Copyright © 1998 by the Genetics Society of America

How Optimized Is the Translational Machinery in *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae*?

Xuhua Xia

Evolutionary Genetics Group, Department of Ecology and Biodiversity, The University of Hong Kong, Hong Kong

Manuscript received September 23, 1997
Accepted for publication January 21, 1998

ABSTRACT

The optimization of the translational machinery in cells requires the mutual adaptation of codon usage and tRNA concentration, and the adaptation of tRNA concentration to amino acid usage. Two predictions were derived based on a simple deterministic model of translation which assumes that elongation of the peptide chain is rate-limiting. The highest translational efficiency is achieved when the codon recognized by the most abundant tRNA reaches the maximum frequency. For each codon family, the tRNA concentration is optimally adapted to codon usage when the concentration of different tRNA species matches the square-root of the frequency of their corresponding synonymous codons. When tRNA concentration and codon usage are well adapted to each other, the optimal content of all tRNA species carrying the same amino acid should match the square-root of the frequency of the amino acid. These predictions are examined against empirical data from *Escherichia coli*, *Salmonella typhimurium*, and *Saccharomyces cerevisiae*.

SYNONYMOUS codon usage differs among different genomes (Grantham *et al.* 1980, 1981; Moriyama and Hartl 1993; Martin 1995; Xia 1996), among different genes within the same genome (Gouy and Gautier 1982; Ikemura 1985, 1992; Sharp and Li 1986, 1987; Sharp *et al.* 1988), and even among different segments of the same gene (Akashi 1994). Three hypotheses have been proposed to account for this variation of synonymous codon usage (or various components of the variation): the mutation bias hypothesis (Martin 1995), the transcription-maximization hypothesis (Xia 1996) and translational efficiency hypothesis (Ikemura 1981; Kimura 1983; Robinson *et al.* 1984; Kurland 1987a,b; Bulmer 1987, 1988, 1991).

Of these three hypotheses, the translational efficiency hypothesis (hereafter referred to as TEH) is the most general and has received the most empirical support. In verbal forms, the hypothesis states that there is strong selection favoring increased rate of protein synthesis and that a coding strategy that increases the rate of translation initiation and peptide elongation (and consequently increases the rate of protein synthesis) is favoured by natural selection. The hypothesis is favored by three independent lines of evidence. First, the frequency of codon usage is positively correlated with tRNA availability (Ikemura 1981, 1982, 1985, 1992; Gouy and Gautier 1982). 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; Ikemura 1985; 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 (Robinson *et al.* 1984; Sorensen *et al.* 1989).

Many models of TEH have been presented that can be called either initiation models or elongation models. Initiation models assume that the initiation of translation is rate-limiting, *e.g.*, Liljenström and von Heijne (1987); Bulmer (1991); Xia (1996), whereas elongation models assume that the elongation of the peptide chain is rate-limiting, *e.g.*, Varenne *et al.* (1984); Bulmer (1987). Empirical data and theoretical considerations suggest that both initiation and elongation are rate-limiting.

The model presented here is strictly a deterministic elongation model, because I think that previous elongation models are not well presented and that expectations are often only vaguely specified. This has resulted in some confusion. For example, Kimura (1983) assumed that the translational efficiency is maximized when the proportion of different synonymous codons matches exactly the proportion of isoaccepting tRNAs. The assumption is unwarranted, and the translational efficiency, given the perfect matching, will be shown later to be the same as the presumably less adaptive scenario when different tRNA species are present in equal amount and codon usage drifts freely in any direction.

Another reason for presenting the model is to relate amino acid usage to the availability of tRNA species carrying different amino acids. From an evolutionary point of view, one would intuitively expect an efficient translational machinery to have more tRNA coding for more frequently used amino acids, but this intuition has not been formally established or rejected.

Address for correspondence: Xuhua Xia, Evolutionary Genetics Group, Department of Ecology & Biodiversity, The University of Hong Kong, Pokfulam Road, Hong Kong. E-mail: xxia@hkusua.hku.hk

Below I present the elongation model, from which a few specific predictions concerning mutual adaptation between tRNA content and codon usage are derived. Also derived is a relationship between tRNA content and amino acid usage. Empirical data from *Escherichia coli*, *Salmonella typhimurium* and *Saccharomyces cerevisiae* were used to test the predictions.

THE ELONGATION MODEL, ITS PREDICTIONS, AND EMPIRICAL TESTS

Consider the time required to translate a single codon coding for amino acid i (AA_i, $i = 1, 2, \ldots, 20$). Designate this codon as SC_{ij} ($j = 1, 2, \ldots, n_i$, where n_i is the number of synonymous codons for AA_i). Let r be the rate of aminoacyl-tRNA diffusing to the A site of the ribosome during translation, P_i be the probability that the arriving aminoacyl-tRNA carries AA_i ($\Sigma_{i=1}^{20} P_i = 1$), and p_{ij} be the conditional probability that the aminoacyl-tRNA recognizes the synonymous codon SC_{ij}, given that the tRNA carries AA_i ($\Sigma_{j=1}^{n_i} p_{ij} = 1$ for each given i). Let t_l be the time spent in linking the right amino acid to the elongating protein chain, and t_r be the time spent in rejecting each wrong aminoacyl-tRNA. Now the total time spent in translating SC_{ij} is

