mol Biol*. 222:265–280.

Stoletzki N, Eyre-Walker A. 2006. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol*. 24:374–381.

Urbonavicius J, Qian Q, Durand JMB, Hagervall TG, Björk GR. 2001. Improvement of reading frame maintenance is a common function for several tRNA modifications. *EMBO J*. 20: 4863–4873.

Vendeix FAB, Dziergowska A, Gustilo EM, Graham WD, Sproat B, Malkiewicz A, Agris PF. 2008. Wobble-position modifications contribute order to tRNA's anticodon for ribosome-mediated codon binding. *Biochemistry* 47:6117–6129.

Weixlbaumer A, Murphy FV, Dziergowska A, Malkiewicz A, Vendeix FAP, Agris PF, Ramakrishnan V. 2007. Mechanism for expanding the decoding capacity of transfer RNAs by modification of uridines. *Nat Struct Mol Biol*. 14:498–502.

Yarian C, Townsend H, Czestkowski W, Sochacka E, Malkiewicz AJ, Guenther R, Miskiewicz A, Agris PF. 2002. Accurate translation of the genetic code depends on tRNA modified nucleosides. *J Biol Chem*. 277:16391–16395.

Yokobori S, Suzuki T, Watanabe K. 2001. Genetic code variantions in mitochondria: tRNA as a major determinant of genetic code plasticity. *J Mol Evol*. 53:314–326.

Yokoyama S, Watanabe T, Murao K, Ishikura H, Yamaizumi Z, Nishimura S, Miyazawa T. 1985. Molecular mechanism of codon recognition by tRNA species with modified uridine in the first position of the anticodon. *Proc Natl Acad Sci U S A*. 82: 4905–4909.