Copyright © 1996 by the Genetics Society of America

Maximizing Transcription Efficiency Causes Codon Usage Bias

Xuhua Xia

Museum of Natural Science and Department of Zoology and Physiology, Louisiana State University, Baton Rouge, Louisiana 70803 and Department of Ecology and Biodiversity, University of Hong Kong, Hong Kong

Manuscript received April 15, 1996
Accepted for publication August 9, 1996

ABSTRACT

The rate of protein synthesis depends on both the rate of initiation of translation and the rate of elongation of the peptide chain. The rate of initiation depends on the encountering rate between ribosomes and mRNA; this rate in turn depends on the concentration of ribosomes and mRNA. Thus, patterns of codon usage that increase transcriptional efficiency should increase mRNA concentration, which in turn would increase the initiation rate and the rate of protein synthesis. An optimality model of the transcriptional process is presented with the prediction that the most frequently used ribonucleotide at the third codon sites in mRNA molecules should be the same as the most abundant ribonucleotide in the cellular matrix where mRNA is transcribed. This prediction is supported by four kinds of evidence. First, A-ending codons are the most frequently used synonymous codons in mitochondria, where ATP is much more abundant than that of the three other ribonucleotides. Second, A-ending codons are more frequently used in mitochondrial genes than in nuclear genes. Third, protein genes from organisms with a high metabolic rate use more A-ending codons and have higher A content in their introns than those from organisms with a low metabolic rate.

GENOMES from distantly related organisms exhibit different patterns of synonymous codon usage (GRANTHAM *et al.* 1980, 1981). In addition to this intergenome difference, there are substantial intergene differences within the same genome (GOUY and GAUTIER 1982; IKEMURA 1985, 1992; SHARP and LI 1986, 1987; SHARP *et al.* 1988). Natural selection for increased translational efficiency has been proposed as the major hypothesis for the intergenome and -gene differences in codon usage (KIMURA 1983; ROBINSON *et al.* 1984; KURLAND 1987a,b; BULMER 1988, 1991). Three lines of evidence appear to support this hypothesis. First, the frequency of codon usage is positively correlated with tRNA availability (GOUY and GAUTIER 1982; IKEMURA 1981, 1982, 1985, 1992). Second, the degree of codon usage bias is related to the level of gene expression, with highly expressed genes exhibiting greater codon bias than lowly expressed genes (BENNETZEN and HALL 1982; SHARP and DEVINE 1989; SHARP *et al.* 1988). Third, mRNA consisting of preferred codons is translated faster than mRNA artificially modified to contain rare codons (SORENSEN *et al.* 1989).

Not only are there differences in codon usage bias among genomes and among genes within the same genome, but there are also differences in codon usage among different regions of the same gene. For example, gene regions of greater amino acid conservation tend to exhibit more dramatic codon usage bias than do regions of less amino acid conservation (AKASHI 1994). This has been proposed as resulting from selection for increased translational accuracy (AKASHI 1994; HARTL *et al.* 1994), because selection for maximum translational efficiency does not seem satisfactory to explain the phenomenon. However, this can be accommodated by the translational efficiency hypothesis if one defines what is maximized as the rate of production of *correctly* translated proteins.

What all these studies have shown is that there is strong selection favoring increased rate of protein synthesis and that a coding strategy that increases the peptide elongation rate (and consequently increases the rate of protein synthesis) is favored by natural selection. However, efficient protein synthesis depends not only on the peptide elongation rate, but also on the initiation rate. Moreover, four lines of evidence support the claim that the initiation of protein synthesis, rather than elongation of the peptide chain, is rate-limiting (BULMER 1991). Thus, if there is selection for increased rate of protein synthesis, then we should expect selection to favor an increase of not only elongation rate, but also initiation rate. The evolutionary consequence of selection for increased elongation rate has been investigated and empirically documented extensively (GOUY and GAUTIER 1982; IKEMURA 1985, 1992; SHARP and LI 1986, 1987; BULMER 1988, 1991; SHARP *et al.* 1988; SHARP and DEVINE 1989). In contrast, the evolutionary consequence of selection for increased initiation rate has not been equally well studied.

The initiation rate is directly proportional to the encountering rate between mRNA molecules and ribosomes, and this encountering rate depends on the con-

Address for correspondence: Xuhua Xia, Department of Ecology and Biodiversity, University of Hong Kong, Pukfulam Road, Hong Kong. E-mail: xxia@hkusua.hku.hk

centration of mRNA and ribosomes. Thus, the initiation rate of protein synthesis can be efficiently increased by increasing mRNA concentration. Both theoretical reasoning and empirical evidence suggest that the number of mRNA copies available is a rate-limiting factor in protein synthesis (XIA 1995). It is conceivable that natural selection should favor increased rate of transcription and that a coding strategy leading to increased transcriptional efficiency should be at a selective advantage. Thus, studying the pattern of codon usage from the perspective of transcription adds one more dimension to our understanding of the evolution of genetic information.

CONSEQUENCES OF MAXIMIZING TRANSCRIPTIONAL EFFICIENCY

I here present an optimality model showing the effect of maximizing transcription rate on codon usage bias. Suppose that an mRNA molecule of length L is composed of A, C, G, and U with frequencies N_A, N_C, N_G and N_U, respectively ($N_A + N_C + N_G + N_U = L$). In terms of a chemical equation,

$$N_A\, A + N_C\, C + N_G\, G + N_U\, U \xrightarrow{k} mRNA \qquad (1)$$

where k is the velocity constant of the transcriptional process. Let C be the concentration of transcribed mRNA, and let C_A, C_C, C_G and C_U be the concentration of A, C, G, and U, respectively, in the cellular matrix surrounding the active transcription site. Then, according to the law of mass action, the rate of transcription is

$$\frac{dC}{dt} = k\, C_A^{N_A}\, C_C^{N_C}\, C_G^{N_G}\, C_U^{N_U} \qquad (2)$$

Evidently, if C_A is greater than C_C, C_G and C_U, then the transcription rate is increased by increasing N_A and decreasing N_C, N_G and N_U with the constraint that $\Sigma N_i = L$, where i = A, C, G, U. Consequently, the maximum transcription rate is reached when $N_A = L$ and $N_C = N_G = N_U = 0$.

Equation (2) links the nucleotide composition of mRNA, *i.e.*, N_A, N_C, N_G and N_U, to the relative nucleotide concentration in the cellular matrix at the transcription site, *i.e.*, C_A, C_C, C_G and C_U. The equation predicts that the most frequently used nucleotide in mRNA molecules should be the same as the most abundant nucleotide in the cellular matrix. This implies that the relative concentration of the four nucleotides in the cellular matrix can affect patterns of synonymous codon usage. This hypothesis will hereafter be referred to as the transcription hypothesis of codon usage (THCU).