$$T_{ij} = \frac{1}{r\,P_i\,p_{ij}} + t_l + \left(\frac{1}{P_i\,p_{ij}} - 1\right)t_r, \qquad (1)$$

where the first term on the right-hand side of the equation is the time needed for an aminoacyl-tRNA carrying the right amino acid and the right cognate anti-codon to arrive at the A site of the ribosome and the third term represents time spent in rejecting all the wrong aminoacyl-tRNA prior to the arrival of the right aminoacyl-tRNA. Similar formulation can be found in Varenne *et al.* (1984) and Bulmer (1987). The total time (T) required to translate L codons (total elongation time) can be shown to be

$$T = \sum_{i=1}^{20} \sum_{j=1}^{n_i} f_{ij} T_{ij} = L(t_l - t_r) + \frac{1 + rt_r}{r} Y, \qquad (2)$$

where

$$Y = \sum_{i=1}^{20} \frac{1}{P_i} \sum_{j=1}^{n_i} \frac{f_{ij}}{p_{ij}} = \sum_{i=1}^{20} \frac{N_i}{P_i} \sum_{j=1}^{n_i} \frac{Q_{ij}}{p_{ij}}. \qquad (3)$$

The term f_{ij} is the frequency of synonymous codon j for amino acid i in the mRNA molecule ($\Sigma_{i=1}^{20} \Sigma_{j=1}^{n_i} f_{ij} = L$), N_i is the number of codons for amino acid i ($\Sigma N_i = L$; $\Sigma_{j=1}^{n_i} f_{ij} = N_i$), and Q_{ij} is the proportion of synonymous codon j for amino acid i in the mRNA molecule. Note that Q_{ij} is a property of the mRNA whereas P_i and p_{ij} are properties of the tRNA pool, with P_i being the proportion of tRNA carrying AA_i, and p_{ij} being the fraction of tRNA that recognizes synonymous codon j among all tRNA species that carry AA_i.

Our objective is to find the condition, *i.e.*, the relationship among Q_{ij}, P_i and p_{ij}, that minimizes T. Because t_l, t_r and L are not dependent on Q_{ij}, P_i and p_{ij}, they are treated as constants. Thus, minimizing T in Equation 2 is equivalent to minimizing Y. Specifically, we are interested in three relationships. First, given the relative availability of different tRNA (P_i and p_{ij}), find what pattern of codon usage (Q_{ij}) in the mRNA would minimize Y. Second, given the pattern of codon usage (Q_{ij}), find what values for P_i and p_{ij} would minimize Y. Third, given amino acid usage, find the distribution of P_i that would minimize Y. Intuitively, we would expect frequently used amino acids to correspond to large P_i values, but the exact relationship has not been derived, let alone tested against empirical evidence.

Adaptation of codon usage to tRNA content: Suppose that an mRNA molecule specifies N residues of the same amino acid with n synonymous codons, and that the associated frequency distribution of synonymous codons is Q_j ($\Sigma Q_j = 1$). For simplicity, we assume that there are also n types of tRNA species for the amino acid, with each type recognizing just one of the n synonymous codons. The proportion of the n types of tRNA species is p_j ($\Sigma p_j = 1$). Now we have Y (which is the term to be minimized) below:

$$Y = \frac{N}{P} \sum_{j=1}^{n} \frac{Q_j}{p_j}. \qquad (3)$$

First consider what values Q_j should take when $p_j = 1/n$. One might intuitively think that, to make full use of the equal availability of the n types of tRNA, Q_j should match p_j and should all be equal to $1/n$. This is false. When p_j values are equal, Y is equal to ($n*N/P$) no matter what value Q_j takes as long as Q_j values sum to 1. Thus, Q_j is a neutral character when p_j values are all equal. I reiterate this point because some confusion has been introduced by Kimura (1983) who wrongly assumed that the highest translational efficiency is achieved when the relative frequencies of synonymous codons exactly match those of the cognate tRNAs.

When p_j values are not equal, then the smallest Y is achieved when the codon recognized by the most abundant tRNA becomes fixed, with the consequent loss of other synonymous codons. To see this more clearly, we re-write Equation 3 as follows:

$$Y = \frac{N}{P} \frac{Q_1 \prod_{j=2}^{n} p_j + Q_2 \prod_{j=1, j\neq 2}^{n} p_j + Q_3 \prod_{j=1, j\neq 3}^{n} p_j + \ldots + Q_n \prod_{j=1, j\neq n}^{n} p_j}{\prod_{j=1}^{n} p_j} \qquad (4)$$

If p_1 is the largest of all p_j values, then the first Π term, *i.e.*, the one associated with Q_1, on the numerator of Equation 4 is the smallest of all Π terms. It is therefore obvious that minimization of Y in Equation 4 requires that Q_1 equal 1 and that all other Q_j values equal zero. This means that whenever the availability of different tRNA species (p_j) for an amino acid is different, the codon usage of this amino acid should evolve towards increasing the frequency of the synonymous codon that is recognized by the most abundant cognate tRNA spe-

