Study of the genetic code adaptability by means of a genetic algorithm

José Santos, Ángel Monteagudo

Study of the genetic code adaptability by means of a genetic algorithm

José Santos, Ángel Monteagudo

PII: S0022-5193(10)00116-5
DOI: doi:10.1016/j.jtbi.2010.02.041
Reference: YJTBI 5894

To appear in: *Journal of Theoretical Biology*

Received date: 3 September 2009
Revised date: 5 January 2010
Accepted date: 23 February 2010

Cite this article as: José Santos and Ángel Monteagudo, Study of the genetic code adaptability by means of a genetic algorithm, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.02.041

Study of the Genetic Code Adaptability by Means of a Genetic Algorithm

José Santos, Ángel Monteagudo,

University of A Coruña, Department of Computer Science, Campus de Elviña s/n, 15071 A Coruña (Spain), Phone: +34 81 167000

Abstract

We used simulated evolution to study the adaptability level of the canonical genetic code. An adapted genetic algorithm (GA) searches for optimal hypothetical codes. Adaptability is measured as the average variation of the hydrophobicity that the encoded amino acids undergo when errors or mutations are present in the codons of the hypothetical codes. Different types of mutations and point mutation rates that depend on codon base number are considered in this study. Previous works have used statistical approaches based on randomly generated alternative codes or have used local search techniques to determine an optimum value. In this work, we emphasize what can be concluded from the use of simulated evolution considering the results of previous works. The GA provides more information about the difficulty of the evolution of codes, without contradicting previous studies using statistical or engineering approaches. The GA also shows that, within the coevolution theory, the third base clearly improves the adaptability of the current genetic code.

Key words: Genetic code theories, error-minimization hypothesis, genetic code evolution, genetic algorithms.

1 Introduction

In this work we use a genetic algorithm as a method to corroborate the adaptability of the standard or canonical genetic code. Although this code is not universal (for example, mitochondrial DNA has variations), it is present in most complex genomes. The genetic code, with the four nitrogenated bases, when grouped in genes, encodes the amino acids that are linked to determine the proteins. This code is redundant given three bases are needed to establish a codon that codifies each of the 20 amino acids that are present in proteins, plus a "stop translation" signal found at the end of every gene. Hence, using this code, most of the amino acids are specified by more than one codon, implying the redundancy of the code.

Nevertheless, as there are 64 possible codons to encode the 21 labels, numerous hypothetical "genetic codes" could be defined, with associations different from those of the standard genetic code. The number of possible codes is $1.4 \cdot 10^{70}$, as Yockey (2005) has calculated, taking into account the amino acid assignments of the modern standard genetic code. The alternative codes have the same number of codons per amino acid as in the standard code: for example, 3 amino acids codified by sets of six codons, 5 amino acids codified by four codons, etc. When the codon set structure of the standard genetic code is unchanged, only considering permutations of the amino acids coded in the 20 sets, there are 20! ($2.43 \cdot 10^{18}$) possible codes. Finally, as indicated by Schönauer and Clote (1997), there are more than $1.51 \cdot 10^{84}$ general codes, without restrictions in the mapping of the 64 codons to the 21 labels.

* Corresponding author: José Santos
Email address: santos@udc.es (José Santos).

The establishment of the genetic code is still under discussion, although the discovery of non-standard genetic codes altered the "frozen accident", as named by Crick (1968). The genetic code could have been an adaptive process by means of natural evolution. This implies that the codes with less harmful effects in the possible errors of the protein synthesis machinery (and in the final proteins) have an evolutionary advantage over codes that present a greater number of harmful effects. An argument in favor of this adaptability of the standard genetic code is that the amino acids with similar chemical properties are coded by similar codons. For example, the codons that share two of the three bases tend to correspond to amino acids that have similar hydrophobicity.

Many authors present results in favor of this option. For instance, Woese (1965) stated that "the codon catalogue has more recently been shown to manifest very definite correlations among the codon assignments for related amino acids", and Goldberg and Wittes (1966) indicated that the pattern of organization of the genetic code decreases to a minimum the phenotypic effects of mutation and of base-pairing errors in protein synthesis. Single base changes, especially transitions, usually cause either no amino acid change or the change to a chemically similar amino acid. In more recent works, Jestin and Kempf (1997) showed that the codon assignment of stop signals optimized the tolerance of polymerase-induced frameshift mutations, that is, most single-base deletions are less deleterious at chain termination codons than at codons encoding amino acids. Furthermore, the work of Ardell and Sella (2002) using a population genetic model of code-message coevolution demonstrated that such coevolution tends to produce structure-preserving codes, and in particular it can reproduce some of the structure-preserving patterns of the standard

genetic code.

Di Giulio (2005), in a review on the theories of the origin and evolution of the genetic code, distinguishes two basic alternatives. The stereochemical theory claims that the origin of the genetic code must lie in the stereochemical interactions between anticodons or codons and amino acids. Then, the physicochemical theory claims that the force defining the origin of the genetic code structure is the one that tends to reduce the deleterious effects of physicochemical distances between amino acids codified by codons differing in one base. There is a third element in the set of hypotheses: the structure of the genetic code reflects the biosynthetic pathways of amino acids through time and the error minimization at the protein level is just a consequence of this process, as indicated, for example, by Di Giulio (2005) and Torabi et al. (2007). This is the so-called coevolution hypothesis (Wong, 1975). This coevolution theory suggests that codons, originally assigned to prebiotic precursor amino acids, were progressively assigned to new amino acids derived from the precursors as biosynthetic pathways evolved. Higgs (2009) proposed a "four column" code as an early state, based on the evidence about which were the earliest amino acids. For the author the driving force during the build-up of the standard code is not the minimization of translational error, and the main factor that influenced the positions in which new amino acids were added is that there should be minimal disruption of the protein sequences that were already encoded. Nevertheless, the code that results is one in which the translational error is minimized.

The works of Freeland and Hurst are significant in the study of the adaptability of the standard genetic code of the physicochemical theory. They corroborate that adaptability by means of a simulation, when they determined that in

a sample of 1 million random hypothetic codes there was only around one hundred with lower error than the one of the standard genetic code (Freeland and Hurst, 1998a). The error is measured as the average variation of the polar requirement property (a measurement of hydrophobicity) that the amino acids undergo when only one letter of each codon is changed, with all the possible variations, in the set of the 64 codons of each hypothetical code. Using a different methodology, Di Giulio measures the optimality of the genetic code in relation to the best possible code that could be reached. This code is obtained in his case with local search techniques (Di Giulio, 1989).

Opposite to that brute-force search of possible codes that are better adapted or to the local search for the best hypothetical code, we used simulated evolution, by means of a standard genetic algorithm (GA) (Goldberg, 1989) adapted to our problem. The GA provides a guided and global search of better adapted hypothetical codes as well as a method to guess the progression and the difficulty in finding such alternative codes.

