Amino Acid Frequencies Freeze the Genetic Code in Its Error Correcting Pattern

Hamed Shateri Najafabadi[†], *Hani Goodarzi*[†], *Reza Kalhor*[†]

† Department of Biotechnology, Faculty of Science, University of Tehran, Enghelab Ave., Tehran, Iran

Correspondence: Hamed Shateri Najafabadi, Department of Biotechnology, Faculty of Science, University of Tehran, Tehran, Iran.

Phone: +98-913-3314580, Fax: +98-21-8040284, E-mail: shateri@khayam.ut.ac.ir.

Abstract

The genetic code is known to be highly efficient in minimizing the consequences of errors during transcription and translation. When considering the average relative frequencies of amino-acids, the genetic code shows even more ability to reduce the effects of errors, suggesting the canonical genetic code as a probable optimum. However, the factors by which the frequencies of occurrence of amino-acids are determined are not completely known. In the work presented, it is suggested that the relative frequencies of amino-acids are optimized so as to minimize the probability of finding a code which can reduce the consequences of errors more efficiently than the canonical genetic code do. Furthermore, it is shown that the less the probability of finding a code better than a certain code, the more hardly that code undergoes further changes, suggesting that the present frequencies of amino-acids freeze the pattern of codon assignments in the genetic code.

Keywords: amino-acid usage; canonical genetic code; load minimization; synonymous codons; translational error

Introduction

It has been proposed that the genetic code has evolved so as to minimize the consequences of errors during transcription and translation (Ardell, 1998; Crick, 1968; Epstein, 1966; Freeland and Hurst, 1998; Freeland et al, 2000a and b; Gilis et al, 2001; Goldberg and Wittes, 1966; Haig and Hurst, 1991; Knight et al, 1999; Pelc, 1965; Sonneborn, 1965; Woese, 1965). To test this hypothesis, some researchers have tried to estimate the percentage of optimal achievement of the natural code by quantifying the cost of single-base changes. Haig and Hurst (1991) and Freeland and Hurst (1998) and Gilis et al (2001) estimated the amount of optimality of the genetic code by comparing the canonical genetic code with randomly generated codes, each by defining a fitness function as a measure of error minimizing feature of code. Supposedly the genetic code is most efficient in load minimization when the fitness function gets close to its minimum.

Haig and Hurst (1991) defined their fitness function as:

(1) $$\varphi = \sum_{c,c'} \left[h\left(a\left(c\right)\right) - h\left(a\left(c'\right)\right) \right]^2 ,$$

where c and c' are all pairs of codons that can be changed to each other by a single-base mutation; $a\left(c\right)$ and $a\left(c'\right)$ are amino-acids coded by c and c', respectively; and $h\left(a\right)$ returns the hydropathy index of amino-acid a. To measure how close the genetic code is to the actual minimum of φ, they generated random genetic codes, and computed the fraction of the random codes that had smaller values of φ than the canonical genetic code. Testing several hydropathy indices, Haig and Hurst (1991) found that single-base changes had the smallest average effect in the canonical code

when using the difference in polarity between amino-acids as a cost measure for amino-acid substitutions.

Consideration of transition/transversion weightings and different probabilities of mistranslation at the three codon positions led to the proposition of a modified fitness function which modeled the probability of translational errors more accurately (Freeland and Hurst, 1998). Freeland and Hurst (1998) defined their new fitness function as:

$$(2) \qquad \varphi = \sum_{c=1}^{64} \sum_{c'=1}^{64} p(c'|c) \left[h(a(c)) - h(a(c')) \right]^2 ,$$

where $p(c'|c)$ stands for the probability of misinterpretation of codon c as c' (Table 1). With their improved fitness function, Freeland and Hurst estimated the fraction of codes that scored better than the natural code to be in the order of 10^{-6}, instead of 10^{-4} that Haig and Hurst (1991) computed.

Table 1. Quantification of translational errors used to measure the relative efficiency of the natural genetic code in terms of mistranslation (Freeland and Hurst, 1998)

	First base	*Second base*	*Third base*
Relative frequency	0.5	0.1	1
Transition weighting	2	5	1
Combined weighting			
For transitions	1	0.5	1
For transversions	0.5	0.1	1

King and Jukes (1969) denoted that there is a correlation between the frequency in which an amino-acid occurs and the number of synonymous codons it possesses. Based on this finding, Gilis et al (2001) highlighted that the amino-acid frequency is

an important parameter in the optimality of the genetic code and should be taken into account in the fitness function φ. They declared a new fitness function as:

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

where $p(a(c))$ is the frequency in which amino-acid $a(c)$ occurs; $n(a(c))$ is an integer standing for the number of synonymous codons that $a(c)$ possesses; $p(c'|c)$ is the same as stated by Freeland and Hurst (1998); and $g(a(c), a(c'))$ is the cost of substitution of amino-acid $a(c)$ by $a(c')$. Using this fitness function and a new matrix designated Mutation Matrix, as the cost measure of amino-acid substitutions, Gilis et al (2001) estimated the fraction of better codes than the natural code to be in the order of 10^{-9}.

Since both amino-acid frequency and the pattern of codon assignment are parameters which are considered in φ^{faa}, it enables us to study the relationships between amino-acid frequencies and the pattern of the genetic code in terms of load minimization. In the work presented, it is shown that although the relative frequency of occurrence of amino-acids possesses a higher ability to minimize the consequences of translational errors than the mean of random frequencies of amino-acids, it is not significant in the statistical point of view. Instead, the relative frequencies of amino-acids are optimized so as to enable the canonical genetic code overmaster other codes, in terms of load minimization.