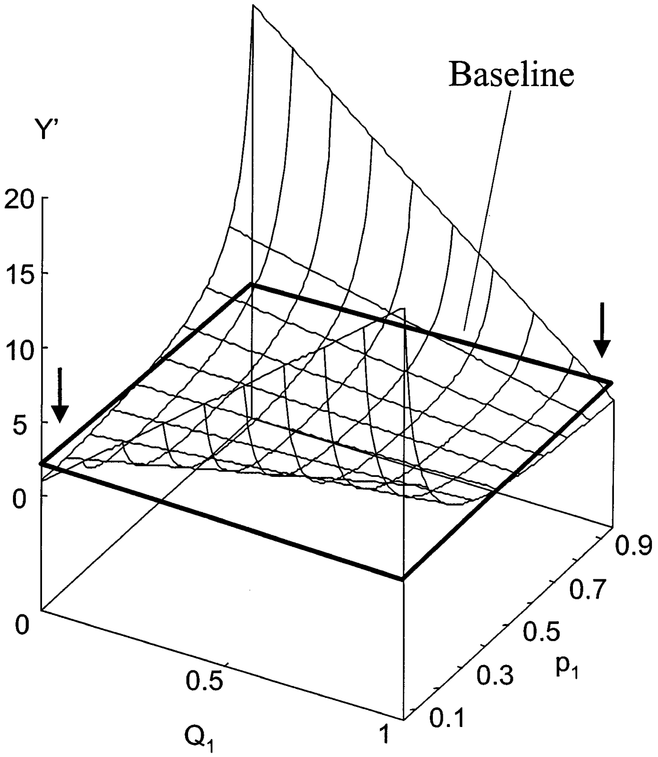

Figure 1.—Change in translational time in relation to Q_1 (the proportion of codon 1 in a two-codon family) and p_1 (the proportion of tRNA species recognizing codon 1). Y′ is the term within the parenthesis in Equation 6. The bolded plane perpendicular to Y′ represents the baseline condition, *i.e.*, the Y′ value for $p_1 = 0.5$. The downward arrows designate areas where Y′ is smaller than it is in the baseline condition.

cies. The minimum of Y achievable through adaptation of codon usage to tRNA content is

$$Y_{min} = \frac{N}{P\,p_1} = \frac{N}{P\,p_M}, \tag{5}$$

where p_M designates the most abundant tRNA species for the amino acid. Y_{min} reaches its minimum value when $p_M = 1$, which requires not only the adaptation of codon usage to tRNA content, but also adaptation of tRNA content to extremely biased codon usage.

For the special case with $n = 2$, Y in Equation 3 can be written as

$$Y = \frac{N}{P}\left(\frac{Q_1}{p_1} + \frac{1 - Q_1}{1 - p_1}\right). \tag{6}$$

The term within the parenthesis is plotted against p_1 and Q_1 (Figure 1). Two conclusions can be drawn. First, when p_j values are all equal to $1/n$, *i.e.*, when $p_1 = 0.5$ in Figure 1 for $n = 2$, then Q_j can take any value between 0 and 1 without affecting translational efficiency, and Y is relatively small. We will call this condition with equal p_j values the baseline condition. For unequal p_j values, *i.e.*, for $p_1 \neq 0.5$ in Figure 1, Y values will be larger than that in the baseline condition whenever Q_j values are smaller than p_j values for $p_j > 1/n$, *e.g.*, when $Q_1 = 0.8$ and $p_1 = 0.9$ in Figure 1, or larger than p_j values for $p_j < 1/n$, *e.g.*, when $Q_1 = 0.9$ and $p_1 = 0.1$, in Figure 1, in which case the reduction in translational efficiency, *i.e.*, the increase in Y, is outstanding (Figure 1). Y will be the same as that in the baseline condition when Q_j exactly matches p_j, *e.g.*, when $Q_1 = p_1$ in Figure 1. The baseline condition therefore seems to guarantee a relatively small Y value over a wide fluctuation of Q_j values. Y will be smaller than the baseline condition only when Q_j values are larger than p_j values for $p_j > 1/n$, *e.g.*, when $Q_1 = 0.9$ and $p_1 = 0.8$ in Figure 1, or smaller than p_j values for $p_j < 1/n$, *e.g.*, when $Q_1 = 0.1$ and $p_1 = 0.2$, in Figure 1.

We have now reached a specific and intuitively appealing prediction, that codon-usage bias should be more extreme than the bias in tRNA content. If p_j is larger than $1/n$, then Q_j should be larger than p_j; if p_j is smaller than $1/n$, then Q_j should be smaller than p_j. If this is not the case, then the translational efficiency is lower than that for the baseline condition.

An empirical test of this prediction has several requirements. First, we need codon families in which a codon will not be recognized by both the common and the rare tRNA, otherwise Q_j would be impossible to calculate in any meaningful way. Among the 23 codon families, *i.e.*, when we split each of the six-member codon families for Leu, Ser, and Arg into two, only six meet this criterion (Table 1). Secondly, we need codon usage of genes that are highly expressed, otherwise we should not expect any mutual adaptations between tRNA content and codon usage bias. Ikemura (1992) compiled codon usage of presumably highly expressed genes in *E. coli*, *S. typhimurium*, and *S. cerevisiae* (five, three and five genes, respectively), which are used to generate Table 1.

