Available online at www.sciencedirect.com
SCIENCE DIRECT®
BioSystems 77 (2004) 163–173
Bio Systems
www.elsevier.com/locate/biosystems

On the optimality of the genetic code, with the consideration of termination codons

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

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

Received 14 February 2004; received in revised form 9 May 2004; accepted 25 May 2004

Abstract

The existence of nonrandom patterns in codon assignments is supported by many statistical and biochemical studies. The canonical genetic code is known to be highly efficient in minimizing the effects of mistranslation errors and point mutations. For example, it is known that when an error induces the conversion of an amino acid to another, the biochemical properties of the resulting amino acid are usually very similar to that of the original. Prior studies include many attempts at quantitative estimation of the fraction of randomly generated codes which, based upon load minimization, score higher than the canonical genetic code.

In this study, we took into consideration both the relative frequencies of amino acids and nonsense mistranslations, factors which had been previously ignored. Incorporation of these parameters, resulted in a fitness function (φ) which rendered the canonical genetic code to be highly optimized with respect to load minimization. Considering termination codons, we applied a biosynthetic version of the coevolution theory, however, with low significance. We employed a revised cost for the precursor-product pairs of amino acids and showed that the significance of this approach depends on the cost measure matrix used by the researcher. Thus, we have compared the two prominent matrices, point accepted mutations 74–100 (PAM_{74-100}) and mutation matrix in our study.
© 2004 Elsevier Ireland Ltd. All rights reserved.

Keywords: Genetic code; Evolution; Optimality; Amino acid frequency

1. Introduction

The canonical genetic code was long thought to be a "frozen accident" (Crick, 1968). However, later studies revealed the existence of small variations in the genetic code, which prompted a search for an applicable theory (Osawa, 1995). Several hypotheses have been presented to explain the evolution of the genetic code to its present form (Crick, 1968; Dillon, 1973; Wong, 1975; Woese, 1965; Pele, 1965; Goldberg and Wittes, 1966; Woese et al., 1966; Szathmary, 1991; Szathmáry and Zintzaras, 1992; Goldman, 1993; Jukes, 1997; Judson and Haydon, 1999; Houen, 1999; Ronneberg et al., 2000; Freeland et al., 2000; Ardell and Sella, 2001; Di Giulio, 2001a,b; Freeland, 2002; Sella and Ardell, 2002; Di Giulio, 2003). One of the latest scenarios, designated "code-metabolism coevolution", postulates that the earliest genetic code used a small number of prebiotically synthesized amino acids, and subsequently expanded to its present form by incorporating codons for novel derivatives of these premordial amino acids as biosynthetic pathways evolved (Wong, 1975, 1976, 1981; Di Giulio, 1999; Di Giulio, 2000; Wong and Bronskill, 1979). However, the statistical significance of this theory was debated by Ronneberg et al. (2000).

* Corresponding author. Tel.: +98 21 8058210; fax: +98 21 8040284.
E-mail address: lucipherase@eml.cc (H. Goodarzi).

0303-2647/$ – see front matter © 2004 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2004.05.031

On the other hand, statistical studies have supported the theory that the genetic code has evolved so, as to minimize the consequences of errors during transcription and translation (Woese, 1965; Sonneborn, 1965; Haig and Hurst, 1991; Freeland and Hurst, 1998; Ardell, 1998; Knight et al., 1999; Freeland et al., 2000). In order to test this hypothesis, attempts have been made to compute the optimality of the genetic code by quantifying the cost of single-base changes as reflected in protein sequences (Alff-Steinberger, 1969; Haig and Hurst, 1991; Ardell, 1998; Freeland and Hurst, 1998; Gilis et al., 2001).

In an improved approach, Haig and Hurst (1991) and Freeland and Hurst (1998) compared the canonical genetic code with randomly generated codes in order to assess the relative efficiency of the natural code in limiting the consequences of transcription and translation errors. Haig and Hurst (1991) considered several fitness functions, φ, based on different physicochemical parameters as cost functions for estimating the efficiency of a code in load minimization. Proposing this function φ supposedly evolved towards a minimum through evolution, they then generated a set of random genetic codes, and computed the fraction of those codes which scored better (i.e. have a smaller value of φ) than the natural code. They found that when using differences in polarity (hydropathy) between amino acids as a cost measure, single base changes in the natural code had a very small average effect (Woese et al., 1966). Thus the following fitness function was introduced to estimate load minimization:

$$\varphi^{\text{Haig and Hurst}} = \sum_{c=1}^{64}\sum_{c'=1}^{64} p(c'|c)(h(a(c)) - h(a(c')))^2 \tag{1}$$

In the equation above, $p(c'|c)$ denotes the probability of codon c being misread as c' (in this case, the probabilities of single base changes at all codon positions and for all types of mutation are considered equal). In the function presented and throughout this text, $a(c)$ is the amino acid coded by codon c and $h(a)$ is the hydropathy index of amino acid a. Calculating φ scores for thousands of randomly generated codes, Haig and Hurst (1991) found the fraction of random codes that beat the natural genetic code to be in the order of 10^{-4}. Furthermore, by taking into account the transition/transversion biases and differences in the effects of changes at the three codon positions on mistranslation, a new fitness function was proposed which more accurately modeled the probability of translational errors (Freeland and Hurst, 1998). With the improved φ function the fraction of random genetic codes that are better than the natural one decreased to 10^{-6}.