Methods

In this work, six different cost measures of amino-acid substitutions are used to calculate the value of φ^{faa}. These measures are Mutation Matrix (Gilis et al, 2001), two different hydrophobicity scales (Nozaki and Tanford, 1971; Engelman et al, 1986), hydropathic character (Kyte and Doolittle, 1982), polar requirement (Woes et al, 1966) and Point Accepted Mutation 74-100 (Benner et al, 1994). Point Accepted Mutation 74-100 is a matrix derived from amino-acid substitution frequencies observed from within highly diverged homologous proteins. Since $PAM_{74\text{-}100}$ is suspected to be biased towards the genetic code, some researchers are not satisfied with its usage as a cost measure for computing the efficiency of the genetic code (DiGiulio, 2001). However, it is used in some previous works (Ardell, 1998; Freeland et al, 2000b; Gilis et al, 2001; Goodarzi et al, 2004).

In the case of Mutation Matrix and $PAM_{74\text{-}100}$, the function g, stated in Equation 3, is defined as:

(4) $$g(a,a') = -MAT(a,a'),$$

where $MAT(a,a')$ indicates the value assigned to the pair a/a' in the used matrix. For other amino-acid substitution cost measures, g is defined as:

(5) $$g(a,a') = [h(a) - h(a')]^2,$$

where $h(a)$ and $h(a')$ are the values in the used index assigned to amino-acids a and a', respectively.

For the sake of comparison, a genetic code reflected matrix is established, designated CGCR Matrix, in which the function g is defined as:

$$(6) \qquad g(a,a') = -\left[\frac{1}{n(a)}\sum_{c}\sum_{c'} p(c'|c) + \frac{1}{n(a')}\sum_{c'}\sum_{c} p(c|c')\right],$$

where $n(a)$ and $n(a')$ are the number of synonymous codons which amino-acids a and a' possess, respectively; c and c' are all codons coding for a and a', respectively; and $p(c'|c)$ and $p(c|c')$ are the same as stated by Freeland and Hurst (1998). This matrix exclusively reflects the structure of the canonical genetic code and is a control for probable tautology resulted from correlation of amino-acid substitution cost measures, enumerated above, with the canonical genetic code. It should be mentioned that the usage of CGCR Matrix in calculation of the value of φ^{faa} results the probability of finding a code better than the canonical genetic code to be in the order of 10^{-29*}.

To avoid any ambiguity, amino-acid usage is defined in this work as:

$$(7) \qquad U = \left\{(a, p(a)) \,\middle|\, a \in A, \sum p(a) = 1\right\},$$

where a is an amino-acid belonging to A, the set of the 20 standard amino-acids; and $p(a)$ is the relative frequency in which amino-acid a occurs. In this work, the natural amino-acid usage represents the relative average frequencies of amino-acids observed in the genomes of a set of organisms including archaea, bacteria and eukaryotes, computed by Gilis et al (2001) and presented in Table 2. A random amino-acid usage

* The program "bc" version 1.06 is used to calculate the probability of finding a code better than the canonical genetic code, using the score distribution of 10^6 random codes and the z-scoring method explained later.

is a set of random p_i's assigned to the twenty amino-acids, so that $0 \le p_i \le 1$ and $\sum_{i=1}^{20} p_i = 1$.

Amino-acid	*p(a) (%)*
Ala	7.8 (2.38)
Arg	5.23 (1.43)
Asp	5.19 (0.81)
Asn	4.37 (1.73)
Cys	1.1 (0.44)
Glu	6.72 (1.24)
Gln	3.45 (1.19)
Gly	6.77 (1.32)
His	2.03 (0.41)
Ile	6.95 (2.16)
Leu	10.15 (0.86)
Lys	6.32 (2.53)
Met	2.28 (0.39)
Phe	4.39 (0.89)
Pro	4.26 (1.01)
Ser	6.46 (1.17)
Thr	5.12 (0.69)
Trp	1.09 (0.25)
Tyr	3.3 (0.63)
Val	7.01 (1.18)

Table 2. The mean frequencies of the individual amino-acids $p(a)$ in the genomes of living organisms (Gilis et al, 2001). The standard deviations of the distributions are given in parentheses. The frequencies $p(a)$ were computed as averages over the frequencies observed in genomes of some archaea, bacteria and eukaryotes.

With these considerations, 10^5 random amino-acid usages were generated, and for each randomly generated amino-acid usage, the value of φ^{faa} was calculated with respect to the canonical genetic code, using the six amino-acid substitution cost measures enumerated above.

To calculate the absolute minimum value of φ^{faa}, φ^{aa} is defined as a measure for the weighted average load of misinterpretation of codons coding for an individual amino-acid:

$$\varphi^{aa}(a) = \frac{1}{n(a)} \sum_{c} \sum_{c'=1}^{64} p(c'|c)\, g(a, a(c')), \tag{8}$$

where c represents all codons coding for amino-acid a; $n(a)$ stands for the number of synonymous codons that a possesses; and $p(c'|c)$ and $g(a, a(c'))$ are the same as stated in Equations 3-5. φ^{faa} can be derived from φ^{aa} using the following equation:

$$\varphi^{faa} = \sum_{i=1}^{20} p(a_i)\varphi^{aa}(a_i), \quad (9)$$

where $p(a_i)$ is the relative frequency of amino-acid a_i. This equation simply shows that φ^{faa} would get into the minimum when the amino-acid which possessed the smallest φ^{aa} through the twenty standard amino-acids got the relative frequency of one (i.e. relative frequencies of other amino-acids were equal to zero). However, this situation is so extreme and implausible. Therefore, we calculated $\varphi^{faa}_{\min}$ as the minimum value of φ^{faa} among the set of amino-acid usages in which $\mu_i - 2\sigma_i \le p_i \le \mu_i + 2\sigma_i$, where μ_i and σ_i are the average relative frequency and the standard deviation of occurrence of amino-acid a_i, respectively (Table 2).