For all three species, the Q_M values are always larger than the p_M values (Table 1). This guarantees that the resulting Y is smaller than that in the baseline condition. The adaptation of codon usage to tRNA content in the highly expressed genes in the three unicellular species is almost perfect (the optimal is when $Q_M = 1$), suggesting that the effect of mutation on codon usage bias must be very weak for these genes. However, if we ignore the expressivity of the genes and pool the codon usage of all genes in the gene bank, then most Q_M values are smaller than the p_M values (data not shown), suggesting that, for most genes, the translational efficiency is lower than that in the baseline condition.

Adaptation of tRNA to codon usage: When Q_j values are fixed, *e.g.*, when codon bias is maintained by mutation bias, the values that p_j should take to minimize Y can be found as follows. We first re-write Y in Equation 3:

$$Y = \frac{N}{P}\left(\frac{Q_1}{1 - \sum_{j=2}^{n} p_j} + \sum_{j=2}^{n} \frac{Q_j}{p_j}\right). \tag{7}$$

The condition that minimizes Y is found by taking par-

TABLE 1

Adaptation of codon usage to tRNA content in *E. coli, S. typhimurium* and *S. cerevisiae*

				E. coli			*S. typhimurium*			*S. cerevisiae*		
AA	$\mathrm{ID_{tRNA}}$	Codon	$\mathrm{N_{tRNA}}$	tRNA	p_M	Q_M	tRNA	p_M	Q_M	tRNA	p_M	Q_M
Gly	3	GGU,GGC	4	1.10	0.815	0.995	0.90	0.750	0.982			
	2	GGA,GGG	1	0.15			0.20					
	1	GGG	1	0.10			0.10					
Ala	1	GCU,GCA,GCG	3	1.00	0.769	0.964	1.00	0.769	0.822			
	2	GCC	2	0.30			0.30					
Arg	2(1)	CGU,CGC,CGA	4	0.90	0.973	1	0.70	0.966	1			
	CGG	CGG	1	0.025[a]			0.025[a]					
Ile	1	AUU,AUC	3	1.00	0.952	1	1.00	0.952	1			
	2	AUA	1	0.05			0.05					
Thr	1 + 3	ACU,ACC	2	0.80	0.800	0.992	0.60	0.750	0.838			
	2	ACG	1	0.10			0.10					
	4	ACA,ACG	1	0.10			0.10					
Gln	2	CAG	2	0.40	0.571	0.954	0.40	0.571	0.869			
	1	CAA	2	0.30			0.30					
Ser		UCU,UCC,UCA								1.17	0.755	0.992
		UCA,UCG								0.38		
Val		GUU,GUC,GUA								1.01	0.871	0.955
		GUG								0.15		

$\mathrm{ID_{tRNA}}$, specific tRNA designation; codon, codons recognized by the corresponding tRNA species; $\mathrm{N_{tRNA}}$, number of tRNA genes; tRNA, tRNA content for the two prokaryotic species (from Table 2 in Ikemura 1992) and the yeast (from Table 3 in Ikemura 1982); p_M, the proportion of the most abundant tRNA among all tRNA species carrying the same amino acids; Q_M, the proportion of the codon(s) recognized by the most abundant tRNA among all synonymous codons. Q_M are based on codon frequency data from Table 1 in Ikemura (1992).

[a] The tRNA content is reported as "minor." I used half of the smallest value.

tial derivatives of Y with respect to p_j, and setting the partial derivatives to zero. This yields

$$\frac{Q_1}{p_1^2} = \frac{Q_2}{p_2^2} = \ldots = \frac{Q_3}{p_n^2}. \tag{8}$$

Expressed in another way, the condition implies

$$\frac{p_j}{p_k} = \sqrt{\frac{Q_j}{Q_k}}, \tag{9}$$

i.e., the bias in tRNA availability for an amino acid should not be as dramatic as that in codon usage. In other words, selection driving tRNA adaptation to codon usage guarantees that tRNA bias will not be as extreme as codon bias. Results similar to Equation 9 have been derived before (Bulmer 1987).

The relationship between p and Q in Equation 9 can also be written as $p = a\sqrt{Q}$, where a is a constant. Ikemura (1992) plotted an equivalent measure of Q versus an equivalent measure of p (Figure 3 in Ikemura 1992) for a highly expressed gene in *E. coli* (*groEL*), and the result confirmed the predicted quadratic relationship between p and Q.

We should now note that the baseline condition depicted in Figure 1 is not stable because, with all p_j values equal to $1/n$, Q_j values can drift to any value without affecting translational efficiency (Equation 3 and Figure 1). When Q_j values differ from $1/n$, there will then be selection favoring adaptation of tRNA content to codon usage (Equation 9), which would drive p_j values away from $1/n$. Note that this selection pressure will not drive p_j values more extreme than Q_j values (Equation 9), otherwise the selection would result in a less efficient translational machinery. The resulting unequal p_j values, in turn, create selection pressure for codon usage adaptation (Equation 5).

