Codon Usage in Mitochondrial Genomes: Distinguishing Context-Dependent Mutation from Translational Selection

Wenli Jia and Paul G. Higgs

Department of Physics and Astronomy, McMaster University, Hamilton, Ontario L8S 4M1, Canada

We analyze the frequencies of synonymous codons in animal mitochondrial genomes, focusing particularly on mammals and fish. The frequencies of bases at 4-fold degenerate sites are found to be strongly influenced by context-dependent mutation, which causes correlations between pairs of neighboring bases. There is a pattern of excess of certain dinucleotides and deficit of others that is consistent across large numbers of species, despite the wide variation of single-nucleotide frequencies among species. In many bacteria, translational selection is an important influence on codon usage. In order to test whether translational selection also plays a role in mitochondria, we need to control for context-dependent mutation. Selection for translational accuracy can be detected by comparison of codon usage in conserved and variable sites in the same genes. We give a test of this type that works in the presence of context-dependent mutation. There is very little evidence for translational accuracy selection in the mitochondrial genes considered here. Selection for translational efficiency might lead to preference for codons that match the limited repertoire of anticodons on the mitochondrial tRNAs. This is difficult to detect because the effect would usually be in the same direction in comparable to codon families and so would not cause an observable difference in codon usage between families. Several lines of evidence suggest that this type of selection is weak in most cases. However, we found several cases where unusual bases occur at the wobble position of the tRNA, and in these cases, some evidence for selection on codon usage was found. We discuss the way that these unusual cases are associated with codon reassignments in the mitochondrial genetic code.

Introduction

For several years, we have been assembling a relational database of complete animal mitochondrial genomes known as OGRe (Jameson et al. 2003), meaning Organellar Genome Retrieval. The current version has over 1,000 species and is available online (http://ogre.mcmaster.ca). This database has provided a foundation for several projects on mitochondrial genome evolution, including studies of mammalian phylogenetics (Hudelot et al. 2003; Gibson et al. 2005), tRNA gene evolution (Higgs et al. 2003), amino acid frequency variation among species (Urbina et al. 2006), genome rearrangements (Xu et al. 2006), and changes in the mitochondrial genetic code (Sengupta et al. 2007). In this paper, we investigate the factors that influence codon usage in mitochondria.

In mitochondrial genomes, mutation pressure causes a wide variation of base frequencies among species at both synonymous and nonsysnonymous sites (Foster et al. 1997; Singer and Hickey 2000; Urbina et al. 2006). Due to the asymmetric nature of the process of replication of mitochondrial genomes, the mutation processes on the two strands of the genome are not equivalent. This leads to differences in base frequencies between strands and means that the frequencies of G and C are not equal and the frequencies of A and U are not equal (Reyes et al. 1998; Bielawski and Gold 2002; Faith and Pollock 2003; Urbina et al. 2006). We expect, therefore, that mutation will have a large influence on codon usage in mitochondria. If there were no selection on 4-fold degenerate (FFD) sites, and if mutations were independent single-site events, the base frequencies at FFD sites would converge to the stationary frequencies of the mutation process. The frequencies would then be the same in each 4-codon family. However, we show here that this is not true because the mutation process is context dependent.

Context-dependent mutation means that the rate of mutation from any one base to any other is influenced by bases at the neighboring sites. Context-dependent rates have been measured, for example, in primate pseudogenes (Blake et al. 1992), chloroplast DNA (Morton 2003), and the maize genome (Morton et al. 2006). A signature of context-dependent mutation is that the frequencies of dinucleotides and trinucleotides differ from their expected frequency if there were no correlation between neighboring bases. This kind of signature has been detected in many different genomes (Karlin and Mrazek 1997; Shioiri and Takahata 2001) and studied specifically in humans (Karlin and Mrazek 1996), *Drosophila* (Antezana and Kreitman 1999; Fedorov et al. 2002), and *Arabidopsis* (Morton and Wright 2007). We will show below that context-dependent mutation also has a major influence on codon usage in mitochondrial genes.

Codon usage can also be influenced by translational selection. The usual type of selection is for translational efficiency, that is, an organism prefers to use codons that are more rapidly translated in order to reduce the time and effort spent on translation. This has been shown in many organisms (Sharp et al. 1988; Akashi 2003; dos Reis et al. 2003) and is most important for organisms with rapid growth rate (Sharp et al. 2005) because the time saving is more significant for such species. Within any genome, translational efficiency is more important in highly expressed genes because a large fraction of the translational effort of the cell can be spent on making large number of copies of a relatively small number of proteins. The presence of translational selection is usually demonstrated by determining the codons that are preferred in highly expressed genes and showing that preferred codons have higher frequencies in highly expressed genes than weakly expressed ones. Unfortunately, the corresponding test in mitochondrial genomes is not possible because expression levels have not

Key words: mitochondrial genomes, codon usage, context-dependent mutation, translational selection, genetic code, codon reassignment.

E-mail: higgsp@mcmaster.ca.

Mol. Biol. Evol. 25(2):339–351. 2008
doi:10.1093/molbev/msm259
Advance Access publication November 28, 2007

Table 1
Codon Usage Table for the 12 Genes on the Plus Strand of the Human mitochondrial Genome

Aa	Codon	Usage	Aa	Codon	Usage	Aa	Codon	Usage	Aa	Codon	Usage
F	UUU	69	S	UCU	29	Y	UAU	35	C	UGU	5
F	UUC	139	S	UCC	99	Y	UAC	89	C	UGC	17
L	UUA	65	S	UCA	81	stop	UAA	4	W	UGA	90
L	UUG	11	S	UCG	7	stop	UAG	3	W	UGG	9
L	CUU	65	P	CCU	37	H	CAU	18	R	CGU	6
L	CUC	167	P	CCC	119	H	CAC	79	R	CGC	26
L	CUA	276	P	CCA	52	Q	CAA	82	R	CGA	28
L	CUG	42	P	CCG	7	Q	CAG	8	R	CGG	0
I	AUU	112	T	ACU	50	N	AAU	29	S	AGU	11
I	AUC	196	T	ACC	155	N	AAC	131	S	AGC	37
M	AUA	165	T	ACA	132	K	AAA	84	stop	AGA	1
M	AUG	32	T	ACG	10	K	AAG	9	stop	AGG	0
V	GUU	22	A	GCU	39	D	GAU	12	G	GGU	16
V	GUC	45	A	GCC	123	D	GAC	51	G	GGC	87
V	GUA	61	A	GCA	79	E	GAA	63	G	GGA	61
V	GUG	8	A	GCG	5	E	GAG	15	G	GGG	19

been measured, and it is presumed that all genes on the mitochondrial genome are essential genes that would have similar expression levels.

When there is more than one type of tRNA for the same amino acid, selection for translational efficiency causes the preferred codons to match the tRNAs that are most abundant in the cell (Ikemura 1981, 1985; Percudani et al. 1997; Duret 2000). This type of selection cannot occur in mitochondria because there is only one tRNA gene for each codon family on the mitochondrial genome. Nevertheless, it is still possible for translational selection to occur between codons translated by the same tRNA, if one codon interacts more effectively than another with the anticodon. It is also possible for selection to act on translational accuracy rather than efficiency, that is, a codon can be preferred because it has a lower probability of mistranslation, rather than because it is more rapidly translated. This has been demonstrated by showing that preferred codons are more frequent at sites where the amino acid is conserved during evolution (Akashi 1994; Stoletzki and Eyre-Walker 2007). It is argued that sites that are critical for protein function have conserved amino acids and that translational accuracy is most important at the critical sites; hence preferred codons should be used at sites that are evolutionary conserved. In contrast, selection for translational efficiency should be equally important at all sites on the gene. The comparison of conserved and variable sites can be done within any gene and does not require knowledge of the expression level. We carry out tests of this type on mitochondrial genes below.

The first object of this paper is to show that context-dependent mutation has a large effect on codon usage in mitochondria. The second object is to test for the possible presence of translational selection that might act on top of the mutational effects. Although many previous methods have been proposed to detect translational selection, none of them seems appropriate for mitochondrial genomes. Calculation of the frequency of optimal codons (Ikemura 1981) or the codon adaptation index (Sharp and Li 1987) can only be done if we know which codons are optimal. In mitochondria, we do not know which codons are optimal or even if there are any optimal codons at all. Another standard method is the measurement of the effective number of codons, N_c (Wright 1990). This does not require prior knowledge of the optimal codons. However, a gene might have a low N_c because of mutational bias or because of translational selection. Measuring N_c would not be useful to distinguish the factors that affect codon usage. Therefore, in the second half of this paper, we develop ways to test for translational selection that are applicable in mitochondria without prior knowledge of expression levels of genes or optimal codons and that control effectively for the complex nature of the mutation process that is occurring in mitochondrial genomes.

