Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bulletin of Mathematical Biology 67 (2005) 1355–1368
Bulletin of Mathematical Biology
www.elsevier.com/locate/ybulm

The impact of including tRNA content on the optimality of the genetic code

Hani Goodarzi[a,*], Hamed Shateri Najafabadi[a], Hamed Ahmadi Nejad[b], Noorossadat Torabi[a]

[a] *Department of Biotechnology, Faculty of Science, University of Tehran, Tehran, Iran*
[b] *Department of Computer Engineering, Sharif University of Technology, Tehran, Iran*

Received 24 January 2005; accepted 4 March 2005

Abstract

Statistical and biochemical studies have revealed nonrandom patterns in codon assignments. The canonical genetic code is known to be highly efficient in minimizing the effects of mistranslational errors and point mutations, since it is known that, when an amino acid is converted to another due to error, the biochemical properties of the resulted amino acid are usually very similar to those of the original one. In this study, we have taken into consideration both relative frequencies of amino acids and relative gene copy frequencies of tRNAs in genomic sequences in order to introduce a fitness function which models the mistranslational probabilities more accurately in modern organisms. The relative gene copy frequencies of tRNAs are used as estimates of the tRNA content. We also altered the rule previously used for the calculation of the probabilities of single base mutation occurrences. Our model signifies higher optimality of the genetic code towards load minimization and suggests the presence of a coevolution of tRNA frequency and the genetic code.

1. Introduction

Gone are the days when the genetic code was thought to be a "frozen accident" (Crick, 1968). Since then, many researches have been conducted on the evolution of the genetic

* Corresponding author. Tel.: +98 21 8058210; fax: +98 21 8040284.
E-mail address: hani.goodarzi@gmail.com (H. Goodarzi).

0092-8240/$30
doi:10.1016/j.bulm.2005.03.002

code and its route towards its contemporary structure (Dillon, 1973; Wong, 1975; Woese, 1965a,b; Pele, 1965; Goldberg and Wittes, 1966; Woese et al., 1966a,b; Amirnovin, 1997; Ardell, 1998; Judson and Haydon, 1999; DiGiulio, 2000; Ronneberg et al., 2000; Ardell and Sella, 2001). The efficiency of the canonical genetic code in minimizing the deleterious effects of errors due to single-base mutations and mistranslations (Load Minimization) is supported (Haig and Hurst, 1991; Freeland and Hurst, 1998a,b; Gilis et al., 2001; Freeland, 2002; Goodarzi et al., 2004), although the hypothesis is debated by some researchers (DiGiulio, 2000, 2001). Haig and Hurst (1991) tried to compare the canonical genetic code with randomly generated ones in order to achieve a quantitative measure of the optimality of the genetic code. Testing several physiochemical parameters, they found that single-base mutations had a very small average effect when applying differences in hydropathy (previously introduced by Woese et al. (1966a)) between the misinterpreted amino acids. Generating several thousands of random codes, they found that only 1 in every 10^4 random codes shows higher optimization towards load minimization.

Considering transition/transversion biases and different probabilities of mistranslation in the three locations of a codon, Freeland and Hurst (1998b) proposed a new model that described translational errors more accurately. Their new model improved the fraction of more optimized codes from 10^{-4} to 10^{-6}.

Gilis et al. (2001) highlighted the importance of another parameter in the optimization of the genetic code, namely the frequency at which different amino acids occur in proteins. Although this frequency differs from protein to protein, a prevailing pattern was recognized in general (King and Jukes, 1969; for further discussions see Gilis et al., 2001). Plotting the number of synonyms codons versus the frequency of the corresponding amino acid, a high correlation between the corresponding variables has been established, emphasizing the significance of this parameter in the evolution of the genetic code (King and Jukes, 1969). In addition, Gilis et al. (2001) brought further improvement to their model by using quantities other than hydropathy to measure the roles of the different amino acids in protein conformation and stability. They devised a cost function designated the "Mutation Matrix" by evaluating, in silico, the change in the folding free energy caused by all possible point mutations in a set of protein structures. As it is driven theoretically, the Mutation Matrix is alleged to be unbiased towards the genetic code, based upon the fact that it compares the impact of single-base mutations on the accurate folding of a protein, which makes it a suitable cost function to use when studying genetic code (Gilis et al., 2001). The corresponding results revealed a lower fraction of random codes which scored better than that of natural code. This fraction was calculated to be 2×10^{-9}, when applying the Mutation Matrix.