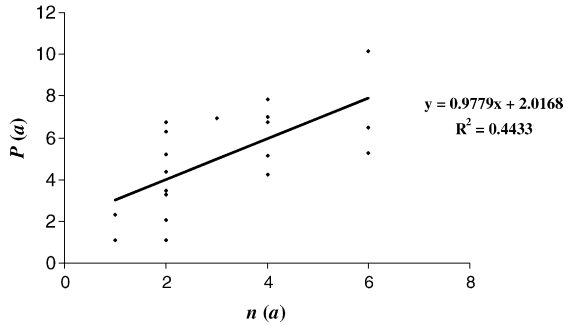

Fig. 1. The relative frequency ($P(a)$) of amino acid a as a function of $n(a)$, the number of synonyms that code for it. The linear regression line (with correlation coefficient of 0.66) is drawn.

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, they recognized a general prevailing pattern (Table 1). Plotting the number of synonymous codons for amino acids versus their frequencies revealed a high correlation between these variables, signifying the importance of this factor in the optimization of genetic code (Fig. 1). They further improved φ by considering properties in addition to polarity to measure the effect of different amino acids on protein conformation and stability. They devised a cost function designated “mutation matrix” by evaluating, in silico, the change in folding free energy caused by all possible point mutations in a set of protein structures. Mutation matrix is alleged to be unbiased towards the genetic code, based on 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 for studying the genetic code (Gilis et al., 2001). The derivation is based on a dataset of 141 protein structures, determined by X-ray crystallography (studied by Wintjens et al., 1996). Each residue in each of these 141 proteins is mutated into the 19 non-wild-type amino acids and for each of these mutations, the change in folding free energy is

Table 1
The mean frequencies (%) of the individual amino acids ($p(a)$) = P(a)/100) in the genomes of living organisms

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)

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 archaea (*Aeropyrum pernix* K1, *Archaeoglobus fulgidus*, *Halobacterium* sp. NRC-1, *Methanococcus jannaschii*, *Methanobacterium thermoautotrophicum*, *Pyrococcus abyssi*, *Pyrococcus horikoshi* and *Thermoplasma acidophilum*), bacteria (*Aquifex aeolicus*, *Bacillus halodurans*, *Bacillus subtilis*, *Borrelia burgdorferi*, *Buchnera aphidicola*, *Campylobacter jejuni*, *Chlamydia trachomatis*, *Deinococcus radiodurans*, *Escherichia coli* K-12, *Haemophilus influenzae*, *Mycobacterium leprae*, *Mycobacterium tuberculosis*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, *Pasteurella multocida*, *Pseudomonas aeruginosa*, *Rickettsia prowazekii*, *Thermotoga maritima*, *Treponema pallidum*, *Ureaplasma parvum*, *Vibrio cholerae* and *Xylella fastidiosa*) and eukaryotes (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* and *Saccharomyces cerevisiae*). This chart was produced by Gilis et al. (2001).

evaluated using the procedure detailed by Gilis et al. (2001). In addition to their own mutation matrix, Gilis et al., also made use of the point accepted mutations 74–100 ($PAM_{74–100}$) scoring matrix (Gilis et al., 2001). The point accepted mutations (PAM) matrices are a family of matrices derived from amino acid substitution frequencies observed within homologous proteins. $PAM_{74\text{-}100}$ is considered the least biased PAM matrix towards genetic code, as it is based on highly diverged sequences and only reflects the physicochemical similarities of the amino acids and not codon proximity (Freeland et al., 2000). For example, when two codons are near each other (in terms of mutation), an amino acid in one protein might mutate to the other simply due to the high probability of this phenomenon and irrespective of similarity between the corresponding amino acids. Yet, when highly diverged proteins are compared, the amino acid replacements occur due to the ability of the corresponding amino acids to compensate each other and not because of randomly occurring mutations. Not withstanding its advantages, a number of researchers are not satisfied with PAM and other similar matrices (DiGiuli, 2001).

Gilis et al. (2001), conducted the classic experiment of generating random genetic codes, while applying their new fitness function,

$$\varphi^{\text{faa}} = \sum_{c=1}^{64} \frac{p(a(c))}{n(a(c))} \sum_{c'=1}^{64} p(c'|c) g(a(c), a(c')) \tag{2}$$

In the above equation and throughout the text, $p(a(c))$ is the frequency of amino acid a coded by codon c ($0 < p(a(c)) < 1$), $n(a(c))$ is an integer standing for the number of synonymous codons of amino acid a. $g(a(c),a(c'))$ is a cost measure function which reflects the deleterious effect of the amino acid substitution resulting from the misreading of codon c as c' (usually PAM or mutation matrices are used). Throughout this text, the functions $p(a)$ and $n(a)$ refer, respectively, to the relative frequency and the number of synonyms of an amino acid. The application of mutation matrix to the natural code and randomly generated codes resulted in a lower fraction of random codes whose value of φ is lower than that of natural code ($f = 2 \times 10^{-9}$). Application of $PAM_{74–100}$ matrix produced the same results.