Statistical Tests for Context-Dependent Mutation

The codon usage numbers for the 12 genes on the plus strand of the human mitochondrial genome are listed in table 1. Similar tables for all mitochondrial genomes can be downloaded from OGRe. We focus on the 8-codon families with FFD third positions. Base frequencies at the FFD sites will be strongly influenced by the mutational process because FFD sites are not subject to selection at the protein level. The simplest assumption is that mutations are independent single-site events and that there is no selection at FFD sites. Relative frequencies of the 4 codons in each 4-codon family should then be the same. We now show that this is not true.

Let n(XYZ) be the number of occurrences of codon XYZ. Let n(YZ) be the total number of occurrences of each doublet YZ at positions 2 and 3, counting only codons where the third position is FFD, that is, $n(\mathrm{UZ}) = n(\mathrm{CUZ}) + n(\mathrm{GUZ})$; $n(\mathrm{CZ}) = n(\mathrm{UCZ}) + n(\mathrm{CCZ}) + n(\mathrm{ACZ}) + n(\mathrm{GCZ})$; $n(\mathrm{GZ}) = n(\mathrm{CGZ}) + n(\mathrm{GGZ})$. The values n(YZ) form a 3×4 contingency table. The null hypothesis that Z is independent of Y can be tested with a χ^2 test with 6 degrees of freedom (DOF). From the human data in table 1 we obtain $\chi^2 = 83.57$ ($P < 0.001$), that is, the frequencies of third-position bases are definitely not independent of the second position. We can also look for more detailed differences in the third base frequencies between the individual 4-codon families. The two families CUN and GUN form

Table 2
Observed and Expected Number of Species in Each Significance Category in χ^2 Tests of Codon Frequencies in 4-Codon Families

	Not Significant ($P > 0.05$)	Significant ($0.001 < P \leq 0.05$)	Highly Significant ($P \leq 0.001$)
Fish (214 species)			
Expected number	203.3	10.5	0.2
UN/CN/GN	0	0	214
CUN/GUN	187	25	2
CGN/GGN	53	101	60
UCN/CCN/ACN/GCN	58	75	81
Mammals (148 species)			
Expected number	140.6	7.3	0.1
UN/CN/GN	0	0	148
CUN/GUN	115	31	2
CGN/GGN	75	58	15
UCN/CCN/ACN/GCN	28	76	44

a 2×4 table with 3 DOF, from which we obtain $\chi^2 = 2.986$ ($P < 0.5$). The 4 families UCN, CCN, ACN, and GCN form a 4×4 table with 9 DOF, from which we obtain $\chi^2 = 14.167$ ($P < 0.1$). Similarly, comparison of CGN and GGN gives $\chi^2 = 8.674$ ($P < 0.05$). These results suggest that there are some differences within the codon families that remain even after the effect of the second base has been accounted for.

Table 2 presents the results of similar χ^2 tests applied to a set of 148 mammals and 214 ray-finned fish (actinopterygii). These form 2 comparable but independent monophyletic groups. All species were used that were available in OGRe at the time the project was begun, although the number of species available has increased considerably since then. Each result was classed as not significant if $P > 0.05$, significant if $0.001 < P \leq 0.05$, and highly significant if $P \leq 0.001$. The probabilities of falling into each of these ranges according to the null hypothesis are 0.95, 0.049, and 0.001, respectively. The expected number of species falling into each category is the total number of species multiplied by these probabilities. Table 2 compares the expected numbers with the observed number of species in each significance category. For the UN/CN/GN tests, every species of mammal and fish analyzed falls in the highly significant category. This shows an extremely strong influence of the second position base on the FFD sites. For the CGN/GGN and UCN/CCN/ACN/GCN tests, there are far more species in the significant and highly significant categories than expected. There are also somewhat more species than expected in the significant category for the CUN/CGN tests, but the effect is weaker than for the other cases.

Context-dependent mutational processes create correlations between neighboring bases. These can be measured by the ratio $R(\mathrm{YZ}) = f_{\mathrm{YZ}}/(f_{\mathrm{Y}} f_{\mathrm{Z}})$, where f_{YZ} is the frequency of dinucleotide YZ in the second and third positions, again with the restriction that only codons in which the third position is FFD are included, and f_{Y} and f_{Z} are the frequencies of the individual bases in second and FFD third positions. Similarly, the ratio $R(\mathrm{ZX}) = f_{\mathrm{ZX}}/(f_{\mathrm{Z}} f_{\mathrm{X}})$ measures correlations between the FFD base Z and the base X at the first position of the following codon. The values of these ratios are listed in table 3. These ratios are far from 1, which indicates nonindependence of neighboring sites. The largest ratio is R(GG) in both data sets and both tables, whereas the smallest is R(CG) in all cases. The frequency ratios for mammals and fish are correlated with one another: Pearson correlation coefficient $r = 0.90$ for the YZ ratios and $r = 0.89$ for the ZX ratios. There is also a correlation between the YZ and ZX ratios (using only the 12 dinucleotides for which R(YZ) can be measured). For the fish data from the top and bottom sections of table 3, $r = 0.74$, and for the mammals, $r = 0.57$. These results suggest that there are context-dependent mutational effects influencing dinucleotide positions in a similar way in both mammals and fish and acting in a similar way in both reading frames.

Relative dinucleotide frequencies in vertebrate mitochondrial genomes are also given by Shioiri and Takahata (2001). They consider all pairs, irrespective of reading frame or whether the sequence is coding or noncoding. They also observe GG is most overrepresented and CG is most underrepresented. The next highest and lowest in their data are CC and GU, respectively. This is similar to what we observe in the YZ frame but not in the ZX frame. We would expect that context-dependent effects on mutation act at the DNA level and are therefore independent of the frame. However, the observed dinucleotide frequencies also depend on selection at nonsynonymous sites, which acts differently at different codon positions. As noted above, there is a reasonable correlation between the frequency ratios in the YZ and ZX frame, but there is no reason why they have to be exactly the same.

It is useful to look at the conditional probability $P(\mathrm{Z}|\mathrm{Y}) = f_{\mathrm{YZ}}/f_{\mathrm{Y}}$ that the FFD base is Z given that the second position base is Y. As an example, figure 1 shows P(U|Y) as a function of f_{U} for all the fish species. There is a big variation in f_{U} from 0.1 to 0.4 in the fish genomes. The conditional probabilities follow linear trends as a function of f_{U}. It can be seen that P(U|U) is consistently higher than f_{U}, whereas P(U|G) is consistently lower, and P(U|C) is slightly less than f_{U}. This agrees with table 3, in that $R(\mathrm{UU}) = 1.250$, $R(\mathrm{GU}) = 0.605$, and $R(\mathrm{CU}) = 0.939$. Figure 1 shows that the direction of the context-dependent mutational effects seems to be the same in all these species, that is, there is a consistent preference for certain dinucleotides, despite the variation of overall single-nucleotide frequencies. The graphs of the conditional probabilities for other base combinations also follow linear trends in both the mammals and the fish.

One process that is known to lead to context-dependent mutational rates is the "CpG effect" that occurs in vertebrate nuclear genomes. The C in this context tends to be methylated and undergoes a rapid deamination to T. This leads to a decrease in CG dinucleotide frequency and an increase in TG and CA. Good data are available in maize, where Morton et al. (2006) find that the rates of CG → TG and CG → CA are the highest 2 rates of dinucleotide changes. Karlin and Mrazek (1997) find that R(CG) is very low in a large range of eukaryotic genomes. They also point out that CG is low in mitochondrial genomes, but that this cannot be due to methylation. Specific mutational processes relevant to mitochondrial genomes are the deamination of C to U and A to hypoxanthine, which occur preferentially

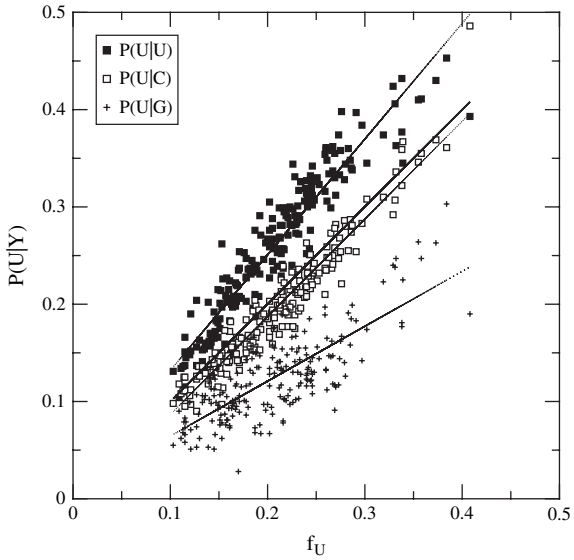

FIG. 1.—Conditional probabilities of the base at the FFD site is U, given that the second position base is U, C, or G. Data are from 214 fish genomes. The dashed lines are the linear regressions to the 3 data sets. The solid line is the equality line ($y = x$).

on the heavy strand (Reyes et al. 1998). This is important because it causes compositional differences between strands, although it is not clear whether this process is affected by the neighboring bases on the same strand. If a specific process, like the CpG effect is thought to be the dominant context-dependent effect in a genome, it is possible to use a mutation rate model specifically for this and to calculate the way this affects dinucleotide frequencies and sequence evolution over time (Arndt and Hwa 2005). In mitochondria, however, we do not have accurate information to propose a detailed rate model, and presumably there are many context-dependent effects, not just one single process. Hwang and Green (2004) have introduced a Bayesian model in which the rates of context-dependent effects can be estimated. Here, however, we make progress using statistical models that predict codon frequencies under a variety of assumptions about mutation and selection but do not use a specific model for the rate matrix.

