BMC Evolutionary Biology

RESEARCH ARTICLE Open Access

Comparison of translation loads for standard and alternative genetic codes

Stefanie Gabriele Sammet[1,3], Ugo Bastolla*[2] and Markus Porto*[1,4]

Abstract

Background: The (almost) universality of the genetic code is one of the most intriguing properties of cellular life. Nevertheless, several variants of the standard genetic code have been observed, which differ in one or several of 64 codon assignments and occur mainly in mitochondrial genomes and in nuclear genomes of some bacterial and eukaryotic parasites. These variants are usually considered to be the result of non-adaptive evolution. It has been shown that the standard genetic code is preferential to randomly assembled codes for its ability to reduce the effects of errors in protein translation.

Results: Using a genotype-to-phenotype mapping based on a quantitative model of protein folding, we compare the standard genetic code to seven of its naturally occurring variants with respect to the fitness loss associated to mistranslation and mutation. These fitness losses are computed through computer simulations of protein evolution with mutations that are either neutral or lethal, and different mutation biases, which influence the balance between unfolding and misfolding stability. We show that the alternative codes may produce significantly different mutation and translation loads, particularly for genomes evolving with a rather large mutation bias. Most of the alternative genetic codes are found to be disadvantageous to the standard code, in agreement with the view that the change of genetic code is a mutationally driven event. Nevertheless, one of the studied alternative genetic codes is predicted to be preferable to the standard code for a broad range of mutation biases.

Conclusions: Our results show that, with one exception, the standard genetic code is generally better able to reduce the translation load than the naturally occurring variants studied here. Besides this exception, some of the other alternative genetic codes are predicted to be better adapted for extreme mutation biases. Hence, the fixation of alternative genetic codes might be a neutral or nearly-neutral event in the majority of the cases, but adaptation cannot be excluded for some of the studied cases.

Background

The origin and universality of the genetic code is one of the biggest enigmas in biology [1]. Soon after the genetic code of *Escherichia coli* was deciphered [2], it was realized that this specific code out of more than 10^{84} possible codes is shared by all studied life forms (albeit sometimes with minor modifications). The question of how this specific code appeared and which physical or chemical constraints and evolutionary forces have shaped its highly non-random codon assignment is subject of an intense debate. In particular, the feature that codons differing by a single nucleotide usually code for either the same or a chemically very similar amino acid and the associated block structure of the assignments is thought to be a necessary condition for the robustness of the genetic code both against mutations as well as against errors in translation [3-13]. This robustness reduces fitness losses due to mutation and mistranslation, which is believed to be a major force in coding sequence evolution [14]. There are three basic theories of the genetic code's nature, origin, and evolution. Whereas the stereochemical theory first proposed by Gamow [15] asserts that the codon assignment was originated by the physicochemical affinity between the amino acid and the codon or anticodon, the adaptive theory posits that the genetic code was shaped under selection for robustness, either against mutations [16,17] or against translation errors [18,19], or against

* Correspondence: ubastolla@cbm.uam.es, porto@thp.uni-koeln.de
[1] Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 8, 64289 Darmstadt, Germany
[2] Centro de Biología Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain
Full list of author information is available at the end of the article

both [18,20,21]; finally, the coevolution theory postulates that the structure of the genetic code reflects the pathways of amino acid biosynthesis [22].

In this work, we address an issue that has received somewhat less attention in this broader context, namely the existing variants of the standard genetic code. These variants are used for instance in the mitochondria of many species, and they consist in the modification of one or several codons of the standard genetic code. Two main mechanisms have been proposed to explain how they may have evolved despite the large fitness cost that is expected to be associated with the modification of a codon [9,23,24]: through an ambiguous intermediate state and through the reassignment of a temporarily unused codon. These mechanisms are not mutually exclusive. The first one assumes that a codon is temporarily recognized both by the current as well as by a mutated tRNA, so that it can result in two different amino acids. Such ambiguity might be preferential in some circumstances and remain present for some time. A codon reassignment occurs if the mutated tRNA finally takes over. The second scenario takes place if one codon disappears from a given genome. This is particularly likely in small genomes with large guanine/cytosine (GC) or adenine/thymine (AT) content, as it is the case for many mitochondrial genomes and nuclear genomes of parasitic microbes. In this case the translation system may change without any cost, and the codon may be 'conquered' by another amino acid. Both scenarios consider that alternative genetic codes are the result of non-adaptive or neutral evolution, even though 'genomic streamlining' (i.e., selective pressure to minimize the genome by eliminating a tRNA) has been proposed as a possible advantage of code changes [25,26]. However, it has not been addressed whether these variants differ from the standard genetic code as far as mutation and translation loads are concerned.

Here, we use computer simulations of neutral protein evolution constrained to maintain the folding and misfolding stability of the native state in order to study the differences between the standard genetic code and seven naturally occurring variants concerning their effects on protein stability. These effects are predicted using a simplified model of protein folding [27], the same that we consistently use in the evolutionary simulations [28-32]. Despite its simplicity, this model is able to predict with similar accuracy as other more complicated models the effects of mutations on folding stability. Due to our simplifying assumption to consider a neutral model, the different genetic codes hardly have any influence on the average unfolding and misfolding stabilities. However, alternative codes yield significantly different mutation and translation loads, in particular for genomes evolving under strong AT or GC mutation bias.

Results and Discussion

In this work, we study how the genetic code influences the fitness consequences of errors (loads) during mutation and translation. This influence may arise because of two mechanisms: (1) Directly, through the change in the rate of occurrence of different amino acid misincorporations in the translated protein; (2) Indirectly, through the evolutionary influence that the genetic code may have on protein stability. We simulated our previously proposed model of protein evolution in order to study this indirect influence as well.

Model

Our model of protein evolution has been presented in previous works [28-32], and it is similar to models used by others [33-40]. It has been successfully used to explain non-Poissonian rates in neutral evolution [28] and the observed site-specific amino acid distributions [31,32], to name two examples. It considers a genetically homogeneous population, i.e. the product of the population size N and the mutation rate μ is assumed to be small. The assumption of a small mutation rate μ is justified when considering an individual protein, but not an entire genome. If we considered a whole evolving genome instead of a single protein, the approximation of very small mutation rate would not be justified, since genomic mutation rates are in a range of 0.003 to 0.004 mutations per genome per generation for DNA-based microbes, including viruses, bacteria, and eukaryotes [41]. In this context, a new interesting effect has to be considered, namely the hitch-hiking effect, which consists in the fixation of mildly disfavorable alleles driven by a positively selected allele present in the same chromosome. However, considering the hitch-hiking effect would make the study much more complicated, and we leave it as a subsequent step. In our model, the fitness of an individual carrying a particular gene depends only on the folding properties of the translated protein, which are estimated through a simple protein folding model. A characteristic of our model that distinguishes it from similar ones is that we consider two types of stability, with respect to misfolding and with respect to unfolding. They are calculated by estimating the normalized energy gap $\alpha(\mathbf{A})$ and the folding free energy $F(\mathbf{A})$, respectively. Misfolding stability is measured through α and unfolding stability is measured through $-F$, which are computed for each protein sequence $\mathbf{A}$ encountered in the simulated evolution. The protein structure is assumed to have been already optimized by natural selection and is kept fixed throughout evolution, as represented by the experimental structure found in the Protein Data Bank (PDB) (see Methods). If the folding stability is too small, the protein will not be stable in its native state; if the misfolding stability is too small, misfolded structures can trap the folding process, and they can expose hydrophobic patches and promote aggregation. In the spirit of Kimura's neutral theory of molecular evolution [42,43], we assume that mutations are either neutral or strongly deleterious. More specifically, all proteins having both unfolding and misfolding stabilities above previously fixed thresholds are regarded as viable and they are assigned the same fitness $\mathcal{F} = 1$ (in arbitrary units) and all proteins for which at least one of the stabilities is below threshold are regarded as unviable and they are assigned fitness $\mathcal{F} = 0$. Mutations to stop codons are considered lethal and receive a fitness $\mathcal{F} = 0$. The two neutral thresholds α_{thr} and F_{thr} are chosen proportional to the values of α_{nat} and F_{nat} of the respective protein in the PDB, multiplied with coefficients slightly smaller than one so that the native protein is above threshold. We present results with both coefficients equal to 0.98, but our results are robust to changing this prefactor in a reasonable range. Note that fitness functions depending continuously on stabilities can be considered, but the resulting non-neutral evolutionary dynamics is significantly more complex due to population size effects [44]. Neutrality of the fitness landscape is assumed here for the sake of simplicity, since otherwise the model would depend on at least two additional parameters, i.e. the smoothness of the fitness landscape and the effective size of the population, making it very difficult to reach clear conclusions about the effect of the genetic code and the mutation bias.