Evolution of tRNA in response to amino acid usage: To my knowledge, none of the TEH models linked tRNA availability to amino acid usage. The 20 amino acids are not used equally in proteins, and we intuitively would expect those frequently used amino acids to be carried by more tRNA than those rarely used amino acids. To better visualize the effect of amino acid usage on P_i, which is the proportion of tRNA species carrying amino acid i in the total tRNA pool, we write Y in Equation 2 in the expanded form:

$$\begin{aligned} Y = {} & \frac{N_1}{P_1}\left(\frac{Q_{1,1}}{p_{1,1}} + \frac{Q_{1,2}}{p_{1,2}} + \ldots + \frac{Q_{1,n1}}{p_{1,n_1}}\right) \\ & + \frac{N_2}{P_2}\left(\frac{Q_{2,1}}{p_{2,1}} + \frac{Q_{2,2}}{p_{2,2}} + \ldots + \frac{Q_{2,n2}}{p_{2,n_2}}\right) + \ldots \\ & + \frac{N_{20}}{P_{20}}\left(\frac{Q_{20,1}}{p_{20,2}} + \frac{Q_{20,2}}{p_{20,2}} + \ldots + \frac{Q_{20,n20}}{p_{20,n_{20}}}\right), \end{aligned} \tag{10}$$

where N_i is the total number of codons for amino acid

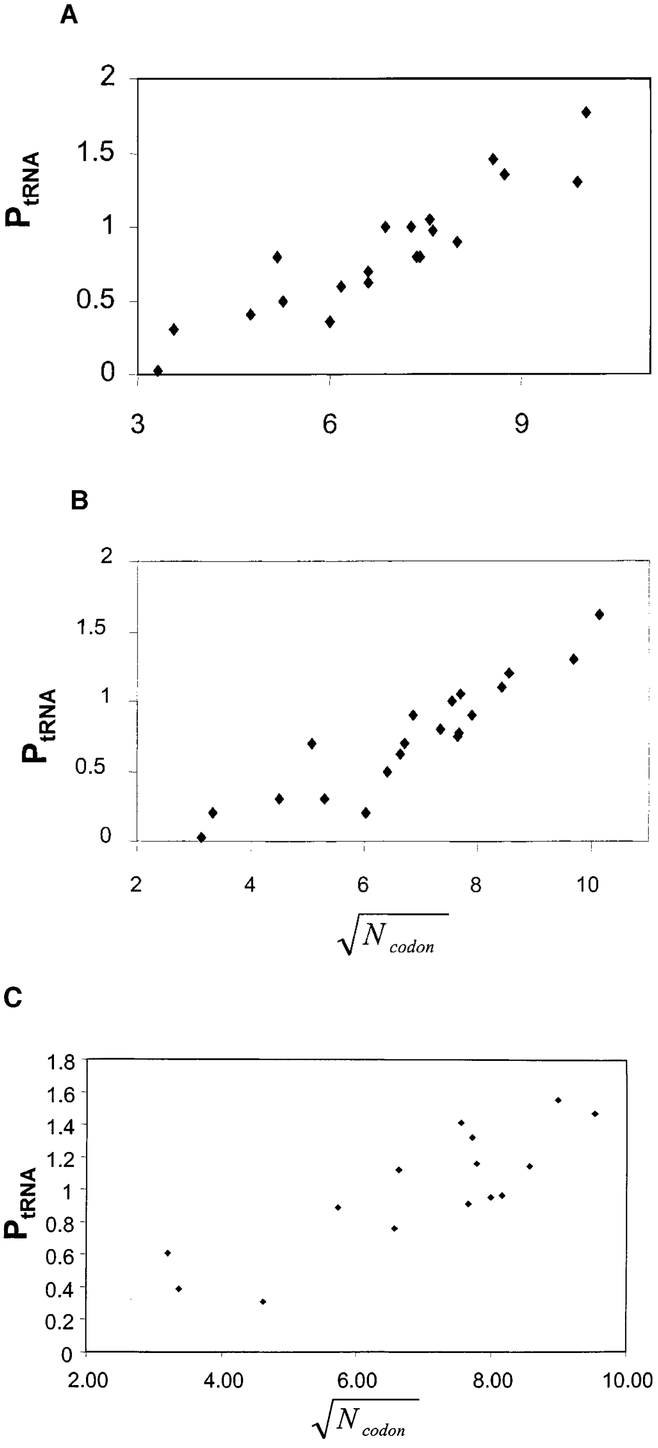

Figure 2.—The availability of tRNA carrying a certain amino acid increases linearly with the square-root of the frequency of the amino acid in (A) *E. coli*, (B) *S. typhimurium* and (C) *S. cerevisiae*. The 20 N_i values are from Table 1 in Ikemura (1992). Corresponding P_i values are from Table 2 in Ikemura (1992) for the two prokaryotic species and Table 3 in Ikemura (1982) for the yeast. N_i values were presented as the number in 1000.

i. When codon usage is perfectly adapted to tRNA availability for each amino acid, which is approximately true based on empirical data in Table 1, Y becomes

$$Y = \frac{N_1}{P_1 p_{M_1}} + \frac{N_2}{P_2 p_{M_2}} + \ldots + \frac{N_{20}}{P_{20} p_{M_{20}}} \tag{11}$$

according to Equation 5. The minimization of Y requires

$$\frac{P_i}{\sqrt{N_i p_{M_j}}} = \frac{P_j}{\sqrt{N_j p_{M_i}}}, \tag{12}$$

where P_i and P_j designate the proportion of tRNA carrying amino acids i and j, respectively; and N_i and N_j are the number of amino acids i and j, respectively. When tRNA concentration for each amino acid is well adapted to codon usage, all p_M values approach 1 and become nearly equal, so that Equation 12 becomes

$$\frac{P_i}{\sqrt{N_i}} = \frac{P_j}{\sqrt{N_j}} \text{ or } P = a\sqrt{N}. \tag{13}$$

This relationship has not been recognized previously.