Likelihood-Based Tests for Context-Dependent Mutation and Translational Selection

The simple tests in the previous section clearly demonstrate the major influence of context-dependent mutation; however, the possibility remains that translational selection also has an effect. We therefore wish to test for the presence of translational selection in a way that controls for context-dependent mutation. A first point to note is that YZ correlations are likely to have a much greater effect on codon usage than ZX correlations, even if context-dependent mutation rates are equivalent in all frames. This is because, for any given codon family, the second base is always the same, whereas the following first-position base is variable. The net influence of the following base is the sum of the different positive and negative influences that would arise from the 4 possible nucleotides at position X; therefore, the effect will be averaged out somewhat. For this reason, we will control for Y before any other factor. This will be done by comparing codons for amino acids with the same Y. The 4-codon families for Ser, Pro, Thr, and Ala (SPTA) all have Y = C. The codons for Tyr, His, Asn, and Asp (YHND) all have Y = A and have only U or C at third position. These 2 sets will be tested separately.

Table 2 already showed significant differences among the frequencies in SPTA codon families for many species; therefore, codon usage is apparently dependent on the amino acid. This could occur if there were selection for codon–anticodon matching that occurred in a different way for the tRNAs for different amino acids. However, it is possible that this is merely due to the influence of the nucleotide at position X. The frequencies of the 4 nucleotides at position X are determined by the amino acid sequences of the proteins. If amino acid sequences were random, the X frequencies would be the same for all 4-codon families; therefore, the influence of X would be the same in each case, and this would not explain the difference in codon frequencies among the 4 families. However, amino acid sequences are determined by the function of the protein. Therefore, it is possible that the frequencies of the nucleotide at position X following a Ser codon are different from those following a Pro codon, for example. In this case, the influence of X could cause an apparent difference in codon usage between codons for different amino acids.

As discussed in the introduction, selection for translational accuracy (but not efficiency) can lead to an increase in frequency of preferred codons at conserved sites in comparison to variable sites in the same gene. The test proposed by Akashi (1994) presumes that the preferred codons are already known by comparison of high- and low-expression genes. For the current mitochondrial sequences, we cannot determine the preferred codons in this way, and we have yet to establish whether there are any preferred codons at all. Nevertheless, in the method given below, it is possible to compare conserved and variable sites without knowing which codons, if any, are preferred.

Furthermore, we note that base frequencies at FFD sites vary systematically with the position along the genome. Base frequencies in each gene depend on D_{ssH}, the amount of time that the gene spends in a single-stranded state during the replication of the mitochondrial genome (Reyes et al. 1998; Faith and Pollock 2003; Urbina et al. 2006). We will also include models that account for this. Both the translational accuracy factor and the D_{ssH} factor might lead to spurious apparent dependence of the codon usage on the amino acid. For example, one amino acid might have a higher frequency than another in sites that are conserved or in genes with low D_{ssH}.

We carried out the following series of tests in order to investigate all these possible effects simultaneously. We chose 19 representative mammals, one from each order, and 21 representative fish, one from each major taxon defined in OGRe, making 40 species in all. These species are listed in the caption to table 4. Amino acid sequences of the 12 plus-strand genes were aligned for the mammals and fish separately. A site was classed as conserved if the same amino acid was present in all 19 mammals or all 21 fish.

Table 3
Dinucleotide Frequency Ratios at Second and Third Positions, *R*(YZ), and at Third and Following First Position, *R*(ZX)

		Mammals				Fish			
R(YZ)		Z				Z			
		U	C	A	G	U	C	A	G
Y	U	0.939	0.743	1.136	1.433	1.250	0.756	1.030	1.274
	C	1.101	1.163	0.906	0.552	0.939	1.205	0.938	0.554
	G	0.763	1.005	1.027	1.654	0.605	0.878	1.145	1.891
R(ZX)		X				X			
		U	C	A	G	U	C	A	G
Z	U	0.855	0.994	1.206	0.856	0.933	0.918	1.096	1.049
	C	1.082	1.363	0.945	0.546	1.162	1.371	0.849	0.609
	A	0.996	0.797	0.974	1.293	0.907	0.739	1.135	1.228
	G	1.115	0.873	0.776	1.369	0.911	0.839	0.758	1.499

Otherwise, it was classed as variable. To test for the variability of base frequencies along the genome, we divided genes into 2 groups according to D_{ssH}. Low D_{ssH} genes are COI, COII, ATP8, ATP6, COIII, and NDIII. High D_{ssH} genes are ND4L, ND4, ND1, ND5, ND2, and CytB. The positions of the genes on the genome are the same for these vertebrate species; therefore, the classification into the low and high D_{ssH} categories is consistent. In fact, base frequencies at FFD sites vary smoothly with D_{ssH} (Urbina et al. 2006), and there is no sharp dividing line between low and high D_{ssH}. However, the two category model below is the simplest model that includes the effect, and it allows a statistical test for the presence of variable base frequencies along the genome that can be done in an analogous way to the test for translational accuracy. In the same way, rates of amino acid substitutions vary in a continuous fashion among sites, and there is no sharp division between conserved and variable sites. Nevertheless, the 2-category model provides a useful test for translational accuracy.

For each of the 40 species, we counted the codon numbers, $n_{abdX}(Z)$. These are the number of times that a codon ending with base Z is used for amino acid a at site-type b in genome position d with following first-position base X. The index a takes 4 values—in one test, SPTA, and in a separate test, YHND. The index b takes 2 values—conserved and variable. The index d takes 2 values—low D_{ssH} and high D_{ssH}. The index X takes 4 values—U, C, A, and G. For the SPTA test, Z can be U, C, A, or G, and for the YHND test, Z can only be U or C.

We consider models that predict the frequencies $P_{abdX}(Z)$ of codons ending in Z for amino acid a at site-type b in genome position d with following base X. The log of the likelihood of observing the data for any one species is

$$\ln L = \sum_a \sum_b \sum_d \sum_X \sum_Z (n_{abdX}(Z) \ln P_{abdX}(Z)). \quad (1)$$

We will compare likelihoods of several models of this form. Akaike's Information Criterion (AIC) is a convenient statistical method of model selection that selects models with high likelihoods but penalizes those with unnecessarily large numbers of parameters. It is defined as

$$\mathrm{AIC} = 2(-\ln\hat{L} + K), \quad (2)$$

where K is the number of free parameters in the model, and the "hat" denotes that the ML value has been obtained by fitting the data. The preferred model is the one with the smallest AIC. The statistical theory of the AIC is described by Burnham and Anderson (1998) and an example of its use in molecular evolution is given by Higgs et al. (2007).