Another important ingredient of our model is the mutation model at the DNA level. We parameterize the mutation model with a single parameter, the AT bias, which represents the equilibrium content of adenine and thymine after a very long evolution under mutation alone (the complementary variable GC bias, expressing the equilibrium content of guanine and cytosine under mutation alone, is sometimes alternatively used). The mutation bias strongly affects the substitution process (i.e., the accepted mutations), biasing the amino acid composition of the protein. Interestingly, the mutation bias also influences the folding properties of the evolving proteins [32,45]. In fact, AT rich codons code for amino acids which are more hydrophobic and the resulting proteins tend to be more stable against unfolding (more negative folding free energy F) but less stable against misfolding (since the set of all potential misfolded protein structures increases their stability faster than the native structure, resulting in a smaller normalized energy gap α), whereas the contrary holds in case of GC bias. This bias at the mutation level produces a bias at the substitution level, both for neutral fitness landscapes [32] and for smooth fitness landscapes [44], such that proteins evolving under higher AT bias will be comparatively more stable against unfolding but less stable against misfolding. Finally, in order to fully specify the mutation model, we have to fix the transition-to-transversion ratio k. Since transitions (such as C to T) tend to conserve the physiochemical properties of the coded amino acid more than transversions (such as C to A or to G), a high transition-to-transversion ratio k usually reduces the mutation load. We used two values of k, $k = 2$, which is suitable for most nuclear sequences [46] and a kind of standard value in molecular evolution simulations, and $k = 20$, a maximal value that has been observed in some mitochondrial genomes [47]. To study the influence of the mutation process, we simulate the evolution of DNA sequences under nine different mutation biases with both transition-to-transversion ratios $k = 2$ and $k = 20$, Our model of protein evolution cannot be treated analytically, so that we have to study it using numerical simulations (see Fig. 1). Point mutations change the DNA sequence (see Methods) and they are accepted if the resulting amino acid sequence, translated from the DNA sequence using the genetic code under consideration, is viable, i.e. both stabilities are above threshold. We simulated the evolution of three different proteins of similar lengths and different secondary structure compositions, (i) the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A, 135 amino acids), (ii) the acyl carrier protein (PDB id. 1hy8, chain A, 76 amino acids), and (iii) the cold-shock protein (PDB id. 1c9o, chain A, 66 amino acids), see Methods for details. We start each simulation with the native amino acid sequence as obtained from the PDB of the chosen structure, from which we construct a corresponding 'native' DNA sequence by randomly choosing codons using the genetic code under consideration with weights determined by the given AT content (inverse translation). The influence of this starting sequence is lost after a relatively short evolutionary trajectory needed for equilibration, after which a stationary situation is reached, in which both stabilities fluctuate around constant values. Statistics is taken only in this stationary state. We compare the standard genetic code with seven naturally occurring variants, see Fig. 2. Out of these seven variants, five are used in mitochondria of many species, and two variants are used by certain species in their nuclear protein production. These variants differ in between one to six codon assignments, and display between one to four stop codons instead of the three stop codons of the standard code, cf. Fig. 2. We follow the naming scheme of the NCBI database http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi, including the `transl_table` numbering). To ease the assessment of our results, we use a consistent color scheme in this work, in which the standard genetic code is shown in black, whereas the naturally occurring variants are color coded according to the following scheme:

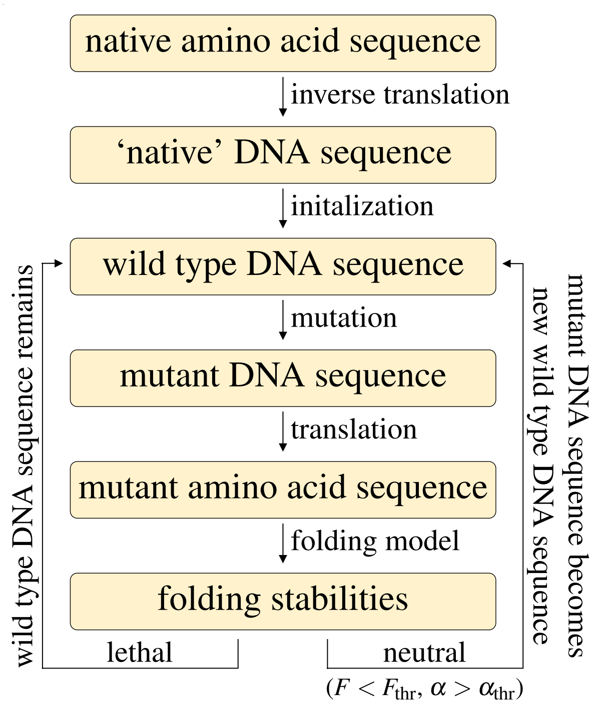

native amino acid sequence
inverse translation
'native' DNA sequence
initalization
wild type DNA sequence
mutation
mutant DNA sequence
translation
mutant amino acid sequence
folding model
folding stabilities
lethal
neutral
($F < F_{\mathrm{thr}}$, $\alpha > \alpha_{\mathrm{thr}}$)
wild type DNA sequence remains
mutant DNA sequence becomes new wild type DNA sequence

Figure 1 Model. Sketch of the model: We start each simulation with the native amino acid sequence as obtained from the Protein Databank (PDB) of the chosen structure, from which we construct a corresponding 'native' DNA sequence by randomly choosing codons using the genetic code under consideration with weights determined by the given AT content (inverse translation). This 'native' DNA sequence is hence as close as possible to the equilibrium with the chosen AT content and becomes the first wild type DNA sequence. Then, at every step, the current wild type DNA sequence is mutated to generate a mutated DNA sequence, which is translated to a mutated amino acid sequence using the genetic code under considerations. The resulting mutated amino acid sequence is evaluated using the folding model with respect to its folding stabilities, based on which the mutated amino acid sequence is either considered as neutral (both unfolding and misfolding stability are above threshold) and the mutated DNA sequence becomes the new wild type DNA sequence, or as lethal (one or both stabilities are below threshold) and the mutated DNA sequence is discarded (the wild type DNA sequence remains as is).

1.Blue: 'The Echinoderm and Flatworm Mitochondrial Code' (NCBI `transl_table = 9`) [48]; Taxonomic range: Asterozoa (starfishes), Echinozoa (sea urchins), Rhabditophora among the Platyhelminthes.

2 Orange: 'The Invertebrate Mitochondrial Code' (NCBI `transl_table = 5`) [48]; Taxonomic range: Nematoda: *Ascaris, Caenorhabditis*; Mollusca: Bivalvia; Polyplacophora. Arthropoda/Crustacea: *Artemia*; Arthropoda/Insecta: *Drosophila; Locusta migratoria* (migratory locust), *Apis mellifera* (honeybee).

3 Magenta: 'The Ascidian Mitochondrial Code' (NCBI `transl_table = 13`) [48]. Taxonomic range: Urochordata: Tunicates.

4.Green: 'The Vertebrate Mitochondrial Code' (NCBI `transl_table = 2`) [49]; Taxonomic range: Vertebrata.

5.Cyan: 'The Yeast Mitochondrial Code' (NCBI `transl_table = 3`) [50]; Taxonomic range: *Saccharomyces cerevisiae, Candida glabrata, Hansenula saturnus, and Kluyveromyces thermotolerans.*

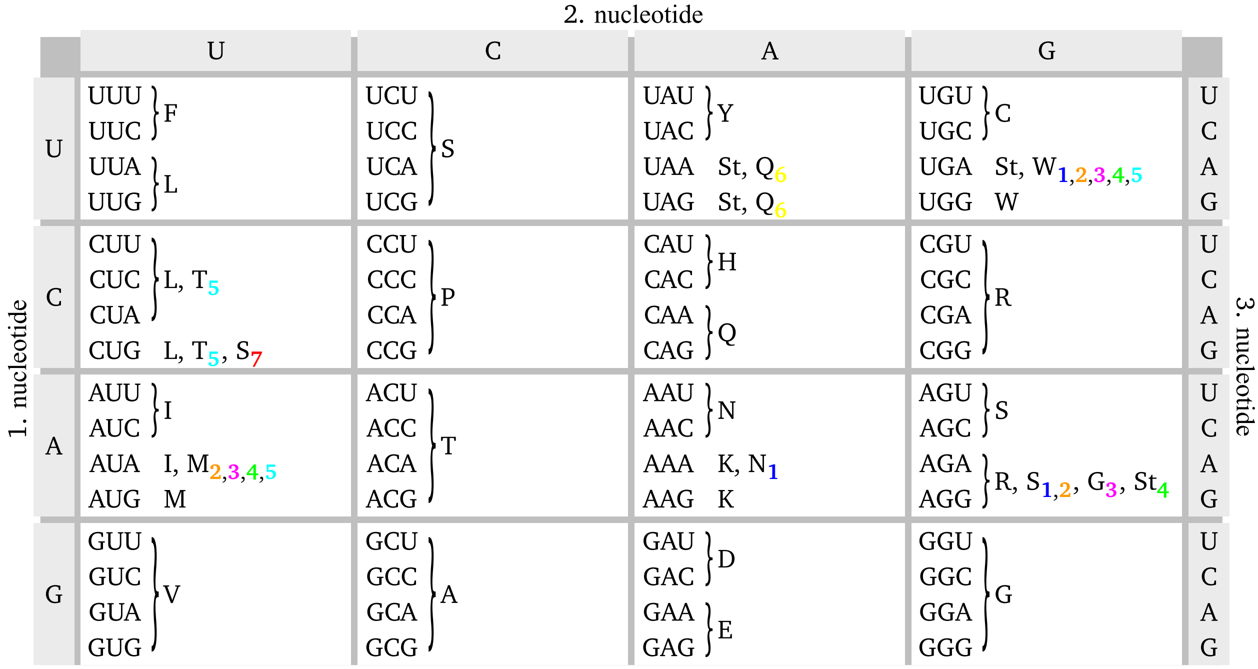
6.Yellow: 'The Ciliate, Dasycladacean and Hexamita Nuclear Code' (NCBI `transl_table = 6`) [51]. Taxonomic range: Ciliata: Oxytricha and Stylonychia, Paramecium, Tetrahymena, Oxytrichidae and probably Glaucoma chattoni. Dasycladaceae: Acetabularia and Batophora. Diplomonadida: *Hexamita inflata, Diplomonadida* ATCC50330 and ATCC50380.

7.Red: 'The Alternative Yeast Nuclear Code' (NCBI `transl_table = 12`) [52]. Taxonomic range: Endomycetales (yeasts): *Candida albicans, Candida cylindracea, Candida melibiosica, Candida parapsilosis, and Candida rugosa* (However, other yeasts, including *Saccharomyces cerevisiae, Candida azyma, Candida diversa, Candida magnoliae, Candida rugopelliculosa, Yarrowia lipolytica*, and *Zygoascus hellenicus*, definitely use the standard (nuclear) code).