Empirical data for testing the above prediction is readily available. The P_i values can be derived from data in Table 2 in Ikemura (1992) for *E. coli*, and *S. typhimurium*, and from Table 3 in Ikemura (1982) for *S. cerevisiae* (whose tRNA data are incomplete). Ikemura (1992) also compiled the codon usage of 937 *E. coli* genes, 130 *S. typhimurium* genes, and 581 *S. cerevisiae* genes, from which one can derive N_i values in Equation 13. The 20 pairs of P_i and $\sqrt{N_i}$ values are plotted on Figure 2, A–C, for *E. coli*, *S. typhimurium*, and *S. cerevisiae*, respectively. The fit is quite remarkable.

Such a seemingly straightforward interpretation, however, has a major difficulty. The argument requires that all p_M values be either approximately one (which should hold only for highly expressed genes), or approximately equal (which we have no reason to expect), so as to cancel each other out. Only a few loci are deemed highly expressed, yet 937 loci from *E. coli*, 130 from *S. typhimurium* and 581 from *S. cerevisiae* were used for Figure 2. Why should lowly expressed genes contribute to the linear relationship? The simplifying assumption, that $p_M \approx 1$, seems unjustified. It is therefore necessary to work out the relationship between P and N when the assumption of $p_M \approx 1$ does not hold.

I propose the following equation, which is more general than Equation 13 and does not require $p_M \approx 1$, to describe the relationship between P and N:

$$P = aN^b. \tag{14}$$

If the parameter b is shown to be $= 1/2$, then Equation 14 is reduced to Equation 13. Given Equation 14, we have

$$\frac{P_i}{P_j} = \frac{aN_i^b}{aN_j^b} = \frac{\sqrt{N_i}}{\sqrt{N_j}} \cdot \frac{\sqrt{p_{M_j}}}{\sqrt{p_{M_i}}}. \tag{15}$$

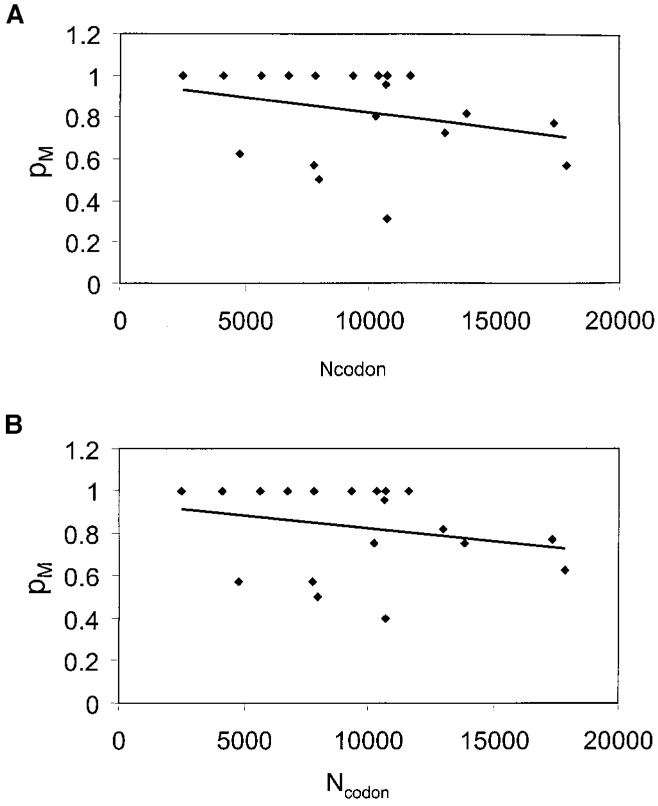

Figure 3.—The slightly negative relationship between p_M (the proportion of the most abundant tRNA among all tRNA species carrying the same amino acids) and N_{codon} (the number of codons for each amino acid). (A) *E. coli*, (B) *S. typhimurium*.

After some algebraic manipulation, we obtain

$$P = aN^{\frac{1+Z}{2}}, \tag{16}$$

where

$$Z = \frac{\ln\frac{p_{M_j}}{p_{M_i}}}{\ln\frac{N_i}{N_j}}. \tag{17}$$

As expected, the relationship between P and N depends on the magnitude of Z, which in turn depends on the relationship between p_M and N. If p_M is independent of N and approaches 1, then $Z = 0$, and $P = aN^{1/2}$, which is Equation 13. If p_M and N are positively correlated, then $Z < 0$. If Z lies within $(-1, 0)$, then P will increase with N at a decreasing rate. If $Z = -1$, then there will be no relationship between P and N, which we know to be false from Figure 2. If $Z < -1$, then P will decrease with N at a decreasing rate, which we also know to be false from Figure 2. If p_M and N are negatively correlated, then $Z > 0$. If Z is between 0 and 1, then P will increase with N at a decreasing rate. If $Z = 1$, then P will increase linearly with N, rather than with the square-root of N as predicted from Equation 13. If $Z > 1$, then P will increase with N at an increasing rate.