Using this method, the percentage of optimality of the natural amino-acid usage is calculated with respect to the six amino-acid substitution cost measures as:

$$opt = \frac{\varphi^{faa}_{natural} - \varphi^{faa}_{mean}}{\varphi^{faa}_{\min} - \varphi^{faa}_{mean}} \times 100\%, \quad (10)$$

where φ^{faa}_{mean} is obtained from 10^5 random amino-acid usages.

Gilis et al (2001) proposed that using the natural frequencies of occurrence of amino-acids in φ^{faa}, the probability of finding a code which scores better than the natural code is smaller than when using random amino-acid usages. However, they tested

only 10^2 amino-acid usages and actually did not compute the fraction of random amino-acid usages that when used in φ^{faa} result in smaller probabilities of finding better codes. This computation is a difficult one because for each amino-acid usage the value of φ^{faa} should be calculated for a great number of codes.

Since the distribution of the fitness function φ^{faa} shows a significantly high proximity to normal distribution, Goodarzi et al (2004) used the z-value scoring method to estimate the probability of finding a code which scores better than the canonical genetic code:

$$(11) \qquad z_{cgc} = \frac{\mu_{\varphi} - \varphi^{faa}_{cgc}}{\sigma_{cgc}},$$

where μ_{φ} is the mean of distribution of the fitness function φ^{faa} obtained from the random set of codes; φ^{faa}_{cgc} is the value of the fitness function for the canonical genetic code; and σ_{cgc} is the standard deviation of distribution of φ^{faa}. Higher values of z_{cgc} indicate smaller probabilities of finding a better code than the natural one. Figure 1 indicates that calculated z_{cgc} may deviate only very slightly from the actual value when the number of random codes exceeds the value of 10^4. Therefore this scoring method may be used when the number of random codes is a restricting parameter in the computation of the fraction of better codes.

With the above considerations, the value of z_{cgc} was computed for 10^5 random amino-acid usages, by generating 10^4 random genetic codes for each random amino-acid usage, with respect to the six amino-acid substitution cost measures and also

CGCR Matrix. The distribution of z_{cgc} was compared with the value of z_{cgc} for the natural amino-acid usage.