Unfolding and misfolding stabilities

We first study the direct effect of the eight different genetic codes on the average unfolding and misfolding stabilities (see Methods). Since we chose a neutral fitness landscape where mutations are either neutral or lethal, we expect that, independent of the mutation rate and the genetic code, the two folding stabilities will be close to the neutral thresholds, i.e. the minimum allowed stability values, which correspond to the maximum number of sequences, while larger stabilities correspond to many fewer sequences. However, large AT content (more than 50% AT) favors unfolding stability at the expense of misfolding stability, whereas small AT content (less than 50% AT) favors misfolding stability at the expense of unfolding stability. Consequently, for large AT content selection mainly acts on misfolding stability, which is expected to be closer to the neutral threshold, whereas unfolding stability is easily obtained and it is above the threshold. Conversely, for small AT content selection mainly acts on unfolding stability, which is expected to be close to its neutral threshold. In general, the smaller of the two stabilities is very close to the neutral threshold and almost independent of the genetic code, whereas the stability favored by the mutation process is above the neutral threshold, although this does not imply any gain in fitness, and it may depend on the genetic code and the mutation bias.

The behavior of the average unfolding stability *-F* for different genetic codes (as summarized in Fig. 2) is exemplified using the epsilon subunit of F1F0-ATP synthase



	2. nucleotide				
1. nucleotide	U	C	A	G	3. nucleotide
U	UUU F	UCU S	UAU Y	UGU C	U
	UUC	UCC	UAC	UGC	C
	UUA L	UCA	UAA St, Q_6	UGA St, $W_{1,2,3,4,5}$	A
	UUG	UCG	UAG St, Q_6	UGG W	G
C	CUU L, T_5	CCU P	CAU H	CGU R	U
	CUC	CCC	CAC	CGC	C
	CUA	CCA	CAA Q	CGA	A
	CUG L, T_5, S_7	CCG	CAG	CGG	G
A	AUU I	ACU T	AAU N	AGU S	U
	AUC	ACC	AAC	AGC	C
	AUA I, $M_{2,3,4,5}$	ACA	AAA K, N_1	AGA R, $S_{1,2}$, G_3, St_4	A
	AUG M	ACG	AAG K	AGG	G
G	GUU V	GCU A	GAU D	GGU G	U
	GUC	GCC	GAC	GGC	C
	GUA	GCA	GAA E	GGA	A
	GUG	GCG	GAG	GGG	G

Figure 2 Standard genetic code and naturally occurring variants. The standard genetic code and the naturally occurring variants studied in this work, written using the RNA alphabet and standard abbreviations for the amino acids ('St' indicates stop codon). Concerning the alternative genetic codes, we follow the naming scheme of the NCBI database (http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi, including the `transl_table` numbering). We give in a following only a very brief description of a given alternative genetic code's systematic range, a more detailed description with references can be found on the NCBI's web page. The standard genetic code is shown in black (3 stop codons), whereas the seven naturally occurring variants studied are shown using the following color scheme: (1, blue): 'The Echinoderm and Flatworm Mitochondrial Code' (NCBI `transl_table=9`), mitochondrial code of Asterozoa, Echinozoa, and Rhabditophora (4 differences, 2 stop codons) [48]; (2, orange): 'The Invertebrate Mitochondrial Code' (NCBI `transl_table=5`), mitochondrial code of Nematoda, Mollusca, Crustacea, and Insecta (4 differences, 2 stop codons) [48]; (3, magenta): 'The Ascidian Mitochondrial Code' (NCBI `transl_table=13),` mitochondrial code of Urochordata (4 differences, 2 stop codons) [48]; (4, green): 'The Vertebrate Mitochondrial Code' (NCBI `transl_table=2`), mitochondrial code of Vertebrata (4 differences, 4 stop codons) [49]; (5, cyan): 'The Yeast Mitochondrial Code' (NCBI `transl_table=3),` mitochondrial code of *Saccharomyces cerevisiae, Candida glabrata, Hansenula saturnus, and Kluyveromyces thermotolerans* (6 differences, 2 stop codons) [50]; (6, yellow): 'The Ciliate, Dasycladacean and Hexamita Nuclear Code' (NCBI `transl_table=6`), nuclear code of Ciliata, Dasycladaceae, Diplomonadida (2 differences, 1 stop codon) [51]; (7, red): 'The Alternative Yeast Nuclear Code' (NCBI `transl_table=12`), nuclear code of *Candida albicans* (1 difference, 3 stop codons) [52]. The color scheme is as in Figs. 3 to 6.

(PDB id. 1aqt, chain A) and the transition-to-transversion ratio $k = 2$, see Fig. 3. The transition-to-transversion ratio k does not affect the results significantly, so that we omit the results for $k = 20$. As expected, there is hardly any influence of the genetic code on the average unfolding stability, except for extremely large AT biases (approx. 80% AT content or more) where unfolding stability is well above threshold and selection mainly acts on misfolding stability. In such cases, most alternative genetic codes display a smaller unfolding stability than the standard code (less negative F), but there are two exceptions, the 'Echinoderm and Flatworm Mitochondrial Code' (blue triangles) and the 'Alternative Yeast Nuclear Code' (red stars), which have slightly better unfolding stability than the standard code, even though the difference is minor and it does not imply any difference in fitness. The two other proteins we studied show a similar behavior (data not shown).

The behavior of the average misfolding stability α for the eight different genetic codes is likewise exemplified using the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A) and the transition-to-transversion ratio $k = 2$, see Fig. 4. Again, no significant difference can be noticed between transition-to-transversion ratio $k = 2$ and $k = 20$, so that we omit the latter. There is hardly any influence of the genetic code on the misfolding stability, significant differences are found only for very small AT content (approx. 20% AT content or less) at which misfolding stability is easily obtained. For such small AT content, most alternative genetic codes display an essentially identical unfolding stability, but there are two exceptions, the 'Yeast Mitochondrial Code' (cyan diamonds) and the 'Alternative Yeast Nuclear Code' (red stars), which yield larger misfolding stability than the standard code, even though the difference is minor. The two other proteins we studied show a similar behavior (data not shown).

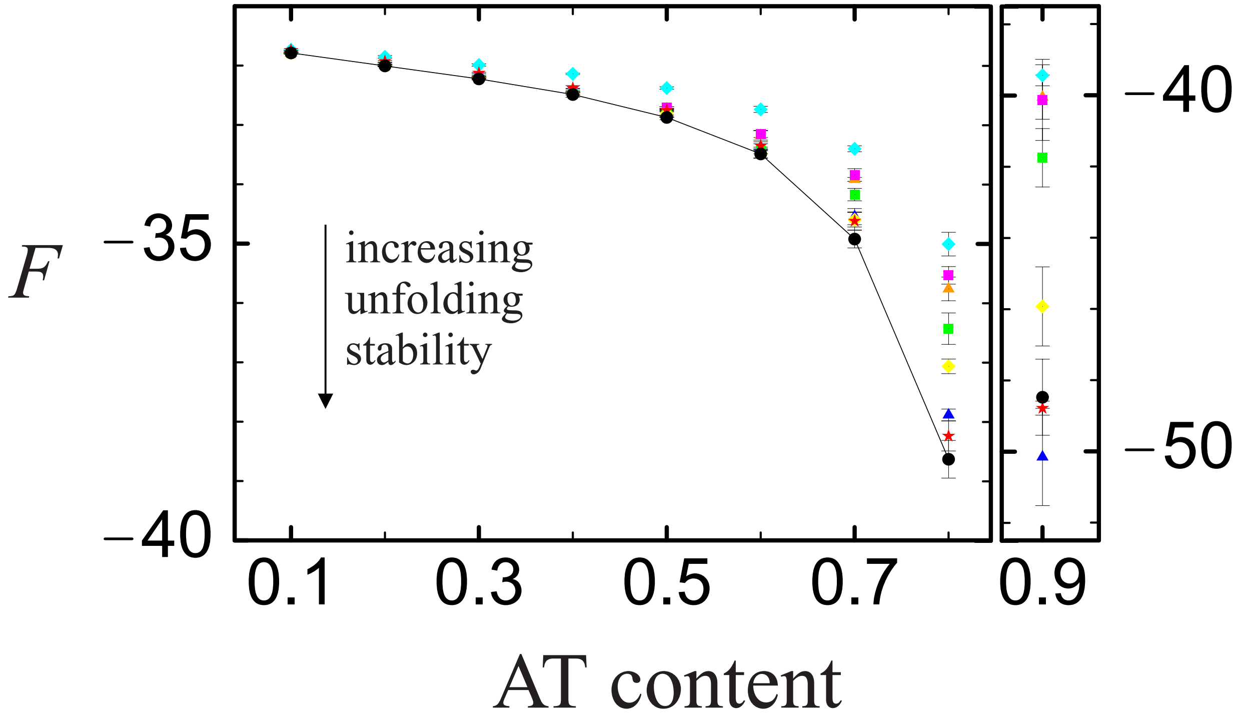


Figure 3 Unfolding stability. Average unfolding stability *F* vs AT content for the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A), exemplified for transition-to-transversion ratio *k* = 2 (the data for AT content 90% is shown using a different scale for better visibility). The standard genetic code is shown as black circle (which are connected by lines for better visibility), whereas the seven naturally occurring variants studied, as listed in Fig. 2, are shown using the following color scheme: (1, blue triangle): 'The Echinoderm and Flatworm Mitochondrial Code' (NCBI`transl_table = 9`) [48]; (2, orange triangle): 'The Invertebrate Mitochondrial Code' (NCBI `transl_table = 5`) [48]; (3, magenta square): 'The Ascidian Mitochondrial Code' (NCBIs `transl_table = 13`) [48]; (4, green square): 'The Vertebrate Mitochondrial Code' (NCBIs `transl_table = 2`) [49]; (5, cyan diamond): 'The Yeast Mitochondrial Code' (NCBIs `transl_table = 3`) [50]; (6, yellow diamond): 'The Ciliate, Dasycladacean and Hexamita Nuclear Code' (NCBIs `transl_table = 6`) [51]; (7, red star): 'The Alternative Yeast Nuclear Code' (NCBIs `transl_table = 12`) [52]. The error bars indicate the mean's standard deviation.