Model 0 is the simplest possible model for the codon frequencies. It is assumed that there is one overall set of base frequencies $P(Z)$, and that $P_{abdX}(Z) = P(Z)$ for all a, b, d, and X. The frequencies must satisfy the constraint that $\sum_Z P(Z) = 1$. Therefore, the number of free parameters in the model is $K = F - 1$, where F is the number of codons in the family ($F = 2$ for YHND and 4 for SPTA). The ML parameter values are

$$P(Z) = n(Z) \Big/ \left(\sum_{Z'} n(Z') \right), \text{ where}$$
$$n(Z) = \sum_a \sum_b \sum_d \sum_X n_{abdX}(Z). \quad (3)$$

Model A assumes that base frequencies depend on the amino acid a but not on the other quantities, that is, $P_{abdX}(Z) = P_a(Z)$ for all b, d, and X. The number of parameters is $K = 4(F - 1)$, as there are 4 amino acids in each of the test groups. The ML parameters are

$$P_a(Z) = n_a(Z) \Big/ \left(\sum_{Z'} n_a(Z') \right), \text{ where}$$
$$n_a(Z) = \sum_b \sum_d \sum_X n_{abdX}(Z). \quad (4)$$

In an analogous way, we can define 3 other single-factor models, B, D, and X, where the frequencies depend on only b, only d, and only X, respectively. The number of

Table 4
Results of the Model Selection Process Applied to 40 Representative Species of Mammals and Fish[a]

	ΔAIC for *Homo sapiens*		Average ΔAIC		Number of Species for Which ΔAIC < 0		Number of Species for Which This Model is the Best of 0 and Single-Factor Models		Number of Species for Which This Model is the Best of All Models		Number of Species for Which This Factor is Included in The Best Model	
Model	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND	SPTA	YHND
0	0.00	0.00	0.00	0.00	—	—	0	3	0	2	—	—
A	−2.02	0.67	−6.96	−2.38	28	21	6	9	3	2	12	15
B	5.26	1.92	2.78	1.13	6	6	0	0	0	0	2	7
D	−3.06	0.16	−2.26	−0.65	27	11	2	5	0	3	8	14
X	−19.08	−9.34	−29.61	−7.51	38	33	32	23	21	8	34	27
AB	12.45	−2.60	3.97	−2.59	15	22	—	—	1	3	—	—
AD	4.30	0.76	−1.62	−0.53	22	16	—	—	2	2	—	—
AX	−6.49	5.54	−19.12	−2.03	33	16	—	—	6	7	—	—
BD	6.85	4.13	2.77	1.11	14	8	—	—	0	0	—	—
BX	−2.24	−3.98	−18.15	−3.71	32	25	—	—	1	3	—	—
DX	−17.50	−7.17	−24.09	−5.75	38	28	—	—	6	9	—	—
ABD	31.09	9.97	19.17	5.74	4	11	—	—	0	1	—	—
ABX	34.15	19.51	26.27	11.47	9	8	—	—	0	0	—	—
ADX	36.23	17.58	21.85	9.95	9	10	—	—	0	0	—	—
BDX	9.54	0.51	−2.48	1.28	20	17	—	—	0	0	—	—
E	141.18	44.94	118.19	34.96	0	1	—	—	0	0	—	—

[a] *Homo sapiens*, *Canis familiaris*, *Sus scrofa*, *Artibeus jamaicensis*, *Cynocephalus variegates*, *Procavia capensis*, *Erinaceus europaeus*, *Oryctolagus cuniculus*, *Elephantulus* sp. VB001, *Equus caballus*, *Manis tetradactyla*, *Elephas maximus*, *Mus musculus*, *Tupaia belangeri*, *Dugong dugon*, *Orycteropus afer*, *Dasypus novemcinctus*, *Didelphis virginiana*, *Ornithorhynchus anatinus*, *Acipenser stellatus*, *Amia calva*, *Polypterus ornatipinnis*, *Lepisosteus oculatus*, *Takifugu rubripes*, *Albula glossodonta*, *Anguilla anguilla*, *Ateleopus japonicus*, *Aulopus japonicus*, *Engraulis japonicus*, *Elops hawaiensis*, *Dallia pectoralis*, *Lampris guttatus*, *Notacanthus chemnitzi*, *Cyprinus carpio*, *Osteoglossum bicirrhosum*, *Gadus morhua*, *Polymixia japonica*, *Salmo salar*, *Diaphus splendidus*, *Gonostoma gracile*.

parameters for these models are $K = 2(F - 1)$, $2(F - 1)$, and $4(F - 1)$, respectively. The ML parameters are given by formulae equivalent to equation (4).

Next, we define 2-factor models AB, AD, AX, BD, BX, and DX where the frequencies depend on 2 factors only. For example, in model AB, $P_{abd\mathrm{X}}(\mathrm{Z}) = P_{ab}(\mathrm{Z})$ for all d and X. The number of parameters is $K = 8(F - 1)$, and the ML parameters are

$$P_{ab}(\mathrm{Z}) = n_{ab}(\mathrm{Z}) \Big/ \left(\sum_{\mathrm{Z}'} n_{ab}(\mathrm{Z}') \right), \text{ where}$$
$$n_{ab}(\mathrm{Z}) = \sum_{d} \sum_{X} n_{abd\mathrm{X}}(\mathrm{Z}). \qquad (5)$$

The other 2-factor models are defined in a similar way. Additionally, there are four 3-factor models in which frequencies depend on 3 of the 4 factors. The ML parameters are defined by obvious analogy to the 2-factor models above. Finally, there is an exact model E, where the probabilities depend on all 4 factors. This is an exact fit of the data because the number of parameters is equal to the number of independent quantities in the data: $K = 64(F - 1)$. The ML parameters are

$$P_{abd\mathrm{X}}(\mathrm{Z}) = n_{abd\mathrm{X}}(\mathrm{Z}) \Big/ \left(\sum_{\mathrm{Z}'} n_{abd\mathrm{X}}(\mathrm{Z}') \right). \qquad (6)$$

Table 4 shows the results of fitting these models to the 40 species. We quote ΔAIC, which is the change in AIC with respect to model 0. A ΔAIC of order 1 denotes a slight preference for one model over another, and a ΔAIC of order 10 is usually considered sufficient to rule out the less well-fitting model. The results from the human mitochondrial genome are given as an example. For both SPTA and YHND, model X has the lowest AIC, followed by model DX. Factor X has a larger influence than any of the other factors, as seen by the large negative ΔAIC for model X. Models A and D also have negative ΔAIC, indicating some apparent influence of factors A and D. However, when these factors are combined with X, there is no significant improvement over factor X alone, that is, models AX and DX have a higher AIC than model X. All the 3-factor models and model E have positive ΔAIC, indicating that these models are over fitting the data.

Rather than quote AIC values for every species separately, we have summarized the results on the 40 species in several ways. Table 4 shows the average ΔAIC for each model. For each species, ΔAIC is measured relative to the AIC for model 0 for that species, and the mean of the ΔAIC is then calculated. For both SPTA and YHND, model X has the lowest average ΔAIC, and model DX has the second lowest. The human example is thus typical of the majority of the species. However, the average values mask considerable variation among the species. The third pair of columns in table 4 gives the number of species for which ΔAIC < 0 for each model (i.e., the number of species for which the model is an improvement over the null assumption). Once again, models X and DX score highest by this criterion. Inclusion of factor X gives an improvement for almost all cases. The 3-factor models and model E perform poorly because they over fit the data for most species.

The fourth pair of columns compares the single-factor models with model 0 by showing the number of species for which each model has the lowest AIC of these 5 models. Clearly, X is the dominant factor for the majority of species, there are a few species for which factors A and D are

dominant, and there are no species for which factor B is dominant. The fifth pair of columns gives the number of species for which each model is best when all models are considered. Once again X and DX score highly, although AX scores almost as highly as DX. In roughly half the cases, a single-factor model is best, and in roughly half the cases a 2-factor model is best. There is only one case where a 3-factor model is best. Finally, we quote the number of species for which each of the 4 factors is included in the best model, either as a single factor or as part of a 2- or 3-factor model.

There are several clear conclusions from table 4. Factor X has a large effect in almost all species. This shows that context-dependent mutation causes correlations between third-position bases and following first-position bases, and hence has a major influence on codon usage in mitochondrial genomes. Factor B has very little effect. There is no example where model B is best and only a small number of cases where factor B is included in the best model as part of a 2- or 3-factor model. We therefore conclude that selection for translational accuracy does not have a significant influence on codon usage in most mitochondrial genomes. Although factors A and D have less importance than factor X, there are still several examples where these factors are included in the best model. The presence of factor D is understandable. It has already been demonstrated that the base frequencies vary systematically along the genome, and this can be understood in terms of the position of the genes with respect to the origin of replication. These results show that the effect of variability along the genome is sufficiently strong to merit the inclusion of extra parameters in the model in at least some species.

The presence of factor A in the best model for several species is difficult to interpret. Ostensibly, this says that codon usage differs among amino acids, but it does not say why. If we looked at model A alone, we would conclude that this factor is important in more than half the examples (28 and 21 for SPTA and YHND). However, a spurious dependence on A can arise because of a correlation between A and other factors. After accounting for the other factors, we find that factor A remains in the best model in many fewer cases (12 and 15 for SPTA and YHND). Thus, to some extent, the apparent dependence on A is due to the real dependence on X and D, both of which are mutational effects. However, there are species for which models AX or AD are better than X or D alone. Can this be explained by translational selection? Although the lack of importance of factor B rules out translational accuracy, it does not rule out translational efficiency.