In the work presented, we devised a set of experiments, introducing a new group of parameters which were previously overlooked. The above studies, made no allowance for mistranslations leading to stop codons. Logically, if the canonical genetic code is optimized in load minimization, it should also be capable of withstanding effects of nonsense mutations as well. In this paper, not only has this been considered, but some of its consequences have been also studied. We have used the average frequencies of codons located near the stop codons (in terms of mutation) as the determining cost of a known mistranslation. We also made a comparison between the

PAM_{74-100} and mutation matrices regarding a biosynthetic version of code-metabolism coevolution theory, both in the presence and absence of nonsense mistranslations. The overall impacts of these parameters resulted in a lower f (the fraction of random codes whose value of φ is lower than that of natural code) and consequently, a reduction in the probability of obtaining, by chance, a code with a higher efficiency.

2. Methods

In the work presented, we first defined a new fitness function, φ^{HH}, on the same basis as Gilis et al. (2001), but in a slightly different manner. We applied the relative frequencies of amino acids, given in Table 1, to form the function below:

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

The variables and functions are the same as those defined in φ^{faa} except for the novel function $f(c)$, which is defined as follows,

$$f(c) = \begin{cases} \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt & y \le 0 \\ 1 - \left(\int_{-\infty}^{y} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt \right) & y > 0 \end{cases},$$

$$y = \frac{(n(a(c))/n(\text{total}) - n(sc)) - p(a(c))}{S\sigma(a(c))} \qquad (4)$$

where n(total) and n(SC) are the total number of codons, and the number of stop codons, respectively. Note that in all our randomly generated codes, n(total) equals 64 and n(SC) returns 3. The value designated S in the y function is a constant value, fixed at 100 in this research.

In Eq. (3), PAM_{74-100} and mutation matrices were used as the substitution matrices of $g(x,y)$ (the cost function). Freeland and Hurst (1998; Gilis, 2001) have chosen the following values for $p(c'|c)$ (the probability of mistranslation at each codon position), which we also use here:

- $p(c'|c)=1/N$ if c and c' differ in the third base;
- $p(c'|c)=1/N$ if c and c' differ in the first base and cause a transition;
- $p(c'|c)=0.5/N$ if c and c' differ in the first base and cause a transversion;
- $p(c'|c)=0.5/N$ if c and c' differ in the second base and cause a transition;
- $p(c'|c)=0.1/N$ if c and c' differ in the second base and cause a transversion;
- $p(c'|c)=0$ otherwise.

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

Rules for generating random genetic codes (Freeland and Hurst, 1998):

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.

In the next step we tried to include nonsense mistranslations along with other types of error, already being considered. We applied this issue based upon the fact that if codons closer to termination codons (in terms of mutation) occur less frequently in DNA, less codons will be likely to mutate to stop codons, thus improving code fitness. The term $p(a)/n(a)$ had already been considered in φ^{HH} as an indication of codon frequency, so we simply defined a variable, K, to alter our cost function in the way that it would include nonsense mistranslations ($g(x, \text{stop codon}) = -K$). Note that higher values of K lead to lower scores of φ^{HH}.

According to the "code-metabolism coevolution" theory, codon assignments coevolved with the invention of biosynthetic pathways for new amino acids, which means a product amino acid being synthesized from a precursor one, usurped codons assigned to its precursor (Ronneberg et al., 2000; Amirnovin, 1997; Osawa et al., 1992). The theory postulates that biosynthetically related amino acids in the genetic code should be positioned near each other in terms of mutation. In this phase, we used the 12 biochemically related precursor-product amino acid pairs introduced by Ronneberg et al. (2000), shown in Table 2. All the pairs are located in the same biochemical branch

Table 2
Precursor-product pairs, indicated by Ronneberg et al. (2000)

Precursor	Product	Precursor	Product
Ser	Trp	Asp	Asn
Ser	Cys	Asp	Met
Phe	Tyr	Asp	Lys
Thr	Ile	Glu	Pro
Gln	His	Asp	Arg
Glu	Gln	Asp	Thr

These are self consistent and biologically relevant set of precursors as the closest direct antecedents to their product amino acid in an energetically favorable non-transamination pathway.

and all conversions are energetically favorable. We dedicated extra weight to these pairs by altering our scoring matrix using the equation:

$$g(a, a') = Inc|g(a, a')| + g(a, a') \text{ where } 0 < Inc < 1. \quad (5)$$

A cost measure is added by a known fraction of itself, if the pair (a,a') is in the set of pairs shown in Table 2, thus, favoring the two amino acids of a pair being positioned near each other. Note that when approaching coevolution theory, our study considers a period of time, when all the twenty amino acids had been included in the genetic code.

In many of our experiments, there was no random code found to score higher than the canonical genetic code. However, since the distributions of our results were very similar to normal distribution (see Figs. 2 and 5), we defined Z-value as a measure of optimality:

$$Z = \frac{\varphi_{\text{cgc}} - \mu}{\sigma} \quad (6)$$

where μ the average score of the generated codes, σ the standard deviation of the scores obtained, and, φ_{cgc} the fitness score of the canonical genetic code.