We see from the two figures that the genetic code may influence the balance between unfolding and misfolding stability. For instance, the 'Yeast Mitochondrial Code' (cyan diamonds) yields systematically lower stability against unfolding (less negative *F*) when compared with other codes, and higher stability against misfolding at low AT, which is an effect similar to the one obtained by decreasing the AT content. Nevertheless, these differences are not relevant, since they are buffered at the level of fitness. In fact, we assume a neutral model in which the fitness is either $\mathcal{F} = 0$ or $\mathcal{F} = 1$ (in arbitrary units), so that differences in stability do not yield differences in fitness. Moreover, notice that the difference of, say, stability against unfolding between different codes is only significant at an AT content at which this stability is anyway high, i.e. it is far from the neutral threshold, so that the selective pressure mainly affects stability against misfolding. These differences might become relevant in a non-neutral fitness landscape where fitness depends smoothly on stability [44].

Mutation and translation load

Next, we study the effect of the different genetic codes on mutation and translation loads (see Methods). These represent the fitness loss due to mutations and translation errors. The two loads differ by the rate at which a given error occurs and by the treatment of stop codons: In the case of mutation load, the rate of a mutation is given by the mutation process we use, which includes both the mutation bias and the transition-to-transversion ratio; mutations to stop codons are equivalent to mutations to sense codons as far as the chemical modification of the DNA sequence is concerned and hence included into the definition of the corresponding load (cf. Eq. (3) in Methods), and the associated fitness is zero. In the case of translation load, all mistranslation to sense codons are assigned equal rate. Since a premature end of translation by misinterpreting a sense codon as a stop codon is caused by release factors and not by tRNAs (and hence by a different mechanism than misinterpreting a sense codon as a another sense codon) which furthermore

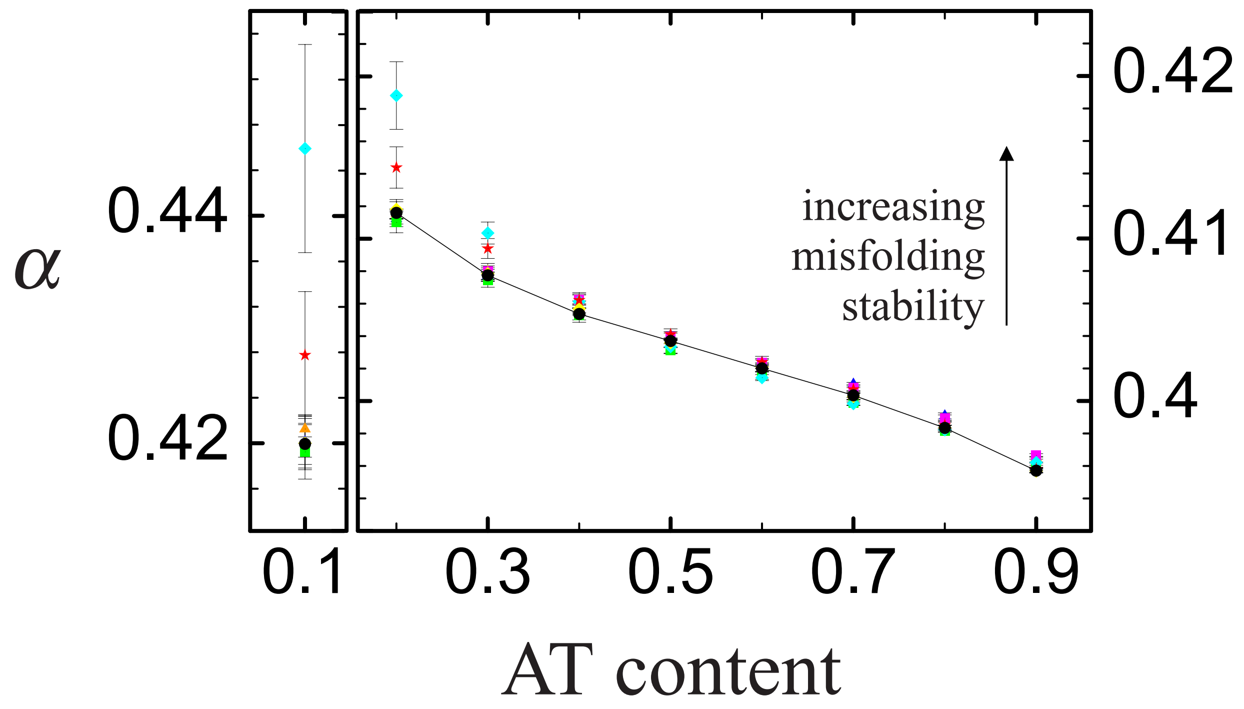


Figure 4 Misfolding stability. Average misfolding stability α vs AT content for the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A), exemplified for transition-to-transversion ratio $k = 2$ (the data for AT content 10% is shown using a different scale for better visibility). Symbols are as in Fig. 3, black circles indicate standard genetic code (which are connected by lines for better visibility), whereas the other different colors indicate the seven naturally occurring variants studied, as listed in Fig. 2. The error bars indicate the mean's standard deviation.

involves neighboring codons [53], for simplicity, we consider its error rate much smaller than the rate of missense errors in translation, and we neglect mistranslations to stop codons. In this way, the translation load is not explicitly influenced by the number of stop codons in the genetic code under consideration (cf. Eq. (4) in Methods).

The average mutation load L_{mut} for different genetic codes is exemplified using the three proteins described above, see Fig. 5. The left panels refer to transition-to-transversion ratio $k = 2$, while the right panels refer to $k = 20$, and the three rows refer to the three proteins. A large transition-to-transversion ratio usually yields smaller mutation loads, except for very small AT content (approx. 20% AT content or less) for which $k = 20$ increases the load considerably. In contrast to the unfolding and misfolding stabilities, different genetic codes show different mutation loads. This is in part due to the different form in which selection acts on stabilities and on loads in our model. Whereas unfolding and misfolding stabilities are strictly constrained in the fitness landscape of neutral or lethal mutations that we modelled, we assume that loads are not explicitly targeted by selection and are free to vary. Most of the alternative genetic codes display mutation loads larger than for the standard code. Nevertheless, some yield consistently smaller mutation loads in some range of mutation bias. For instance, the 'Yeast Mitochondrial Code' (cyan diamonds) yields a smaller mutation load for very small AT content (approx. 20% AT content), although it has a rather large load for large AT content which is characteristic of mitochondria genomes of Yeast (typically 75% to 85% AT content). Two alternative genetic codes, the 'Ciliate, Dasycladacean and Hexamita Nuclear Code (yellow diamonds) and the 'Echinoderm and Flatworm Mitochondrial Code' (blue triangles), display a smaller mutation load for large AT content.

The behavior of the average translation load L_{trans} for different genetic codes is likewise exemplified using the three proteins, see Fig. 6. Again, the left panels refer to transition-to-transversion ratio $k = 2$, while the right panels refer to $k = 20$, and the three rows refer to the three proteins. The different genetic codes show significantly different translation loads, and even different dependences on the mutation bias (decreasing or increasing translation load with increasing AT content). Notice that the dependence of the translation load on the bias is not due to how the error rates depend on the bias, as in the case of the mutation load, but it is due to how the mutation bias influences protein stabilities. Most alternative genetic codes yield a larger translation load than the stan-