In many bacteria, there are clear differences in codon usage among amino acids that correlate with the anticodons of the tRNAs, and the bias is stronger in highly expressed genes. This is an obvious indication of selection for efficient codon–anticodon matching. However, the anticodons of the mitochondrial tRNAs in the SPTA set all have wobble position U and middle position G and differ only in the third anticodon position (which pairs with the first codon position). Similarly, the mitochondrial tRNAs for YHND all have wobble position G and middle position U and differ only at the third anticodon position. Thus, the only way that selection for codon–anticodon matching could explain the relevance of factor A is if the selection on the third codon base caused by the wobble base were dependent on the bases at the first codon and third anticodon position. Nevertheless, it is not impossible that this could occur. For example, the stacking free energy of the codon–anticodon interaction, following the usual energy rules used in RNA folding algorithms (Higgs 2000), depends on the bases at all 3 positions. If the selection coefficients for the alternative codons depended on these stacking free energies, then there could be different selective effects on the codons for different amino acids. It is important that the coefficients should not simply depend on the difference in stacking free energy, however, because the difference in free energies between the different codons would be independent of the base at the first codon position. Thus, if codon–anticodon selection were to be responsible for differences in codon usage among these codon families, it would have to operate in a very subtle way that is not simply a function of the wobble base. We therefore feel that this is an unlikely explanation.

With the above in mind, it seems that a more plausible explanation for the relevance of factor A is, once again, context-dependent mutation. The models above control for the middle position base (by separating amino acids into groups with the same Y) and for the following first-position base (by including X in the model), but they do not consider longer range correlations. If the mutation process depends on nearest neighbors, this automatically sets up correlations between a site and its neighbors, and the neighbors of the neighbors, and so on. Thus, it is quite possible to create a significant correlation between a third-position base and the first-position base in the same codon. This is exactly what is described by model A.

Evolution of the Wobble Base in tRNAs

In the previous section, we argued against the possibility that selection for codon–anticodon pairing was responsible for differences in codon usage among SPTA codon families and among YHND codon families. This still leaves the possibility that selection could act the same way in each family. For example, a C-ending codon for each YHND codon family might be selected because it best matches the wobble G of the corresponding anticodon. Similarly, an A-ending codon for each SPTA codon family might be preferred because it best matches the wobble U of the corresponding anticodon. None of the models above would distinguish such an effect. Lim and Curran (2001) have reviewed what is known about the codon–anticodon interaction and the variant base pairs that can occur at the wobble position. It is not fully clear how to predict the strength of selection acting on codons that would arise from differences in pairing at the wobble position. In particular, there may be differences between the mitochondrial and bacterial translation systems. The preference for C in YHND is well established in bacteria (Sharp et al. 2005), where there are many examples with the same tRNAs present as in mitochondria. It is likely that the C codon would also be preferred for these amino acids if translational selection were significant in mitochondria. The situation for SPTA in bacteria is less easy to compare

with mitochondria because most bacteria have more than one tRNA for these amino acids—typically a wobble-U and a wobble-G tRNA and occasionally a wobble-C tRNA also. The wobble-C tRNA seems to be a luxury that can easily be managed without, because it pairs with only the G-ending codon, and this role can also be carried out by the wobble-U tRNA. The wobble-U and wobble-G tRNAs are both required in the vast majority of cases, however (Mark and Grosjean 2002; Rocha 2004). The preferred codons in bacteria depend on which tRNA genes are present, and these can be different for each amino acid.

Selection for codon–anticodon matching is strongest in those bacteria that are fastest growing (Rocha 2004; Sharp et al. 2005) and tends to be negligible in very slow-growing species. The latter group includes internal parasites and endosymbionts with reduced genome sizes, which resemble mitochondria to some extent. The fast-growing bacteria with strong translational selection have multiple copies of tRNA genes for each amino acid, whereas the slow-growing parasites and endosymbionts usually have minimal numbers of tRNAs. Although usually at least one wobble-U tRNA gene and one wobble-G tRNA gene are retained on the genome for a 4-codon family, there are a few interesting examples where bacteria manage with only one wobble-U tRNA. These examples also occur principally in parasites and endosymbionts. For alanine, there is only one tRNA in *Wolbachia* and *Blochmannia*. For valine, there is only one in *Rickettsia*. For proline, there is only one in *Buchnera*, *Blochmannia*, *Wigglesworthia*, *Wolbachia*, and *Rickettsia*, all of which have reduced genome sizes, and also a few species with larger genomes, like *Campylobacter* and *Haemophilus*. In all these cases, a wobble-G tRNA has been lost from the genome and a wobble-U tRNA has been retained. This is precisely what has occurred in mitochondrial genomes as well.

The wobble position is always U for the tRNAs for all eight 4-codon families for all the fish and mammal species considered here in "Statistical Tests for Context-Dependent Mutation." To be more precise, in the set of 214 fish and 148 mammals analyzed in this paper, there are $8 \times (214 + 148) = 2{,}896$ tRNAs for 4-codon families. Among these, we observed only 3 exceptions to the rule that the wobble position is U. In each case, it was apparently C, which makes us suspect a sequencing error. There were a further 4 cases where the anticodon differed at one of the other 2 positions. These are almost certainly sequencing problems because the tRNAs would pair with codons of a different amino acid if the sequences were correct. In addition, there are 2 fish where the tRNA-Pro is absent because this part of the genome was not completely sequenced and 4 marsupials in which the tRNA-Pro is a pseudogene or absent. None of these exceptions is of interest in the present context. Hence, we can say that in vertebrate mitochondrial tRNAs, the wobble position is essentially fixed to be U for 4-codon tRNAs.

Thus, for 8 amino acids, it is the wobble-U tRNA that has been retained in the very small genome of vertebrate mitochondria. We interpret this as an indication that U is strongly selected at the wobble position because U is the base that can pair most effectively with all 4 third-position bases, that is, U is the most versatile base at the wobble position. The relaxed wobble pairing rules in mitochondria mean that the translation system is able to get away with only one tRNA for 4 codons. However, this is not the same as saying that the third position is irrelevant for codon–anticodon pairing, and that "two out of three will do." If the interaction at the third codon position were unimportant, then mutations at the wobble position in the tRNAs would be neutral and any base could occur at this position. On the contrary, the wobble position is one of the most strongly conserved sites in the tRNA gene.

If selection on the wobble position is strong, does this mean that selection on the third codon positions is also strong? We think not. Selection on the wobble position is likely to be orders of magnitude stronger than on any one-third codon position because a change at the wobble position affects the translation of every codon for that amino acid, whereas a change at a third codon position affects the translation of only one codon. Thus, it is quite possible for selection to strongly influence the wobble position while having negligible effect at third codon positions in comparison to mutation and drift. The wobble base and the third codon base are also not equivalent in terms of the structure of the codon–anticodon interaction. Therefore, a nonstandard base pair like UC might be tolerated when U is the wobble base, but the reverse pair CU could be completely rejected if C is the wobble base. Another point worth noting that might explain why translational selection is apparently much stronger in bacteria than mitochondria is that the effective population size for bacteria is likely to be much larger.

An alternative argument for evolution of the wobble position in mitochondrial tRNAs has been proposed. Xia (2005) notes that in a large number of animal mitochondrial genomes, A is the most frequent base at FFD sites, presumably because the mutation process is biased toward A. He then argues that the U base at the wobble position is selected because it matches the most frequent codon. Thus, the hypothesis is that the wobble position adapts to pair effectively with the most frequent codon. It is clear that this hypothesis is not true for the cases we have considered so far in this paper. Of the 214 fish and 148 mammals we considered, the number of species with U, C, A, and G as the most frequent FFD base are 8, 108, 246, and 0, respectively. If the tRNAs adapted to match the most frequent codon, the tRNAs for the 108 species where C is the most frequent FFD base would have wobble-G tRNAs, for example. We already saw that, apart from the small number of questionable exceptions noted above, all the tRNAs are wobble U. The exceptions do not correlate with the frequencies of bases at FFD sites. A better explanation of all these observations is that U is the only base that will interact effectively with all 4 codons. All 4 bases usually occur at FFD sites, even if their frequencies are very unequal; thus, there is selection to maintain the wobble base as U because it is versatile, irrespective of the frequencies of the bases at FFD sites.

Is There Any Detectable Influence of the Wobble Base on Codon Usage?