Higher Z-values imply lower probabilities of obtaining a better random code by chance.

3. Results

Using the PAM_{74-100} scoring matrix as a cost function, we applied our φ^{HH} to 2×10^9 random genetic

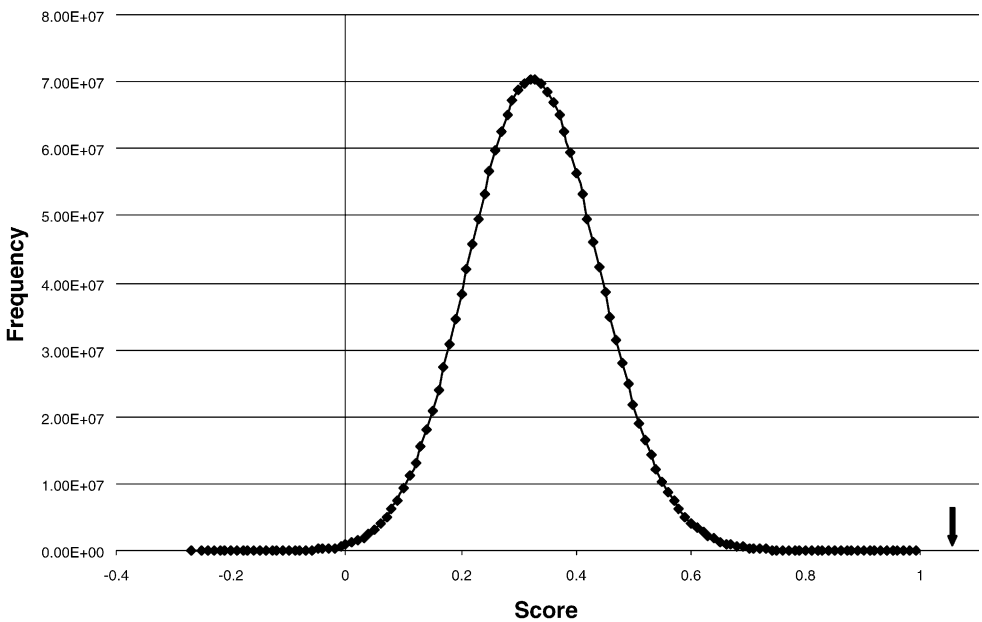

Fig. 2. The score distribution of 2×10^9 randomly generated codes, applying φ^{HH}. The score of the canonical genetic code is indicated by an arrow.

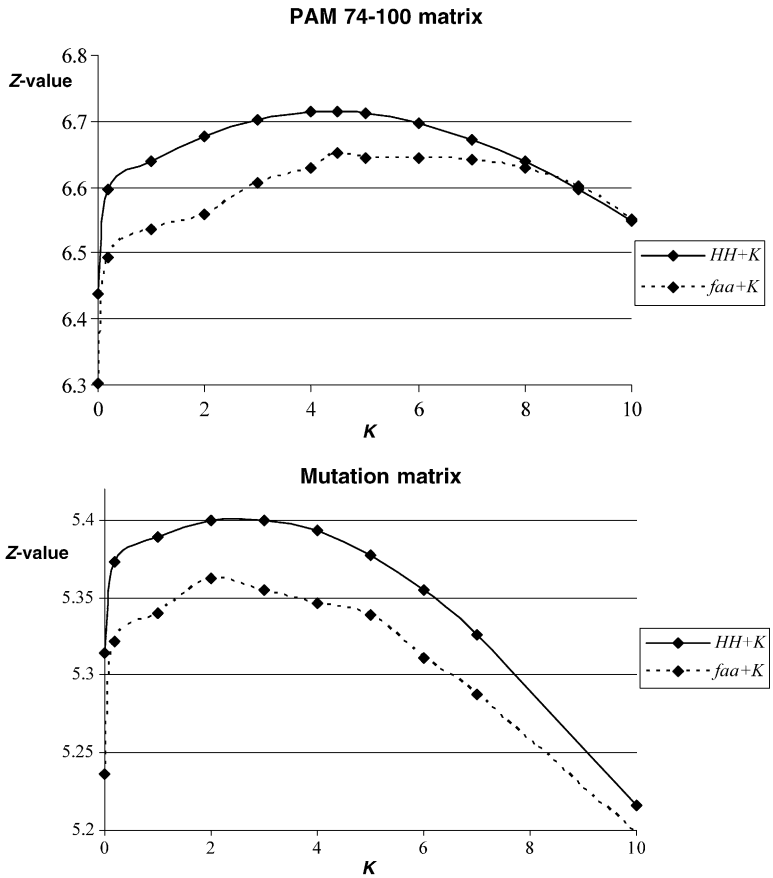

Fig. 3. Z-values, computed versus different values of K, reflecting different cost measures appointed to the nonsense mistranslations ($g(a, \text{stop codon}') = -K$). Z-value is maximum when $K = 4.5$ and PAM_{74-100}matrix are applied; whereas, the optimal value of K is 3.0 for mutation matrix. Note that these values are relatively lower than other values in the corresponding matrices which are in accordance with the high deleterious effects of nonsense mutations. In case of φ^{faa} (presented by faa + K in the graph) the plots behave approximately the same as φ^{HH} (presented by HH + K); however, with lower significance.

codes. The results were astonishing, as no random genetic code scored higher than the canonical one (compare this result with an average of 2 better codes in every 10^9 ones, obtained by Gilis et al., 2001). The distribution of scores is shown in Fig. 2, where the score of the canonical genetic code is indicated by an arrow.