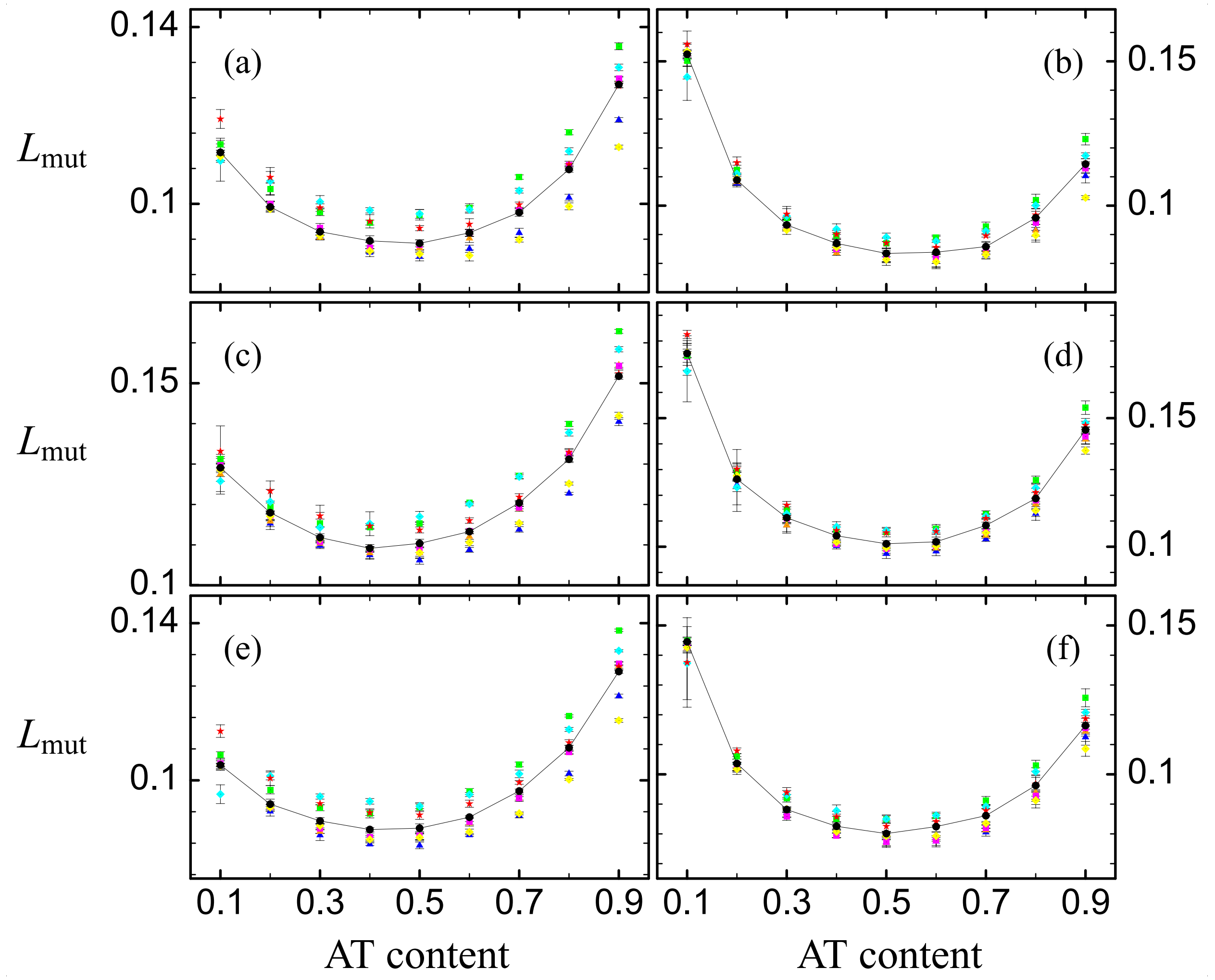


Figure 5 Mutation load. Average mutation load L_{mut} vs AT content for (a),(b) for the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A) (c),(d) for the acyl carrier protein (PDB id. 1hy8, chain A) and (e),(f) for the cold-shock protein (PDB id. 1c9o, chain A) as well as for (a),(c),(e) transition-to-transversion ratio $k = 2$ and (b),(d),(f) $k = 20$. Symbols are as in Fig. 3, black circles indicate standard genetic code (which are connected by lines for better visibility), whereas the other different colors indicate the seven naturally occurring variants studied, as listed in Fig. 2. The error bars indicate the mean's standard deviation.

dard code. Nevertheless, the 'Yeast Mitochondrial Code' (cyan diamonds) yields a smaller translation load for very small AT content (approx. 20% AT content or less), but this result does not hold for very large transition-to-transversion ratio and large AT content, both being characteristic of mitochondria genomes of Yeast. One alternative genetic code, the 'Echinoderm and Flatworm Mitochondrial Code' (blue triangles), results in a smaller translation load for most mutation biases, in particular for large ones, and independent of the transition-to-transversion ratio, and is hence preferential to the standard genetic code.

It is interesting to note that there seems not to be any trivial dependence of the average mutation load on the number of stop codons in the genetic code. Since the mutation load is calculated including mutations to stop codons, one would expect genetic codes containing more stop codons to have a larger mutation load than genetic codes containing fewer stop codons. This is, however, not the case, as one sees in Fig. 5 by comparing the 'Echinoderm and Flatworm Mitochondrial Code' (blue triangles) and the 'Yeast Mitochondrial Code' (cyan diamonds), which both have two stop codons, with the standard genetic code, which has three stop codons. Additionally, as our definition of translation load excludes mistranslations corresponding to stop codons (as misinterpreting a sense codon as a stop codon is caused by release factors and not by tRNAs and hence by a different mechanism

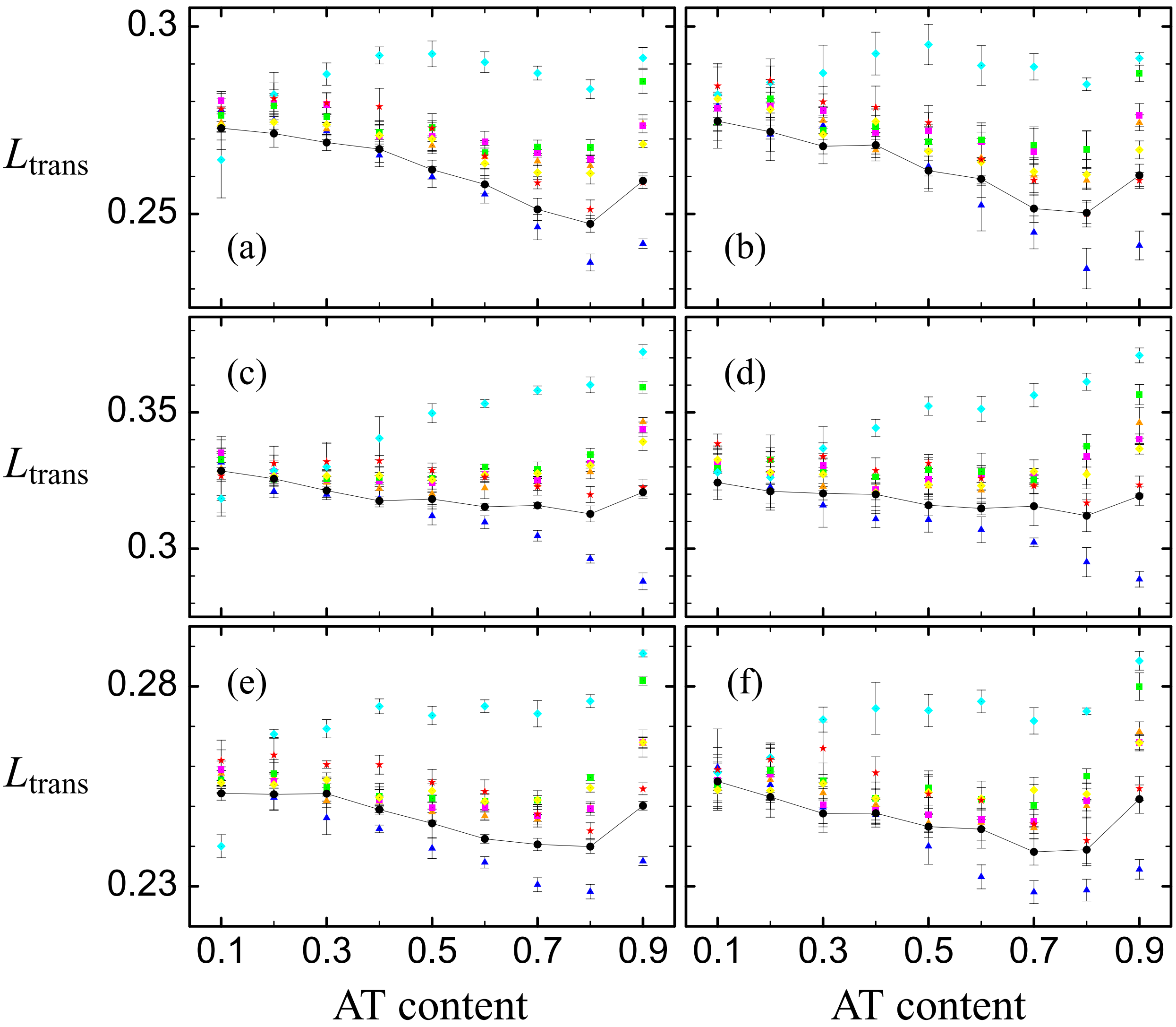


Figure 6 Translation load. Average translation load L_{trans} vs AT content for (a),(b) for the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A) (c),(d) for the acyl carrier protein (PDB id. 1hy8, chain A) and (e),(f) for the cold-shock protein (PDB id. 1c9o, chain A) as well as for (a),(c),(e) transition-to-transversion ratio $k = 2$ and (b),(d),(f) $k = 20$. Symbols are as in Fig. 3, black circles indicate standard genetic code (which are connected by lines for better visibility), whereas the other different colors indicate the seven naturally occurring variants studied, as listed in Fig. 2. The error bars indicate the mean's standard deviation.

than misinterpreting a sense codon as another sense codon), the lower translation load of the 'Echinoderm and Flatworm Mitochondrial Code' (blue triangles) in comparison to the standard genetic code seen in Fig. 6 is not trivially due to the fact that this alternative genetic code has two instead of three stop codons.

Assumptions and empirical observations

Like all mathematical models of evolution, our model depends on several assumptions and parameters. An important assumption is that the population is genetically homogeneous, i.e. the product $N\mu$ of population size times mutation rate is small. This assumption is considered approximately valid for eukaryotic and bacterial populations when considering an individual protein, in particular if population size is small. Strict validity of this assumption would imply that the number of different alleles at a typical locus is not larger than two. Despite that this is not the case, the number of alleles at a typical locus is usually small, so that the assumption is at least approximately valid. The high mutation rates of RNA viruses violate this assumption, and in this case recent work [54,55] has shown that even the neutral model should be re-formulated in the framework of the quasi species theory [56]. If we considered a whole evolving genome instead of a single protein, the approximation of very small mutation rate would not be justified, and a new interesting effect has to be considered, namely the hitch-hiking effect, which consists in the fixation of mildly disfavorable alleles driven by a positively selected allele present in the same chromosome. The mutation process was modelled using two parameters, the mutation bias and the transition-to-transversion ratio. While this parameterization might appear too simplified, it has the merit to focus on two variables whose relevance has been pointed out by a large number of experimental studies, statistical analysis, and models.