This paper is organized as follows. Section 2 summarizes previous works performed in this field. Section 3 briefly explains the definitions of the hypothetical codes used in our simulations, whereas Section 4 presents the implementation details of the genetic algorithm in our application in the search for better adapted codes. Section 5 explains the measurements used to determine the level of adaptation of the genetic code. Section 6 expounds the first results of evolution when searching for alternative codes, Section 7 shows the results with a bias in the transition/transversion probability ratio of mutations whereas Section 8 presents the results with the introduction of errors as a function of the base position in the codon. Finally, Section 9 presents some conclusions.

2 Previous work

Several authors have studied the adaptability of the genetic code using different simulations. In a first computational experiment, Haig and Hurst (1991) determined an adaptability level of the canonical genetic code by means of a simple simulation. They found that of 10,000 randomly generated codes, only 2 performed better at minimizing the effects of errors, when polar requirement was taken as the amino acid property. Thus, they estimated that the probability that a code as conservative as the standard genetic code arose by chance was 0.0002, and therefore, they concluded that the standard genetic code was a product of natural selection for load minimization. To quantify the efficiency of each possible code they used a measurement that considers the changes in a basic property of the amino acids when all the possible mutations are considered in a generated code (see Section 4.3 for more details). The property used by the authors was the polar requirement, which may be considered a measurement of hydrophobicity, as the one that gave the most significant evidence of load minimization from an array of four amino acid properties (hydropathy, molecular volume, isoelectric point and the polar requirement).

Freeland and Hurst (1998a) in their refinement of the previous estimate used a larger sample consisting of 1,000,000 possible codes. The criteria for creating plausible alternative codes in these works (Haig and Hurst, 1991; Freeland and Hurst, 1998a) are summarized in Section 3. Their alternative codes maintain the same synonymous block structure of the canonical code. The authors found 114 better codes (a proportion of 0.000114), indicating, according to the authors' results, a refinement of the previous estimate for relative code efficiency such that the code was even more conservative. When the authors

also took into account restrictions about the biosynthetic pathways (Taylor and Coates, 1989), they found 284 better codes. The authors indicate that "for the most part historical features do not explain the load-minimization property of the natural code" (Freeland and Hurst, 1998b).

In addition, the same authors extended this work to investigate the effect of weighting transition errors differently from transversion errors and the effect of weighting each base differently, depending on reported mistranslation biases (Section 7 explains these types of mutations). When they used weightings to allow for biases in translation, they found that only 1 in every million randomly generated codes was more efficient than the standard genetic code. In (Ronneberg et al., 2001) the authors presented a graphical tool for testing the adaptive nature of the genetic code under those assumptions about patterns of genetic error and the nature of amino acid similarity. Additionally, as the adaptability of the genetic code does not hold for other amino acid properties other than polar requirement, in (Freeland et al., 2000b) the authors applied point accepted mutation (PAM) 74-100 matrix data, which derives from frequently observed substitution patterns of amino acids in naturally occurring pairs of homologous proteins. The matrix used by the authors was built solely from evolutionary diverged proteins. Hence, as the authors indicate, the matrix provides a direct measurement of amino acid similarity in terms of protein biochemistry. The results indicate that the standard genetic code is close to the global optimum of all codes with regard to error rates. However, Di Giulio has questioned this work, as the title of the work "the origins of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analysis tautologous" clearly explains (Di Giulio, 2001).

Gilis et al. (2001) extended Freeland and Hurst's work by studying the frequency at which different amino acids occur in proteins. Their results indicate that the fraction of random codes that beat the standard genetic code decreases. In addition, they used a new function of error measurement that evaluates *in silico* the change in folding free energy caused by all possible point mutations in a set of protein structures, this being a measurement of protein stability. With that function the authors estimated that around two random codes in a billion (10^9) are fitter than the standard code.

Other works (Goodarzi et al., 2006; Torabi et al., 2007) have taken into consideration both relative frequencies of amino acids and relative gene copy frequencies of tRNAs in genomic sequences to use a fitness function which models the mistranslational probabilities more accurately in modern organisms. The relative gene copy frequencies of tRNAs are used as estimates of the tRNA content. The methodology is the same as that used in previous works, but now the aim is to find better assignments of amino acids for the two main families of aaRSs (aminoacyl-tRNA synthetases). These enzymes are responsible for charging tRNAs with their cognate amino acids. The probability of mischarging is based on the difference of molecular volume between the incorrectly charged amino acid and the correct one. In addition, they used the frequencies of amino acid concentration and the mentioned frequencies of tRNAs to measure the fitness of the alternative assignments. For example, in one of the tests, the relative frequency of more optimal classifications hardly exceeded 0.03. Their model signifies higher optimality of the genetic code towards load minimization and suggests, according to the authors, the presence of a coevolution of tRNA frequency and the genetic code.

Previous studies have used search techniques to obtain the best possible codes.

Di Giulio (1989) estimated that the standard genetic code has achieved 68% minimization of polarity distance, by comparing the standard code with random block respecting codes (the codes obtained by relabeling the 20 amino acids in the canonical table by a permutation thereof). The value of the best possible code (necessary in the comparison) was obtained using a simulated annealing technique to solve a constrained minimization problem. The minimization percentage was defined by the Di Giulio as $(\triangle_{mean} - \triangle_{code})/(\triangle_{mean} - \triangle_{low})$, where $\triangle_{mean}$ is the average error value (Section 5), obtained by averaging over many random block respecting codes, and $\triangle_{low}$ is the best (or approximated) $\triangle$ value. The author indicated that the percentage distance minimization (p.d.m.) can be interpreted as the optimization level reached during genetic code evolution. In addition, the author criticized the works of Freeland and Hurst (1998a). As an argument, Di Giulio stated that with the probability of 10^{-6}, there are $2.4 \cdot 10^{12}$ codes that should display a better value than that of the genetic code (Di Giulio, 2000). Therefore, the author preferred the use of the p.d.m. in his analysis. In Section 5 this measurement is commented again. Novozhilov et al. (2007) also employed a local greedy search which used as elementary evolutionary step a swap of the amino acids assignments with alternative codes that possessed the same block structure and the same degree of degeneracy as the standard code. One of the main authors´ conclusions was that the standard code is much closer to its local minimum (fitness peak) than most of the random codes with similar levels of robustness.

In (Di Giulio and Medugno, 1999) the authors extended the work to consider the evolution of the genetic code under the coevolution theory. In the stages of evolution considered where the code codifies less than 13 amino acids, which

implies a not very high number of permutations, the authors conduct an exhaustive search to obtain the best possible code, while they use the simulated annealing technique for codes codifying for 13 or more amino acids. The minimization percentage decreased in the early and intermediate stages of genetic code evolution, an observation, as the authors remark, that cannot be explained easily under the hypothesis of the reduction of the deleterious effects of translation errors.