The same conclusion can be derived from a deterministic model with more explicit assumptions. Consider the time required to transcribe a single nucleotide i. Let r be the rate of nucleotides diffusing to the transcription site and P_i be the probability that the arriving nucleotide is nucleotide i. Note that P_i (where i = A, C, G, U) simply represents the relative availability of the four nucleotides. Let t_l be the time spent in linking this nucleotide to the elongating mRNA chain, and t_r be the time spent in rejecting each of the wrong nucleotides that diffuse to the transcription site prior to the arrival of the nucleotide i. Now the total time spent in transcribing nucleotide i is

$$T_i = \frac{1}{r\, P_i} + t_l + \left(\frac{1}{P_i} - 1\right) t_r \qquad (3)$$

where the first term on the right-hand side of the equation is the time needed for the correct nucleotide to arrive at the transcription site, and the third term represents time spent in rejecting the wrong nucleotides prior to the arrival of the correct nucleotide. The total time (T) required to transcribe L nucleotides (total elongation time) can be shown to be:

$$T = \sum_{i=1}^{4} N_i\, T_i = L\,(t_l - t_r) + \frac{1 + r\, t_r}{r}\, Y \qquad (4)$$

where

$$Y = \sum_{i=1}^{4} \frac{N_i}{P_i} = \frac{N_A}{P_A} + \frac{N_C}{P_C} + \frac{N_G}{P_G} + \frac{N_U}{P_U}$$

Note that N_i is a property of the mRNA, whereas P_i is a property of the cellular matrix.

Our objective, then, is to find the conditions that minimize T. Because t_l, t_r and L are not dependent on N_i and P_i, they are treated as constants. Thus, minimizing T in Equation (4) is equivalent to minimizing Y. We rewrite Y as:

$$Y = \frac{N_A\, P_C\, P_G\, P_U + N_C\, P_A\, P_G P_U + N_G\, P_A\, P_C P_U + N_U\, P_A\, P_C P_G}{P_A\, P_C\, P_G\, P_U} \qquad (5)$$

If P_A is the largest of the four, then ($P_C\, P_G\, P_U$) is smaller than either ($P_A\, P_G\, P_U$), ($P_A\, P_C\, P_U$) or ($P_A\, P_C\, P_G$). It is therefore obvious that minimization of Y in Equation (5), given that P_A is the largest of the four, requires an increase in N_A and a decrease in N_C, N_G and N_U, with the minimum of Y reached when $N_A = L$ and $N_C = N_G = N_U = 0$. The general prediction from the optimality model, therefore, states that whenever different nucleotides in the cellular matrix differ in relative availability, the codon usage of the mRNA should evolve toward increasing the frequency of the most abundant nucleotide in the cellular matrix. Thus, we reached the same conclusion as that from the law of mass action.

Most synonymous codons differ at the third codon site. According to the general prediction above, we expect that, within each codon family, a codon ending with a nucleotide that is the most abundant in the cellular medium should be used the most frequently. This leads to three testable predictions (predictions 1–3 be-

TABLE 1

Codon usage bias in the mitochondrial genome (mtDNA) of the cow, *Bos taurus*

Codon	AA	*N*	RSCU	Codon	AA	*N*	RSCU
GCU	A	52	0.84	**CAA**	**Q**	**79**	**1.82**
GCC	A	91	1.47	CAG	Q	8	0.18
GCA	**A**	**103**	**1.66**	CGU	R	7	0.44
GCG	A	2	0.03	CGC	R	11	0.70
GAA	**E**	**78**	**1.64**	**CGA**	**R**	**42**	**2.67**
GAG	E	17	0.36	CGG	R	3	0.19
GGU	G	29	0.53	UCU	S	51	1.11
GGC	G	62	1.13	UCC	S	65	1.42
GGA	**G**	**97**	**1.77**	**UCA**	**S**	**99**	**2.16**
GGG	G	31	0.57	UCG	S	5	0.11
AAA	**K**	**90**	**1.78**	**UAA**	*	**8**	**3.20**
AAG	K	11	0.22	UAG	*	1	0.40
UUA	**L**	**110**	**1.11**	**AGA**	*	**1**	**0.40**
UUG	L	16	0.16	AGG	*	0	0.00
CUU	L	62	0.62	ACU	T	44	0.57
CUC	L	95	0.95	ACC	T	96	1.25
CUA	**L**	**285**	**2.86**	**ACA**	**T**	**153**	**1.99**
CUG	L	29	0.29	ACG	T	15	0.19
AUA	**M**	**218**	**1.66**	GUU	V	40	0.84
AUG	M	44	0.34	GUC	V	48	1.01
CCU	P	42	0.87	**GUA**	**V**	**87**	**1.83**
CCC	P	63	1.31	GUG	V	15	0.32
CCA	**P**	**85**	**1.76**	**UGA**	**W**	**92**	**1.77**
CCG	P	3	0.06	UGG	W	12	0.23

Values are based on all protein-coding genes with a total of 3800 codons in the complete mitochondrial sequence (GenBank accession No. J01394). AA, one-letter code for amino acid; *N*, number of codons; RSCU, relative synonymous codon usage (calculated as the observed frequency of a codon divided by its expected frequency under the assumption of equal codon usage, SHARP *et al.* 1986); *, stop codons. Note the excess of A-ending codons (shown in bold type) in all synonymous codon families. The probability that RSCU for A-ending codons is not greater than 1 is <0.0001. The pattern is similar for rat, rabbit, sheep, human, and macaque.

low). In addition, because introns are also transcribed and should be subject to selection maximizing transcription efficiency, we expect a nucleotide species to be used more frequently in introns when the concentration of that nucleotide species increases in the cellular medium (prediction 4 below).

PREDICTIONS AND EMPIRICAL TESTS

Prediction 1: A-ending codons should be more frequent than alternative synonymous codons in mitochondrial protein genes: The concentration of cellular ATP is much higher than that of the other three nucleotides (C, G, and U), and the ATP concentration in mitochondria is still higher than that in cytosol (BRIDGER and HENDERSON 1983, pp. 4–5). The high ATP concentration in mitochondria might be caused by many factors, and one of these factors is that mitochondria have an efficient transport system to bring ADP into mitochondria for ATP production, but the transport system does not carry nonadenine nucleotides (BRIDGER and HENDERSON 1983, p. 59–60; OLSON 1986). Given that ATP concentration is higher than that of the other three nucleotides in mitochondria, we should expect synonymous codon usage to be biased toward A-ending codons in mitochondria to facilitate transcription, according to THCU. This expectation is born out with empirical evidence (Table 1).

Prediction 2: The proportion of A-ending codons in each synonymous codon family should be smaller in nuclear protein genes than in mitochondrial protein genes: Whereas ATP concentration should be much higher than the concentration of non-ATP nucleotides in mitochondria for reasons stated in the previous paragraph, the difference in concentration between ATP and non-ATP nucleotides should be relatively small in the nucleus because ATP concentration is lower in nucleus than in mitochondria. ATP concentrations in rat liver cytosol and mitochondria were 6.2 ± 0.63 and 7.5 ± 0.73 μmol/(ml water), respectively (BRIDGER and HENDERSON 1983, p. 5). The actual difference is expected to be greater because mitochondrial preparation was not absolutely free of cytosol contamination and vice versa. It is believed that little difference exists in ATP concentration between nucleus and cytoplasm (BRIDGER and HENDERSON 1983, p. 5), *i.e.*, the difference in ATP concentration between mitochondria and cytosol is also the difference between mitochondria and nucleus.