As evolutionary model, we adopt a model in spirit Kimura's neutral theory of molecular evolution [42,43], in which mutations are either neutral or lethal. The assumption of neutrality prevents us to study how the genetic code affects the fitness that can be achieved in evolution (in the neutral model the fitness of a viable sequence is equal to one in arbitrary units by definition), however it allows to study its influence on the mutation and translation load without any further assumptions concerning the shape of the fitness landscape. Whether a mutation is neutral or lethal is decided based on unfolding and misfolding stabilities of the resulting amino acid sequence.

The ingredient of our model that seems more debatable is the genotype to phenotype mapping, which is based on predicted unfolding and misfolding stabilities. While we do not claim that our predictions are accurate in individual cases, our experience suggests that they are statistically correct, so that they are able to discover statistical trends, which is what we address here. However, a limitation of our approach consists in that the mitochondrial proteome mainly contains membrane proteins, whereas our predictions of unfolding and misfolding stabilities are only valid for soluble (i.e., non-membrane) proteins, hence implying caution about the interpretation of the deleterious effect of mitochondrial codes. Besides of that, we note that the statistical potentials that we use here are quite general, as they have been optimized based on all soluble globular proteins in the protein data bank, and they are not limited to a particular organism or protein family. An alternative way to derive empirical potentials for protein evolution consists in fitting the potentials to maximize the likelihood of the observed sequences, which provides an improved fit in an evolutionary context [57].

Another important point concerns the choice of the neutral thresholds α_{thr} and F_{thr}.We have tested in previous studies that changing the neutral thresholds within reasonable limits (approx. 25% in both directions) does not significantly affect the results of neutral simulations [28-32].

Furthermore, we assume in our model that all synonymous mutations that do not change the amino acid sequence are neutral. Nevertheless, it is known that the use of alternative codons has important phenotypic effects on the translation dynamics [58,59], and it can affect the rate of translation errors [60]. Modelling these effects, however, would require assumptions on the abundance of different tRNA molecules and the dynamics of the ribosome that are outside the scope of our model. Therefore, selection on optimal codon usage is a way to reduce the load due to translation error that is complementary to the influence of the genetic code that we investigate here.

Finally, an interesting empirical observation that might be related with our protein evolution model is the finding that long genes tend to have lower codon usage bias [61]. One of us previously observed that longer proteins have contact interaction energies that are less optimized than for shorter proteins. This finding has a simple neutral explanation, since the number of contact interaction per protein is larger in longer proteins, whereas the conformation entropy loss per residue that these interactions have to compensate does not depend on protein length [62]. Therefore, contact interactions are subject to weaker selective constraints in long proteins. If this is true, and if the selective forces on codon bias are mainly due to the advantage of reducing folding problems in mistranslated proteins, which is now considered the prevailing view [14], one would expect that the selective forces on codon bias are also reduced for longer proteins, consistent with the empirical observation by Duret and Mouchiroud [61]. This subject may be addressed in the framework of an improved model in which fitness takes into account both the two protein folding stabilities and the translation load.

Conclusions

Our results show that the standard genetic code is generally preferable to naturally occurring variants, in the sense that it typically yields smaller mutation and translation loads. This finding is consistent with the view that the standard genetic code is well adapted for reducing the consequences of translation errors on protein folding stability, as expected in the framework of the adaptive theory on the origin of the genetic code, and that the fixation of alternative genetic codes is either a slightly deleterious event that is mutationally driven or, if it brings some selective advantage, this is through the reduction of the number of tRNA needed.

Nevertheless, we found one alternative genetic code (the 'Echinoderm and Flatworm Mitochondrial Code') that seems to be better at reducing mutation and translation loads, and particularly yields better translation load except for very small AT content (approx. 20% AT content or less). As translation load is a very important constraint in protein evolution [14], this difference might result in small but significant fitness differences and hence be subject to positive selection. Therefore, although our model confirms the view that code changes are slightly deleterious events in the majority of the cases, it also suggests that adaptation cannot be excluded for one of the studied cases.

Methods

Unfolding and misfolding stability

As in our previous work [28-32], the unfolding free energy $F(\mathbf{A})$ of a protein with sequence $\mathbf{A} = \{A_1 ... A_L\}$ and contact matrix $C_{ij}= 1$ if the minimal interatomic distance between residues i and j is below 4.5 Å, 0 otherwise, is defined as

$$F(\mathbf{A}) = \sum_{i<j} C_{ij} U(A_i, A_j), \tag{1}$$

where $U(a, b)$ is the contact interaction matrix determined in Ref. [27]. Although rather simple, this model is accurate enough to allow quantitative predictions of the folding free energy of small proteins that fold with two-state thermodynamics (the correlation coefficient between experimental and predicted free energy is r = 0.92 over a representative test set of 20 proteins, UB, unpublished result) and of the stability effect of mutations (correlation coefficient r = 0.72 over a set of 195 mutations, UB, unpublished result). This is comparable to state-of-the-art programs such as Fold-X [63]. However, the computational simplicity of the model makes it affordable to use it for simulating very long evolutionary trajectories, which would not be possible using other tools. The unfolding free energy should also take into account the loss of conformation entropy upon folding, which we modelled in other works as sL with s being the chain entropy per amino acid and L the protein length. However, this term only induces a constant shift sL in the unfolding free energy, $F'(\mathbf{A}) = \ _{i<j} C_{ij} U(A_i, A_j) + sL$ and its effect is just to shift the neutral threshold F_{thr} in the same direction, without influencing the results.

The normalized energy gap $\alpha(\mathbf{A})$ measures the (positive) energy difference between alternative compact conformations and the native conformation, and it is defined using the random energy model [64,65] as

$$\alpha(\mathbf{A}) = \frac{1}{1-q0} \left(1 - \frac{\langle e \rangle_{\mathbf{A}} N_c - \sigma_{e,\mathbf{A}} \sqrt{2N_c(AL+B)}}{\sum_{i<j} C_{ij} U(A_i, A_j)} \right), \tag{2}$$

with $A = 0.1$, $B = 4$, $q_0 = 0.1$, and $N_c = \ _{i<j} C_{ij}$. $7e8_{\mathbf{A}}$ and $\sigma_{e,\mathbf{A}}$ are the mean and standard deviation of the interaction energy of both native and non-native contacts in sequence **A.**

Mutation process

Mutations are modelled through the HKY process [66], in which the mutation rate from nucleotide n to n', $T(n, n')$, is $\mu f(n')$ if $n\mathrm{T}n'$ is a transition, and $\mu k f(n')$ if it is a transversion. The transition-to-transversion ratios used in this work are k = 2 and k = 20, suitable for nuclear and mitochondrial DNA, respectively [46,47]. The microscopic rate μ is assumed to be very small and it does not affect the results. We further assume π(A) = π(T) and π(C) = π(G) (Chargaff second parity rule), so that the only parameter of the mutation model is the stationary AT frequency, AT = π(A) + π(T).

Simulation of the evolutionary process

Our model of protein evolution cannot be treated analytically, so that we have to study it using numerical simulations (see Fig. 1). We start each simulation with the native amino acid sequence obtained from the Protein Databank (PDB) of the chosen structure, from which we construct a corresponding 'native' DNA sequence by randomly choosing codons using the genetic code under consideration with weights determined by the given AT content (inverse translation). The simulation thus starts as close as possible to the equilibrium with the chosen AT content. The initial part of the trajectory is discarded to ensure that relevant quantities are sampled at the stationary state. To do so, we visually verified that the stabilities had reached the stationary state for simulations with the standard code, and then discarded the same transient part of the trajectory for all alternative codes.

The simulations are performed as follows. At every step, we randomly select one DNA site j with probability dependent on the nucleotide n_j occupying it, $P_j \propto \sum_{n' \neq n_j} T_\mu(n_j, n')$ and we extract the mutated nucleotide $n' \neq n_j$ with probability proportional to $T_\mu(n_j, n')$. The mutated DNA is then translated to an amino acid sequence, whose unfolding and misfolding stabilities are computed through Eqs. (1) and (2). The mutation receives fitness $\mathcal{F}$ = 1 (in arbitrary units) and is accepted if both $F < F_{\text{thr}}$ and $\alpha > \alpha_{\text{thr}}$, or gets a fitness $\mathcal{F}$ = 0 and is rejected otherwise, hence assuming a model in spirit of Kimura's neutral theory of molecular evolution. Mutations to stop codons are considered lethal and receive a fitness $\mathcal{F}$ = 0. As the mutation process is continuous, the waiting time until a new mutation arises is a Poissonian variable with mean μ^{-1}. Instead of drawing an explicit waiting time for each mutation to arise, we assign each mutation the mean time μ^{-1} (this is equivalent to performing an average over possible realizations of waiting times). In case several mutations occur before one gets fixed, the weighting of the sequence before the accepted mutation is increased accordingly. The simulation is run for a large number of 10^6 mutations to obtain long evolutionary trajectories which are used to calculate the averages. The neutral thresholds F_{thr} and α_{thr} are calculated for each simulated protein and kept fixed during the simulations. We set $F_{\text{thr}} = \gamma F_{\text{nat}}$ and $\alpha_{\text{thr}} = \gamma\alpha_{\text{nat}}$, where F_{nat} and α_{nat} are the unfolding free energy and misfolding stability of the respective native amino acid sequence. The factor γ is chosen as $\gamma = 0.98$, so that the native sequence is considered viable. Changing γ within reasonable limits (approx. 25% in both directions) does not significantly effect the results.