3 Generation of variant genetic codes

In the works mentioned, different possibilities were used to generate alternative codes. In the works of Haig and Hurst (1991) and of Freeland and Hurst (1998a), when hypothetical codes were generated, two restrictions were considered:

(1) The codon space (64 codons) was divided into 21 nonoverlapping sets of codons observed in the standard genetic code, each set comprising all codons specifying a particular amino acid in the standard code. Twenty sets correspond to the amino acids and one set to the 3 stop codons.

(2) Each alternative code is formed by randomly assigning each of the 20 amino acids to one of these sets. The three stop codons remain invariant in position for all the alternative codes. Moreover, these three codons are the same stop codons of the standard genetic code (UAA, UAG and UGA).

This conservative restriction, which maintains the pattern of synonymous coding found with the standard genetic code, controls, as indicated by Freeland

(2002), for possible biochemical restrictions on code variation and for the level of redundancy inherent in the canonical code (Freeland and Hurst, 1998b).

In addition, the authors, in a more restrictive generation of possible codes, divided the 20 synonymous blocks into four groups, each group comprising the synonymous codon sets described by Taylor and Coates (1989), which share a common base identity at the first codon position (UNN, CNN, ANN and GNN, N refers to any nucleotide). In this case, the codon assignments of synonymous blocks are allowed to vary within groups, but not between groups. These restrictions incorporate the general observation of biosynthetic relatedness which exposes that amino acids which share a biosynthetic pathway tend to also share the same nucleotide identity in the first base position of their corresponding codons.

We will use the first method to generate alternative codes, but without considering this last and more restrictive generation, as the results between the two alternatives are difficult to analyze. As the authors stated, "the biological significance of the small discrepancies between the two alternatives is uncertain" (Freeland and Hurst, 1998b). Moreover, we will use a freer evolution with one restriction: we only impose three codons for the stop signal. The aim of the introduction of this last possibility, also used by Di Giulio et al. (1994), is a comparison between the restrictive and non-restrictive hypothetical codes in terms of optimal values that can be obtained and in terms of evolution difficulty.

4 Genetic algorithm adapted to the problem

Evolutionary computing methods are global search methods based on a population of solutions to a problem. The individuals of the population encode in a genotypic representation those solutions. The methods are variations of a general process whereby each generation of individuals from a population is evaluated, and these individuals procreate according to their fitness. The fitness represents how well an individual resolves a problem such as a typical computational optimization problem, and it is associated with the level of adaptation in the natural world. A selection operator defines what individuals are selected to procreate and pass their genetic material to the next generation, through different operators such as the crossover operator. Selection methods have the common property of a higher probability of selection of an individual with higher fitness. Later on, the population undergoes mutation processes and a new population is somehow selected from the old one to continue with the process. These methods include Genetic Algorithms (GAs), evolutionary strategies, genetic and evolutionary programming and coevolution, although in this work we only used GAs (Goldberg, 1989).

4.1 Encoding

Each individual of the genetic population must encode a hypothetical code. In our solution, in the case of non-restrictive codes, each individual has 64 positions, which correspond to the 64 codons, and each position encodes the particular amino acid associated with the codon. In the case of restrictive codes, each individual has 20 positions, which correspond to the 20 codon

sets, and each position encodes the particular amino acid associated with the codon set. In both cases, a basic procedure ensures that the individuals of the initial population encode, at least in one position, the 20 amino acids. As in (Haig and Hurst, 1991) a fixed number of three codons are used for the stop label. The genetic operators require that a given individual should always encode the 21 labels.

4.2 Genetic operators

We used a mutation operator and a swap operator. A mutation changes the amino acid encoded in each of the 64 positions, by a mutation probability, to a different one. This operator is only applied with the unrestrictive codes. The mutation does not operate if the amino acid to mutate is the only one in the whole code. These mutations simulate the possible errors in the transcription process from DNA to RNA and in the translation process when incorrect transfer RNAs join a given codon of the messenger RNA. From our application point of view, it is the operator that varies the number of codons associated with a particular amino acid.

The other genetic operator is the swap operator, hardly ever used in GA applications, although it is appropriate for the present problem. The operator interchanges the contents of two genes, that is, once two genes are randomly selected, the amino acids codified by the two respective codons (or codon sets) are swapped. Figure 1 shows how these genetic operators work in the case of unrestrictive codes. The two operators guarantee that the 20 amino acids are always represented in the individuals. Other operators, such as the classical crossover operator, do not guarantee this important restriction.

Finally, as selection operator we used tournament selection. The operator selects the best individual in a window of randomly selected individuals from the population. Hence, the size of the window determines the required selective pressure. Moreover, we used elitism of the best individual; that is, this individual is kept in the next generation without changes.

4.3 Fitness function

We used as fitness function the measurement applied, for example, by Haig and Hurst (1991) and Freeland and Hurst (1998a) to quantify the relative efficiency of any given code. The measurement calculates the mean squared (MS) change in an amino acid property resulting from all possible changes to each base of all the codons within a given code. Any one change is calculated as the squared difference between the property value (polar requirement) of the amino acid coded for by the original codon and the value of the amino acid coded for by the new (mutated) codon. The changes from and to "stop" codons are ignored, while synonymous changes (the mutated codon encodes the same amino acid) are included in the calculation. Figure 2 summarizes the error calculation, when the first base of the codon UUU is mutated, taking into account the new values of the polar requirement of the new coded amino acids. The final error is an average of the effects of all the substitutions over the whole code.

Many other alternative types of weighting are imaginable, the best model relating chemical distance to code fitness being difficult to know, as commented by the previously cited authors. In the MS measurement we can consider the MS1, MS2 and MS3 values that correspond to all single-base substitutions

in the first, second and third codon positions, respectively, of all the codons in a given genetic code. The MS value (or any of the components) defines the fitness value of a given code and the evolutionary algorithm will try to minimize it.

5 Statistical vs. engineering analysis

Knight et al. (1999) indicate that two criteria can be used to assess if the genetic code is in some sense optimal. The first one is the "statistical approach" (Freeland et al., 2000a), applied by these authors (Haig and Hurst, 1991; Freeland and Hurst, 1998a; Knight et al., 1999), which compares the standard genetic code with many randomly generated alternative codes. Comparing the error values of the standard genetic code and alternative codes indicates, according to the authors, the role of selection. The main conclusion of the authors with this approach is that the genetic code conserves amino acid properties far better than expected from a random code.

The second one is the "engineering approach", which compares the standard genetic code with the best possible alternative. As mentioned previously, this approach is taken by Di Giulio (Di Giulio et al., 1994; Di Giulio, 2000), although it is also used in (Freeland et al., 2000b). This approach tends to indicate that the genetic code is still far from optimal. They used the p.d.m. measurement, which determines code optimality on a linear scale. The measurement is calculated as the percentage in which the canonical genetic code is in relation to the randomized mean code and the most optimized code, as defined in Section 2. We also used a measurement called "improvement", related with percentage minimization, but defined as the improvement obtained

in fitness by the best evolved individual, measured as the final fitness with respect to the corresponding MS value of the standard genetic code.