Given the lower concentration of ATP in nucleus than in mitochondria, we should expect A-ending co-

TABLE 2

Proportion of A-ending codons (N_A/N) in each synonymous codon family on the nuclear and mitochondrial genomes in the cow, *Bos taurus*

	N_A/N		
Amino acids	NUC	mtDNA	*E. coli*
Ala	0.1733	0.4153	0.22
Arg	0.1593	0.6667	0.06
Gln	0.2136	0.9080	0.31
Glu	0.3383	0.8211	0.70
Gly	0.2130	0.4429	0.09
Leu	0.0899	0.6616	0.14
Lys	0.3469	0.8911	0.76
Pro	0.2235	0.4404	0.20
Ser	0.1906	0.4500	0.12
Thr	0.2211	0.4968	0.12
Val	0.0766	0.4579	0.17

The last column is *E. coli* DNA. N_A, number of A-ending codons, N, total number of codons. Data from mitochondrial DNA were based on data in Table 1. Data for Arginine is limited to CGN codons (*i.e.*, excluding AGA and AGG, which are stop codons in mtDNA). Note the relative deficiency of A-ending codons in the nuclear genome relative to the mitochondrial genome in the cow ($P < 0.0001$ based on paired *t*-test). The pattern is similar for rat, rabbit, cow, sheep, human, and macaque. Data for cow nuclear DNA and *E. coli* DNA were derived from 261 and 681 genes, respectively, found in GenBank 63. NUC, nuclear; mtDNA, mitochondrial.

dons to be less frequent in the nuclear genome than in the mitochondrial genome, which is also true (Table 2). The data are derived from a compilation by J. M. CHERRY (cherry@frodo.mgh.harvard.edu) with the GCG program CodonFrequency. The original compilation, together with a description of the compilation procedure, is available at the FTP site ftp.bio.indiana.edu. Thus, the difference in synonymous codon usage between the nuclear genome and the mitochondrial genome appears to be explained, at least partially, by the difference in relative ATP concentration between the nuclear medium and the mitochondrial medium.

An alternative explanation for the difference between mtDNA and nuclear DNA in Table 2 is that the prokaryotic ancestor of mtDNA had a high frequency of A-ending codons and that this high frequency of A-ending codons has been maintained through evolutionary inertia rather than through any optimization process suggested by THCU. If this is true, then we would expect prokaryotic genomes, which presumably share the same ancestor with the mitochondrial genome, also to exhibit a high frequency of A-ending codons. This expectation is clearly not fulfilled (Table 2). The frequency of A-ending codons in *Escherichia coli* genome is significantly smaller ($P < 0.0001$, Table 2) than that of the mitochondrial genome in the cow. The pattern in Table 2 holds true if the cow in Table 2 is replaced by other eukaryotic organisms such as rat, rabbit, sheep, human, Macaca, Saccharomyces, or Drosophila. Note that the mtDNA and nuclear genome have diverged a very long time. So an explanation of evolutionary inertia is perhaps unnecessary in the first place.

Prediction 3: The proportion of A-ending codons should be greater in organisms with a high weight-specific metabolic rate (SMR) than in organisms with a low SMR: Different animal species differ greatly in SMR (measured as O_2 consumption in unit of milliliters $\cdot$ hour$^{-1}\cdot$ grams^{-1}). In mammals, SMR is inversely correlated with body size, with the mouse having much higher SMR than the cow (ALTMAN and DITTMER 1972: 1613–1616). Differences in SMR among animals of different body sizes are correlated with the number of mitochondria per unit volume of tissue, with mammals of high SMR having more mitochondria per unit volume of tissue than mammals of small SMR (SMITH 1956; MATHIEU *et al.* 1981; EKERT and RANDALL 1983, pp. 698–699). According to WEIBEL's (1984) authoritative review, the cell's potential for ATP production is proportional to the volume density of its mitochondria. This explains the rapid decrease of maximum sustainable metabolic rate with increasing body weight (decreasing volume density of mitochondria) in mammalian species (HOCHACHKA 1991). In light of all these related lines of evidence, I think it reasonable to assume that nucleotide production is more ATP-biased in small mammalian species with a high SMR than in large mammalian species with a low SMR. In other words, the availability of cellular ATP (relative to the other three nucleotides, C, G, and U) is greater in small mammalian species with a high SMR than in large mammalian species with a low SMR.

If the inference above is correct, then we should expect a greater proportion of A-ending codons in small mammals, such as the mouse with SMR equal to 1.59, than in large mammals, such as the cow and sheep with SMR equal to 0.127 and 0.206, respectively (ALTMAN and DITTMER 1972: 1613–1616). This expectation is confirmed with erythropoietin gene from the mouse, cow and sheep (Table 3). A-ending codons are used significantly more frequently in the mouse gene than in the cow and sheep gene. Complete DNA sequence is also available for the erythropoietin receptor gene from the mouse and human (accession numbers J04843 and M60459), with the mouse gene having a significantly greater RSCU values for A-ending codons than the human gene. SMR for the human is 0.228, which is much smaller than that for the mouse.

The test in Table 3 is weak because of the lack of phylogenetic control (FELSENSTEIN 1985, 1988; HARVEY and PAGEL 1991). The mouse differs from the cow and sheep not only in metabolic rate, but also in many other ways, each of which could potentially be responsible for the difference in the usage of A-ending codons. A more rigorous comparative analysis (FELSENSTEIN 1985, 1988; HARVEY and PAGEL 1991) is therefore needed to test the third prediction. Data for such an analysis are sum-

TABLE 3

Relative synonymous codon usage (RSCU) in the cow, sheep and mouse erythropoietin gene (complete sequence)

Codon	AA	RSCU		
		Bos	Ovis	Mus
UUA	L	0.00	0.00	0.34
CUA	L	0.19	0.18	0.51
AUA	I	0.00	0.43	0.75
GUA	V	0.00	0.00	0.33
UCA	S	0.50	0.50	1.09
CCA	P	2.22	2.18	2.13
ACA	T	0.44	0.44	1.00
GCA	A	0.80	0.70	0.57
UAA	*	0.00	0.00	0.00
CAA	Q	0.00	0.00	0.29
AAA	K	0.67	0.67	1.00
GAA	E	0.83	0.83	1.20
UGA	*	3.00	3.00	3.00
CGA	R	1.13	1.13	0.86
AGA	R	0.38	0.38	1.29
GGA	G	0.67	0.67	0.80

The GenBank accession numbers for the cow, sheep and mouse genes are L41354, Z24681, and M12482, respectively (The mouse erythropoietin gene M12930 exhibits identical codon usage as M12482). RSCU values based on 193 codons for the cow gene, 195 codons for the sheep gene and 193 codons for the mouse gene. RSCU for A-coding codons in the mouse gene is significantly greater than that in the cow gene ($P = 0.0056$, paired t-test) and that in the sheep gene ($P = 0.0041$). There is no difference between the cow gene and the sheep gene ($P = 0.5451$, paired t-test). Bos, cow; Ovis, sheep; Mus, mouse.

marized in Table 4, with globin genes from six different mammalian species.

A comparative analysis involves constructing a phylogenetic tree of the focal taxa, partitioning the differences in the focal variables among the taxa (in our case SMR and $\bar{N}_A$ in Table 4) along the branches to obtain independent contrasts, and testing the covariation of the two focal variables during evolutionary history (FELSENSTEIN 1985, 1988; HARVEY and PAGEL 1991). The phylogenetic tree for the six species (Figure 1) was taken from NOVACEK *et al.* (1988).