Next, we calculated Z-value, while including nonsense mistranslations. As it will be discussed later, it is alleged that, the lower the mean frequency of a codon, the less reduction it would cause in the total fitness due to its nonsense mistranslations. In other words, we have, in effect, changed φ^{HH} to:

$$\varphi^{\text{HH}} = \sum_{c=1}^{64} \left(\frac{p(a(c))}{n(a(c))} f(c) \sum_{c'=1}^{64} p(c'|c) g(a(c), a(c'))\right) + \sum_{c'=1}^{3} p(\text{sc}|c) g(a(c), \text{sc})) \quad (7)$$

Note that ‘sc’, stands for “stop codon”.

Setting $g(a, \text{stop codon}) = -K$, we plotted Z-values versus values of K, aiming to determine an appropriate function as the cost measure of the nonsense mistranslations. In each case, 4×10^7 random

codes were generated before calculating Z-value for both $PAM_{74–100}$ and mutation matrix. As depicted in Fig. 3, Z is maximized when K equals 4.5 in the case of $PAM_{74–100}$, and 3.0 when applying mutation matrix, suggesting that the average cost of a nonsense mistranslation, when using these two matrices, roughly equals −4.5 and −3.0, respectively. Comparing these values with other cost measures indicated in $PAM_{74–100}$ and mutation matrices, reveals the fact that very few of the allowed mistranslations (single base) are as costly as the reported values (−4.5 and −3.0), which is in accordance with our classic intuition of the high deleterious effect of a nonsense mutation. For the sake of comparison, the plots of Z-value versus K while applying nonsense mistranslations to the function φ^{faa} have also been included in Fig. 3.

In our third set of experiments, we again plotted Z-value, but this time against different values of Inc, treating K as a constant. Both matrices were applied, and as shown in Fig. 4, the results were controversial in each case. The canonical genetic code shows no optimization while applying $PAM_{74–100}$; whereas, mutation matrix proves to be more promising in the case of biochemically related amino acid pairs. For the sake of comparison, we plotted φ^{HH} ignoring nonsense mistranslations against increasing values of Inc, providing extra weights to the biochemically related amino acids. In both cases, Z-values were decreased at approximately the same rate, subject to increasing values of Inc.

Ultimately 1.2×10^9 random genetic codes were generated and φ^{HH} s were calculated, applying $K = 4.5$ and 3.0, for both $PAM_{74\text{-}100}$ and mutation matrices, respectively, as cost functions (results shown in Fig. 5).

Table 3
Comparison of Z-values obtained, while using different φ functions and cost functions

Applied fitness function	Z-value, using $PAM_{74–100}$	Using mutation matrix
φ^{faa} by Gilis et al. (2001)	6.3018	5.2359
φ^{HH}	6.4373	5.3148
φ^{HH} after applying $K = 4.5$ and 3.0	6.7139	5.4004

Results include both $PAM_{74–100}$ and mutation matrix.

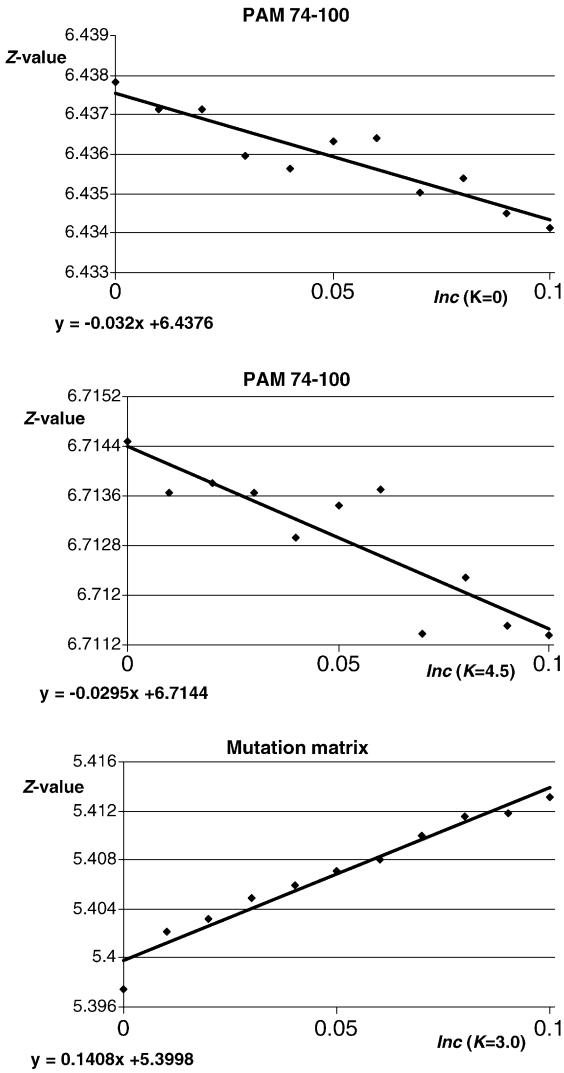

Fig. 4. Z-values calculated from increasing values of Inc (a parameter which controls the value by which the cost measure between two biochemically related amino acids is altered), while considering different values for the occurrences of nonsense mistranslations ($K = 0$; $K = 4.5$; $K = 3.0$). The results showed lower optimality of the canonical genetic code for $PAM_{74–100}$; whereas, mutation matrix made improvements in the Z-value with increasing Inc. Linear regressions have been included along with their equations.