From all the previous considerations, it is likely that the influence of the wobble base on codon usage in

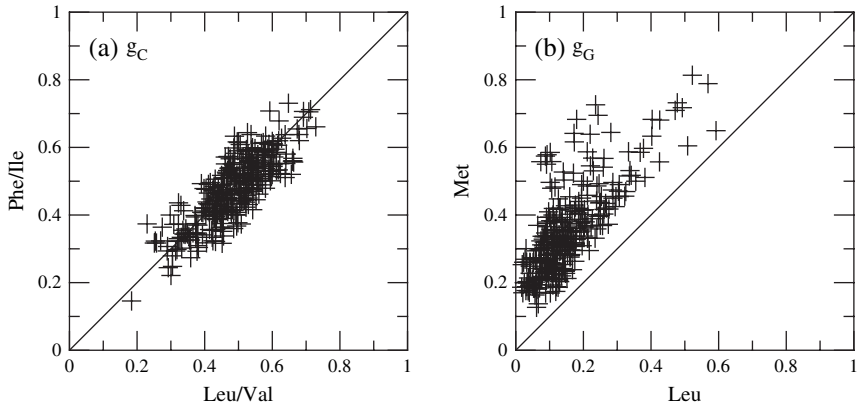

FIG. 2.—(*a*) Relative frequencies of third-position C bases in Phe/Ile codons versus Leu/Val codons. (*b*) Relative frequencies of third-position G bases in Met codons versus Leu(UUR) codons. Data are from 326 current fish genomes.

mitochondria is weak. In this section, we set out to test this by looking for cases where an effect should show up, if there were one. We define $g_C = n(\mathrm{C})/(n(\mathrm{C}) + n(\mathrm{U}))$, the relative frequency of C-ending codons with respect to the sum of C and U codons. This can be measured in Phe and Ile codons (2-codon families with wobble-G tRNAs) and in Leu and Val codons (4-codon families with wobble-U tRNAs). All these codon families have U at middle position; therefore, the strong effect of the middle position base is controlled for. If there were a preference for the C-ending codon in the 2-codon families (as there is in bacteria), we would expect g_C to be larger for Phe/Ile than for Leu/Val. Figure 2*a* shows that this is not the case. Although g_C varies a lot among species, it is roughly equal for the 2 groups of codons (points are close to the line in fig. 2*a*).

It is interesting to compare this with the situation in Met codons. Met is coded by 2 AUR codons (R = A or G) in the vertebrate mitochondrial genetic code. Two-codon families with A and G codons usually have a wobble-U tRNA, e.g., Gln, Lys, Glu, and Leu(UUR). However, the Met tRNA has a C at the wobble position in the gene, which is modified to f^5C in the tRNA molecule. This is a relic of the reassignment of the AUA codon from Ile to Met that occurred in most mitochondrial genomes. In genomes where the standard code is used, the wobble position is an unmodified C, which pairs only with the G-ending codon. Modification of this base allows it to pair with both codons after the codon reassignment. The mechanisms by which codons are reassigned in mitochondrial genetic codes are discussed by Sengupta and Higgs (2005) and Sengupta et al. (2007). We define $g_G = n(\mathrm{G})/(n(\mathrm{A}) + n(\mathrm{G}))$, the relative frequency of G-ending codons with respect to the sum of A and G codons. We compare Met codons with Leu(UUR) codons because they have the same second base. Figure 2*b* shows that g_G is larger for Met than for Leu(UUR) for all 326 species of actinopterygian fish in the current OGRe database. Thus, we conclude that the f^5C base exerts a preference for the G codon, even though it is capable of translating the A codon too. Translational selection is elevating the frequency of G-ending codons for Met due to the unusual nature of the tRNA-Met. This was pointed out previously by Xia X (personal communication and Xia 2005).

The case of the nonstandard wobble base in tRNA-Met is the only example in vertebrate mitochondria where we are convinced that there is a significant influence of the wobble position on codon usage. If we extend our analysis beyond the vertebrates, several other interesting cases arise. The first of these is the Arg CGN family. By analysis of the tRNAs from the OGRe database, we find that in both platyhelminthes and nematodes, there are some species in which the wobble position base is A, whereas there are others in which it is U. In all other animal phyla, the wobble base is always U. We will investigate whether there is a preference for the U-ending codon in the species where the wobble base is A. We define $g_U = n(\mathrm{U})/(n(U) + n(C) + n(A) + n(G))$, the relative frequency of U-ending codons with respect to the sum of the codons in a 4-codon family. Appropriate comparisons for the Arg CGN family are the Gly GGN and Ser AGN and families, which have the same second base, and for which the wobble base is always U in these phyla. (Note that all 4 AGN codons are assigned to Ser in these phyla.) When discussing vertebrates, we considered only the 12 genes on the plus strand because we know the strands are different and because the gene order is conserved. In this example, and all subsequent examples using invertebrates, we will sum codons on both strands because the gene orders vary considerably. Also we are attempting to detect effects of tRNAs, which should act on both strands anyway. The comparison is shown in figure 3. Species with wobble A have high values of g_U for both Arg and Ser/Gly, whereas those with wobble U have much lower values. Thus, there is a clear correlation between the frequency of U at third position and the wobble base. Furthermore, for all 11/11 nematodes and 6/6 platyhelminthes with wobble A, g_U is larger for Arg than for Ser/Gly (points are above the line), which suggests an additional preference for the U-ending codon in the Arg codons of species with a wobble A. A caveat is that 7/9 nematodes and 7/9 platyhelminthes with wobble U are also above the line, which might suggest that there is some effect from the following first-position base, which is not controlled for. However, most of the wobble A species are further above the line than the wobble-U species. Therefore, we interpret this as a real effect of the tRNA, albeit a weak one.

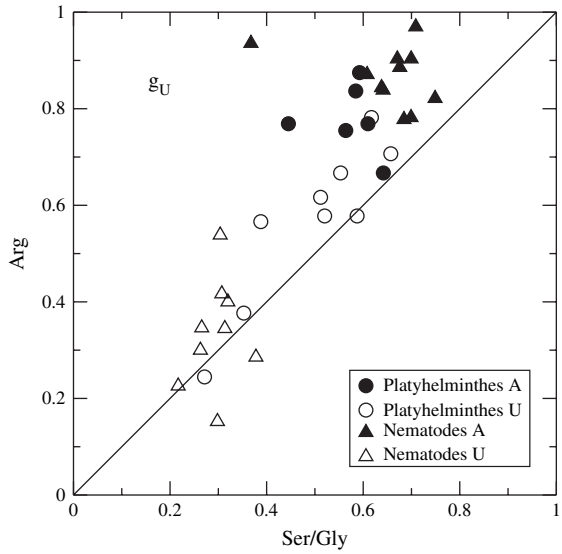

FIG. 3.—Relative frequency of U in Arg versus Ser and Gly codons in nematodes and platyhelminthes where the tRNA-Arg wobble position is A, U, or G.

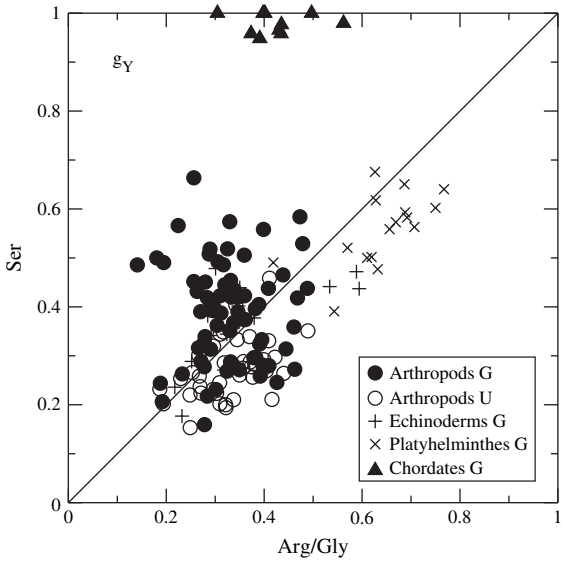

FIG. 4.—Relative frequency of pyrimidines in Ser(AGN) versus Arg and Gly codons in phyla where the tRNA-Ser wobble position is G in at least some species.

Another example on nonstandard wobble base occurs with Lys. In the usual case, a tRNA with wobble U pairs with both AAA and AAG codons. In some animal phyla, including echinoderms and platyhelminthes, the AAA codon is reassigned to Asn, and there is a change to C at the wobble position of tRNA-Lys so that it pairs only with the G-ending codon (Sengupta et al. 2007). In arthropods, however, AAA remains a Lys codon. Surprisingly, therefore, there are some arthropods that have a wobble C on the tRNA-Lys. As the genetic code is not changed in these species, the wobble-C tRNA apparently must translate both codons. Possibly there is a modification of the C, similar to the case of the tRNA-Met discussed previously. We compared g_G in Lys codons with those for Gln and Glu. We found 51/83 species with wobble C have g_G larger for Lys than for Gln/Glu, whereas this is true for only 9/43 species with wobble U. This again suggests a weak preference for the G-ending codon in the wobble-C species, although the significance is difficult to assess because species are evolutionarily related and because the effect of the following first position is not controlled for.