In addition to the Mutation Matrix, Gilis et al. (2001) also made use of the Point Accepted Mutations 74–100 ($PAM_{74–100}$) scoring matrix. The PAM (Point Accepted Mutations) matrices are a family of matrices derived from amino acid substitution frequencies observed within homologous proteins. $PAM_{74–100}$ is known to be the least biased PAM matrix towards the genetic code, marked with the fact that it includes highly diverged sequences and only reflects the physicochemical similarities and not mutations due to codon proximity (Benner et al., 1994). Thus, when highly diverged proteins are compared, amino acids are replaced due to their ability to compensate each other and not randomly occurring mutations. Consequently, $PAM_{74–100}$ is the chosen matrix for many

studies (Freeland et al., 2000; Gilis et al., 2001) including the one at hand. However, it should be mentioned that a number of researchers are not satisfied with the usage of PAM and similar matrices, proposing that they are biased towards the genetic code as they are obtained empirically (DiGiulio, 2001; Goodarzi et al., 2005). PAM_{74-100} reflects the genetic code itself and statistical inferences cannot be made based solely on this matrix; yet, we have recalculated all the parameters and results using this matrix in order to compare them with the results obtained from the Mutation Matrix. Furthermore, these calculations enable us to compare our work with previous studies (Gilis et al., 2001; Goodarzi et al., 2004).

Recently, Goodarzi et al. (2004) introduced a new model which incorporated the impacts of nonsense mistranslations and a new aspect of including amino acid frequencies. They also used a z-value as a new measure for optimality of the genetic code and found that their model resulted in a higher z-value than the previous works.

2. Theoretical background

Haig and Hurst (1991) had to define a fitness score corresponding to any given code in order to be capable of comparing the canonical genetic code quantitatively with randomly generated ones. Thus, they defined a fitness function which estimated the efficiency of a code towards minimization of the differences between the mutating amino acids due to a hydropathy index:

$$\varphi^{\text{Haig and Hurst}} = \sum_{c,c'} [h(a(c)) - h(a(c'))]^2 \tag{1}$$

where c and c' are all pairs of codons that can be changed to each other by a single-base mutation (stop codons are excluded); $a(c)$ and $a(c')$ are amino acids coded by c and c', respectively; and $h(a)$ returns the hydropathy index of amino acid a. They tested several physicochemical properties and found that when using polarity, the canonical genetic code best minimizes the effects of single-base changes.

Freeland and Hurst (1998b) further improved this function by incorporating the probability of each mutation (stop codons are excluded):

$$\varphi^{\text{Freeland and Hurst}} = \sum_{c=1}^{64} \sum_{c'=1}^{64} p(c' \mid c)[h(a(c)) - h(a(c'))]^2 \tag{2}$$

where $p(c' \mid c)$ stands for the probability of misinterpretation of codon c as c'. They chose $p(c' \mid c)$ to have the following values:

- $p(c' \mid c) = \frac{1}{N}$ if c and c' differ only in the third base;
- $p(c' \mid c) = \frac{1}{N}$ if c and c' differ only in the first base and cause a transition;
- $p(c' \mid c) = \frac{0.5}{N}$ if c and c' differ only in the first base and cause a transversion;
- $p(c' \mid c) = \frac{0.5}{N}$ if c and c' differ only in the second base and cause a transition;
- $p(c' \mid c) = \frac{0.1}{N}$ if c and c' differ only in the second base and cause a transversion;
- $p(c' \mid c) = 0$ otherwise.

(N is the normalization factor selected so that $\sum_{c'} p(c' \mid c) = 1$).

Gilis et al. (2001) defined their fitness function as (stop codons are excluded):

$$\varphi^{faa} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} \sum_{c'=1}^{64} p(c' \mid c) \cdot g(a(c), a(c')), \tag{3}$$

where $p(a(c))$ is the frequency in which amino acid $a(c)$ occurs (Gilis et al., 2001); $n(a(c))$ is an integer standing for the number of synonymous codons that $a(c)$ possesses; $p(c' \mid c)$ is the same as stated by Freeland and Hurst (1998b); and $g(a(c), a(c'))$ is the cost of substitution of amino-acid $a(c)$ by $a(c')$.

Goodarzi et al. (2004) further improved φ^{faa} to become:

$$\varphi^{HH} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} f(c) \cdot \sum_{c'=1}^{64} p(c' \mid c) \cdot g(a(c), a(c')); \tag{4}$$

$$f(c) = 1 - \int_0^{|y|} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \mathrm{d}\, t, \tag{5}$$

where

$$y = \frac{1}{S} \cdot \frac{n(a(c))/(n(\text{total}) - n(\text{term})) - p(a(c))}{\sigma(a(c))} \tag{6}$$

and n(total) and n(term) are the total number of codons and the number of termination codons, respectively. Note that in all the randomly generated codes, n(total) equals 64 and n(term) returns 3. $\sigma(a(c))$ stands for the standard deviation of the frequency of the amino acid a coded by codon c (the frequencies and standard deviations are the same as those introduced by Gilis et al., 2001). The value designated S was defined as a constant value, fixed at 100 (Goodarzi et al., 2004). The function $f(c)$ highlights the correlation between the frequency of an amino acid and the number of its synonymous codons, resulting in a lower score for codes in which some highly frequent amino acids have possessed a small number of synonymous codons and vice versa (for further discussion see Goodarzi et al., 2004).