For the sake of comparison, we calculated Z-values using different φ functions, generating 1.98×10^9 random codes in each case. The results have been tabulated in Table 3, where Z-values obtained from mutation matrix and $PAM_{74–100}$ are both included.

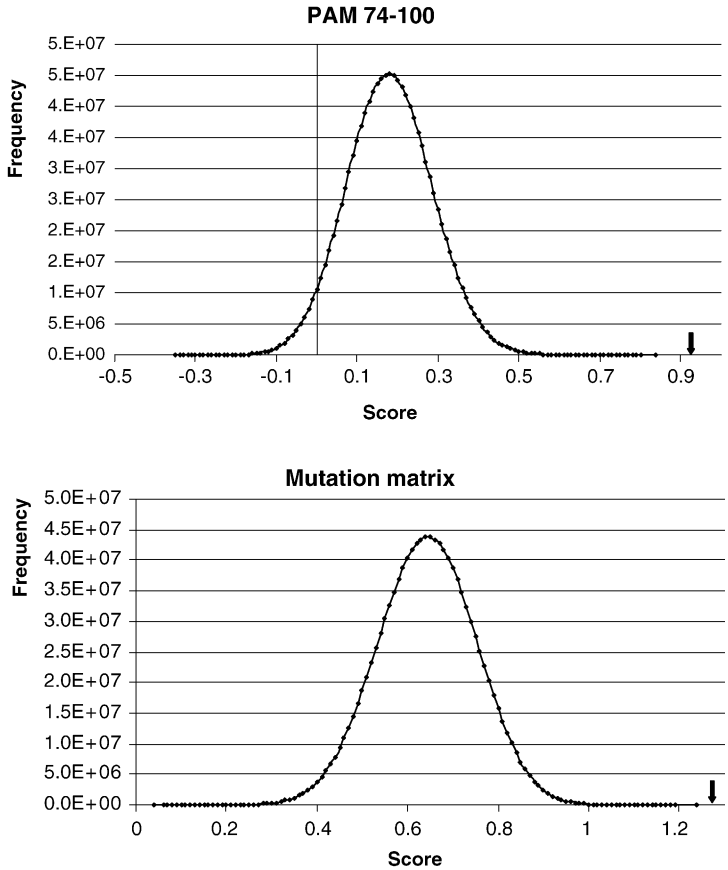

Fig. 5. The score distribution of 1.2×10^9 random codes when applying nonsense mistranslations ($K = 4.5$ for PAM_{74-100} and $K = 3.0$ for mutation matrix). In each case, the score of the canonical genetic code is indicated by an arrow.

4. Discussion and conclusion

The hypothesis that natural selection has shaped the genetic code to minimize the effects of mutation and mistranslation has been established and supported by researches and experiments prior to ours. But this becomes even more apparent when we consider other constraints whose influence might have played some role in the genetic code's evolution to its final state.

First, we modified φ^{faa} (introduced by Gilis et al., 2001) in order to add the effects of nonsense mistranslations to the other types of errors previously included. We started with the assumption that observed amino acids in proteomes are selected and optimal, and that genetic codes were selected in part to produce those frequencies (Gilis et al., 2001). The term $p(a(c))/n(a(c))$ in φ^{faa} was included to reflect the average frequency of codons in the genome of living organisms, based upon the assumption that the $n(a)$ codons of amino acid a have an average frequency of $p(a)/n(a)$. For example, suppose that in a randomly generated code, Leucine is assigned the single codon AUG. In this case, an average of 10.15% of all codons in the genome of living organisms should be AUG, in order to satisfy Leucine's empirical frequency of occurrence in protein sequences. Such a deviation from the expected value of 1/61 (1.64%) requires unusual evolutionary pressure over the genetic code and seems only probable under highly restricted conditions. This favors a high correlation between the number of codons assigned to an amino acid and its frequency, as previously debated by King and Jukes