6 Equal transition/transversion bias

In this first analysis, we tested the capability and difficulty of simulated evolution to find better adapted codes than the canonical code, taking into account the MS value previously explained, and with an equal probability of mutations in the three bases and the two types of mutations. As we commented before, Haig and Hurst (1991) only found two alternative codes with lower MS than the standard genetic code in a set of 10,000 randomly generated codes, and Freeland and Hurst (1998a) refined the probability with a larger sample of 1,000,000, where they found only 114 better codes. Both results were obtained using the codes defined with the restrictions mentioned previously and using the statistical approach.

6.1 GA parameters

We tested the implemented GA, searching for alternative codes, with the two definitions of codes presented in Section 3. Figure 3 shows the evolution of the MS across 150 generations of the genetic algorithm. The quality of the best individual and the average quality of the population are the result of an average of 10 evolutions with different initial populations. The population size was 1,000 individuals for the different tests. Larger populations do not improve the results, as we will discuss later. The other evolutionary parameters were a mutation probability of 0.01 and a swap probability of 0.5. The mutation

operator is only used in the non-restrictive model, as the restrictive model only needs the swap operator to interchange the 20 amino acids among the 20 sets of codons.

For non-restrictive codes, the best (minimum) MS found in one of the evolutions was 1.784, whereas the minimum was around 3.48 for the restrictive codes. These values can be compared with the best value found by Freeland and Hurst (~ 4.7) and the value of the standard genetic code (5.19).

There is not a general rule to set the parameter values of the different genetic operators, since the results with a particular value set depend on the application. As Mitchell (1997) indicates the parameters typically interact with one another nonlinearly, so they cannot be optimized one at a time. There is a great deal of discussion on parameter settings and approaches to parameter adaptation in the evolutionary computation literature, but there are not conclusive results on what is the best.

We selected the rates mentioned after experimentation with incremental values in the intervals [0.001, 0.1] for the mutation operator and [0.01, 0.8] for the swap operator. Mutation rates around the highest tested value (0.1) imply a high level of exploration and give worse results. For instance, with a mutation rate of 0.1 the best value was 2.593 in the non-restrictive model, whereas a value of 0.001 showed slightly worse values than the ones obtained with a mutation rate of 0.01, and with very slow quality evolutions. These results were obtained when the values of the other parameters were set to their final selected values. The swap operator shows less sensitivity to the values of the tested interval, the worst results being those with low values (~ 0.01), as expected since this operator is necessary to swap the amino acids in both

code models.

We used tournament selection with a window size of 3% of the population. The size determines the number of randomly selected individuals from the population, from which the best in the window is selected. We experimented with incremental window sizes in the interval [1%, 10%], the best result being that obtained with the selected window size. This size imposes a low selective pressure, which makes the evolutions less dependent on the appearance of good sub-optimal solutions. With low selective pressure values the evolutions are quite similar, and they begin to worsen with a value of around 8% in the window size. For instance, with a window size of 10%, the best value was only 2.757 in the unrestrictive model, with a population of 1,000 individuals and the mutation and swap rates selected. Finally, we used the commented elitism of the best individual through generations.

A population of 1,000 individuals provides a good trade-off between efficiency (computing time) and good results. For instance, in the case of non-restrictive codes, with a population of 1,000 individuals, the mean value of the best codes obtained in the ten different tests (different initial populations) was 1.853 with a standard deviation of 0.025. With a smaller population, 100 individuals, the mean value of the best codes obtained in ten different executions of the GA was 2.443, with a greater standard deviation of 0.149. With larger populations, over 1,000 individuals, the best values improved very little. For instance, with 10,000 individuals, the mean value of the best codes obtained was 1.849, with a similar standard deviation (0.027) to that obtained for 1,000 individuals. The same conclusions can be obtained with the restrictive codes. With a population of 1,000 individuals the mean value of the best codes was 3.508, with a standard deviation of 0.017. Again, the low standard deviations, together

with the fact that there is not practically any improvement in the best values obtained with larger populations, justifies the use of a population of 1,000 individuals.

6.2 Percentage distance minimization results

The percentage distance minimization, using the best values obtained by the GA, was 67% with the non-restrictive codes, while 71% with the restrictive codes, a higher value since the restrictive model also restricts the optimum value. Di Giulio et al. (1994) report p.d.m. values of 72.7% in the case of codes with only amino acid permutations in the 20 sets of codons and 64% with codes with the same degeneracy of the genetic code, although using the absolute value of the differences of the polarity values instead of the squared differences and not considering synonymous changes. They explain that "this percentage decreases as the number of codes considered increases", which is in accordance with our results. The same can be inferred with the improvement measurement, as its value was 33% with the restrictive codes model and 66% in the non-restrictive case.

The MS values of each sample of codes in each generation form a probability distribution against which the standard genetic code MS value may be compared. The upper part of Figure 3 also shows the histograms of the initial population and at the end of the evolution with the non-restrictive codes and in one of the tests. In the histograms, the x-axis gives a particular range of categories of MS values whereas the y-axis indicates the number of individuals with an MS in that category. The histogram of the initial population presents a similar distribution as the ones of Freeland and Hurst (1998a), as

the population is random. A better code (better than the canonical code) was not found by chance in that initial population. At the end of the evolutionary process, the situation changed radically, where all the individuals showed a better MS than the one of the standard genetic code.

An analysis of the amino acids encoded in the codons of a variety of the best non-restrictive codes indicates two considerations: there is a great variety of better codes, with very different assignments; and there are not clear coincidences between the best codes with the assignments of the standard code. When evolution works without restrictions (except for the stop signal), five amino acids appear in most of the codons: alanine (Ala), proline (Pro), serine (Ser), threonine (Thr), and also asparagine (Asn) with a lower number, while the rest are codified by only one or two codons. These amino acids are codified in the standard genetic code with six or four codons (except Asn), although the other two amino acids codified by six codons in the standard code (Arg and Leu) are codified by only one codon in most of the best evolved codes. There is no biological reason for these assignments, as we are working with non-restrictive codes. The amino acids that appear in most of the codons have an intermediate value of polar requirement (between 6.6 in Pro and 7.5 in Ser). This helps minimize the error, when the majority of changes due to mutations are among the intermediate values.

If restrictions are taken into account, there are few coincidences in the amino acid assignments in the 20 sets of codons between the standard code and the best codes obtained, such as the one shown in Figure 4. The assignments of the amino acids to the codons in the standard genetic code as well as in the best code obtained with restrictions are shown. The polar requirement values are also shown, associated with their amino acids of the best evolved

code. This last result coincides with the observations of Freeland and Hurst (1998a), although their better codes correspond with those obtained by chance in numerous samples. Nevertheless, the best evolved codes, like the one in Figure 4, have the same property as that of the standard genetic code: amino acids that share the two first bases have similar values of polar requirement.