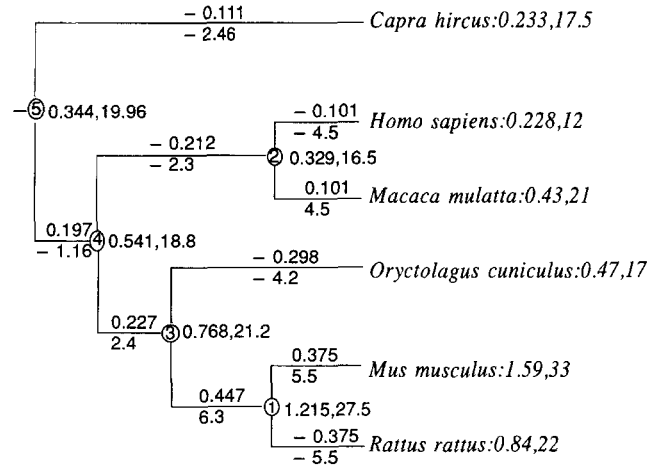

FIGURE 1.—Partition of the differences in weight-specific metabolic rate (SMR) and the average number of A-ending codons ($\bar{N}_A$) along branches of the phylogenetic tree of the six mammalian species. The two numbers following each species name are observed SMR and $\bar{N}_A$values, respectively. The two numbers at each numbered internal node are inferred SMR and $\bar{N}_A$ values. The two numbers above and below each branch represent changes in SMR and $\bar{N}_A$ from the ancestral taxon to the descendant taxon.

The differences in the two focal variables were partitioned along the branches by using a computer program C.A.I.C, with the option of setting all branches

TABLE 4

Locus names of α-globin and β-globin genes of six mammalian species

Species	Locus name		$\bar{N}_A$	SMR
	α-globin	β-globin		
Capra hircus	GOTHBAI GOTHBAII	GOTHBBEI GOTHBBEII	17.5	0.233
Homo sapiens	HUMHBA4 (α2) HUMHBA4 (α1)	HUMBETGLOA HUMBETGLOB HUMBETGLOC	12	0.228
Macaca mulatta	MACHBA	MACHBGA1 MACHBGA2	21	0.43
Oryctolagus cuniculus	RABHBAP	RABBGLOB	17	0.47
Mus musculus	MUSHBA	MUSHBBH0 MUSHBBH1	33	1.59
Rattus norvegicus	RATHBAM	RATHBBM	22	0.84

The average number of A-ending codons ($\bar{N}_A$) was calculated as the sum of the average number of A-ending codons for α-globin (143 codons) and that for β-globin (148 codons). For example, *Capra hircus* has two sequences of the α-globin gene, with five A-ending codons for each sequences. It also has two sequences of the β-globin gene, with the number of A-ending codons equal to 11 and 14, respectively. So $\bar{N}_A$ for *Capra hircus* is $(5 + 5)/2 + (11 + 14)/2 = 17.5$. SMR is weight-specific metabolic rate in unit of (ml $O_2 \cdot hr^{-1} \cdot g^{-1}$). When there are more than one SMR value reported for each species, the average value is used.

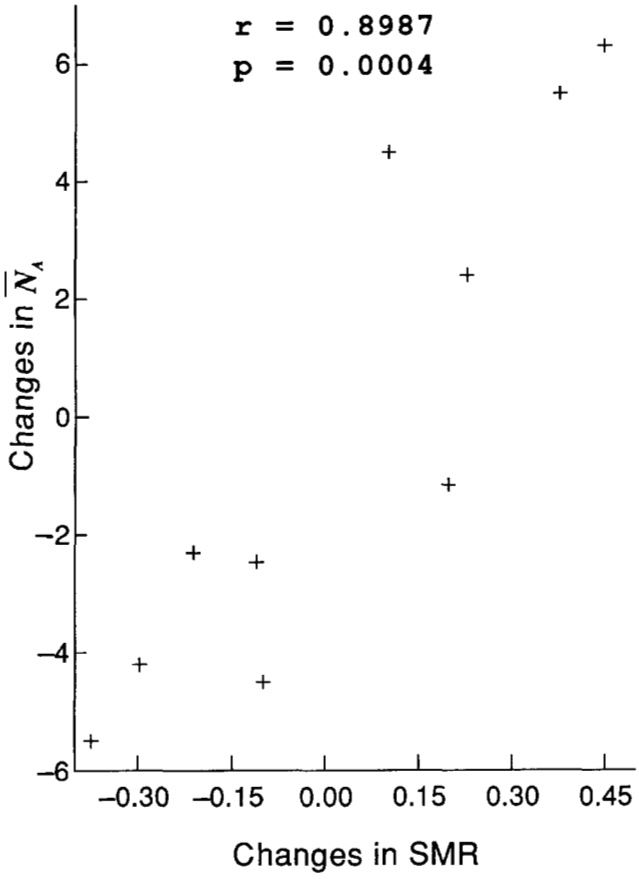

FIGURE 2.—Heuristic illustration of the effect of changes in weight-specific metabolic rate (SMR) on $\bar{N}_A$. The 10 points are taken from the 10 branches in Figure 1.

to the same value (PURVIS and RAMBAUT 1994). The independent contrasts show that an increase in SMR is associated with an increase in $\bar{N}_A$ (Figures 1 and 2), which is consistent with the third prediction.

There are two problems in the above test. First, the phylogenetic tree in Figure 1 is not universally agreed upon. Second, the test is based on globin genes only. One cannot make generalizations based on one or few genes. I have addressed these problems in several ways. First, I have compared codon usage between the mouse (SMR = 1.59) and the rat (SMR = 0.84) based on 877 and 833 genes, respectively, from the mouse and the rat (Table 5). The mouse genes used A-ending codons significantly more frequently than the rat genes (P = 0.0006, Table 5). These data strongly supported the third prediction.

I have also compared codon usage between the rabbit (SMR = 0.47) and the cow (SMR = 0.13) based on 133 and 261 genes, respectively, from the rabbit and the cow. There are 36417 and 58199 codons for the rabbit and the cow, respectively, of which 7614 and 11762 codons, respectively, are A-ending codons. A chi-square test showed that rabbit genes contained significantly more A-ending codons than the cow genes (χ^2 = 6.699, d.f. = 1, P = 0.0096), which is again consistent with the third prediction. A similar comparison between the human (1952 genes) and the macaque (19 genes) did not show any significant difference, which is perhaps attributable to the small number of macaque genes and to the fact that the difference in SMR between the two species are not as great as that between the rat and the mouse or between the rabbit and the cow. In short, when two species differ much in SMR, they also differ in the use of A-ending codons in the direction predicted by THCU; when two species differ little in SMR, they also have similar codon usage.

Prediction 4: The A-content of introns should be greater in organisms with a high weight-specific metabolic rate (SMR) than in organisms with a low SMR: This prediction is interesting in two aspects. First, its confirmation would strengthen THCU. Second, it helps to distinguish between the transcription hypothesis and translational hypothesis concerning codon usage. The transcription hypothesis predicts that both introns and coding sequences should show the predicted "nucleotide usage bias", whereas the translation hypothesis predicts that only coding sequences should exhibit nucleotide usage bias (or codon usage bias).

The test of prediction 4 can be illustrated with the cytoplasmic β-actin gene, which has been sequenced for the human and the rat, with GenBank LOCUS names HUMACCYBB and RATACCYB, respectively. The gene from both the human and the rat contain five introns, which are spliced out, joined, and the percentage of A nucleotide calculated. Because the rat has a much high metabolic rate (0.84) than the human (0.23), prediction 4 would be supported if the introns of the rat gene have a higher percentage of A nucleotide than those of the human gene. Such a test should be applied to many genes to increase the generality of the test results.

I retrieved DNA sequences from GenBank by using proteins listed in Table 1 of chapter 4 in LI and GRAUR (1991) as keywords, which resulted in a total of 756 DNA sequences for various mammalian species. Most protein gene "loci" in GenBank do not contain sequence information on introns. Some genes have intron sequences for only one species, which are useless for our comparative purpose (which requires intron information from at least two species differing in metabolic rate, SMR). Some genes contain only partial intron sequences, which are discarded. Also discarded are those intron sequences with long stretches of unresolved sites, *i.e.*, marked by "nnnnn . . .". For the few genes that do contain complete intron sequences from multiple species, only five (skeletal α-actin, cytoplasmic β-actin, growth hormone, α- and β-globin genes) can have their exons and introns aligned properly as shown in Figure 3.