Unlike previous works, Goodarzi et al. (2004) included the effect of nonsense mistranslations by appointing a constant as the cost measure of these misinterpretations. This value, designated K, was chosen to be -3.0 in the case of the Mutation Matrix and -4.5 for PAM_{74-100}. Furthermore, Goodarzi et al. (2004) used the z-value as a measure of the optimality of the genetic code, defined as $z = \frac{\varphi_{cgc} - \mu}{\sigma}$, where φ_{cgc}, μ and σ are the value of the fitness function for the canonical genetic code, the mean value of the fitness function obtained from randomly generated codes and the standard deviation of the same values, respectively.

In the work presented, we have weighed the value of an amino acid by the number of the tRNAs its corresponding codons possess. It has also been considered that there is no tRNA capable of distinguishing U from C at the third codon position (Ronneberg et al., 2000). The corresponding results of this model suggest that the gene copy numbers of tRNAs and the structure of the genetic code have coevolved so as to minimize the effects of single-base mistranslations.

Table 1
The relative frequencies of tRNAs for different amino acids

Amino acid	$\tau(a) \times 100$	$n(a)$	$\tau(a)/n(a)$
Ala	6.72	4	0.01680
Arg	8.73	6	0.01455
Asp	3.88	2	0.01940
Asn	3.25	2	0.01625
Cys	2.08	2	0.01040
Glu	4.09	2	0.02045
Gln	3.81	2	0.01905
Gly	7.34	4	0.01835
His	2.35	2	0.01175
Ile	3.95	3	0.01317
Leu	10.11	6	0.01685
Lys	4.36	2	0.02180
Met	7.48	1	0.07480
Phe	2.49	2	0.01245
Pro	4.71	4	0.01178
Ser	7.76	6	0.01293
Thr	6.23	4	0.01558
Trp	2.01	1	0.02010
Tyr	2.77	2	0.01385
Val	5.89	4	0.01473

See text for description of terms.

3. Methods

The gene copy number of tRNAs of 29 eubactrial organisms have been studied in this work. tRNA gene copy numbers are derived from the genomic tRNA database containing tRNA identifications made by the program tRNAscan-SE (http://rna.wustl.edu/tRNAdb). The relative gene copy frequency for each tRNA species was calculated for each organism and the relative frequencies were averaged over the 29 organisms (Fig. 1).

3.1. A new fitness function

Based on the previous works (Haig and Hurst, 1991; Freeland and Hurst, 1998b; Gilis et al., 2001; Goodarzi et al., 2004), a new fitness function was established which incorporates the relative gene copy frequencies of tRNAs:

$$\varphi^{fta} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} f(c) \cdot \sum_{c'=1}^{64} p(c' \mid c) \cdot [\omega(a(c), a(c'))]^{\lambda} \cdot g(a(c), a(c')), \tag{7}$$

where

$$\omega(a_1, a_2) = \mathrm{e}^{\frac{\tau(a_2)}{n(a_2)} - \frac{\tau(a_1)}{n(a_1)}}. \tag{8}$$

In Eq. (8), $\tau(a)$ returns the sum of relative gene copy frequencies of tRNA species related to amino acid a (Table 1) and $n(a)$ returns the number of synonym codons that code for amino acid a.

	U	C	A	G	
U	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	U
	2.484 ± 0.557	2.058 ± 0.572	2.592 ± 0.787	2.089 ± 0.593	C
	2.532 ± 0.637	2.412 ± 0.885	0.000 ± 0.000	0.583 ± 0.983	A
	1.827 ± 0.756	1.329 ± 1.356	0.000 ± 0.000	2.017 ± 0.594	G
C	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	2.559 ± 1.848	U
	1.954 ± 0.705	1.272 ± 1.112	2.339 ± 0.476	0.634 ± 1.127	C
	2.333 ± 0.697	2.568 ± 0.637	3.179 ± 1.274	0.733 ± 1.189	A
	1.591 ± 1.434	0.938 ± 1.051	0.544 ± 0.960	1.362 ± 0.925	G
A	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	U
	3.871 ± 1.510	2.035 ± 0.661	3.114 ± 1.160	2.178 ± 0.468	C
	0.000 ± 0.000	2.817 ± 1.107	3.444 ± 1.292	2.017 ± 0.594	A
	7.201 ± 1.335	1.385 ± 1.174	0.871 ± 1.118	1.540 ± 1.019	G
G	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	U
	1.728 ± 0.847	1.860 ± 0.868	3.745 ± 1.388	3.956 ± 1.490	C
	3.501 ± 1.509	4.219 ± 1.743	3.822 ± 1.661	2.511 ± 0.796	A
	0.532 ± 0.939	0.619 ± 0.999	0.389 ± 0.849	0.718 ± 1.005	G