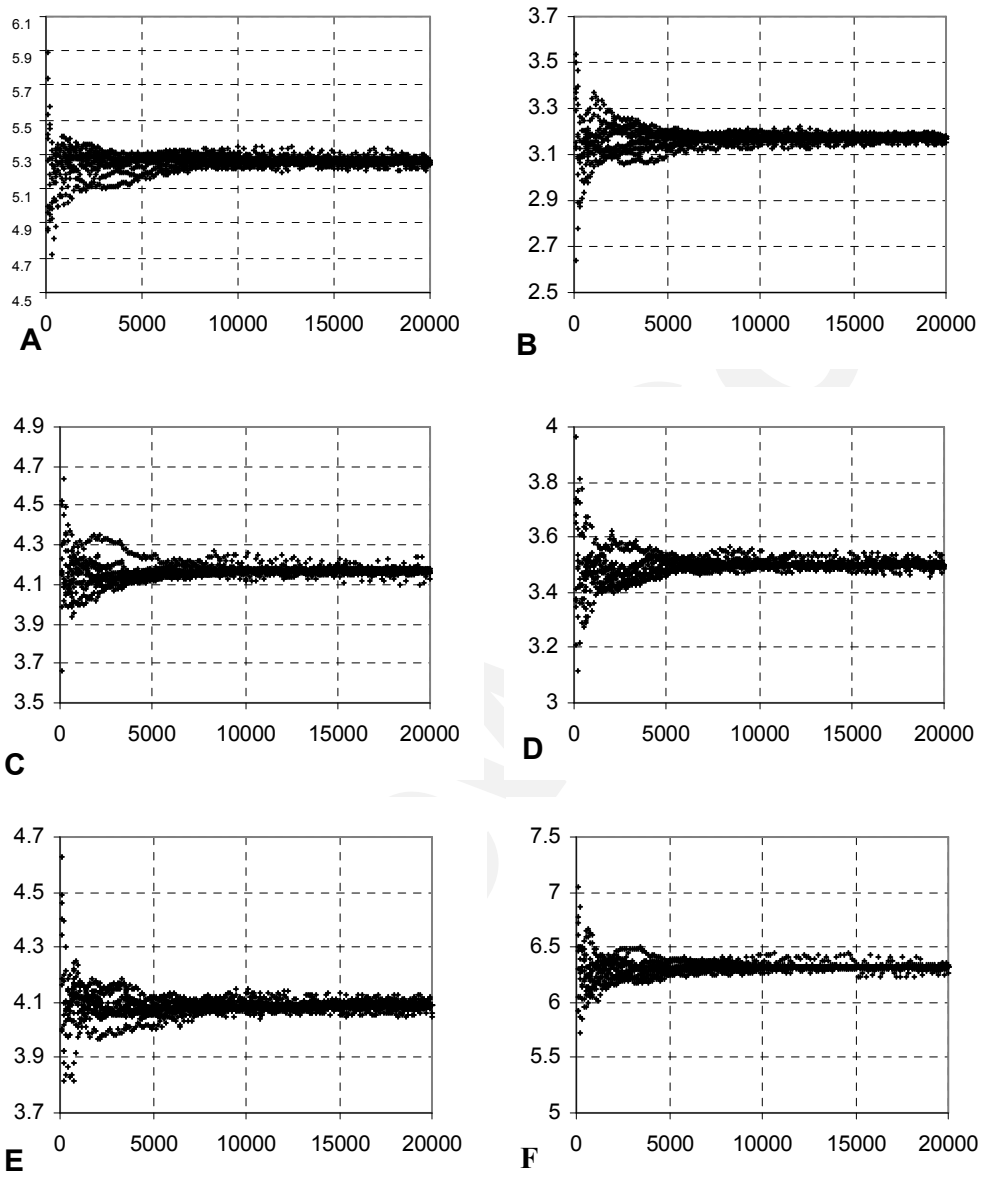

Figure 1. z_{cgc} plotted against the number of random codes generated in different experiments, using six different cost measures: **(A)** Mutation Matrix (Gilis et al, 2001), **(B)** hydrophobicity scale (Nozaki and Tanford, 1971), **(C)** hydrophobicity scale (Engelman et al, 1986), **(D)** hydropathic character (Kyte and Doolittle, 1982), **(E)** polar requirement (Woese et al, 1966) and **(F)** $PAM_{74\text{-}100}$ (Benner et al, 1994).

Results

As shown in Figure 2, when the present frequencies of occurrence of amino-acids (Table 2) are used in φ^{faa}, the resulted value is lower than the mean of distribution of φ^{faa} obtained from 10^5 random amino-acid usages for all the six amino-acid substitution cost measures. However, as Table 3 shows, the percent of random amino-acid usages which possessed lower values of φ^{faa} than the natural amino-acid usage is at least 3.772. This value is achieved when using Mutation Matrix as the cost measure for amino-acid substitutions and is the only significant value ($P<0.05$). Furthermore, as presented in Table 4, the percentage of optimality of the natural amino-acid usage is 63.26 in maximum, which is again reached when Mutation Matrix is used.

Table 3. Percent of random amino-acid usages that possessed lower values of φ^{faa} than the natural frequencies of amino-acids, using the six amino-acid substitution cost measures. See text for description of terms.

Cost measure for amino-acid substitutions	***Percent of random amino-acid usages with lower φ^{faa}***
Mutation Matrix (Gilis et al, 2001)	3.772
hydrophobicity scale (Nozaki and Tanford, 1971)	9.307
hydrophobicity scale (Engelman et al, 1986)	16.095
hydropathic character (Kyte and Doolittle, 1982)	17.509
polar requirement (Woese et al, 1966)	15.899
PAM 74-100 (Benner et al, 1994)	19.245

Table 4. Percentage of optimality of the natural amino-acid usage with respect to load minimization, using the six amino-acid substitution cost measures. See text for description of terms.

Cost measure for amino-acid substitutions	$\varphi^{faa}_{natural}$	φ^{faa}_{mean}	$\varphi^{faa}_{\min}$	$opt.\%$
Mutation Matrix (Gilis et al, 2001)	-2.80	-2.33	-3.08	63.26
hydrophobicity scale (Nozaki and Tanford, 1971)	2.38	3.34	1.57	54.08
hydrophobicity scale (Engelman et al, 1986)	4.01	4.75	3.18	47.36
hydropathic character (Kyte and Doolittle, 1982)	1.97	2.26	1.56	41.76
polar requirement (Woese et al, 1966)	3.40	4.28	2.29	44.33
PAM 74-100 (Benner et al, 1994)	-2.13	-1.94	-2.39	41.98

Figure 3 shows the distribution of z_{cgc} obtained from 10^5 random amino-acid usages. As presented in Table 5, z_{cgc} value of the natural amino-acid usage is significantly higher than the mean of distribution of z_{cgc} obtained from random amino-acid usages for all the six amino-acid substitution cost measures ($P<0.05$). More significantly, in the case of Mutation Matrix and $PAM_{74\text{-}100}$ no amino-acid usages with higher values of z_{cgc} than the natural amino-acid usage were found. Therefore, the probability of finding an amino-acid usage with higher z_{cgc} than the natural amino-acid usage was estimated, as explained in Appendix A.

Table 5. Fraction of random amino-acid usages that possessed higher values of z_{cgc} than the natural frequencies of amino-acids, with respect to the six amino-acid substitution cost measures. See text for description of terms.

Cost measure for amino-acid substitutions	***Fraction of random amino-acid usages with higher z_{cgc}***
Mutation Matrix (Gilis et al, 2001)	4.0×10^{-10}*
hydrophobicity scale (Nozaki and Tanford, 1971)	1.4×10^{-2}
hydrophobicity scale (Engelman et al, 1986)	8.3×10^{-3}
hydropathic character (Kyte and Doolittle, 1982)	2.1×10^{-2}
polar requirement (Woese et al, 1966)	2.2×10^{-2}
PAM 74-100 (Benner et al, 1994)	1.3×10^{-7}*
CGCR Matrix	1.6×10^{-1}

* Since no amino acid usages with higher z_{cgc} were found through the 10^5 randomly generated amino acid usages, the fraction of amino acid usages with higher values of z_{cgc} was estimated using the method explained in Appendix A.