6.3 Other error measurements and amino acid properties

Finally, we also applied other measurements, besides MS. An evolution with a linear fitness function, using the absolute value of the difference of the polar requirement property instead of the mean squared difference, gave similar conclusions. The p.d.m. in the non-restrictive case was 65%, very similar value (64%) to that obtained by Di Giulio et al. (1994), using the same linear fitness definition. Nevertheless, the best codes obtained have few similarities when the two fitness functions were used: only amino acids Pro and Thr were codified by more than three codons with the two fitness definitions.

Moreover, when different amino acid properties were used, the results indicated that polar requirement is the property that provides the most significant evidence of error minimization. These properties were hydropathy index, isoelectric point and molecular volume, the same as those used by Haig and Hurst (1991). For instance, in the non-restrictive case, the p.d.m. was 53% with the hydropathy property, 42% with molecular volume and 23% when the isoelectric point was used in the MS calculation of the fitness. The GA parameters were the same as those of the previous experiments. The standard genetic code even presents a good level of optimization for another property of hydrophobicity (hydropathy), but a poor level for the isoelectric point prop-

erty.

We have also used the alternative distance matrix used by Higgs (2009) to measure the distance between amino acids. The matrix is derived from observations of rates of substitutions in proteins, and uses different weights assigned to 9 different properties (including polar requirement). The mean distance between pairs of non-identical amino acids is 100. In this case, the numerator of the fitness function uses directly the distances of the matrix when there is a change of the codified amino acid after a change in each base. The p.d.m. was 60.4% with the non-restrictive codes, a lower value with respect to the use of only polar requirement (65%), but better with respect to the other properties commented before. The author indicates that the distance matrix is "a more realistic amino acid distance measurement that is derived from maximum likelihood fitting of real protein sequence data" (Higgs, 2009), resulting in a level of optimality close to the value when polar requirement was only used.

7 Introduction of a transition/transversion bias and evolution of the individual bases

In nature, transition errors tend to occur more frequently than transversion mutations, because of the unequal chemical similarity of the four nucleotides to one another (Freeland, 2002). A transition error is the substitution of a purine base (A, G) into another purine, or a pyrimidine (C, U/T) into another pyrimidine (i.e., $C \leftrightarrow T$ and $A \leftrightarrow G$), whereas a transversion interchanges pyrimidines and purines (i.e., $C, U \leftrightarrow A, G$).

We can use the MS values for each code calculated at different weightings

of transition/transversion bias (WMS), and turning the MS measurements into WMS measurements, as previously done by Freeland and Hurst (1998a). Thereby, at a weighting of 1, all possible mutations are equal when calculating the MS values for each position of each codon. And, for example, at a weighting of 2, the differences in amino acid attribute resulting from transition errors are weighted twice as heavily as those resulting from transversion errors.

Freeland and Hurst (1998a) investigated the effect of weighting the two types of mutations differently. The main conclusion of their work was the dramatic effect of transition/transversion bias on the relative efficiency of the second codon base; that is, the number of better alternative codes (regarding WMS2) decreases almost six fold as the transition/transversion bias increases from 1 to 5. Nevertheless, even at higher transition/transversion bias the second base remains an order of magnitude less relatively efficient than the first and third bases.

Another aspect was that the individual bases combine in such a way that the overall relative efficiency of the standard code (measured by WMS) increases with increasing transition/transversion bias up to a bias of approximately 3. Moreover, this effect is clearer with WMS1. As commented by the authors, this observation coincides quite well with typical empirical data, which reveal general transition/transversion biases between 1.7 and 5. In addition, as the causes of this natural bias are physiochemical, basically size and shape of purines and pyrimidines, "it seems reasonable to suppose that the biases observed now were present to a similar extent during the early evolution of life" (Freeland and Hurst, 1998a).

We used the GA to determine the difficulty of evolution in obtaining better

codes by simulated evolution taking into account the three bases separately. In addition, we used a weight of 1 and a weight of 3, with the intention of testing the conclusions previously mentioned by Freeland and Hurst (1998a), where the first and third bases of the canonical code showed a better adaptability level.

Figure 5 shows the evolution of the three individual components of the MS. This indicates, for example, that the evolution with MS1 uses a fitness that only considers errors in that first base. The evolutions in that Figure were with a population of 1,000 individuals and with the restrictions mentioned in the generation of alternative codes. The top graphs correspond to a weight of 1 and the bottom graphs were obtained with a weight of 3. The graphs include the improvement in fitness obtained by the best evolved individual, measured as the final fitness with respect to the corresponding MS value of the standard genetic code. In addition, the graphs show the percentage distance minimization obtained in each case, where the value of the best code obtained by the GA is used for the calculation of the percentage.

If we consider Freeland and Hurst's analysis of random individuals of the first generation, the second base is clearly the worst adapted in the standard genetic code, since in the first generation there are random codes with better MS2. The p.d.m. is 19%, which indicates that the second base in the canonical code is quite far from the possible best value. The first base is better adapted in the standard code (p.d.m. 62%) with lower number of random individuals with better MS1 than the canonical MS1 value (4.88). This can be guessed by the relation, in the first generation, of the fitness of the best individual and the mean value of the population with respect to the value of the canonical code. Finally, the third base is the best adapted in the standard code (p.d.m.

96%), as there are not random codes with lower values than the MS3 of the standard genetic code (0.14). Table 1 shows the mean values of the randomly generated populations, together with the best values obtained, that were used in the calculation of the measurements in Figure 5.

The evolution of the qualities in the evolutionary processes gives us similar conclusions. In the case where the genetic algorithm takes into account only the optimization of the second base, the average quality of the genetic population, in less than five generations, obtains a better value than the MS2 value of the canonical code. This is obtained even with low selective pressure (tournament size of 3% of the population). In simulated evolution, it is more difficult for the average population quality to have an MS1 that is better than the value of the genetic code. Finally, with the evolution of the third base, the genetic algorithm requires a few generations to obtain better individuals, and it is not able to obtain an average quality with a lower value than the MS3 value of the standard genetic code.

The values of the improvement measurement indicate, especially in the MS2 case, that we must consider in the analysis the best values that can be obtained. The best MS2 value the GA can reach is worse than the optimal value of MS1 and, especially, MS3. The imposed restrictions in the definition of alternative codes are the reason for the different values obtained in the MS individual components. In other words, the second base is clearly the worst adapted, but the value that could be reached is also worse than the value of the other two cases.

The bottom graphs in Figure 5 are the same evolutionary processes with a bias of 3. No appreciable differences exist in the curves of quality evolution,

except that the p.d.m. increases in the three cases. In the second base, the improvement of p.d.m. is quite high, small with MS1 and lower with MS3. These improvements are in accordance with the results previously mentioned of Freeland and Hurst (1998a) with random populations.

7.1 Codes with two-base codons

We studied the level of optimality of the first two bases within the coevolution hypothesis. This hypothesis maintains that early on in the genetic code few precursor amino acids were codified. As the other amino acids arose biosynthetically from these precursors, part or all of the codon domain of the precursor amino acids was passed to the product amino acids (Wong, 1975).