There seems to be a slightly negative relationship between p_M and N for data from the two prokaryotic species (Figure 3), which is not statistically significant. It is not possible to obtain good p_M values for *S. cerevisiae* because some of its tRNA species remain unquantified. Based on the relationship between p_M and N for the two prokaryotic species, we expect Z in Equations 16 and 17 to be slightly larger than 0. Consequently, the coefficient b in $P = aN^b$ should be slightly larger than 1/2. The b values that provide the best fit to the data points in Figure 2, A–C are 1.10, 0.99 and 0.88, respectively, *i.e.*, $Z = 1.20$, 0.98 and 0.76, respectively. Equation 14, however, does not fit the empirical data significantly better than Equation 13.

Translational efficiency and translational accuracy: Translational accuracy has recently been suggested to be an important factor related to codon usage bias (Bulmer 1991; Akashi 1994). This proposal received empirical substantiation from a study of protein-coding genes in Drosophila that revealed 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 lower amino acid conservation (Akashi 1994).

Translational efficiency and translational accuracy are inextricably coupled in their effect on codon usage bias. To reduce translational error, one needs to reduce the number of wrong aminoacyl-tRNA species that have to be rejected before the arrival of the right aminoacyl-tRNA. Equation 1 shows this number to be

$$N_{wrong} = \left(\frac{1}{P_i p_{ij}} - 1\right) \tag{18}$$

for each codon translated. To translate an mRNA with L codons, the total number of wrong aminoacyl-tRNA species that translational machinery needs to reject is

$$N_{L.wrong} = Y - L. \tag{19}$$

where Y is exactly the same as the Y in Equation 2. To minimize the number of translational errors, we minimize Y, which leads to exactly the same predictions that we have already attributed to TEH. The rationale for separating the effect of maximizing translational accuracy on codon usage bias from that of maximizing translational efficiency is discussed later.

DISCUSSION

Validity of the model: Protein synthesis is a multi-step process including initiation of transcription, elongation of mRNA chain, initiation of translation, and elongation of the peptide chain. Opinions differ concerning which step might be rate-limiting. Xia (1996) argued that the rate of protein synthesis depends much on the rate of initiation of translation. He reasoned that the rate of initiation depends on the encountering rate between ribosomes and mRNA, which 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. He presented a model predicting 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 several lines of evidence. That the initiation step is rate-limiting has also been suggested by other studies, *e.g.*, Liljenström and von Heijne 1987; Bulmer 1991.

While not denying the possibility that initiation of translation may be rate-limiting, the model presented here explicitly assumes that the elongation of the peptide chain is rate-limiting. There is a substantial amount of empirical evidence supporting this assumption (Pedersen 1984; Bonekamp *et al.* 1985; Bonekamp and Jensen 1988; Williams *et al.* 1988). In particular, mRNA consisting of preferred codons is translated faster than mRNA artificially modified to contain rare codons (Robinson *et al.* 1984; Sorensen *et al.* 1989). That elongation is a rate-limiting process has also been suggested on the basis of theoretical considerations (Liljenström and von Heijne 1987).

Bulmer (1991), however, argued that initiation rather than elongation is rate-limiting. He reasoned that, for elongation to be rate-limiting, there should be so many ribosomes that would bind to all free mRNA molecules as soon as the latter become available for binding. Since ribosomes form the largest part of the protein translational machinery (and are therefore likely to be costly and time-consuming to make), it would be inefficient to saturate the system with them. He summarized empirical evidence that seems to suggest that ribosomes are far from saturating the system. For example, there are an average of 225 bases per ribosome in a polysome (Ingraham *et al.* 1983), and each ribosome covers only about 30 bases (Kozak 1983). This Bulmer (1991) interprets to mean that it is very rare for more than one ribosome to compete for the free binding site of the mRNA. Thus, there is no need for the ribosome to travel down the length of the mRNA in a hurry, *i.e.*, there is little benefit associated with more efficient elongation.

There are two weaknesses in such arguments. First, Kozak's (1983) study does not necessarily mean that a ribosome needs clear only 30 bases to free the initiation site for the binding of the next ribosome. Second, even if the ribosome needs to move only 30 bases to free the initiation site, there is still some probability for more than one ribosomes to arrive at the free initiation site. Only one of the arriving ribosomes would have a chance to bind to the initiation site, while the rest would have to be turned away. Increased elongation rate would reduce the occurrence of such events.

In addition to the assumption that elongation is rate-limiting, the model also assumes that either *r*, *i.e.*, the rate of aminoacyl-tRNA diffusing to the A site of the ribosome during translation, is not extremely large, or t_r, *i.e.*, the time spent in rejecting each wrong aminoacyl-tRNA, is not negligibly small. These seem to be reasonable assumptions, although Bilgin *et al.* (1988) suggested that t_r might indeed be very small.

Relative importance of translational efficiency and accuracy on codon usage bias: Although the model of maximizing translational accuracy and that of maximizing translational efficiency produce the same set of predictions, it is still possible to separate the effect of maximizing translational accuracy on codon usage bias from that of maximizing translational efficiency. For example, a protein gene could have arginine codons in different domains of different functional importance. Being in the same protein gene, these arginine codons are subject to the same selection pressure exerted by maximizing translational efficiency, and consequently should have the same codon usage bias according to the model of maximizing translational efficiency. However, those arginine codons located in the functionally important domains are subject to greater selection pressure exerted by maximizing translational accuracy than those located in the functionally unimportant domains. Consequently, the former codons will be more biased towards using the optimal codon than the latter. Some preliminary findings along this line of reasoning have already been reported (Akashi 1994).