(1969). When the theoretical expected frequency of amino acid a is equal to its empirical expected frequency, then the relative degeneracy of the amino acid in the code is optimal. The universal genetic code exhibits such a correlation, as expected (Fig. 1). In order to consider this issue in our fitness function, we introduced $f(c)$, which reflects the probability of obtaining $n(a)$ codons for the amino acid a by conjecturing a z-score function named y. y is defined to reflect the probability that a genetic code retains the frequency $p(a)$ for the amino acid a, if its relative codon degeneracy equals $n(a(c))/(n(\text{total}) - n(\text{sc}))$. Thus, according to the central limit theorem, y roughly exhibits a normal distribution with a mean value of 0, and a variance of 1. Yet, since the correlation between code and amino acid is not a precise one, y is divided by a constant (S), in order to increase the standard deviation of amino acid frequency. Note that a value of 100 is used for S to be as conservative as possible. Literally, $f(c)$ is a derivative of normal probability function with the assumption that y has a z-score distribution, indicating the probability that a random genetic code, by chance, might be able to show the observed amino acid frequencies if it possesses a known amino acid degeneracy. In short, this function results in a lower score for codes in which some codon frequencies are too high or too low. As previously reported, the inclusion of $f(c)$ in previously defined φ^{faa} resulted in an improvement of the computed genetic code optimization, far more than what had been previously attained. We should hereby mention that amino acid target frequencies have been assumed to be homogenous throughout code evolution and are equal to present-day empirical frequencies, however, there is evidence that amino acid frequencies have not been homogenous over time and if confirmed, this should be also taken into account in future studies (Brooks et al., 2002).

When considering the nonsense mistranslations, we hypothesized that positioning the codons with lower average frequencies, in a shorter codon distance from the stop codons should increase the fitness score, as there are fewer opportunities for such mutations to occur. This pattern can be observed in the canonical genetic code, as many of the amino acids near the stop codons are both rare and low in degeneracy (like Cys, Trp, Tyr). $p(a)/n(a)$ is introduced to be an indication of codon frequency, although rather roughly. This is equivalent to stating that in an optimized genetic code, we need a relatively low $p(a)/n(a)$ for the amino acids located in the vicinity of stop codons (in terms of mutation). The term $p(a(c))/n(a(c))$ had already been considered in our φ^{HH}, so we simply assigned a value of -1 as a cost measure of mutations to stop codons ($g(x, \text{stop codon}) = -1$), and the corresponding results showed an increase of about 1.7 percent in the Z-value, which was quite impressive. Despite the improvements achieved, it was not satisfactory due to the fact that, when assigning this cost measure, mistranslations leading to stop codons were occasionally more favorable than many other amino acids, which does not seem rational. Thus we tried to determine an appropriate value as the cost measure of such mistranslations (K), assuming that the code, and the value of K, is co-optimized ($g(x, \text{stop codon}) = -K$). Hence, we altered the cost measure of nonsense mistranslations from -1 to $-K$ in our scoring matrices (both PAM_{74-100} and mutation), and determined the most optimized K as previously mentioned in Section 3.

As a test case for determining the accuracy of our fitness function (φ^{HH}), we compared the scores of the canonical genetic code and the vertebrate mitochondrial genetic code, applying different φ functions. The vertebrate mitochondrial genetic code has been shaped under specific circumstances with many constraints. In an environment forcing the same limitations, the mitochondrial genetic code should score higher than the canonical one, but in normal organisms the canonical genetic code should illustrate a higher fitness. In the case of φ^{faa} and φ^{HH}, when using PAM_{74-100}, the mitochondrial genetic code scores higher than the canonical one, but when applying $K = 4.5$, the universal genetic code shows a significantly better fitness score (Table 4).

Some years ago, Ronneberg et al. (2000), debated the significance of code-metabolism coevolution theory, based on the biochemical pathways and determined a value of 62% as the probability of the genetic code having shaped itself by chance. In accordance with this statement, we hereby suggest that any assertion on the basis of the biosynthetic relations should be made using a special cost function, as different cost measures reveal controversial behaviors. In case of PAM_{74-100}, no better Z-values were obtained when overweighting the cost measures between biochemically related amino acids. On the other hand, applying mutation matrix exhibits the exact opposite. As

Table 4
The value of $(\varphi_{\text{mtlcode}} - \varphi_{\text{realcode}})/\varphi_{\text{mtlcode}}$ obtained using different fitness functions

Applied fitness function	$PAM_{74\text{-}100}$ (%)	Mutation matrix (%)
φ^{faa} by Gilis et al. (2001)	1.77	−0.50
φ^{HH}	1.70	−0.45
φ^{HH} after applying $K = 4.5$ or 3.0	−3.57	−4.68

In this table we have calculated the relative differences between the scores of the canonical genetic code and the vertebrate mitochondrial one. Note that, in case of PAM_{74-100}, only in the last function does the canonical genetic code score higher than the vertebrate mitochondrial one. Besides, in case of Mutation matrix, the canonical genetic code performs more efficiently, especially in the presence of nonsense mistranslations.

Table 5
Precursor-product pairs, indicating those which might have caused the difference in the behavior between the PAM_{74-100} and mutation matrices when overweighting the cost measure of the biochemically related amino acid pairs introduced by Ronneberg et al. (2000)

Amino acid pairs	
Ser	Trp
Phe	Tyr
Asp	Met

to which one is more accurate, further studies should be employed. Here we have included the main pairs of amino acids which might have caused such differences, as their values show great differences in the two studied matrices (Table 5).

Since the initial shape of the genetic code has had a great deal of influence on its final state, we should take into account all possible factors that might have affected the evolution of the genetic code, however, to their appropriate extent. In this paper, we have increased the known significance of the canonical genetic code by considering the two new factors discussed above (nonsense mistranslations and relative frequencies of amino acids). However, there are still many factors to be discussed, those with which we can more accurately model the constraints and conditions being present at the dawn of life on this planet.