Yockey (2005) has considered an initial scenario with codons with only two letters. However, other authors reject this possibility because, as Higgs indicates, the evolution of a code with codons of length two it is unlikely "because the evolution of such a code to a triplet code would require the complete reinvention of the mechanism by which the ribosome shifts along the mRNA and the complete rewriting of all the gene sequences that were written in the two-base or one-base code" (Higgs, 2009). From the point of view of the evolutionary simulated search there is no significant difference if we use the possibility of codons with two bases or codons with three letters where the third base is completely degenerate. The only difference in the fitness function between the two possibilities is because the synonymous changes of the degenerate third base, changes that are considered in the calculation.

Figure 6 shows the evolution of MS1 and MS2 when codes with two bases are

considered. The possible original code proposed by Yockey (2005) is used as reference (Figure 7). This code codifies 14 amino acids and uses two codons for the stop signal. Again, evolution shows that the first base of the possible precursor code is clearly better adapted than the second one, as also indicated by the values of the improvement measurement. Evolution needs several generations to reach the situation in which the average fitness of the population improves the MS1 value of the reference code, whereas in the first generation the average fitness overcomes the MS2 value. The p.d.m. value in the second base cannot be measured (the value would be negative), as the mean value of the random distribution is even better than the MS2 value of the reference code. When simulated evolution works with the combined MS as fitness (without considering the possible degenerate third base), the p.d.m. was 55%. This value indicates that when the code expanded to the current code with three-letter codons, there was a clear improvement in the adaptability level, as the MS value was 71% in that case (Section 6).

This increase in the p.d.m. value in the transition from the code with two-base codons to the code with three-base codons is compatible with the results of the work of Di Giulio and Medugno (1999). As mentioned before, these authors considered the evolution of the genetic code under the coevolution theory, although without an explicit consideration of a code with two-base codons. The authors tested 10 evolutionary stages through which genetic code organization might have passed prior to reaching its current form. The minimization percentage decreased in the early and intermediate stages of genetic code evolution. However, the authors found that the "real increase in minimization percentages seems to have taken place only in the final four stages of code evolution, and therefore, the physicochemical hypothesis is unable to

explain code evolution in the early and intermediate stages". The final four stages begin after the transition from a stage of 13 amino acids (plus 3 labels to the stop signal) to a stage of 15 amino acids.

The stages described in (Di Giulio and Medugno, 1999) and in (Yockey, 2005) basically differ in the presence of the Asp amino acid, considered in the first stage in (Di Giulio and Medugno, 1999), seen as one of the first amino acids to be assigned codons in Wong (1975) and not in the previous code considered in (Yockey, 2005). This last author arguments that eight of the amino acids considered in his set of original amino acids have "cylinder" codons (any third base considered in the extension of the code codifies the same amino acid), as well as that most of the remaining amino acids are not present in ferredoxins, very ancient proteins present in ancestral organisms soon after the origin of life. Figure 7 shows the best code obtained by the GA and the reference code consisting of two-letter codons considered in (Yockey, 2005).

8 Errors as a function of the base position in the codon

The previous experiments assumed that mistakes are equally likely to be made at any of the three codon positions. Freeland and Hurst (1998a) indicate that this assumption is correct when we consider point mutations in the DNA sequence and that these are accurately translated via mRNA into an erroneous amino acid. However, the assumption must be reconsidered if we take into account mistranslation of mRNA. The translation machinery acts upon mRNA reading bases in triplets (codons), and that translation accuracy varies according to the base position of the codon. The rules from (Freeland and Hurst, 1998a), used to consider the empirical data, were applied and are summarized

as:

A Mistranslation of the second base is much less frequent than the other two positions, and mistranslation of the first base is less frequent than the third base position.

B The mistranslations at the second base appear to be almost-exclusively transitional in nature.

C At the first base, mistranslations appear to be fairly heavily biased toward transitional errors.

D At the third codon position, there is very little transition bias.

The MS calculation can be modified to take into account these rules, weighting the errors accordingly (tMS). The left part of Figure 8 shows the quantification of the mistranslation used in (Freeland and Hurst, 1998a) as well as in our work to weight the relative efficiency of the three bases in the tMS calculation. Freeland and Hurst (1998a), with their 1 million randomly generated codes, found only 1 with a lower tMS value. Now the probability of a code as efficient as or more efficient than the standard genetic code evolving by chance falls until 10^{-6}.

The right part of Figure 8 shows the evolution of tMS with the two cases previously considered, with and without the introduction of Freeland and Hurst's restrictions. The p.d.m. values were 85% and 84% with the non-restrictive and restrictive genetic codes models, respectively. The values are similar, but in any case the two values are better with respect to the same p.d.m. values when the MS fitness was used (p.d.m. values of 67% and 71%, Section 6). Therefore, the p.d.m. measurement, that considers the distance between the mean value of random codes and the best possible value, again indicates that the

canonical genetic code is better adapted with the considerations incorporated in the tMS calculation.

In addition, the improvement measurement gives extra information. The similar values of p.d.m. with the tMS calculation are differentiated with that measurement. In both cases the improvement in fitness quality was similar with respect to an equal probability of errors in the three bases: the improvement measurement value (decrease in fitness of the best individual with respect to the canonical code) in tMS is 63% without restrictions and 37% with restrictions. This improvement measurement considers the best value that could be obtained with the rules that define the alternative codes. The value of 37% with restrictive codes indicates that the imposed restrictions imply a worse optimum value. Moreover, with such restrictions the evolution with restrictive codes improves the tMS value of the canonical code with fewer generations (regarding the best and average values) with respect to the evolution with unrestrictive codes. Nevertheless, the best final value corresponds to the unrestrictive codes.

9 Discussion and conclusions

Yockey (2005) criticized the idea of evolution of the genetic code in the sense of minimization of the effects of mutations. As argument he stated that the $1.4 \cdot 10^{70}$ possible codes could not have been tested in the $8 \cdot 10^8$ years between the event of Earth formation ($4.6 \cdot 10^9$ years ago) and the origin of life in it ($3.8 \cdot 10^9$ years ago). We do not consider this statement as a correct argument, as neither natural evolution has to check all possible codes to minimize the deleterious effects of mutations, nor simulated evolution tests all the possibil-

ities in the search space. Thanks to the selection of best individuals, natural and simulated evolution concentrates only on the most promising areas of the search landscape, sampling it to search for optimal solutions in a computational problem or discovering the genetic material that provides better adaptation to an organism.

Nevertheless, the fact that the GA easily finds better codes than the standard genetic code does not imply that there was not any adaptive evolution of the genetic code. Two considerations must be taken in this regard. Firstly, we only have considered one (important) property. Knight et al. (2004) stated that "the average effect of amino acid changes in proteins is unlikely to be perfectly recaptured by a single linear scale of physical properties". Secondly, we agree with the authors that the code could be trapped in a local, rather than global, optimum. The authors continued by saying that "The fact that the code is not the best of all possible codes on a particular hydrophobicity scale does not mean that it has not evolved to minimize changes in hydrophobicity under point misreading".