Fig. 1. The mean frequencies (%) of the tRNAs of 29 selected organisms and the corresponding standard deviations. The frequencies were computed over the genomes of *Haemophilus influenzae* Rd, *Mycoplasma genitalium*, *Mycoplasma pneumoiae*, *Helicobacter pylori* 26695, *Helicobacter pylori* J99, *Escherichia coli* K-12, *Bacillus subtilis*, *Borrelia burgdorferi*, *Aquifex aeolicus*, *Mycobacterium tuberculosis*, *Treponema pallidum*, *Rickettsia prowazekii*, *Chlamydia pneumoniae* CWL029, *Chlamydia trachomatis* D/UW-3/Cx, *Thermotoga maritima*, *Inococcus radiodurans*, *Campylobacter jejuni*, *Neisseria meningitidis* MC58, *Neisseria meningitidis* Z2491, *Ureaplasma urealyticum*, *Vibrio cholerae*, *Xylella fastidiosa*, *Pseudomonas aeruginosa*, *Buchnera* sp., *Bacillus halodurans*, *Mycobacterium leprae*, *Caulobacter crescentus*, *Pasteurella multocida*, *Staphylococcus aureus* N315, as described in text. Data were downloaded from http://rna.wustl.edu/tRNAdb/.

Since the relationship between ω and φ^{fta} may be other than linear, a power factor, designated λ, is incorporated. The exact value of λ (i.e., the exact relationship between

ω and φ^{fta}) is unknown. Therefore, we optimized λ so as to result in the maximum z-value, testing different values in the range 0–35 and plotting the z-value against them (for a description of the z-value see Goodarzi et al., 2004). The nonsense mutations are considered in the calculation of φ^{fta} in the same way as described in Goodarzi et al. (2004).

For the sake of comparison, ω has been included in φ^{faa} (Eq. (3)). Also the effect of ω in φ^{fta} has been studied when nonsense mistranslations have been excluded.

3.2. A modified method of obtaining $p(c' \mid c)$

No known tRNA anticodon base modification, natural or engineered, can discriminate third-base pyrimidines within any codon, despite the large number of nonstandard codes now recognized and diverse experimental manipulation of coding components (Ronneberg et al., 2000). So, codon XYU is not distinguishable from XYC and vice versa. Thus, XYU can mutate to any codon which XYC is able to mutate to and XYC can mutate to any codon which XYU is able to mutate to. For example, UUC can be mutated to UGU, in the same way as UUU does.

Therefore we have chosen $p(c' \mid c)$ to have the following values:

- $p(c' \mid c) = \frac{1}{N}$ if c and c' differ only in the third base;
- $p(c' \mid c) = \frac{1}{N}$ if c and c' differ only in the first base and cause a transition, or differ in the first base and cause a transition while the third codon position bears a transition in a pyrimidine;
- $p(c' \mid c) = \frac{0.5}{N}$ if c and c' differ only in the first base and cause a transversion, or differ in the first base and cause a transversion while the third codon position bears a transition in a pyrimidine;
- $p(c' \mid c) = \frac{0.5}{N}$ if c and c' differ only in the second base and cause a transition, or differ in the second base and cause a transition while the third codon position bears a transition in a pyrimidine;
- $p(c' \mid c) = \frac{0.1}{N}$ if c and c' differ only in the second base and cause a transversion, or differ in the second base and cause a transversion while the third codon position bears a transition in a pyrimidine;
- $p(c' \mid c) = 0$ otherwise.

(N is the normalization factor selected so that $\sum_{c'} p(c' \mid c) = 1$).

3.3. Rules for generating random codes (Freeland and Hurst, 1998b)

1. The “codon space” is divided into 21 non-overlapping sets of codons observed in the canonical code, each set specifying an amino acid in the natural genetic code (one set consists of stop codons).
2. Each alternative code is obtained by randomly assigning each of the 20 amino acids to one of these sets. All three stop codons remain invariant in position for all alternative codes.

It should be noted that applying these constraints on the generation of random codes is not necessarily optimal, as previously discussed by Luo and Li (2002) and Chechetkin (2003). Furthermore, Archetti (2004) has postulated that the frequency of the codes that

are more efficient in minimizing errors is relatively higher near the canonical genetic code, whereas, in this standard method the randomly generated codes are distant from the canonical one. Thus, the results obtained from this method are not one hundred per cent accurate, and we suggest devising other random code generating functions in future studies.

3.4. A genetic algorithm based program

A heuristic program was developed in order to obtain the code which best maximizes φ. A set of eight random codes was generated as the initial population. In each generation, each member code of population reproduced seven mutant forms of itself, in which two amino acids, chosen randomly, swapped their codons. From within the parent codes and the mutant child codes, the first eight codes that had the largest values of φ were chosen to make the next generation. 10^6 generations were passed before determining the ultimate code and the process was repeated 100 times, in order to bypass the local minimums as much as possible, if any occurred. This approach finds a highly efficient code in maximizing φ after many generations.