As a final example, we consider AGN codons. In the standard genetic code, AGY is Ser and AGR is Arg. However, the tRNA-Arg that translated AGR codons was lost from the mitochondrial genome in the common ancestor of protostomes and deuterostomes (Sengupta et al. 2007). In most animal phyla, all 4 AGN codons are Ser. In chordates, there has been a subsequent reassignment of AGR to stop. Here, we are interested in the phyla for which there is a 4-codon AGN family for Ser. We find the wobble position is variable between phyla. Brachiopods, annelids, and nematodes all have wobble U, as we would expect. Arthropods and molluscs have a mixture of species with wobble U and wobble G. Echinoderms and platyhelminthes always have wobble G. Chordates also have wobble G, but this is to be expected because only 2 codons code for Ser. We define $g_Y = (n(U) + n(C))/(n(U) + n(C) + n(A) + n(G))$, the relative frequency of pyrimidine codons with respect to the sum of the codons in a 4-codon family. We compare Ser codons with those for Arg and Gly. We might expect g_Y to be larger in Ser than Arg/Gly for species in which there is a wobble G because G usually pairs with pyrimidine codons only. In figure 4, we see that 50 out of 69 wobble-G arthropods are above the line (some of them by a long way), but only 11 out 43 wobble-U arthropods are above the line. This seems to confirm selection in favor of Y-ending codons in the wobble-G species. However, for echinoderms and platyhelminthes, which are all wobble G, there is no observable effect. This means that the wobble G is translating all 4 codons effectively in these species. As a comparison, we also included a few randomly chosen chordates in figure 4. All these have g_Y very close to 1 for Ser because the AGR codons have been reassigned to stop. This makes it clear that codon reassignment induces much larger effects on codon usage than does translational selection for effective codon–anticodon pairing. It also suggests that there must be some modification that allows the wobble G to work with all 4 codons in the echinoderms and platyhelminthes.

Discussion

Mutation is a major force in mitochondrial genomes because mutation rates are large. This is often attributed to the generation of free radicals by the respiration process inside mitochondria in close proximity to the mitochondrial DNA. Rates of evolution of animal mitochondrial genes are usually much faster than nuclear genes, presumably because of the higher mutation rate. It is therefore not surprising that mutation should be dominant in determining codon usage in mitochondria. The mutation process in mitochondrial genomes is complex: rates are different on the 2 strands (as is apparent from simple comparison of the base

frequencies between strands) and are context dependent (as is apparent from the statistical analysis above).

Observation of the frequencies of the dinucleotides in a genome gives only limited information about the mutation rates. On the other hand, if we knew the mutation rates, it would be possible to calculate the equilibrium dinucleotide frequencies for the specified rates. A full specification of a mutation rate matrix that accounts for nearest neighbor context-dependent effects would require $3 \times 4 \times 4 = 48$ rates to be measured (because each base can mutate to 3 other types, and there are 4 possible types for the neighbor on each side). Rate matrices like this have been estimated at least partially for some genomes (Blake et al. 1992; Morton 2003; Morton et al. 2006) but are not available for mitochondrial genomes. The relative rates of different mutation processes differ a lot among species as is evident from the fact that the base frequencies in mitochondrial genomes vary widely between species, especially at FFD sites (Urbina et al. 2006). It was shown above that the same dinucleotides seem to be preferred or avoided in mammals as in fish, and that the preference or avoidance of a dinucleotide seems to be consistent, even though the mean single-nucleotide frequencies vary tremendously among species in each of these groups. This suggests that the processes causing the context-dependent biases may be operating similarly in all these species. Experimental evidence will be required to determine exactly what the chemical processes are that cause these biases.

Although the presence of translational selection in rapidly growing microorganisms such as yeast and many bacteria has long been recognized, it has been much more difficult to distinguish between selection and mutation effects in metazoa (Duret 2002), where selection may be weaker and where differences in base composition occur within the genome (isochores). A particular type of mutational bias that seems to be relevant in this case is biased gene conversion, which may occur during meiosis. Detection of a relationship between expression level and codon usage is one of the principal ways of showing the existence of translation selection. However, a potential confounding factor is that the choice of synonymous codons can affect the degradation rate of mRNAs and hence the gene expression level. Selection might thus influence codon usage even if this has nothing to do with translational efficiency (Duan and Antezana 2003).

The above examples are not specifically relevant to mitochondria, but they show that detection of weak translational selection is not straightforward. In mitochondria, it is not straightforward because there is no easy comparison of high- and low-expression genes. However, the test for translational accuracy used above does not require this, and we showed that this can be done carefully in a way that controls for the context-dependent mutation. We found that there was no evidence for translational accuracy selection in almost all the mitochondrial genomes considered. Evidence has previously been given for the importance of translational accuracy in some genomes other than mitochondria (Akashi 1994; Stoletzki and Eyre-Walker 2007). The analysis in these papers did not control for context-dependent mutation, which is surely a factor in many genomes in addition to mitochondria. Therefore, it would be interesting to know if the evidence for translational accuracy selection would stand if context-dependent mutations were also allowed for in the other genomes.

Selection for translational efficiency could, in principle, lead to selection of codons that best match the limited set of tRNAs still present on mitochondrial genomes. The rule that the wobble position is U for 4-codon families, U for A + G families, and G for U + C families applies for almost all codon families in almost all the animal phyla. It is difficult to explicitly test for this kind of selection because it would act in the same way on different codon families and therefore not lead to an observable difference in codon usage between families. Our interpretation of the results above is that these standard wobble bases do not exert any significant preference on codon usage, and that mutational effects are sufficient to explain the observed patterns. The standard bases at the wobble positions are usually strongly conserved among species. This suggests that these bases are selected because they are the most versatile bases for pairing with all members of the codon family. Selection is orders of magnitude stronger on a wobble base than on a third-position base in a codon; therefore, conservation of the wobble position is compatible with the absence of apparent selection on codon usage.

We also considered cases of nonstandard wobble-position bases, which are of interest precisely because they are rare. All these cases were found to be linked in some way to changes in the mitochondrial genetic code (Sengupta et al. 2007). The most frequently occurring change is the reassignment of UGA from Stop to Trp, which has occurred at least 12 times independently in different lineages of mitochondria. One of these changes is in the common ancestor of all metazoa and their closest protist relatives. The current paper considers data from OGRe, which contains only metazoa; hence, for all the species considered in the present paper, Trp has a 2-codon A + G family. The tRNA-Trp has wobble U, following the standard rule for A + G families in mitochondria. This is worthy of note here because in species that use the canonical genetic code (where UGG is the only Trp codon), the tRNA-Trp has wobble C. Thus, the mutation of the wobble position from C to U is a key aspect of the codon reassignment. This emphasizes the fact that a wobble-C tRNA will not translate an A-ending codon.

This situation is comparable to that for the AUA codon, which is Ile in the canonical code, and is reassigned to Met in most metazoan mitochondrial genomes. The reassignment involves the deletion of the tRNA-Ile that formerly translated AUA and the gain of function of the tRNA-Met so that it can translate both AUA and AUG (Sengupta et al. 2007). Before the codon reassignment, the tRNA-Met has wobble C (analogous to the tRNA-Trp in the previous paragraph). We might expect that this would mutate to U when the genetic code changes. However, in all vertebrate mitochondria, this position remains as C in the tRNA gene sequence. In a few species, we know from experiment that this C is posttranscriptionally modified to f^5C, and it is presumed that this modification occurs in all vertebrates. In figure 2*b*, we showed that there is a preference for the G-ending codon for Met. The f^5C is doing the job of translating both codons, but it is doing

it less well than would a wobble-U tRNA. This raises the question of why the C does not simply mutate to a U in the gene sequence, making the f^5C modification unnecessary. It is likely that this has something to do with the special role of AUG as an initiator codon. In bacteria, there are distinct initiator and elongator tRNA-Met genes, both with C at the wobble position. This is also true in some nonmetazoan mitochondria where AUG is the only Met codon (the case of fungi is discussed in detail in the supplementary data of Sengupta et al. 2007). The single tRNA-Met in metazoan mitochondria must be adapted to do both jobs. We suggest (although we have no evidence) that there is a constraint that a C (or modified C) is required at the wobble position specifically when the tRNA acts as an initiator, and that a wobble-U tRNA would not work as an initiator.

The case of the Arg CGN codons shown in figure 3 is linked to another peculiarity in the genetic code. We showed that the wobble base is often A in nematodes and platyhelminthes. It is intriguing that Arg is the only amino acid for which there is a tRNA gene with a wobble A in bacteria (Rocha 2004). The A is modified to I in the tRNA molecule and this translates U-, C-, and A-ending codons. A second tRNA with wobble C translates the G-ending codon. The wobble-C tRNA is not present in any mitochondrial genome that we know of. In many protists and fungi, the mitochondrial genome has a single tRNA-Arg with wobble A. We presume that this is modified to I (or some other base), and that the resulting tRNA can deal with all 4 codons. This system switched to a single wobble-U tRNA in the ancestor of the metazoa and also in several independent lineages of fungi and protists (Sengupta et al. 2007). The examples of wobble A in nematodes and platyhelminthes that we considered here are therefore reversions to the old system of translation.