Mutation and translation load

The mutation load per site $L_{\text{mut}}(\mathbf{n})$ of a DNA sequence **n** translated to amino acid sequence $\mathbf{A}[\mathbf{n}]$ is defined as

$$L_{\text{mut}}(\mathbf{n}) = \frac{1}{\mu N_{\text{mut}}(\mathbf{n})} \sum_{\mathbf{n}'} \Delta\mathcal{F}(\mathrm{A}[\mathbf{n}] \to \mathrm{A}[\mathbf{n}']) \times R_{\text{mut}}(\mathbf{n} \to \mathbf{n}'), \tag{3}$$

where $\Delta\mathcal{F}(\mathbf{A}[\mathbf{n}]\to\mathbf{A}[\mathbf{n}'])$ is the fitness difference between amino acid sequence $\mathbf{A}[\mathbf{n}']$, as translated from the mutated DNA sequence $\mathbf{n}'$, and amino acid sequence $\mathbf{A}[\mathbf{n}]$, and $R_{\text{mut}}(\text{nTn}')$ is the rate of a mutation from **n** to $\mathbf{n}'$, which is calculated according to our mutation process with the mutation rate μ and only single nucleotide mutations are considered (i.e. only terms linear in μ, ignoring the higher order terms which have rates proportional to μ^2 and μ^3 and are hence much smaller than μ). Since we study a neutral model, so that the fitness of a viable sequence is $\mathcal{F} = 1$ (in arbitrary units) and $\mathcal{F} = 0$ otherwise, the fitness difference $\Delta\mathcal{F}(\mathbf{A}[\mathbf{n}]\to\mathbf{A}[\mathbf{n}'])$ can only take the values 0 if $\mathbf{A}[\mathbf{n}']$ is a viable sequence as well, and 1 if $\mathbf{A}[\mathbf{n}']$ is not a viable sequence. As we restrict ourselves to those $\mathbf{n}'$ which differ from **n** by a single nucleotide, the sum in Eq. (3) contains $9L$ terms, and hence the normalization $N_{\text{mut}}(\mathbf{n})\ \mathcal{F}\ {}_{\mathbf{n}'}\Theta[R_{\text{mut}}(\mathbf{n}\ \mathrm{T}\ \mathbf{n}')]$, where $\Theta[R_{\text{mut}}(\mathbf{n} \to \mathbf{n}')] = 1$ if $R_{\text{mut}}(\mathbf{n}\mathrm{T}\mathbf{n}') > 0$ and 0 otherwise, yields $N_{\text{mut}}(\mathbf{n}) = 9L$ independent of **n**. Consequently, in the extreme case of all DNA sequence $\mathbf{n}'$ which differ from **n** by a single nucleotide (one point mutation) being viable and hence $\Delta\mathcal{F}(\mathbf{A}[\mathbf{n}]\to\mathbf{A}[\mathbf{n}']) = 0$ for all these $\mathbf{n}'$, the mutational load is $L_{\text{mut}}(\mathbf{n}) = 0$. If the transition-to-transversion ratio was $k = 1$ so that $R_{\text{mut}}(\mathbf{n}\to\mathbf{n}') = \mu$ for all $\mathbf{n}'$ differing from **n** by a single nucleotide, then $\mu N_{\text{mut}} = {}_{\mathbf{n}'} R_{\text{mut}}(\mathbf{n} \to \mathbf{n}')$ and the mutation load $L_{\text{mut}}(\mathbf{n})$ would be the fraction of lethal sequences $\mathbf{n}'$. Due to our choice for the normalization, there is no explicit dependence of $L_{\text{mut}}(\mathbf{n})$ on sequence length.

The translation load per site $L_{\text{trans}}(\mathbf{n})$ of a DNA sequence **n** translated to amino acid sequence $\mathbf{A}[\mathbf{n}]$ is similarly defined as

$$L_{\text{trans}}(\mathbf{n}) = \frac{1}{\nu N_{\text{trans}}(\mathbf{n})} \sum_{\mathbf{n}'} \Delta\mathcal{F}(\mathrm{A}[\mathbf{n}] \to \mathrm{A}[\mathbf{n}']) \times R_{\text{trans}}(\mathbf{n} \to \mathbf{n}'), \tag{4}$$

where $R_{\text{trans}}(\mathbf{A}[\mathbf{n}] \to \mathbf{A}[\mathbf{n}'])$ is the rate of a translation error resulting in amino acid sequence $\mathbf{A}[\mathbf{n}']$ instead of $\mathbf{A}[\mathbf{n}]$ and ν the rate of single nucleotide mismatches. For simplicity, we assume that $R_{\text{trans}}(\mathrm{n}\to\mathrm{n}') = \nu$ if nucleotide sequence $\mathbf{n}'$ resulted from **n** by a single nucleotide mismatch and does not contain any stop codon and 0 otherwise (i.e. only terms linear in ν are considered, as for the mutation load), so that our definition of the translation load does not depend on the error rate of translation, which is approximately 10^{-4} per translated mRNA codon [67] but may differ from species to species. We exclude nucleotide sequences $\mathbf{n}'$ containing a stop codon for the computation of the translation load since a premature end of translation by misinterpreting a sense codon as a stop codon is caused by release factors and not by tRNAs (and hence by a different mechanism than misinterpreting a sense codon as another sense codon) which furthermore depends on neighboring codons [53]. For simplicity, we neglect here this error rate with respect to the rate of missense errors in translation. In this way, the translation load does not explicitly dependent on the number of stop codons, and the normalization $N_{\text{trans}}(\mathbf{n}) = {}_{\mathbf{n}'}\Theta[R_{\text{trans}}(\mathbf{n} \to \mathbf{n}')]$, where $\Theta[R_{\text{trans}}(\mathbf{n} \to \mathbf{n}')] = 1$ if $R_{\text{trans}}(\mathbf{n}\ \mathrm{T}\ \mathbf{n}') > 0$ and 0 otherwise, does dependent on **n**. Even though we use the above general definition, Eq. (4), in analogy to the mutation load, note that with our choice for $R_{\text{trans}}(\mathbf{n} \to \mathbf{n}')$, $\nu N_{\text{trans}} = {}_{\mathbf{n}'} R_{\text{trans}}(\mathbf{n} \to \mathbf{n}')$ and the translation load $L_{\text{trans}}(\mathbf{n})$ is the fraction of lethal sequences among all $\mathbf{n}'$ differing from **n** by a single nucleotide and not containing a stop codon. Due to our choice for the normalization, there is neither an explicit dependence of $L_{\text{trans}}(\mathbf{n})$ on sequence length nor on the number of stop codons in the genetic code considered.

Protein list

We studied the three following proteins structures: (i) the epsilon subunit of F1F0-ATP synthase (PDB id. 1aqt, chain A, $\alpha + \beta$ protein, 135 amino acids, GenBank CBG36944.1), (ii) the acyl carrier protein (PDB id. 1hy8, chain A, all-β protein, 76 amino acids, GenBank BAA10975.1 and CAB13465.1), and (iii) the cold-shock protein (PDB id. 1c9o, chain A, all-β protein, 66 amino acids, GenBank CAA51790.1).

Authors' contributions

MP and UB designed research; SGS performed research; SGS, UB, and MP analyzed data; MP and UB wrote the paper. All authors read and approved the final manuscript

Acknowledgements

We acknowledge contributions of Andreas Buhr in early stages of this work. UB acknowledges financial support from the Spanish Science and Innovation Ministry through the Ramon y Cajal program and through the projects BIO2008-04384 and CSD2006-00023. Our collaboration was facilitated through the program "Acciones Integradas España-Alemania" of the Spanish Science and Innovation Ministry, project HA2006-0044, and of the Deutscher Akademischer Austauschdienst project D/06/12848.

Author Details

[1]Institut für Festkörperphysik, Technische Universität Darmstadt, Hochschulstr. 8, 64289 Darmstadt, Germany, [2]Centro de Biología Molecular 'Severo Ochoa', (CSIC-UAM), Cantoblanco, 28049 Madrid, Spain, [3]Karlstr. 23, 64283 Darmstadt, Germany and [4]Institut für Theoretische Physik, Universität zu Köln, Zülpicher Str. 77, 50937 Köln, Germany