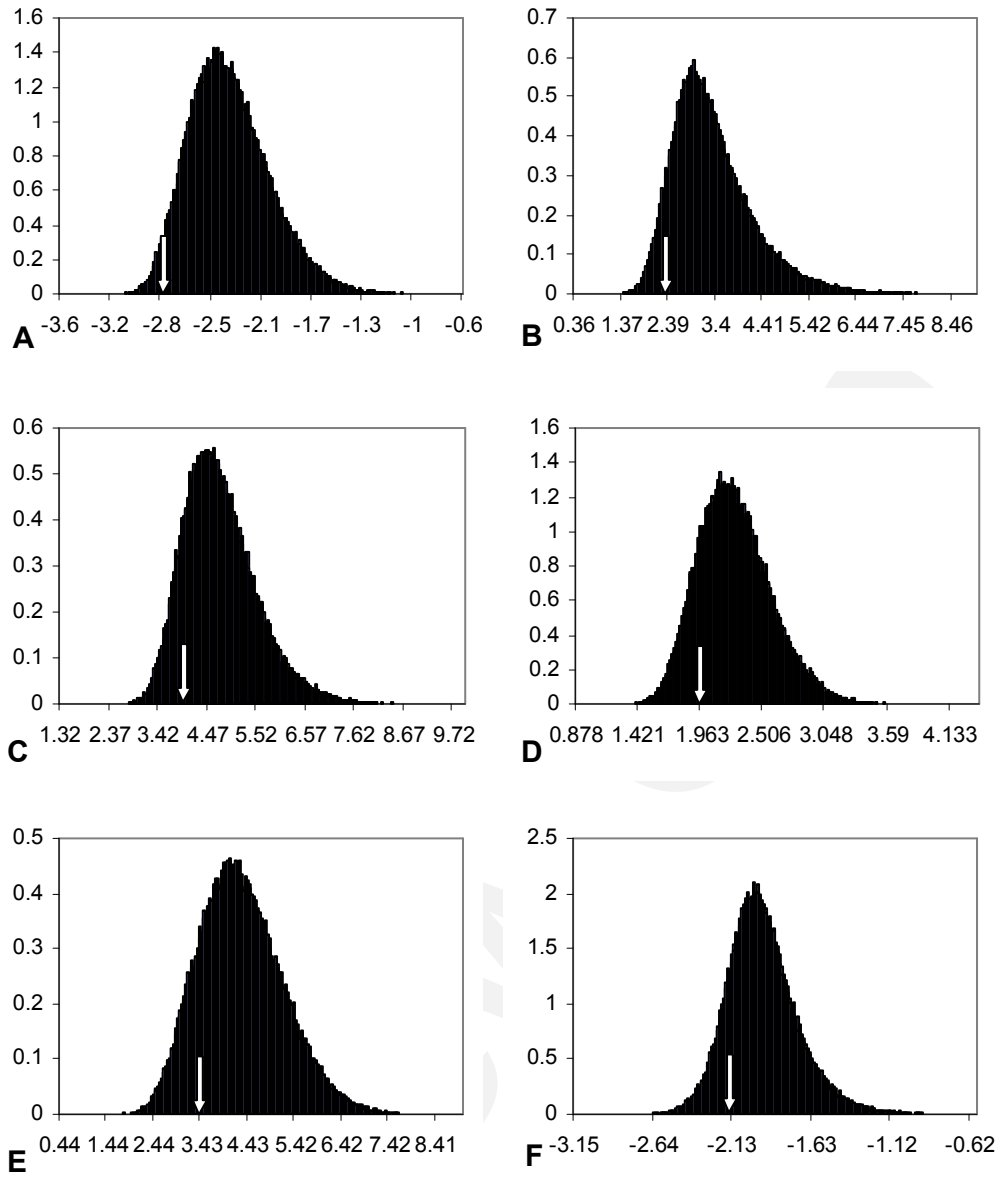

Figure 2. The distribution of φ^{faa} obtained from random amino-acid usages using six different cost measures: **(A)** Mutation Matrix (Gilis et al, 2001), **(B)** hydrophobicity scale (Nozaki and Tanford, 1971), **(C)** hydrophobicity scale (Engelman et al, 1986), **(D)** hydropathic character (Kyte and Doolittle, 1982), **(E)** polar requirement (Woese et al, 1966) and **(F)** $PAM_{74\text{-}100}$ (Benner et al, 1994). The natural amino-acid usage is shown by arrows. See text for description of terms.