The percentage of A nucleotide in introns for each of the five genes representing multiple mammalian species was displayed in Table 6, together with the corresponding SMR values. Although the data are limited,

TABLE 5

Comparison of codon usage between the mouse and the rat based on 877 mouse genes and 833 rat genes

		Mouse			Rat		
AA	CF	N_{cod}	P_A	P_T	N_{cod}	P_A	P_T
Gly	GGN	17077	0.2724	0.1778	16008	0.2508	0.1778
Glu	GGR	15886	0.4070		16173	0.3955	
Val	GTN	14710	0.1054	0.1615	15063	0.1032	0.1587
Ala	GCN	16160	0.2142	0.3033	16860	0.2109	0.3134
Arg	AGR	5879	0.5086		5397	0.5006	
Lys	AAR	14331	0.3725		14543	0.3521	
Ile	ATH	10820	0.1335	0.3372	11406	0.1198	0.3275
Thr	ACN	13700	0.2738	0.2485	13338	0.2725	0.2411
End	TAR	481	0.5904		445	0.5775	
Leu	TTR	4042	0.2853		3938	0.2819	
Ser	TCN	11969	0.2338	0.3112	11112	0.2231	0.3081
Arg	CGN	7323	0.2058	0.1632	7104	0.2051	0.1724
Gln	CAR	10918	0.2806		9940	0.2499	
Pro	CCN	14777	0.2843	0.2997	13186	0.2872	0.2922

AA, amino acid; CF, codon family; N_{cod}, total number of codons; P_A and P_T, proportion of A- and T-ending codons. P_A for mouse is significantly larger than P_A for rat (paired t-test, $P = 0.0006$). Data are based on a compilation by J. M. CHERRY (cherry@frodo.mgh.harvard.edu) with the GCG program CodonFrequency. The original compilation, together with a description of the compilation procedure, is available at the FTP site ftp.bio.indiana.edu.

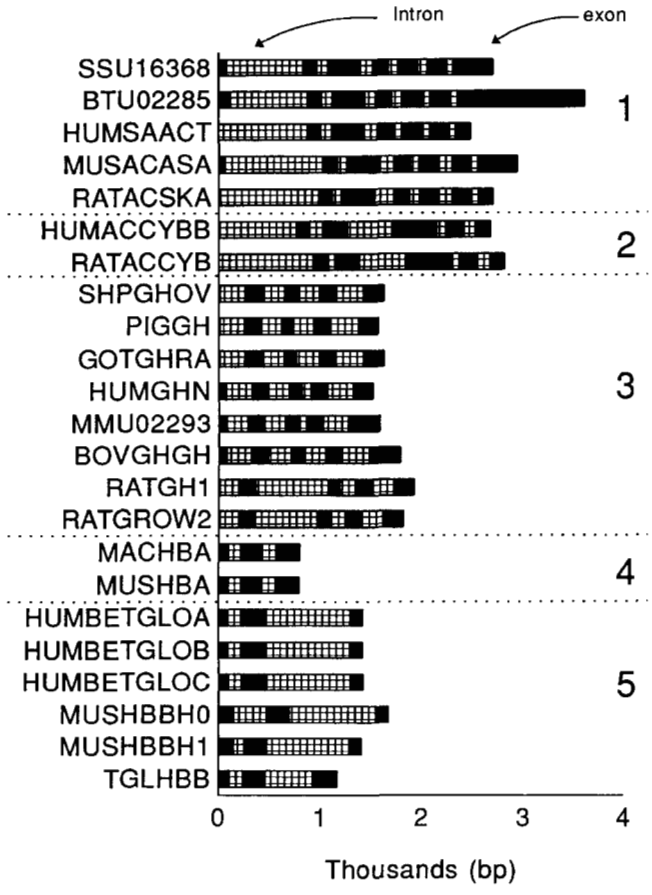

FIGURE 3.—Aligned exons and introns for five protein genes from various mammalian species. The numbers designate genes: 1, skeletal α-actin gene; 2, cytoplasmic β-actin gene; 3, growth hormone gene; 4, α-globin gene; 5, β-globin gene. GenBank LOCUS names are on the y axis. See Table 6 for the common name of the species.

we do find a consistent pattern for each of the five genes that the A-content of introns is greater in organisms with a high weight-specific metabolic rate (SMR) than in organisms with a low SMR (Table 6). For example, rodents have higher P_A than the human and the ungulates.

AN ALTERNATIVE EXPLANATION

One alternative hypothesis for the patterns shown in Tables 1–6 and Figures 1 and 2 is that of mutation bias. For example, a greater mutational pressure favoring A against G in the mitochondrial genomes than in the nuclear genomes would result in a greater proportion of A-ending codons in the mitochondrial genes than in the nuclear genes. MARTIN (1995) has argued that organisms of high metabolic rate should experience higher mutation rate favoring A than organisms of low metabolic rate.

The mutation hypothesis can be distinguished from THCU because the two hypotheses have different predictions. Let us first focus on the consequence of mutation favoring A against G. Suppose a protein gene with equal number of A, C, G, and T distributed randomly on both template and nontemplate strands (*i.e.*, the original sequence in Figure 4). When five Gs are replaced by five As through mutation on the template strand, five Cs will consequently be replaced by five Ts on the nontemplate strand. Because mutation occurs randomly on both template and nontemplate strand of the gene, we also expect five Gs to be replaced by five As on the nontemplate strand and five Cs to be consequently replaced by five Ts on the template strand. The net result is that on either template or nontemplate

TABLE 6

Data from five protein genes for testing prediction 4

Species	Locus	SMR	P_A	P_C	P_G	P_T	N_{total}
Skeletal muscle α-actin gene							
Pig	SSU16368	0.126	0.1609	0.3451	0.3073	0.1866	1243
Cow	BTU02285	0.127	0.1579	0.3607	0.3070	0.1744	1267
Human	HUMSAACT	0.228	0.1539	0.3428	0.3182	0.1851	1345
Rat	RATACSKA	0.84	0.1954	0.3072	0.2612	0.2363	1566
Mouse	MUSACASA	1.59	0.2029	0.2837	0.2710	0.2423	1498
Cytoplasmic β-actin gene							
Human	HUMACCYBB	0.228	0.1099	0.3174	0.3555	0.2172	1547
Rat	RATACCYB	0.84	0.1533	0.2763	0.3053	0.2651	1690
Growth hormone gene							
Pig	PIGGH	0.126	0.1726	0.2729	0.3538	0.2006	927
Cow	BOVGHGH	0.127	0.1715	0.2793	0.3439	0.2053	974
Sheep	SHPGHOV	0.200	0.1801	0.2774	0.3327	0.2098	977
Goat	GOTGHRA	0.233	0.1859	0.2727	0.3330	0.2084	979
Human	HUMGHN	0.228	0.2204	0.2648	0.2993	0.2155	812
Macaque	MMU02293	0.43	0.2212	0.2552	0.3063	0.2173	764
Rat	RATGH1	0.84	0.2643	0.2533	0.2502	0.2322	1169
Rat	RATGROW2	0.84	0.2575	0.2566	0.2472	0.2387	1275
α-Globin gene							
Macaca	MACHBA	0.43	0.1226	0.4176	0.3180	0.1418	261
Mouse	MUSHBA	1.59	0.1992	0.2695	0.3008	0.2305	256
β-Globin gene							
Human	HUMBETGLOA	0.228	0.2776	0.1673	0.1622	0.3929	980
Human	HUMBETGLOB	0.228	0.2776	0.1663	0.1633	0.3929	980
Human	HUMBETGLOC	0.228	0.2776	0.1663	0.1633	0.3929	980
Mouse	MUSHBBH0	1.59	0.3016	0.1665	0.2048	0.3271	1177
Mouse	MUSHBBH1	1.59	0.3031	0.1411	0.2243	0.3315	914
Echidna	TGLHBB	0.225[a]	0.2242	0.2144	0.3191	0.2422	611