Using the best codes obtained from our genetic algorithm based program, the percentages of optimality of the canonical genetic code have been calculated as (Haig and Hurst, 1991; Freeland et al., 2000; Gilis et al., 2001):

$$\text{optimality\%} = \frac{\varphi_{cgc} - \varphi_{\text{mean}}}{\varphi_{\max} - \varphi_{\text{mean}}}, \tag{9}$$

where φ_{mean} is the mean of distribution for each fitness function obtained from 10^9 randomly generated codes. $\varphi_{\max}$ is assumed to be the value of the fitness function for the best code achieved using the genetic algorithm based program.

4. Results

4.1. Determining the value of λ

Different values of λ, in the range 0–35, were tested to compute the corresponding z-values, applying φ^{faa}, φ^{fta} excluding nonsense mistranslations, and φ^{fta} including nonsense mistranslations. In this step the classic method of obtaining mistranslational probabilities was used. Both the Mutation Matrix and PAM_{74-100} were used as the cost measures for the amino acid substitutions. For each value of λ, 10^6 random codes were generated in order to calculate the z-values. As shown in Fig. 2, the canonical genetic code best outscored randomly generated codes when λ was set to 11.5 in the case of the Mutation Matrix and 18.4 in the case of PAM_{74-100}. For all the values of λ, φ^{faa} with nonsense mistranslations revealed the maximum z-value compared to other fitness functions.

4.2. Modified method of obtaining mistranslational probabilities

Table 2 compares the z-values obtained when the classic method of obtaining mistranslational probabilities and the modified one were applied. For all different fitness

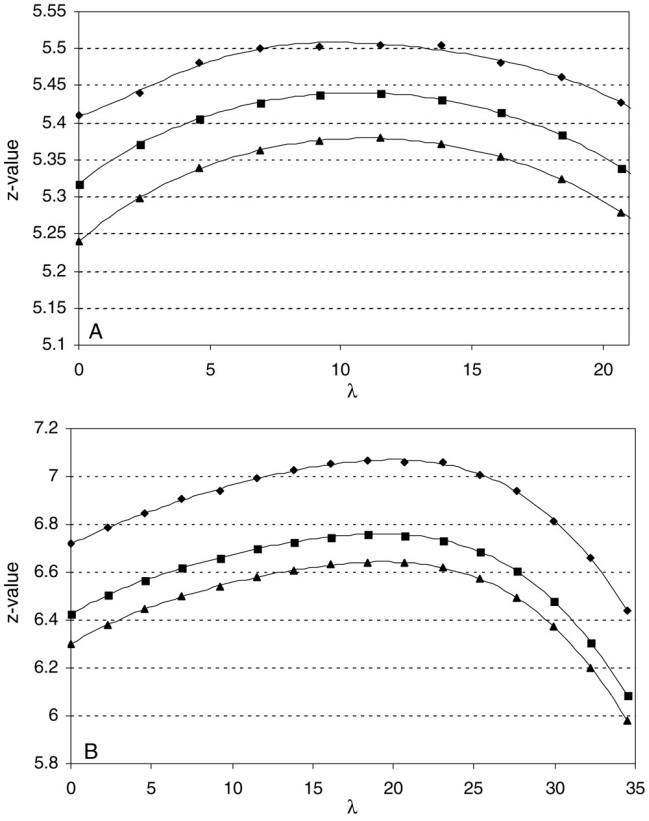

Fig. 2. z-value plotted against λ, measured for A, the Mutation Matrix and B, PAM_{74-100}, using φ^{faa} (closed triangles), φ^{fta} without nonsense mistranslations (closed squares), and φ^{fta} (closed circles).

Table 2
z-value calculated for φ^{faa}, φ^{HH}, φ^{HH} with the consideration of termination codons and φ^{fta}

	Mutation Matrix		PAM_{74-100}	
	Classic	Modified	Classic	Modified
φ^{faa}	5.25961	5.27573	6.33037	6.62666
φ^{HH}	5.35733	5.37943	6.49312	6.85906
φ^{HH}, K	5.44091	5.47639	6.81581	7.1094
φ^{fta}	5.4915	5.54155	7.03011	7.4132

Using both the Mutation Matrix and PAM_{74-100}, higher optimality of the canonical genetic code was achieved when applying the modified version of codon boxes.

functions and both the Mutation Matrix and PAM_{74-100}, a higher z-value was obtained when applying the modified method.

Table 3
The values of $(\varphi_{mt} - \varphi_{cgc})/\varphi_{mt}$ computed using both the classic and the modified methods of determining single-base mutations, for φ^{faa}, φ^{HH}, φ^{HH} with the consideration of termination codons and φ^{fta}

	Mutation Matrix		PAM_{74-100}	
	Classic	Modified	Classic	Modified
φ^{faa}	−0.00505	−0.02853	0.01766	0.010516
φ^{HH}	−0.00448	−0.02763	0.017046	0.009542
φ^{HH}, K	−0.04647	−0.08480	−0.03569	−0.05985
φ^{fta}	−0.04184	−0.07737	−0.0415	−0.06482

4.3. Optimum codes found by the genetic algorithm based program

Fig. 3 shows the best codes which were achieved using the genetic algorithm based program for φ^{faa}, φ^{HH}, φ^{HH} including nonsense mistranslations and φ^{fta}. The modified mistranslational probabilities were used.