Acknowledgements

We are grateful to Stephen J. Freeland and Dimitri Gilis for their useful comments.

References

Alff-Steinberger, 1969. The genetic code and error transmission. Proc. Natl. Acad. Sci. U.S.A. 64, 584–591.
Amirnovin, R., 1997. An analysis of the metabolic theory of the origin of the genetic code. J. Mol. Evol. 44, 473–476.
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.
Brooks, D.J., Fresco, J.R., Lesk, A.M., Singh, M., 2002. Evolution of amino acid frequencies in proteins over deep time: inferred order of introduction of amino acids into the genetic code. Mol. Biol. Evol. 19, 1645–1655.
Crick, F.H., 1968. The origin of genetic code. J. Mol. Biol. 38, 367–379.
Di Giulio, M., 2000. The origin of the genetic code. Trends Biochem. Sci. 25, 44–47.
Di Giulio, M., 2001a. The non-universality of the genetic code: the Universal Ancestor was a Progenote. J. Theor. Biol. 209, 345–349.
Di Giulio, M., 2001b. 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.
Di Giulio, M., 2003. The early phases of genetic code origin: conjectures on the evolution of coded catalysis. Origins Life Evol. Biosphere 33 (4), 479–489.
Dillon, L.S., 1973. The origins of the genetic code. Bat. Rev. 39, 301–345.
Freeland, S.J., 2002. The genetic code: an adaptation for adapting? J. Genet. Prog. Evolvable Mach. 3 (2), 113–127.
Freeland, S.J., Hurst, L.D., 1998. Load minimization of the genetic code: history does not explain the pattern. Proc. Ray. Sac. Land. B 266, 2111–2119.
Freeland, S.J., Hurst, L.D., 1998. 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.
Goldman, N., 1993. Further results on error minimization in the genetic code. J. Mol. Evol. 37, 662–664.
Haig, D., Hurst, L.D., 1991. A quantitative measure of error minimization on the genetic code. J. Mol. Evol. 33, 412–417.
Houen, G., 1999. Evolution of the genetic code: the nonsense, antisense, and antinonsense codes make no sense. Biosystems 54, 39–46.
Judson, O.P., Haydon, D., 1999. The genetic code: what is it good for? J. Mol. Evol. 49, 539–550.
Jukes, T.H., 1997. Neutral changes and modifications of the genetic code. Theor. Popul. Biol. 49, 143–145.
King, J.L., Jukes, T.H., 1969. Non-Darwinian evolution. Science 164, 788–798.

Knight, R.D., Freeland, S.J., Landweber, L.F., 1999. Selection, history, and chemistry: the three faces of the genetic code. Trends Biochem. Sci. 24, 241–247.

Osawa, S., Jukes, T.H., Watanabe, K., Muto, A., 1992. Recent evidence for evolution of genetic code. Microbiol. Rev. 56, 229–264.

Osawa, S., 1995. The evolution of the genetic code. Oxford University Press, Oxford, UK.

Pele, S.R., 1965. Correlation between coding-triplets and amino acids. Nature 207, 597–599.

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. U.S.A. 97, 13690–13695.

Sella, G., Ardell, D.H., 2002. The impact of message mutation on the fitness of a genetic code. J. Mol. Evol. 54 (5), 638–651.

Sonneborn, T.M., 1965. Degeneracy of the Genetic Code: Extent, Nature, and Genetic Implications, Evolving Genes and Proteins, Academic Press, New York, pp. 97–166.

Szathmary, E., 1991. Codon swapping as a possible evolutionary mechanism. J. Mol. Evol. 32, 178–182.

Szathmáry, E., Zintzaras, E., 1992. A statistical test of hypotheses on the organization and origin of the genetic code. J. Mol. Evol. 35, 185–189.

Wintjens, R.T., Rooman, M.J., Wodak, S.J., 1996. Automatic classification of αα-turn motifs in proteins. J. Mol. Biol. 255, 235–253.

Woese, C.R., 1965. On the evolution of the genetic code. Proc. Natl. Acad. Sci. U.S.A. 54, 1546–1552.

Woese, C.R., 1965. Order in the genetic code. Proc. Natl. Acad. Sci. U.S.A. 54, 71–75.

Woese, C.R., Dagre, D.H., Dagre, S.A., Kondo, M., Saxinger, W.C., 1966. On the fundamental nature and evolution of genetic code. Coldspring Harbw Symp. Quant. Biol. 31, 723–736.

Woese, C.R., Dugne, D.H., Saxinger, W.C., Dayre, S.A., 1966. The molecular basis for the genetic code. Proc. Natl. Acad. Sci. U.S.A. 55, 966–974.

Wong, J.T., 1975. A co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. U.S.A. 72, 1909–1912.

Wong, J.T., 1976. The evolution of a universal genetic code. Proc. Natl. Acad. Sci. U.S.A. 73, 2336–2340.

Wong, J.T., Bronskill, P.M., 1979. Inadequacy of pre-biotic synthesis as the origin of proteinaceous amino acids. J. Mol. E 13, 115–125.

Wong, J.T., 1981. Coevolution of genetic code and amino acid biosynthesis. Trends Biochem. Sci. 6, 33–36.