LOCUS designates LOCUS name in GenBank. P_A, P_C, P_G, and P_T are the proportion of nucleotides A, C, G, and T, respectively. N_{total} is the total number of codons. Phylogenetically similar species are grouped next to each other. Note that species with higher SMR values tend to have higher A-content (P_A). All SMR (weight-specific metabolic rate) values are taken from ALTMAN and DITTMER (1972: 1613–1616).

[a] The reported value is a range (0.20–0.25).

strand, the increment in the number of A nucleotides (five in our fictitious case) is matched by the increment in the number of T nucleotides (also five in our fictitious case). In other words, $\Delta N_A = \Delta N_T$ on both template and nontemplate strands (Figure 4), so that A- and T-ending codons will be used equally frequently, and both used more frequently than G- and C-ending codons.

In contrast to the mutation hypothesis, THCU predicts that, with ATP more readily available than other nucleotides, the protein gene should evolve toward maximizing the use of A in mRNA (*i.e.*, maximizing the number of A on the nontemplate strand of the coding sequence, Figure 4). This will result in an increase in the number of A, and a decrease in the number of T in the nontemplate strand of the gene (Figure 4). In short, although both THCU and the mutation hypothesis would predict that A-ending codons should be much more frequent than G-ending codons, the two hypotheses differ in that THCU predicts A-rich and T-poor on the nontemplate strand, whereas the mutation hypothesis (*e.g.*, with mutation favoring A against G) predicts that both strands should be AT-rich and GC-poor, with As and Ts distributed equally on the two strands (Figure 4).

The mutation hypothesis seems to explain satisfactorily the pattern of codon usage in Drosophila mitochondrial DNA. The number of codons ending with A, C, G, and T in *Drosophila yakuba* is 1052, 107, 45, and 1092, respectively, for protein genes on the H strand, and is 403, 6, 31, and 428, respectively, for protein genes on the L strand. Thus, Drosophila mtDNA is AT-rich, with A- and T-ending codons used roughly equally, and both used much more frequently than C- and G-ending codons. These fulfill the prediction based on mutation hypothesis (Figure 4, top). In neither strand do we observe A-richness and T-poorness expected from THCU (Figure 4).

Additional evidence confirming that codon usage in Drosophila is mainly controlled by mutations favouring A or T comes from an AT-rich region flanking the origin of replication. This region spans 1.0–5.1 kb and is homologous in various Drosophila species (GODDARD and WOLSTENHOLME 1980; FAURON and WOLSTENHOLME 1980a,b). The region exhibits extensive se-

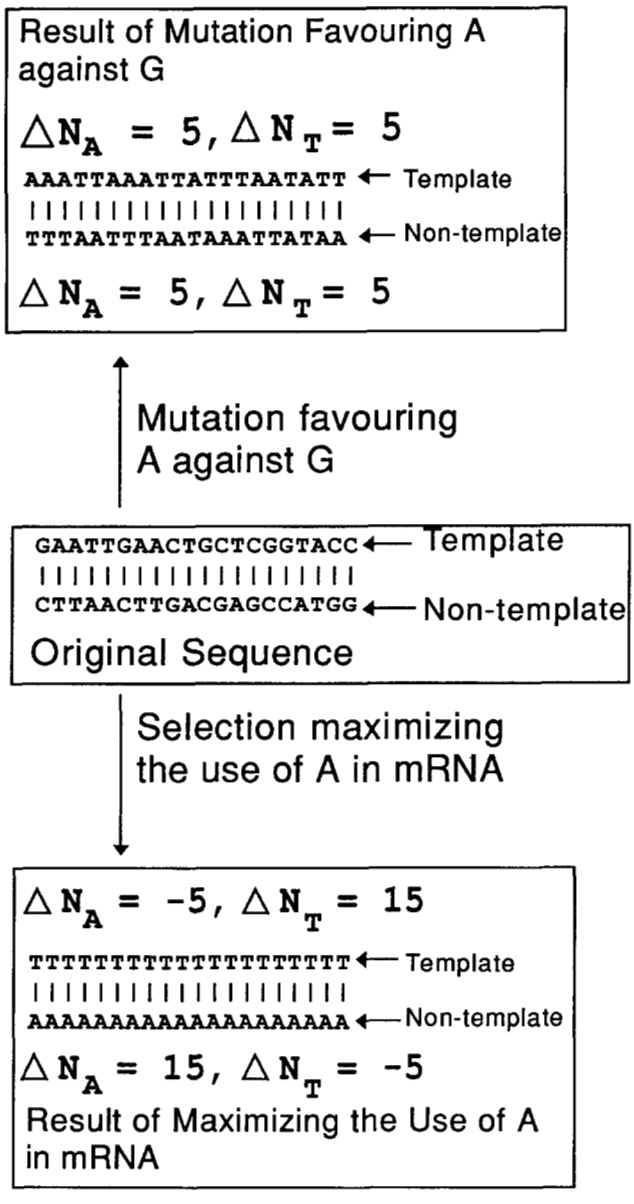

FIGURE 4.—Contrasting predictions from the mutation hypothesis (with mutation favoring A against G) and THCU (transcription hypothesis of codon usage). THCU predicts A-richness and T-poorness in the nontemplate strand (bottom), and the opposite in the template strand. Δ designates increment. Because mutation favoring A against G is expected to occurred equally frequently on both DNA strands, the mutation hypothesis expects both DNA strands to accumulate equal number of As and, consequently, equal number of Ts, so that both strands will be AT-rich and GC-poor.

quence divergence, suggesting that the nucleotide sequence is mainly under the control of mutation bias (GODDARD *et al.* 1982). The fact that the region is made of almost entirely of AT pairs implies that the mutation spectrum in Drosophila is strongly AT-biased, and that the preponderance of A- and T-ending codons in Drosophila mtDNA can be explained as a consequence of the mutation bias.

Another DNA region that appears to be strongly affected by mutation bias is the D-loop of mammalian mtDNA. GODDARD *et al.* (1982) has suggested that the D-loop is homologous to the highly variable AT-rich region in Drosophila mtDNA mentioned above. Like the AT-rich region in Drosophila, the D-loop also flanks the replication origin, is also highly variable in nucleotide sequences (AVISE 1994), and is not transcribed except for perhaps a few bases. Thus, the nucleotide composition of the D-loop should reflect the mutation spectrum in the mammalian mtDNA. The number of A, C, G, and T in the mouse D-loop is 258, 104, 218, and 299, respectively. This is consistent with what we would expect if the D-loop is under mutation bias favoring A or T (Figure 4, top).

The mutation hypothesis, however, fails in explaining the pattern of codon usage in mammalian mtDNA. The data in Table 1 shows that A-ending codons are always much more frequently used than T (or U)-ending codons in bovine mtDNA, in contrast to what we see in Drosophila mtDNA where A- and T-ending codons are used equally frequently and also in contrast to the D-loop region where T is more frequent than A.