4.4. Comparison of the canonical genetic code and the mitochondrial genetic code

Table 3 contains the relative differences of the canonical genetic code and the mammalian mitochondrial genetic code, which is defined as:

$$D = \frac{\varphi_{mt} - \varphi_{cgc}}{\varphi_{mt}}, \tag{10}$$

where φ_{mt} stands for the value of the fitness function applied to the mammalian mitochondrial genetic code, and φ_{cgc} is the value of the fitness function applied to the canonical genetic code. φ_{mt} is calculated using the same parameters as φ_{cgc}. One expects φ_{mt} to be smaller than φ_{cgc} when the natural forces on the error minimizing feature of the canonical genetic code is considered, resulting in a negative value for D. Thus, D can be used as a reference to check whether a fitness function is defined rationally. Note that in the case of PAM_{74-100}, the value of D is positive when φ^{faa} and φ^{HH} are used as the fitness function (Table 3). Therefore, it can be taken into interpretation that φ^{fta} and φ^{HH}, K model the error minimization of the genetic code more accurately.

5. Discussion

In the previous decade, many researchers have tried to highlight the optimality of the canonical genetic code in minimizing the deleterious effects of single base mutations. The results maintain little debate on this hypothesis (Goodarzi et al., 2004). In this work, we have tried to show that codons differ from one another, putting aside the alphabetical symbols used for representing them. An amino acid with higher content of related tRNA species in the cell has a lower probability of being substituted, as there are more accurate tRNAs available. Thus, it can be inferred that, if the mistranslation of a distinct amino acid has a relatively high deleterious effect, it should have a higher number of tRNAs in order to translate its codons properly. When the ratio of tRNA gene frequency to codon number (i.e., the average of tRNA gene frequency for each codon) for amino acid a_2 is higher than for