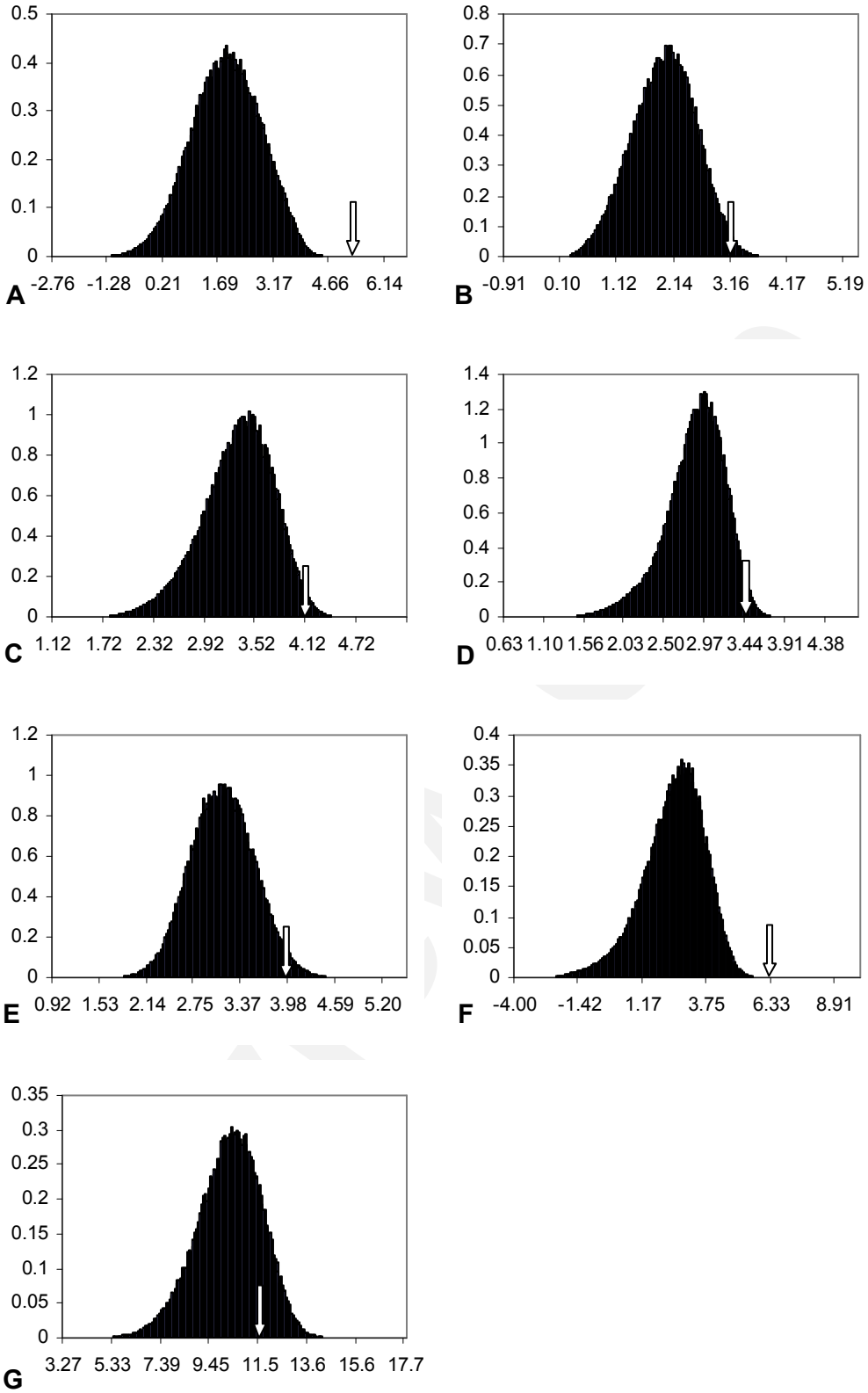

Figure 3. The distribution of z_{cgc} through random amino-acid usages using six different cost measures: **(A)** Mutation Matrix (Gilis et al, 2001), **(B)** hydrophobicity scale (Nozaki and Tanford, 1971), **(C)** hydrophobicity scale (Engelman et al, 1986), **(D)** hydropathic character (Kyte and Doolittle, 1982), **(E)** polar requirement (Woese et al, 1966), **(F)** $PAM_{74\text{-}100}$ (Benner et al, 1994) and **(G)** CGCR Matrix. The natural amino-acid usage is shown by arrows. See text for description of terms.

Discussion

The distribution of φ^{faa} through randomly generated amino-acid usages shows that there are lots of amino-acid usages that have smaller values of φ^{faa} than the natural amino-acid usage, and therefore can result in more accurate translation with respect to load minimization. Although the value of φ^{faa} for the natural frequencies of amino-acids is always below the mean of distribution, this latter result is only significant when Mutation Matrix is used and since this significance is not strong, it may be simply due to chance. So load minimization may not be the case that shapes amino-acid frequencies (especially supported by data presented in Table 4) and there are other factors that override load minimization. However, these other factors have some kind of association with load minimization that makes the natural amino-acid usage have lower value of φ^{faa} than the mean value.

Figure 3 and data presented in Table 5 suggest that enabling the canonical genetic code to beat other codes may be a factor that determines the frequencies of amino-acids. The natural amino-acid usage is arranged so as to result in the minimum probability of finding a code better than the canonical genetic code. Therefore the present frequencies of amino-acids restrict the rate in which the genetic code changes, if minimizing the effects of errors during transcription and translation is the goal for which the pattern of the genetic code is shaped, as we believe it is. This interpretation would be agreeable if it was accepted that the better adapted a code became, the slower further improvement would be, as Freeland et al (2000a) suggested. Although this criterion is obvious from the bell-shaped distribution of possible codes' error values (Freeland, 2002), since it does not satisfy some researchers (DiGiulio, 2000), a

genetic based algorithm was developed (Appendix B) to show that the less the probability of obtaining a better code than the present code became, the less likely the code would be improved by further changes (Figure 4).

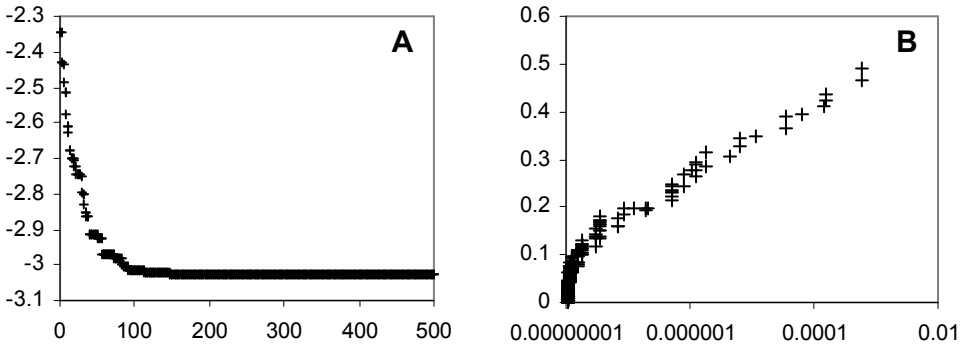

Figure 4. (A) The minimum value of φ^{faa} observed in different steps through a simplified simulation of genetic code improvement (Appendix B), plotted against the number of generations passed. **(B)** The fraction of mutant codes that took place in the next generation, plotted against the probability of finding a code better than the best code of that generation (logarithmic scale).

However, like any stable system, the rate-limiting effect of amino-acid usage on the changes of the genetic code can be fixed when the canonical genetic code has a reciprocal limiting effect on the alteration rate of the natural amino-acid usage, directly or indirectly. It should be confessed that no proper direct mechanism was found by authors' knowledge for rate-limiting effect of the genetic code on natural amino-acid usage alteration.