The two approaches for measuring the adaptability level of the genetic code were valid. Within the statistical approach, the conclusions inferred are similar to those obtained by the GA. When a bias in the transition/transversion mutations is included and when the errors as a function of the base position in the codons are considered, GA evolution shows a better adaptability level of the canonical genetic code. Nevertheless, does simulated evolution provide more information than the statistical approach or the engineering approach? The answer is that the GA gives us more information about the difficulty of evolution of codes.

For instance, if we compare the evolution of the quality of the best individual in the case of an equal transition/transversion bias and in the case of inclusion of errors as a function of base position (same evolutionary parameters, Figures 3 and 8), we can observe the greater difficulty of simulated evolution with tMS to discover an individual that overcomes the value of the standard genetic code, in the two models considered, with and without the conservative restrictions. In the second case (tMS), practically twice as many generations are needed to obtain better values, which indicates the better adaptability level of the standard genetic code under these assumptions. Regarding the adaptability level of the individual bases, the evolution of the average quality of the genetic population reflects the different adaptability level of the three bases. For example, to obtain an average quality with a better value than the corresponding one of the standard genetic code, the required number of generations in the first base is more than double that required in the second base (Figure 5). The third base shows a clear adaptability level as the GA is not able for the average quality of the population to overcome the MS3 value of the canonical code.

Within the engineering approach, the use of the percentage distance minimization and improvement measurements gives a clear view of the adaptability level in relation to the best codes that could be reached. In the case of restrictive codes, their own definition imposes a better level of adaptability of the third base, whereas the second one is the worst adapted.

Within the coevolution theory, we have considered an initial possible scenario with codons of two bases and 14 codified amino acids and the final scenario of the current canonical code. Although a more detailed analysis can incorporate intermediate scenarios in the simulation, our analysis shows that the

third base clearly improves the adaptability of the current genetic code, which is in agreement with the error minimization hypothesis. Nevertheless, our simulations cannot establish if that minimization is mainly due to the selective pressure to minimize the effects of the errors that the physicochemical theory claims or to the main factor of the coevolution theory, that is, "the mechanism which concedes codons from the precursor amino acids to the product amino acids is the primary factor determining the evolutionary structuring of the genetic code" (Di Giulio and Medugno, 1999). If the codons of the product amino acids are assigned to physicochemical similar amino acids, it implies also an error minimization.

Although the best possible values for the p.d.m. measurement in the engineering approach can be calculated with local search procedures, simulated evolution tries to mimic the process of natural selection with the competition among the individuals of the genetic population, operating in that case over the genetic code. Hence, the simulation of evolution can provide a more realistic view of the adaptability of the genetic code, within the physicochemical theory, in the fight for survival of the codes in the evolution in the RNA World.

As final conclusion, we remark that the simulated evolution of codes indicates that the canonical code is better adapted (in terms of p.d.m.) when we consider the restrictive codes. This could be in favor of the coevolution theory, since the restrictive codes incorporate the block structure of the canonical code and because the freer evolution of unrestrictive codes could obtain a better level of adaptability considering the physicochemical theory. As we have mentioned, both approaches (statistical and engineering) to quantify the canonical code's susceptibility to error are valid to us. We do not agree with Freeland and coworkers (Freeland and Hurst, 1998a; Freeland et al., 2000b) when they favor

the statistical approach because, as they emphasize, the approach considers that the possible codes form a Gaussian distribution of error values. However, even when beginning with those Gaussian distributions of random codes in the initial genetic populations, the GA simulations indicate that is easy to improve the adaptability level of the standard genetic code, with low selective pressure and in few generations. Our results are in agreement with the work of Novozhilov et al. (2007) when they used a local and greedy search to find the shortest evolutionary trajectory from a given starting code to its local minimum of the error cost function. The authors concluded that "the standard genetic code appears to be a point on an evolutionary trajectory from a random point (code) about half the way to the summit of the local peak. Moreover, this peak appears to be rather mediocre, with a huge number of taller peaks existing in the landscape". In our simulations with the evolutionary computing methodology, the situation of the canonical code in the graphs of evolution indicates that the canonical code is adapted, but it is clearly in a local minimum.

10 Acknowledgements

This paper has been funded by the Ministry of Science and Innovation of Spain through project TIN2007-64330.

References

Ardell, D.H., Sella, G., 2002. No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. Phil. Trans. Royal Soc.

London B, Biol. Sci., Vol. 357, pp. 1625-1642.

Crick, F., 1968. The origin of the genetic code. Journal of Theoretical Biology 38, pp. 367–379.

Di Giulio, M., 1989. The extension reached by the minimization of the polarity distances during the evolution of the genetic code. Journal of Molecular Evolution 29, pp. 288-293.

Di Giulio, M., 2000. The origin of the genetic code. Trends in Biochemical Sciences 25(2), pp. 44.

Di Giulio, M., 2001. The origins of the genetic code cannot be studied using measurements based on the PAM matrix because this matrix reflects the code itself, making any such analysis tautologous. Journal of Theoretical Biology 208(2), pp. 141-144.

Di Giulio, M., 2005. The Origin of the genetic code: theories and their relationship, a review. Biosystems 80, pp. 175–184.

Di Giulio, M., Capobianco, M.R., Medugno, M., 1994. On the optimization of the physicochemical distances between amino acids in the evolution of the genetic code. Journal of Theoretical Biology 168, pp. 43–51.

Di Giulio, M., Medugno, M., 1999. Physicochemical optimization in the genetic code origin as the number of codified amino acids increases. Journal of Molecular Evolution 49(1), pp. 1–10.

Freeland, S.J., 2002. The Darwinian genetic code: an adaptation for adapting?. Genetic Programming and Evolvable Machines, Kluwer Academic Publishers 3, pp. 113–127.

Freeland, S.J., Hurst, L.D., 1998a. The genetic code is one in a million. Journal of Molecular Evolution 47(3), pp. 238–248.

Freeland, S.J., Hurst, L.D., 1998b. Load minimization of the genetic code: History does not explain the pattern. Proceedings of the The Royal Society

265, pp. 2111–2119.

Freeland, S.J., Knight, R.D., Landweber, L.F., 2000a. Measuring adaptation within the genetic code. Trends in Biochemical Sciences 25(2), pp. 44–45.

Freeland, S.J., Knight, R.D., Landweber, L.F., Hurst, L.D., 2000b. Early fixation of an optimal genetic code. Mol. Biol. Evol. 17(4), pp. 511–518.

Gilis, D., Massar, S., Cerf, N.J. and Rooman, M., 2001. Optimality of the genetic code with respect to protein stability and amino-acid frequencies. Genome Biology 2(11).

Goldberg, D.E. 1989. Genetic algorithms in search, optimization and machine learning. Addison-Wesley Longman Publishing Co., Inc.1, Boston, MA, USA.