The data from mouse mtDNA further highlight the inadequacy of the mutation hypothesis. The number of codons ending with A, C, G, and T in the mouse mtDNA is 1677, 1000, 117, and 825, respectively, with A-ending codons far outnumbering not only G-ending codons but also T-ending codons. This pattern is the same as what we see in Table 1 for the cow and is expected from THCU, but not from the mutation hypothesis. (Note that there are more NNY codons than NNR codons in mammalian mtDNA, with the difference >300. So the observed excess of A-ending codons and deficiency of T-ending codons in mammalian mtDNA is not a consequence of protein genes made of mostly NNR codons).

Although the pattern of codon usage in mtDNA is more satisfactorily explained by THCU than by the mutation hypothesis, one can still argue that the difference in codon usage between mtDNA and nuclear DNA (Table 2) is attributable to mutations in mtDNA more biased in favor of A than mutations in nuclear genome, which could result in more A-ending codons in mtDNA than in nuclear genome. A new finding summarized below appears to favor THCU.

ZISCHLER *et al.* (1995) discovered a segment (540 bp) of the human mitochondrial D-loop to have been inserted into the nuclear genome, and that the inserted sequence has presumably existed as nonfunctional DNA. The nucleotide frequencies of the insert for A, C, G, and T are 30.7%, 32.6%, 13.9%, and 22.8%, respectively. The equivalent values for the homologous 540-bp D-loop segment (from LOCUS HUMMTCG in GenBank) are 30.4%, 32.8%, 14.1%, and 22.8%, respectively. If mutations are more biased in favor of A in mtDNA than in nuclear genome, then we should expect

a reduction of the proportion of A in the insert, which is not true.

ZISCHLER *et al.* (1995) also sequenced the nuclear DNA sequences flanking the insert. The two flanking regions add up to a total of 385 bp, with the nucleotide frequencies being 41.3%, 18.2%, 13.5%, and 27.0%, respectively, for A, C, G, and T. Thus, the A-content of the nonfunctional DNA of nuclear origin appears to be in excess rather than in deficiency in comparison with the equivalent values in mitochondrial D-loop. This suggests that mutations in mtDNA is not more biased in favor of A than those in the nuclear genome. In short, the larger proportion of A-ending codons in mtDNA relative to nuclear DNA is not due to mutation bias favoring A in mtDNA.

It is much more difficult to distinguish THCU from the mutation hypothesis regarding the differences in codon usage among mammalian species of different metabolic rates (Tables 3–5). For example, although the proportion of A-ending codons (P_A in Table 5) is greater for mouse genes than for rat genes, the proportion of T-ending codons (P_T) also seems to be greater for the mouse genes than for the rat genes. This concurrent increase in both P_A and P_T in animals of higher metabolic rate (*i.e.*, the mouse) is compatible with the mutation hypothesis (MARTIN 1995) invoking mutation bias favoring A or T in animals of higher metabolic rate (SRM). However, for the nine codon families with both A- and T-ending codons in Table 5, P_A for the mouse genes is significantly larger than P_A for the rat genes ($P = 0.017$, paired t-test, one-tailed), whereas the difference in P_T between the mouse genes and the rat genes is not significant ($P = 0.507$). This suggests that THCU is a plausible alternative to the mutation hypothesis.

The data of introns (Table 6) are almost entirely compatible with the mutation hypothesis in that a concurrent increase in both A- and T-content is observed in genes from mammalian species with a high metabolic rate relative to those from mammalian species with a low metabolic rate. The only exception involves comparisons between human and mouse for the β-globin gene. The mouse introns for the β-globin gene show higher A content and lower T content than human introns. This is expected under THCU, but not under the mutation hypothesis. However, such a single case should not be taken as a rejection of the mutation hypothesis, which to me remains a plausible hypothesis in many other cases.

I conclude that THCU is a sufficient, and perhaps unique, explanation for the biased codon usage favoring A-ending codons in mammalian mtDNA (Table 1) and the differences in codon usage between the mitochondrial genomes and the nuclear genomes (Table 2). My results further suggest that THCU is a plausible hypothesis in explaining the differences in codon usage in nuclear genomes among mammalian species of different metabolic rates (Table 3–6 and Figures 1 and 2).

DISCUSSION

The prevailing hypothesis on the evolution of codon usage suggests that the pattern of synonymous codon bias is a consequence of adaptation of codon usage to relative availability of tRNAs in the cellular matrix (reviewed by IKEMURA 1992). A more relaxed hypothesis invokes the mutual adaptation of codon usage and tRNA availability (BULMER 1988). According to this second hypothesis, there are three elements in the system determining the evolution of codon usage: mutation bias, tRNA availability, and random genetic drift (BULMER 1991). Random genetic drift could lead to biased codon usage and unequal availability of different tRNAs in the absence of natural selection. If a synonymous codon that drifts to high frequency happens to be the one recognized by the most abundant tRNA or if a tRNA that drifts to high abundance happens to be the one that recognizes the most frequently used codon, then these genetic drifts would result in increased translational efficiency and accuracy, and would therefore be favored by natural selection. This would ultimately result in the most frequently used synonymous codon being recognized by the most abundant tRNAs, which has been documented by GOUY and GAUTIER (1982) and IKEMURA (1981, 1982, 1985, 1991). In short, the second hypothesis suggests that mutation bias, tRNA availability and random genetic drift form a self-contained system such that the interaction among the three elements is sufficient to explain the pattern of codon usage.

The results in this paper indicate that this second hypothesis is too restrictive because some features of codon usage, such as the usage of A-ending codons, depend on factors that are not contained in the system of the three elements specified in that hypothesis. Specifically, our optimality model of the transcriptional process predicts that the pattern of synonymous codon usage should depend on the relative concentration of nucleotides in the cellular medium. This is consistent with the findings that the mitochondrial genome has a greater proportion of A-ending codons than the nuclear genome and that the nuclear genome in organisms with a high metabolic rate has a greater proportion of A-ending codons than the nuclear genome in organisms with a low metabolic rate. Thus, a more complete theory of the evolution of codon usage should consider the relative availability of ribonucleotides in the cellular matrix.

One potential misunderstanding concerning THCU and its predictions on biased usage of ATP in transcription is that, because ATP and GTP were used as energy sources in cellular processes, the use of ATP would tend to deplete available energy sources. The benefit of using ATP to enhance the transcription would consequently be offset by the cost of depleting the available energy sources. This argument arises from a misunderstanding

that CTP and UTP can come free without spending ATP to synthesize them. It is in fact energetically more efficient to use ATP directly to fill a nucleotide site than to use ATP to synthesize an alternative NTP and then use that alternative NTP to fill in the nucleotide site. In other word, using ATP directly in transcription not only speeds up transcription, it also *conserves* available energy sources.

I should admit here that, although predictions from the model appear consistent with empirical data, the construction of the model itself is not vigorous because of simplifying assumptions. Protein synthesis is a multistep process including initiation of transcription, elongation of mRNA chain, initiation of translation, and elongation of the peptide chain. By assuming that the rate of transcription rate is limiting, we have reduced the multistep process to a one-step process, which obviously is a distortion of the reality. This criticism can also be levelled against those studies focusing on the translation component of the process. However, the recognition that even a very simple model could account for a substantial amount of variation in codon (nucleotide) usage would help to reduce the mystique surrounding the operation of natural selection on the biochemical systems in the living cell.