When CGCR Matrix was used for computation of φ^{faa}, the fraction of random amino-acid usages which possessed higher values of z_{cgc} than the natural amino-acid usage was 0.16408. This value, being compared to the values obtained from the six amino-acid substitution cost measures used in this work (Table 5), ensures that the results are not achieved tautologically due to the correlation of the used amino-acid substitution

cost measures with the canonical genetic code, as was suspected for $PAM_{74\text{-}100}$. It implies that $PAM_{74\text{-}100}$ can be used to indicate the optimality of the amino-acid usage in minimizing the probability of finding a code better than the canonical genetic code. Still the usage of $PAM_{74\text{-}100}$ for measuring the optimality of the canonical genetic code in minimizing the effects of translational errors is the subject of suspicion; especially that CGCR Matrix indicates a correlation coefficient of about 0.49 with $PAM_{74\text{-}100}$ which supports that $PAM_{74\text{-}100}$ reflects the structure of the genetic code. However, it should mentioned that CGCR Matrix indicates a correlation coefficient of about 0.37 with Mutation Matrix though, owing to the method by which this matrix is achieved (Gilis et al, 2001), it is independent of code's structure. This latter description argues with the confidence of interpretations which are based solely on the correlation coefficient between CGCR Matrix or any similar matrix and $PAM_{74\text{-}100}$.

Figure 5 explains why the value of φ^{faa} for the natural amino-acid usage is below the mean of distribution of φ^{faa} through random amino-acid usages. For all the six amino-acid substitution cost measures used in this work, there is a significant negative correlation between the value of φ^{faa} for an amino-acid usage and the value of z_{cgc} which that amino-acid usage possesses, so that lower values of φ^{faa} are associated with higher values of z_{cgc} and vice versa. This correlation suggests that the value of φ^{faa} for the natural amino-acid usage is below the mean value not due to evolutionary forces acting directly on φ^{faa}; instead, it is because of evolutionary forces that act on z_{cgc}, and φ^{faa} is affected because of its association with z_{cgc}. Since the correlation of φ^{faa} and z_{cgc} is not a complete one, z_{cgc} takes a maximum value while φ^{faa} is only below the mean value.