The reasoning above leaves one question unanswered. Why is it necessary to invoke translational efficiency to account for codon usage bias? Can't we attribute all the codon usage bias to the effect of maximizing translational accuracy and forget about translational efficiency? The answer is that the effect of maximizing translational accuracy is insufficient to account for the observed codon usage bias. For example, highly expressed genes exhibit greater codon bias than lowly expressed genes, but the former are not necessarily more conservative than the latter (greater conservativeness presumably implies greater demand for accuracy). We can rank protein genes according to their conservativeness, or rank them according to their expressivity, and find out which ranking explains codon usage bias better. Preliminary results (unpublished) suggest that the expressivity is the more important of the two.

It should be noted that the within-gene variation in codon usage bias found in Drosophila (Akashi 1994) does not seem to be general. For example, it is not observed in *E. coli* and *S. typhimurium* (Hartl *et al.* 1994). More empirical studies are needed to assess the effect of maximizing translational accuracy on codon usage bias.

How optimized are the translational machinery? From our results, we can say that codon usage in those highly expressed genes is almost as optimal as possible, with the Q_M values larger than p_M values and almost equal to one. However, for the majority of genes, the Q_M

values are smaller than p_M (data not shown), which implies that the translational efficiency for the majority of the genes is less than in the seemingly less adaptive scenario when different tRNA species are present in equal amounts and codon usage drifts freely in any direction.

We should note that selection for codon adaptation to tRNA content operates on individual genes, whereas selection for the adaptation of tRNA content to codon usage operates at the genome level. Thus, although Equation 5 suggests that the optimal condition is when both the most abundant tRNA and its cognate codon become fixed, Equation 9 shows that selection for tRNA adaptation to codon usage will always lag behind codon usage bias.

The most remarkable feature from the model is the prediction relating amino acid usage (N_i) to tRNA content (P_i), which is strongly supported by empirical evidence (Figure 2). A more extensive study is underway to confirm the generality of the relationship.

I thank L. Choy and A. K. F. Lo for assistance and members of University of Hong Kong Ecology and Evolutionary Genetics Group (Y. Sadovy, in particular) for discussion and comments. Part of the work was done when I was at the Museum of Natural Science of Louisiana State University. I benefited greatly from referees' comments. This project is supported by a Research Grant Council grant (HKU 7233/98M) from Hong Kong Government and a Committee on Research and Conference Grants grant (335/023/0022) from The University of Hong Kong to X. Xia, and a National Science Foundation grant (DEB95-27583) to Mark S. Hafner and X. Xia.

LITERATURE CITED

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

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

Bilgin, N., M. Ehrenberg and C. Kurland, 1988 Is translation inhibited by noncognate ternary complexes? FEBS-LETT. **233:** 95–99.

Bonekamp, F., and K. F. Jensen, 1988 The AGG codon is translated slowly in *E. coli* even at very low expression levels. Nucleic Acids Res. **16:** 3013.

Bonekamp, F., H. D. Andersen, T. Christensen and K. F. Jensen, 1985 Codon-defined ribosomal pausing in *Escherichia coli* detected by using the *pyrE* attenuator to probe the coupling between transcription and translation. Nucleic Acids Res. **13:** 4113–4123.

Bulmer, M., 1987 Coevolution of codon usage and transfer RNA abundance. Nature **325:** 728–730.

Bulmer, M., 1988 Codon usage and intergenic position. J. Theor. Biol. **133:** 67–71.

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

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.

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. M. Pirtle. CRC Press, Boca Raton, FL.

Ingraham, J. L., O. Maaløe and F. C. Neidhardt, 1983 *Growth of the bacterial cell.* Sinauer, Sunderland, Mass.

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

Kozak, M., 1983 Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47:** 1–43.

Kurland, C. G., 1987a Strategies for efficiency and accuracy in gene expression. 1. 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.

Liljenström, H., and G. von Heijne, 1987 Translation rate modification by preferential codon usage: intragenic position effects. J. Theor. Biol. **124:** 43–55.

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

Moriyama, E. N., and D. L. Hartl, 1993 Codon usage bias and base composition of nuclear genes in Drosophila. Genetics **134:** 847–858.

Pedersen, S., 1984 *Escherichia coli* ribosomes translate in vivo with variable rate. EMBO J. **3:** 2895.

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., 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.

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

Varenne, S., J. Bug, R. Lloubes and C. Lazdunski, 1984 Translation is a nonuniform process: effect of tRNA availability on the rate of elogation of nascent polypeptide hains. J. Mol. Biol. **180:** 549–576.

Williams, D. P., D. Rigier, D. Akiyoshi, F. Genbauffe and J. R. Murphy, 1988 Design, synthesis and expression of a human interleukin-2 gene incorporating the codon usage bias found in highly expressed Escherichia coli genes. Nucleic Acids Res. **16:** 10453–10467.

Xia, X., 1996 Maximizing transcription efficiency causes codon usage bias. Genetics **144:** 1309–1320.

Communicating editor: A. G. Clark