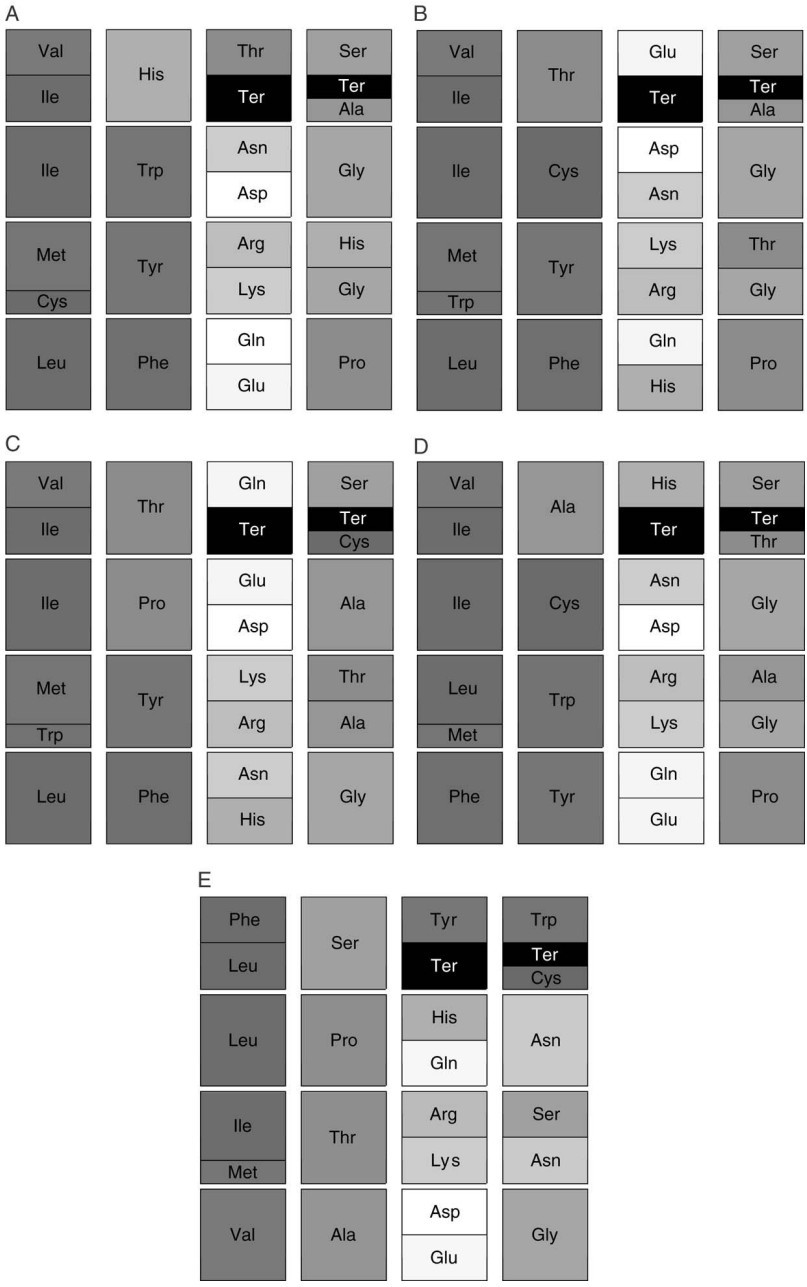

Fig. 3. The codes with highest scores resulted from the genetic algorithm based program while applying different fitness functions: A, φ^{faa}; B, φ^{HH}; C, φ^{HH}, K; and D, φ^{tfa}. The canonical genetic code is represented in E. Hydrophobic amino acids are shown in dark gray, according to hydropathy (Woese et al., 1966a).

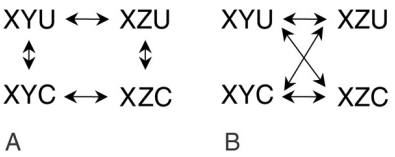

Fig. 4. A, The classic method of determining $p(c' \mid c)$. B, The modified method of determining $p(c' \mid c)$ as described in text. The lines stand for the mutations allowed in each method.

Table 4
Percentage of optimality of the canonical genetic code, with respect to φ^{faa}, φ^{HH}, φ^{HH} with the consideration of termination codons and φ^{fta}

	φ^{faa}	φ^{HH}	φ^{HH}, K	φ^{fta}
Percentage of optimality (%)	86.09	87.01	92.99	93.58

In each case, the value of fitness function for the best code obtained using the genetic algorithm based program is assumed as the maximum fitness. The mean values are obtained from a set of 10^9 randomly generated codes.

amino acid a_1, the value of $\frac{\tau(a_2)}{n(a_2)} - \frac{\tau(a_1)}{n(a_1)}$ would be greater than zero. This results in a value greater than one for $\omega(a_1, a_2)$, which means a greater probability of a codon of amino acid a_1 being misinterpreted as a codon of amino acid a_2. On the other hand, the value of $\frac{\tau(a_1)}{n(a_1)} - \frac{\tau(a_2)}{n(a_2)}$ would be negative, resulting in a value smaller than one for $\omega(a_2, a_1)$ and a lower probability of a codon of amino acid a_2 being misinterpreted as a codon of amino acid a_1.

Various studies have shown that the tRNA contents of *E. coli*, *B. subtilis* and *S. cerevisiae* are highly correlated with respective tRNA gene copy numbers (Dong et al., 1996; Ikemura, 1981; Kanaya et al., 1999; Percudani et al., 1997), suggesting that the tRNA contents of other species may also be estimated by their tRNA gene copy numbers (Kanaya et al., 1999). Based on this conjecture, we have introduced the ω function which uses tRNA gene copy numbers as an estimation of tRNA concentration.

In the previous fitness functions, only single-base mutations were allowed. Yet, as previously mentioned by some researchers (Ronneberg et al., 2000), we are dealing with a genetic code incapable of distinguishing U and C at the third codon positions. In other words, U and C at the third codon position are interchangeable, not only without the change of sense but also with the same tRNAs. Thus, we reformatted $p(c' \mid c)$ to cover this fact, by allowing the occurrence of double-base mutations as long as one of them is a transition between U and C at the third codon position. The visualization of the new method for obtaining translational probabilities is represented in Fig. 4.

Table 4 shows the percentage of optimality of the canonical genetic code in the case of each fitness function, calculated using the value of the corresponding fitness function for the best code obtained from our genetic based algorithm. φ^{fta} revealed the highest percentage of optimality amongst the studied fitness functions. Meanwhile, the pattern of amino acids in the best code found by the program highly resembled that of the canonical genetic code, in terms of hydrophobicity, especially in the case of φ^{fta}. Significantly, Met takes its original position regarding the canonical genetic code, whereas in the case of all the other fitness functions Met has swapped its position usually with Ile.

6. Conclusion

In the work presented, the impact of tRNA content on the level by which the genetic code minimizes the consequences of errors during translation was studied. The results showed that the canonical genetic code acts more efficiently in load minimization when the tRNA content is considered in the modern eubacteria. This suggests that the tRNA content might have been coevolved with the genetic code in the way that enhances the reduction of the effects of mistranslational errors. However, further studies should be conducted on the impact of tRNA content on the optimality of the genetic code in eukaryotic organisms and archeae.

Different codon–anticodon interactions result in diverse free energy changes (several studies have been conducted on the RNA–RNA interactions; see Fink and Crothers, 1972; He et al., 1991; Mathews et al., 1999). However, in this study, all correct interactions are assumed to be the same regarding their free energy changes, which is only a crude estimation and needs to be evaluated.

Since tRNA gene copy numbers differ in different species, the tRNA based fitness function introduced in this work may potentially be used for predicting variations from the standard genetic code in different species (reviewed in Knight et al., 2001) based directly on their genomic sequences.