I thank M. S. HAFNER, A. ZHARKIKH, T. SPRADLING, J. DEMASTES, and D. W. FOLTZ for their stimulating discussion and constructive comments on various versions of the manuscript. The two anonymous reviewers and JODY HEY suggested the analysis on introns and pointed to additional sources of data relevant to the hypothesis presented in this paper. This project is supported by National Science Foundation grant DEB95–27583 to M. S. HAFNER and X.X. and by the University of Hong Kong.

LITERATURE CITED

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

ALTMAN, P. L., and D. S. DITTMER, 1972 *Biology Data Book,* Vol. III. Ed. 2. Federation of American Societies for Experimental Biology, Bethesda, MD.

AVISE, J. C., 1994 *Molecular Markers, Natural History and Evolution.* Chapman and Hall, New York.

BENNETZEN, J. L., and B. D. HALL, 1982 Codon selection in yeast. J. Biol. Chem. **257:** 3026–3031.

BRIDGER, W. A., and J. F. HENDERSON, 1983 *Cell ATP.* Wiley, New York.

BULMER, M., 1988 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728–730.

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

EKERT, R., and D. RANDALL, 1983 *Animal Physiology.* Ed. 2. Freeman, New York.

FAURON, C. M. R., and D. R. WOLSTENHOLME, 1980a Extensive diversity among Drosophila species with respect to nucleotide sequences within the adenine + thymine-rich region of mitochondrial DNA molecule. Nucleic Acids Res. **8:** 2439–2452.

FAURON, C. M.-R., and D. R. WOLSTENHOLME, 1980b Intraspecific diversity of nucleotide sequences within the adenine+thymine-rich region of mitochondrial DNA molecules of *Drosophila mauritiana, Drosophila melanogaster* and *Drosophila simulans.* Nucleic Acids Res. **8:** 5391–5410.

FELSENSTEIN, J., 1985 Phylogenies and the comparative method. Am. Nat. **125:** 1–15.

FELSENSTEIN, J., 1988 Phylogenies and quantitative methods. Annu. Rev. Ecol. Syst. **19:** 445–471.

GODDARD, J. M., and D. R. WOLSTENHOLME, 1980 Origin and direction of replication in mitochondrial DNA molecules from the genus Drosophila. Nucleic Acids Res. **8:** 741–757.

GODDARD, J. M, FAURON, C. M.-R. FAURON and D. R. WOLSTENHOLME, 1982 Nucleotide sequences within the A+T-rich region and the large-rRNA gene of mitochondrial DNA molecules of *Drosophila yakuba,* pp. 99–103 in *Mitochondrial Genes,* edited by P. SLONIMSKI, P. BORST and G. ATTARDI. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

GOUY, M., and C. GAUTIER, 1982 Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. **10:** 7055–7064.

GRANTHAM, R., C. GAUTIER, M. GOUY, R. MERCIER and A. PAVE, 1980 Codon catalog usage and the genome hypothesis. Nucleic Acids Res. **8:** 49–79.

GRANTHAM, R., C. GAUTIER, M. GOUY, M. JACOBZONE and R. MERCIER, 1981 Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. **9:** 43–79.

HARTL, D. L., E. N. MORIYAMA and S. A. SAWYER, 1994 Selection intensity for codon bias. Genetics **1138:** 227–234.

HARVEY, P. H., and M. D. PAGEL, 1991 *The Comparative Method in Evolutionary Biology.* Oxford University Press, Oxford.

HOCHACHKA, P., 1991 Design of energy metabolism, pp. 353–436 in *Environmental and Metabolic Animal Physiology,* edited by C. L. PROSSER. Wiley-Liss, New York.

IKEMURA, T., 1981 Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system. J. Mol. Biol. **151:** 389–409.

IKEMURA, T., 1982 Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. **158:** 573–597.

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

IKEMURA, T., 1992 Correlation between codon usage and tRNA content in microorganisms, pp. 87–111 in *Transfer RNA in Protein Synthesis,* edited by D. L. HATFIELD, B. J. LEE, and R. PIRTLE. CRC Press, Boca Raton, Fla.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge.

KURLAND, C. G., 1987a Strategies for efficiency and accuracy in gene expression. I. The major codon preference: a growth optimization strategy. Trends Biochem. Sci. **12:** 126–128.

KURLAND, C. G., 1987b Strategies for efficiency and accuracy in gene expression. 2. Growth optimized ribosomes. Trends Biochem. Sci. **12:** 169–171.

LI, W.-H., and D. GRAUR, 1991 *Fundamentals of Molecular Evolution.* Sinauer Associates, Sunderland, MA.

MARTIN, A. P., 1995 Metabolic rate and directional nucleotide substitution in animal mitochondrial DNA. Mol. Biol. Evol. **12:** 1124–1131.

MATHIEU, O., R. KRAUER, H. HOPPELER, P. GEHR, S. L. LINDSTEDT *et al.,* 1981 Design of the mammalian respiratory system. VI. Scaling mitochondrial volume in skeletal muscle to body mass. Respir. Physiol. **44:** 113–128.

NOVACEK, M. J., A. R. WYSS and M. MCKENNA, 1988 The major groups of eutherian mammals, pp. 31–71 in *The Phylogeny and Classification of the Tetrapods,* Vol. 2, edited by M. J. BENTON. Clarendon Press, Oxford.

OLSON, M. S., 1986 Bioenergetics and oxidative metabolism. pp 212–260 in *Text Book of Biochemistry with Clinical Correlations,* Ed. 2, edited by T. M. DEVLIN. John Wiley & Sons, New York.

PURVIS, A., and A. RAMBAUT, 1994 *Comparative Analysis by Independent Contrasts* (CAIC, version 2). Oxford University.

ROBINSON, M., R. LILLEY, S. LITTLE, J. S. EMTAGE, G. YAMAMOTO *et al.,* 1984 Codon usage can effect efficiency of translation of genes in *Escherichia coli.* Nucleic Acids Res. **12:** 6663–6671.

SHARP, P. M., and K. M. DEVINE, 1989 Codon usage and gene expression level in *Dictyostelium discoideum*: highly expressed genes do "prefer" optimal codons. Nucleic Acids Res. **17:** 5029–5038.

SHARP, P. M., and W. H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

SHARP, P. M., and W. H. LI, 1987 The codon adaptation index - a

measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15:** 1281–1295.

SHARP, P. M, M. F. TUOHY and K. R. MOSURSKI, 1986 Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14:** 5125–5143.

SHARP, P. M., E. COWE, D. G. HIGGINS, D. C. SHIELDS, K. H. WOLFE *et al.*, 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.

SMITH, R. E., 1956 Quantitative relations between liver mitochondria metabolism and total body weight in mammals. Ann. NY Acad. Sci. **62:** 403–422.

SORENSEN, M. A., C. G. KURLAND, and S. PEDERSEN, 1989 Codon usage determines translation rate in *Escherichia coli.* J. Mol. Biol. **207:** 365–377.

WEIBEL, E. R., 1984 *The Pathway for Oxygen: Structure and Function in the Mammalian Respiratory System.* Harvard Univ. Press, Cambridge.

XIA, X., 1995 Body temperature, rate of biosynthesis and evolution of genome size. Mol. Biol. Evol. **12:** 834–842.

ZISCHLER, H., H. GEISERT, A. VON HAESELER, and S. PÄÄBO, 1995 A nuclear "fossil" of the mitochondrial D-loop and the origin of modern humans. Nature **378:** 489–492.

Communicating editor: W-H. LI