Received: 7 August 2009 Accepted: 14 June 2010
Published: 14 June 2010

References

1. Koonin E, Nozozhilov A: **Origin and Evolution of the Genetic Code: The Universal Enigma.** *IUBMB Life* 2009, **61:**99-111.
2. Nirenberg M, Jones W, Leder P, Clark B, Sly W, Pestka S: **On the coding of genetic information.** *Cold Spring Harbor Symp Quant Biol* 1963, **28:**549-557.
3. Haig D, Hurst L: **A quantitative measure of error minimization in the genetic code.** *J Mol Evol* 1991, **33:**412-417.
4. Freeland S, Hurst L: **The genetic code is one in a million.** *J Mol Evol* 1998, **47:**238-248.
5. Freeland S, Knight R, Landweber L, Hurst L: **Early fixation of an optimal genetic code.** *Mol Biol Evol* 2000, **17:**511-518.
6. Gilis D, Massar S, Cerf N, Rooman M: **Optimality of the genetic code with respect to protein stability and amino-acid frequencies.** *Genome Biology* 2001, **2:**11.
7. Zhu C, Zeng X, Huang W: **Codon usage decreases the error minimization within the genetic code.** *J Mol Evol* 2003, **57:**533-537.
8. Goodarzi H, Nejad HA, Torabi N: **On the optimality of the genetic code, with the consideration of termination codons.** *BioSystems* 2004, **77:**163-173.
9. Santos M, Moura G, Massey S, Tuite M: **Driving change: The evolution of alternative genetic codes.** *Trends in Genetics* 2004, **20:**95-102.
10. Sella G, Ardell D: **The coevolution of genes and genetic code: Crick's frozen accident revisited.** *J Mol Evol* 2006, **63:**297.
11. Hohn M, Park H, O'Donoghue P, Schnitzbauer M, Söll D: **Emergenze of the universal genetic code imprinted in an RNA record.** *Proc Natl Acad Sci USA* 2006, **103:**18095-18100.
12. Freeland S, Wu T, Keulmann N: **The case for an error minimizing standard genetic code.** *Origins of Life and Evolution of the Biosphere* 2003, **33:**457-477.
13. Itzkovitz S, Alon U: **The genetic code is nearly optimal for allowing additional information within protein-coding sequences.** *Genome Research* 2007, **17:**405-412.
14. Drummond D, Wilke C: **Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.** *Cell* 2008, **134:**341-352.
15. Gamow G: **Possible relation between deoxyribonucleic acid and protein structure.** *Nature* 1954, **173:**318.
16. Sonneborn T: **Degeneracy of the genetic code: Extent, nature, and genetic implications.** In *Evolving Genes and Proteins* Edited by: Bryson V, Vogel H. New York: Academic Press; 1965:377-397.
17. Epstein C: **Role of amino acid "code" and selection for confirmation in the evolution of proteins.** *Nature* 1966, **210:**25-28.
18. Woese C: **On the evolution of the genetic code.** *Proc Natl Acad Sci USA* 1965, **54:**1546-1552.
19. Goldberg A, Wittes R: **Genetic code: Aspects of organization.** *Science* 1966, **153:**240.
20. Davies J, Gilbert W, Gorini L: **Streptomycin, suppression, and the code.** *Proc Natl Acad Sci USA* 1964, **51:**883-890.
21. Friedman S, Weinstein I: **Lack of fidelity in the translation of ribopolynucleotides.** *Proc Natl Acad Sci USA* 1964, **52:**988-996.
22. Wong J: **Coevolution theory of the genetic code.** *Proc Natl Acad Sci USA* 1975, **72:**1909-1912.
23. Sengupta S, Higgs P: **A unified model of codon reassignment in alternative genetic codes.** *Genetics* 2005, **170:**831-840.
24. Sengupta S, Yang X, Higgs P: **The mechanisms of codon reassignments in mitochondrial genetic codes.** *J Mol Evol* 2007, **64:**662-688.
25. Andersson S, Kurland C: **Genomic evolution drives the evolution of the translation system.** *Biochem Cell Biol* 1995, **73:**775-787.
26. Andersson S, Kurland C: **Reductive evolution of resident genomes.** *Trends Microbiol* 1998, **6:**263-268.
27. Bastolla U, Farwer J, Knapp E, Vendruscolo M: **How to guarantee optimal stability for most representative structures in the Protein Data Bank.** *Proteins* 2001, **44:**79-96.
28. Bastolla U, Porto M, Roman H, Vendruscolo M: **Lack of self-averaging in neutral evolution of proteins.** *Phys Rev Lett* 2002, **89:**208101.
29. Bastolla U, Porto M, Roman H, Vendruscolo M: **Connectivity of neutral networks, overdispersion and structural conservation in protein evolution.** *J Mol Evol* 2003, **56:**243-254.
30. Bastolla U, Porto M, Roman H, Vendruscolo M: **Statistical properties of neutral evolution.** *J Mol Evol* 2003, **57:**S103-S119.
31. Porto M, Roman H, Vendruscolo M, Bastolla U: **Prediction of site-specific amino acid distributions and limits of divergent evolutionary changes in protein sequences.** *Mol Biol Evol* 2005, **22:**630-638.
32. Bastolla U, Porto M, Roman H, Vendruscolo M: **A protein evolution model with independent sites that reproduces site-specific amino acid distributions from the Protein Data Bank.** *BMC Evol Biol* 2006, **6:**43.
33. Babajide A, Hofacker I, Sippl M, Stadler P: **Neutral networks in protein space.** *Fol Des* 1997, **2:**261-269.
34. Bussemaker H, Thirumalai D, Bhattacharjee J: **Thermodynamic stability of folded proteins against mutations.** *Phys Rev Lett* 1997, **79:**3530-3533.
35. Govindarajan S, Goldstein R: **On the thermodynamic hypothesis of protein folding.** *Proc Natl Acad Sci USA* 1998, **95:**5545-5549.
36. Tiana G, Broglia R, Roman H, Vigezzi E, Shakhnovich E: **Folding and misfolding of designed proteinlike chains with mutations.** *J Chem Phys* 1998, **108:**757-761.
37. LA Mirny VA, Shakhnovich E: **How evolution makes proteins fold quickly.** *Proc Natl Acad Sci USA* 1998, **95:**4976-4981.
38. Bornberg-Bauer E, Cha H: **Modeling evolutionary landscapes: Mutational stability, topology, and superfunnels in sequence space.** *Proc Natl Acad Sci USA* 1999, **96:**10689-10694.
39. Dokholyan N, Shakhnovich E: **Understanding hierarchical protein evolution from first principles.** *J Mol Biol* 2001, **312:**289-307.
40. Parisi G, Echave J: **Structural constraints and emergence of sequence patterns in protein evolution.** *Mol Biol Evol* 2001, **18:**750-756.
41. Drake J: **Avoiding dangerous missense: Thermophiles display especially low mutation rates.** *PLoS Genetics* 2009, **5:**e1000520.
42. Kimura M: **Evolutionary rate at the molecular level.** *Nature* 1968, **217:**624-626.
43. Graur D, Li W: **The neutral theory of molecular evolution.** Cambridge University Press, Cambridge; 1983.
44. Mendez R, Fritsche M, Porto M, Bastolla U: **Mutation bias favors protein folding stability in the evolution of small populations.** *PLoS Comp. Biol* 2010, **6:**e1000767.
45. Bastolla U, Moya A, Viguera E, van Ham R: **Genomic determinants of protein folding thermodynamics.** *J Mol. Biol* 2004, **343:**1451-1466.
46. Graur D, Li W: **Fundamentals of molecular evolution.** Sinauer, Sunderland; 2000.
47. Belle E, Piganeau G, Gardner M, Eyre-Walker A: **An investigation of the variation in the transition bias among various animal mitochondrial DNA.** *Gene* 2005, **355:**58-66.
48. Yokobori S, Suzuki T, Watanabe K: **Genetic code variations in mitochondria: tRNA as a major determinant of genetic code plasticity.** *J Mol Evol* 2001, **53:**314-326.
49. Anderson S, Bankier A, Barrell B, de Bruijn M, Coulson A, Drouin J, Eperon I, Nierlich D, Roe B, Sanger F, Schreier P, Smith A, Staden R, Young I: **Sequence and organization of the human mitochondrial genome.** *Nature* 1981, **290:**457-465.

50. Clark-Walker G, Weiller G: **The structure of the small mitochondrial DNA of kluyveromyces thermotolerans is likely to reflect the ancestral gene order in fungi.** *Biochimica et Biophysica Acta* 1995, **1228**:1-27.
51. Hoffman D, Anderson R, DuBois M, Prescott D: **Macronuclear gene-sized molecules of hypotrichs.** *Nucl Ac Res* 1995, **23**:1279-1283.
52. Santos M, Keith G, Tuite M: **Non-standard translational events in candida albicans mediated by an unusual seryl-tRNA with a 5'-cag-3' (leucine) anticodon.** *The EMBO Journal* 1993, **12**:607-616.
53. Freistroffer D, Kwiatkowski M, Buckingham R, Ehrenberg M: **The accuracy of codon recognition by polypeptide release factors.** *Proc. Natl. Acad. Sci. USA* 2000, **97**:2046-2051.
54. van Nimwegen E, Crutchfield J, Huynen M: **Neutral evolution of mutational robustness.** *Proc Natl Acad Sci USA* 1999, **96**:9716-9720.
55. Wilke C: **Molecular clock in neutral protein evolution.** *BMC Genetics* 2004, **5**:25.
56. Eigen M: **Selforganization of matter and the evolution of biological macromolecules.** *Naturwissenschaften* 1971, **58**:465-523.
57. Kleinman C, Rodrigue N, Lartillot N, Philippe H: **Statistical potentials for improved structurally constrained evolutionary models.** *Mol Biol Evol* 2010 in press.
58. Cannarozzi G, Schraudolph N, Faty M, von Rohr P, Friberg M, Roth A, Gonnet P, Gonnet G, Barral Y: **A role for codon order in translation dynamics.** *Cell* 2010, **141**:355-67.
59. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, Pan T, Dahan O, Furman I, Pilpel Y: **An evolutionarily conserved mechanism for controlling the efficiency of protein translation.** *Cell* 2010, **141**:344-54.
60. Hershberg R, Petrov D: **Selection on codon bias.** *Annu Rev Genet* 2008, **42**:287-99.
61. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in Caenorhabditis, Drosophila, and Arabidopsis.** *Proc Natl Acad Sci USA* 1999, **96**:4482-7.
62. Bastolla U, Demetrius L: **Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds.** *Protein Eng Des Sel* 2005, **18**:405-415.
63. Guerois R, Nielsen JE, Serrano L: **Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations.** *J Mol Biol* 2002, **320**:369-387.
64. Derrida B: **Random Energy Model: an exactly solvable model of disordered systems.** *Phys Rev B* 1981, **24**:2613-2626.
65. Shakhnovich E, Gutin A: **Formation of unique structure in polypeptide chains. Theoretical investigation with the aid of a replica approach.** *Biophys Chem* 1989, **34**:187-199.
66. Hasegawa M, Kishino H, Yano T: **Dating the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22**:160-174.
67. Kurland C: **Translational accuracy and the fitness of bacteria.** *Annu Rev Genet* 1992, **26**:29-50.

doi: 10.1186/1471-2148-10-178
Cite this article as: Sammet *et al.*, Comparison of translation loads for standard and alternative genetic codes *BMC Evolutionary Biology* 2010, **10**:178