Acknowledgements

We are grateful to Stephen J. Freeland and Elahe Elahi for their useful comments.

References

Amirnovin, R., 1997. An analysis of the metabolic theory of the origin of the genetic code. J. Mol. Evol. 44, 473–476.

Archetti, M., 2004. Codon usage bias and mutation constraints reduce the level of error minimization of the genetic code. J. Mol. Evol. 59 (2), 258–266.

Ardell, D.H., 1998. On error-minimization in a sequential origin of the standard genetic code. J. Mol. Evol. 47, 1–13.

Ardell, D.H., Sella, G., 2001. On the evolution of redundancy in genetic codes. J. Mol. Evol. 53, 269–281.

Benner, S.A., Cohen, M.A., Gonnet, G.H., 1994. Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng. 7, 1323–1332.

Chechetkin, V.R., 2003. Block structure and stability of the genetic code. J. Theor. Biol. 222, 177–188.

Crick, F.H., 1968. The origin of genetic code. J. Mol. Biol. 38, 367–379.

DiGiulio, M., 2000. The origin of the genetic code. Trends Biochem. Sci. 25, 44–47.

DiGiulio, M., 2001. The origin of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analyses tautologous. J. Theor. Biol. 208, 141–144.

Dillon, L.S., 1973. The origins of the genetic code. Bot. Rev. 39, 301–345.

Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. 260, 649–663.

Fink, T.R., Crothers, D.M., 1972. Free energy of imperfect nucleic acid helices. J. Mol. Biol. 66, 1–12.

Freeland, S.J., 2002. The genetic code: an adaptation for adapting? J. Genet. Program. Evol. Mach. 3 (2), 113–127.

Freeland, S.J., Hurst, L.D., 1998a. Load minimization of the genetic code: history does not explain the pattern. Proc. R. Soc. Lond. B 266, 2111–2119.

Freeland, S.J., Hurst, L.D., 1998b. The genetic code is one in a million. J. Mol. Evol. 47, 238–248.
Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000. Early fixation of an optimal genetic code. Mol. Biol. Evol. 17, 511–518.
Gilis, D., Massar, S., Cerf, N.J., Rooman, M., 2001. Optimality of the genetic code with respect to protein stability and amino acid frequencies. Genome Biol. 2 (11), 49.1–49.12.
Goldberg, A.L., Wittes, R.E., 1966. Genetic code: aspects of organization. Science 153, 420–424.
Goodarzi, H., Nejad, H.A., Torabi, N., 2004. On the optimality of the genetic code with the consideration of termination codons. Biosystems 77 (1–3), 163–173.
Goodarzi, H., Najafabadi, H.S., Hassani, K., Nejad, A.H., 2005. On the optimality of the genetic code, with the consideration of coevolution theory by comparison of prominent cost measure matrices. J. Theor. Biol. 235 (3), 318–325.
Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization on the genetic code. J. Mol. Evol. 33, 412–417.
He, L., Kierzek, R., SantaLucia, J., Walter, A.E., Turner, D.H., 1991. Nearest-neighbor parameters for G.U mismatches. Biochemistry 30, 11124–11132.
Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. Nucleic Acids Res. 28, 3517–3523.
Judson, O.P., Haydon, D., 1999. The genetic code: what is it good for? J. Mol. Evol. 49, 539–550.
Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.
King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798.
Knight, R.D., Freeland, S.J., Landweber, L.F., 2001. Rewiring the keyboard: evolvability of the genetic code. Nat. Rev. Genet. 2, 49–58.
Luo, L., Li, X., 2002. Coding rules for amino acids in the genetic code: the genetic code is a minimal code of mutational deterioration. Orig. Life 32, 23–33.
Mathews, D.H., Sabina, J., Zuker, M., Turner, D.H., 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J. Mol. Biol. 288, 911–940.
Pele, S.R., 1965. Correlation between coding-triplets and amino acids. Nature 207, 597–599.
Percudani, R., Pavesi, A., Ottonello, S., 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol. 268, 322–330.
Ronneberg, T.A., Landweber, L.F., Freeland, S.J., 2000. Testing a biosynthetic theory of the genetic code: fact or artifact? Proc. Natl. Acad. Sci. USA 97, 13690–13695.
Woese, C.R., 1965a. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA 54, 1546–1552.
Woese, C.R., 1965b. Order in the genetic code. Proc. Natl. Acad. Sci. USA 54, 71–75.
Woese, C.R., Dagre, D.H., Dagre, S.A., Kondo, M., Saxinger, W.C., 1966a. On the fundamental nature and evolution of genetic code. Cold Spring Harb. Symp. Quant. Biol. 31, 723–736.
Woese, C.R., Dugne, D.H., Saxinger, W.C., Dayre, S.A., 1966b. The molecular basis for the genetic code. Proc. Natl. Acad. Sci. USA 55, 966–974.
Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72, 1909–1912.