The case of the AAA codon is similar. The wobble position of tRNA-Lys switches to C in some arthropods and nematodes. This is the same change that happens independently in echinoderms and platyhelminthes where the AAA codon is reassigned from Lys to Asn. We are not aware of any evidence that the code has been reassigned in the wobble-C species of arthropods and nematodes, although this observation suggests that a change might occur in the future in these species. It could be that the AAA is ambiguously translated as both Lys and Asn in those species at present, although we have no direct evidence for this. There seems to be something that predisposes this change to happen, whereas no change occurs in the apparently equivalent cases of the Gln CAA and Glu GAA codons.

For the AGN codons (fig. 4), there have been a number of genetic code changes. The initial cause in this case seems to have been the deletion of the tRNA-Arg that normally translates AGR codons in the canonical code. This occurred early in metazoan evolution (Sengupta et al. 2007). In bacteria, and in nonmetazoan mitochondria, this gene is always present and is presumably required. However, it is clear that the metazoan mitochondria can manage without this gene. Immediately after the deletion of the tRNA-Arg, the tRNA-Ser would still have had a wobble G, as would be usual for a 2-codon U + C family. By default, this tRNA had to deal with all 4 AGN codons. Figure 4 shows that it could do this reasonably well: there is no evidence for avoidance of AGR codons in echinoderms and platyhelminthes where the tRNA-Ser still has a wobble G, and there is only a fairly weak effect in the wobble-G arthropods. However, this situation does not seem to be stable in the long term. It has been replaced by 3 other systems: in brachiopods, annelids, and nematodes, the wobble G has mutated to U, so AGN becomes a normal 4-codon family; in chordates, AGR has been reassigned to stop and AGY remains a normal 2-codon family, and in urochordates, extra tRNA-Gly has appeared in the genome that translates AGY as Gly. These 3 alternatives seem like more permanent solutions to the problem created by the deletion of the original tRNA-Arg.

We conclude that in unusual cases where the wobble position is different from the base expected in the standard mitochondrial translation system, there is evidence for weak selection preferring codons that match the anticodons. However, translational efficiency does not seem to be the cause of the changes in the wobble positions. The reasons that the tRNA wobble positions change in these few cases are linked to factors that also cause codon reassignments, such as changes to the base modification pattern or responses of one tRNA to deletion of another tRNA from the genome. Therefore, our main result is that in contrast to bacteria, where there are many species in which translational selection has an important influence on codon usage, in mitochondria, codon usage patterns seem to be determined principally by complex context-dependent mutational effects.

Acknowledgments

This work was supported by the Canada Research Chairs program and the Natural Sciences and Engineering Research Council of Canada.

Literature Cited

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics. 136:927–935.

Akashi H. 2003. Translational selection and yeast proteome evolution. Genetics. 164:1291–1303.

Antezana MA, Kreitman M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. J Mol Evol. 49: 36–43.

Arndt PF, Hwa T. 2005. Identification and measurement of neighbor-dependent nucleotide substitution processes. Bioinformatics. 21:2322–2328.

Bielawski JP, Gold JR. 2002. Mutation patterns of mitochondrial H- and L-strand DNA in closely-related cyprinid fishes. Genetics. 161:1589–1597.

Blake RD, Hess ST, Nicholson-Tuell J. 1992. The influence of nearest neighbours on the rate and pattern of spontaneous point mutations. J Mol Evol. 34:189–200.

Burnham KP, Anderson DR. 1998. Model selection and inference: a practical information-theoretic approach. New York: Springer-Verlag.

dos Reis M, Wernisch L, Savva R. 2003. Unexpected correlations between gene expression and codon usage bias

from microarray data for the whole *Escherichia coli* K-12 genome. Nucleic Acids Res. 31:6976–6985.
Duan J, Antezana MA. 2003. Mammalian mutation pressure, synonymous codon choice, and mRNA degradation. J Mol Evol. 57:694–701.
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.
Duret L. 2002. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev. 12:640–649.
Faith JJ, Pollock DD. 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. Genetics. 165:735–745.
Fedorov A, Saxonov S, Gilbert W. 2002. Regularities of context-dependent codon bias in eukaryotic genes. Nucleic Acids Res. 30:1192–1197.
Foster PG, Jermiin LS, Hickey DA. 1997. Nucleotide compositional bias affects amino acid frequencies in proteins coded by animal mitochondria. J Mol Evol. 44:282–288.
Gibson A, Gowri-Shankar V, Higgs PG, Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Mol Biol Evol. 22:251–264.
Higgs PG. 2000. RNA secondary structure: physical and computational aspects. Quart Rev Biophys. 33:199–253.
Higgs PG, Hao W, Golding GB. 2007. Identification of selective effects on highly expressed genes. Evol Bioinform. 2:1–13.
Higgs PG, Jameson D, Jow H, Rattray M. 2003. The evolution of tRNA-Leucine genes in animal mitochondrial genomes. J Mol Evol. 57:435–445.
Hudelot C, Gowri-Shankar V, Jow H, Rattray M, Higgs PG. 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol Phyl Evol. 28:241–252.
Hwang DG, Green P. 2004. Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci USA. 101:13994–14001.
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. 151: 389–409.
Ikemura T. 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol. 2: 13–34.
Jameson D, Gibson AP, Hudelot C, Higgs PG. 2003. OGRe: a relational database for comparative analysis of mitochondrial genomes. Nucleic Acids Res. 31:202–206.
Karlin S, Mrazek J. 1996. What drives codon choices in human genes? J Mol Biol. 262:459–472.
Karlin S, Mrazek J. 1997. Compositional differences within and between eukaryotic genomes. Proc Natl Acad Sci USA. 94:10227–10232.
Lim VI, Curran JF. 2001. Analysis of codon:anticodon interactions within the ribosome provides new insights into codon reading and the genetic code structure. RNA. 7: 942–957.
Mark C, Grosjean H. 2002. tRNomics: Analysis of tRNA genes from 50 genomes of Eukarya, Archaea and Bacteria reveals anticodon-sparing strategies and domain-specific features. RNA. 8:1189–1232.
Morton BR. 2003. The role of context-dependent mutations in generating compositional and codon usage bias in grass chloroplast DNA. J Mol Evol. 56:616–629.
Morton BR, Bi IV, McMullen MD, Gaut BS. 2006. Variation in mutation dynamics across the maize genome as a function of regional and flanking base composition. Genetics. 172:569–577.
Morton BR, Wright SI. 2007. Selective constraints on codon usage of nuclear genes from *Arabidopsis thaliana*. Mol Biol Evol. 24:122–129.
Percudani R, Pavesi A, Ottonello S. 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J Mol Biol. 268:322–330.
Reyes A, Gissi C, Pesole G, Saccone C. 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. Mol Biol Evol. 15:957–966.
Rocha EPC. 2004. Codon usage from the tRNA's point of view: redundancy, specialization, and efficient decoding for translational optimization. Genome Res. 14:2279–2286.
Sengupta S, Higgs PG. 2005. A unified model of codon reassignment in alternative genetic codes. Genetics. 170:831–840.
Sengupta S, Yang X, Higgs PG. 2007. The mechanisms of codon reassignments in mitochondrial genetic codes. J Mol Evol. 64:662–688.
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.
Sharp PM, Cowe E, Higgins DG, Shields DC, Wolfe KH, Wright F. 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within species diversity. Nucleic Acids Res. 16:8207–8211.
Sharp PM, Li WH. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15:1281–1295.
Shioiri C, Takahata N. 2001. Skew of mononucleotide frequencies, relative abundance of dinucleotides and DNA strand asymmetry. J Mol Evol. 53:364–376.
Singer GAC, Hickey DA. 2000. Nucleotide bias causes a genome-wide bias in amino acid composition of proteins. Mol Biol Evol. 17:1581–1588.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: Selection for translational accuracy. Mol Biol Evol. 24:374–381.
Urbina D, Tang B, Higgs PG. 2006. The response of amino acid frequencies to directional mutational pressure in mitochondrial genome sequences is related to the physical properties of the amino acids and to the structure of the genetic code. J Mol Evol. 62:340–361.
Wright F. 1990. The effective number of codons used in a gene. Gene. 87:23–29.
Xia XH. 2005. Mutation and selection on the anticodon of tRNA genes in vertebrate mitochondrial genomes. Gene. 345:13–20.
Xu W, Jameson D, Tang B, Higgs PG. 2006. The relationship between the rate of molecular evolution and the rate of genome rearrangement in animal mitochondrial genomes. J Mol Evol. 63:375–392.

Jeffrey Thorne, Associate Editor

Accepted November 6, 2007