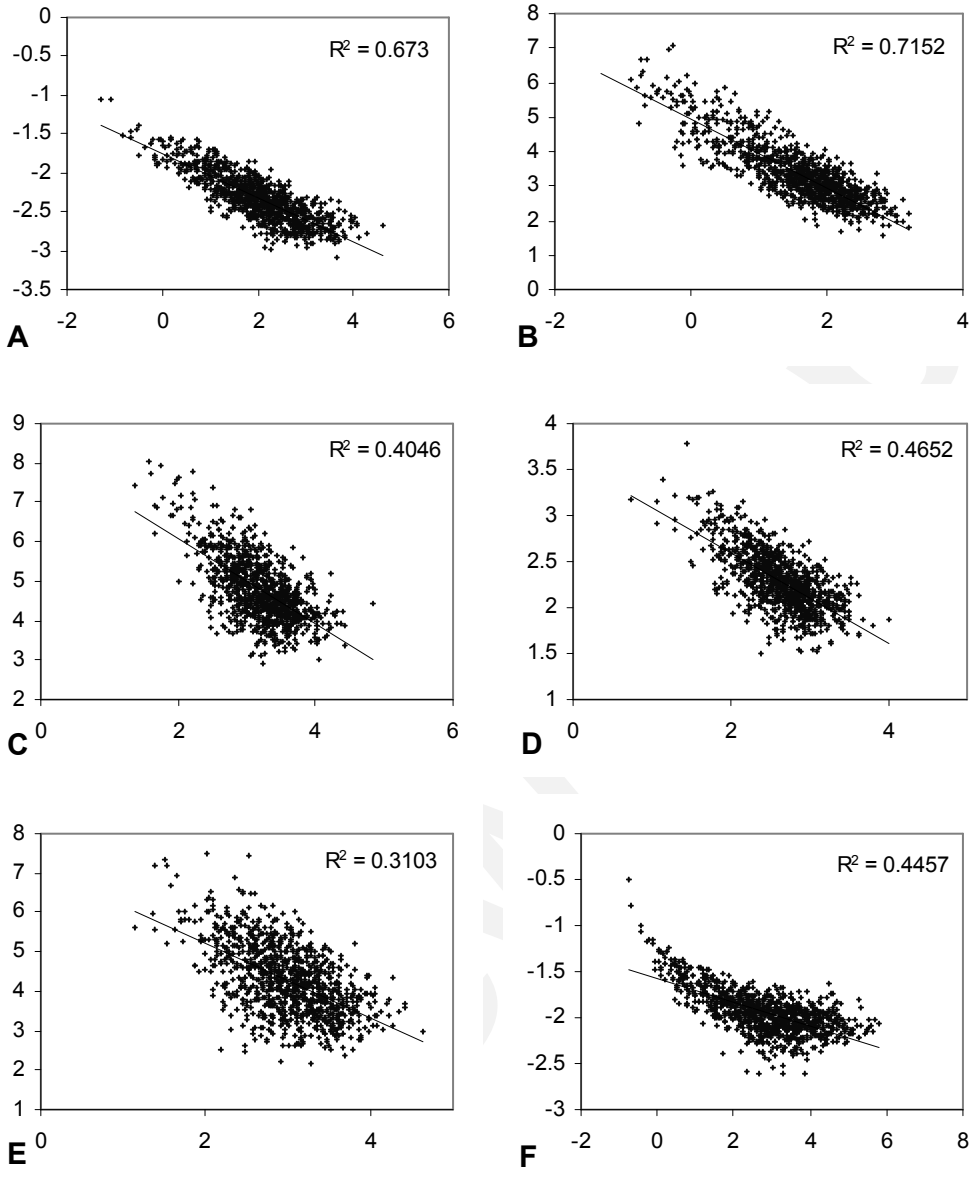

Figure 5. φ^{faa} plotted against z_{cgc} for 10^3 random amino-acid usages using six different cost measures: **(A)** Mutation Matrix (Gilis et al, 2001), **(B)** hydrophobicity scale (Nozaki and Tanford, 1971), **(C)** hydrophobicity scale (Engelman et al, 1986), **(D)** hydropathic character (Kyte and Doolittle, 1982), **(E)** polar requirement (Woese et al, 1966) and **(F)** $PAM_{74\text{-}100}$ (Benner et al, 1994). See text for description of terms.

Appendix A

To estimate the probability of finding an amino-acid usage with higher z_{cgc} than the natural amino-acid usage when no random amino-acid usages were found to beat it, the following procedure was used. From the values of z_{cgc} obtained within randomly generated amino-acid usages, the probability function $\pi(z_{cgc})$ was computed in the range of observed z_{cgc}'s, as the fraction of random amino-acid usages that had higher values of z_{cgc} than a given z_{cgc}. A polynomial of degree six was fitted to $\log(\pi(z_{cgc}))$ and this curve was extrapolated up to the value of z_{cgc} for the natural amino-acid usage. A similar approach was used when the probability of finding a code with lower φ^{faa} than a certain value (e.g. the value of φ^{faa} for the canonical genetic code) was required (Gilis et al, 2001). The probability function $\pi(\varphi^{faa})$ was computed from a set of 10^6 random codes.

Appendix B

A genetic based algorithm was developed to show that the fitter a code became, the less likely the code would be improved by further changes. A set of 10^3 random codes was generated as the initial population. In each generation, each member code of population reproduced a mutant form of itself, in which two amino-acids, chosen randomly, swapped their codons. From within the parent codes and the mutant child codes, the first 10^3 codes that had the smallest values of φ^{faa} were chosen to make the next generation. In each generation, the minimum value of φ^{faa} and the fraction of mutant codes that took place in the next generation were computed. The probability of finding a code better than the best code of each generation was calculated by comparing the minimum value of φ^{faa} in that generation with the distribution of φ^{faa}, obtained from 10^6 randomly generated codes, using the method explained in Appendix A when needed. Mutation Matrix is used as a representative measure for cost of amino-acid substitutions.

Acknowledgements

Authors are grateful to Elahe Elahi for her useful comments and everlasting support.

References

1. Ardell DH (1998). On Error Minimization in a Sequential Origin of the Standard Genetic Code. J Mol Evol 47:1–13
2. Benner SA, Cohen MA, Gonnet GH (1994). Amino acid substitution during functionally constrained divergent evolution of protein sequences. Protein Eng 7(11):1323-32.
3. Crick FHC (1968). The origin of the genetic code. J Mol Biol 38:367-379.
4. Di Giulio M (2001). The origin of the genetic code can not 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.
5. Di Giulio M (2000). Genetic Code Origin and the Strength of Natural Selection. J theor Biol 205:659-661.
6. Epstein CJ (1966). Role of the amino-acid "code" and of selection for conformation in the evolution of proteins. Nature 210:25-28.
7. Engelman DM, Steitz TA, Goldman A (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Chem 15:321–353.
8. Freeland SJ (2002). The Darwinian Genetic Code: An Adaptation for Adapting? Genetic Programming and Evolvable Machines 3: 113–127.

9. Freeland SJ, Hurst LD (1998). The genetic code is one in a million. J Mol Evol 47: 238-248.

10. Freeland SJ, Knight RD, Landweber LF (2000a). Measuring adaptation within the genetic code. Trends Biochem Sci 25:44-45.

11. Freeland SJ, Knight RD, Landweber LF, Hurst LD (2000b). Early Fixation of an Optimal Genetic Code. Mol Biol Evol 17(4):511–518.

12. Gilis D, Massar S, Cerf NJ, Rooman M (2001). Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biology 2(11):research0049.1–0049.12.

13. Goldberg AL, Wittes RE (1966). Genetic code: aspects of organization. Science 153:420-424.

14. Goodarzi H, Nejad HA, Torabi N (2004). On the optimality of the genetic code with the consideration of the termination codons. Biosystems J ,in press.

15. Haig D, Hurst LD (1991). A quantitative measure of error minimization on the genetic code. J Mol Evol 33: 412-417.

16. King JL, Jukes TH (1969) Non-Darwinian evolution. Science 164:788-798.

17. Knight RD, Freeland SJ, Landweber LF (1999). Selection, history, and chemistry: the three faces of the genetic code. Trends Biochem Sci 24:241-247.

18. Kyte J, Doolittle RF (1982). A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132.

19. Nozaki Y, Tanford C (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishments of a hydrophobicity scale. J Mol Chem 246:2111

20. Pelc SR (1965). Correlation between coding-triplets and amino acids. Nature 207:597-599.

21. Sonneborn TM (1965). Degeneracy of the genetic code: extent, nature and the genetic implications. In Evolving Genes and Proteins. Edited by Bryson V, Vogel HJ. New York: Academic Press; 377-397.

22. Woese CR (1965). On the evolution of the genetic code. Proc Natl Acad Sci USA 54:1546-1552.

23. Woese CR, Dugre DH, Dugre SA, Kondo M, Saxinger WC (1966). On the fundamental nature and evolution of the genetic code. Cold Spring Habor Symp Quant Biol 31:723–736.