Goldberg, A.L., Wittes, R.E. 1966. Genetic code: aspects of organization. Science 153. pp. 420-424

Goodarzi, H., Najafabadi, H.S., Nejad, H.A., Torabi, N., 2006. The impact of including tRNA content on the optimality of the genetic code. Bulletin of Mathematical Biology 67(6), pp. 1355–1368.

Haig, D., Hurst, L.D, 1991. A Quantitative measure of error minimization in the genetic code. Journal of Molecular Evolution 33, pp. 412–417.

Higgs, P.G., 2009. A four-column theory for the origin of the genetic code: tracing the evolutionary pathways that gave rise to an optimized code. Biology Direct, pp. 4–16.

Jestin, J-L., Kempf, A. 1997. Chain termination codons and polymerase-induced frameshift mutations, FEBS Letters 419, pp. 153-156.

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, pp. 241–247.

Knight, R.D., Freeland, S.J., Landweber, L.F., 2004. Adaptive evolution of the

genetic code. The Genetic Code and the Origin of Life, Vol. 80, Lluís Ribas de Pouplana (Ed.), Kluwer Academic/Plenum Publishers, pp. 175–184.

Mitchell, M., 1997. An Introduction to genetic algorithms, MIT Press.

Novozhilov, A.S., Wolf, Y.I., Koonin, E.V. 2007. Evolution of the genetic code: partial optimization of a random code for robustness to translation error in a rugged fitness landscape. Biology Direct, pp. 2–24.

Ronneberg, T.A., Freeland, S.J., Landweber, L.F., 2001. Genview and Gencode: a pair of programs to test theories of genetic code evolution. Bioinformatics 17(3), pp. 280–281.

Schönauer, S., Clote, P., 1997. How optimal is the genetic code?. Computer Science and Biology, German Conference on Bioinformatics (GCB 97), Frishman, D. and Mewes, H. (Eds.), pp. 65–67.

Taylor, F.J.R., Coates, D., 1989. The code within the codons. BioSystems 22, pp. 177–187.

Torabi, N., Goodarzi, H., Najafabadi, H.S., 2007. The case for an error minimizing set of coding amino acids. Journal of Theoretical Biology 244(4), pp. 737-744.

Woese, C.R., 1965. On the evolution of the genetic code. Proc. Natl. Acad. Sci. USA, Vol. 54, pp. 1546–1552.

Wong, J.T., 1975. A Co-evolution theory of the genetic code. Proc. Natl. Acad. Sci. USA 72, pp. 1909–1912.

Yockey, H.P., 2005. Information theory, evolution, and the origin of life. Cambridge University Press, NY.

Table 1
Mean and best values in the MS1, MS2 and MS3 evolutions of Figure 5.

		MS1	MS2	MS3
Mean value	W=1	12.027	12.680	3.522
	W=3	12.027	12.455	2.340
Best value	W=1	0.536	1.586	0.013
	W=3	0.421	1.586	0.009

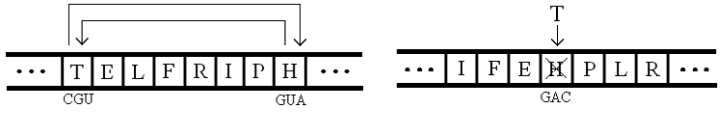

Fig. 1. Swap operator (left) and mutation operator (right).

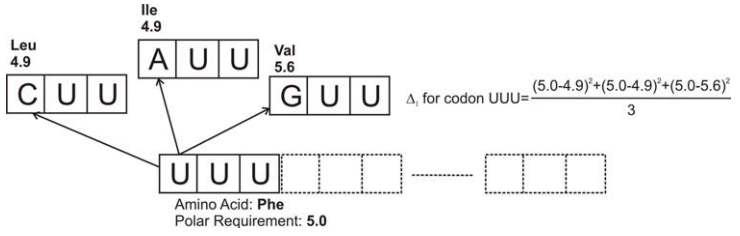

Fig. 2. Calculation of the mean squared (MS) error value of a code to define the fitness function.

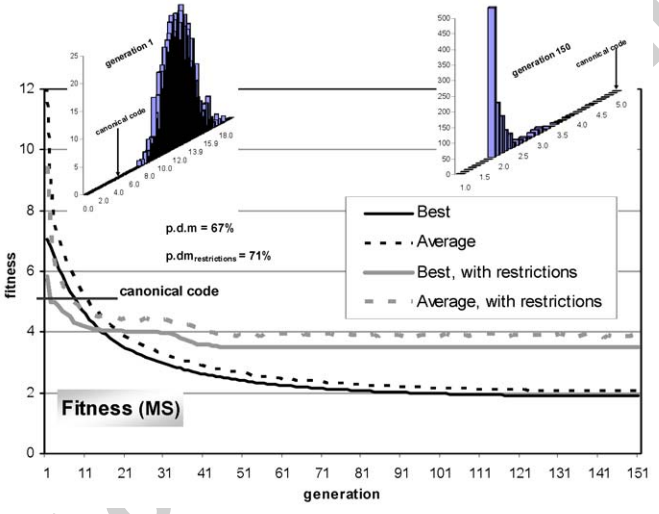

Fig. 3. Evolution of the MS in codes without restrictions (except the number of stop codons) and in codes with restrictions. The histograms shown on the top of this Figure correspond to the population at the beginning and end of the evolutionary process in the non-restrictive case. In these histograms, the x-axis gives a particular range of categories of MS values whereas the y-axis indicates the number of individuals with an MS in that category. The arrows in the histograms indicate the category into which the MS value of the canonical code falls.

Codon set	Best code		Canonical code
CGA CGC CGG CGU AGA AGG	Gly	7.9	Arg
CUA CUC CUG CUU UUA UUG	Gln	8.6	Leu
UCA UCC UCG UCU AGC AGU	Ser	7.5	Ser
ACA ACC ACG ACU	Pro	6.6	Thr
CCA CCC CCG CCU	Thr	6.6	Pro
GCA GCC GCG GCU	Val	5.6	Ala
GGA GGC GGG GGU	Ala	7.0	Gly
GUA GUC GUG GUU	His	8.4	Val
AAA AAG	Met	5.3	Lys
AAC AAU	Trp	5.2	Asn
CAA CAG	Leu	4.9	Gln
CAC CAU	Phe	5.0	His
GAA GAG	Cys	4.8	Glu
GAC GAU	Ile	4.9	Asp
UAC UAU	Tyr	5.4	Tyr
UGC UGU	Lys	10.1	Cys
UUC UUU	Asn	10.0	Phe
AUA AUC AUU	Arg	9.1	Ile
AUG	Glu	12.5	Met
UGG	Asp	13.0	Trp
UAA UAG UGA	Stop		Stop

Fig. 4. Best code obtained with restrictive codes. The Table shows the 21 sets of codons considered in the restrictive model and the encoded amino acids in the best evolved code, together with their polar requirement values used in the fitness calculation. The last column shows the associations in the standard code.

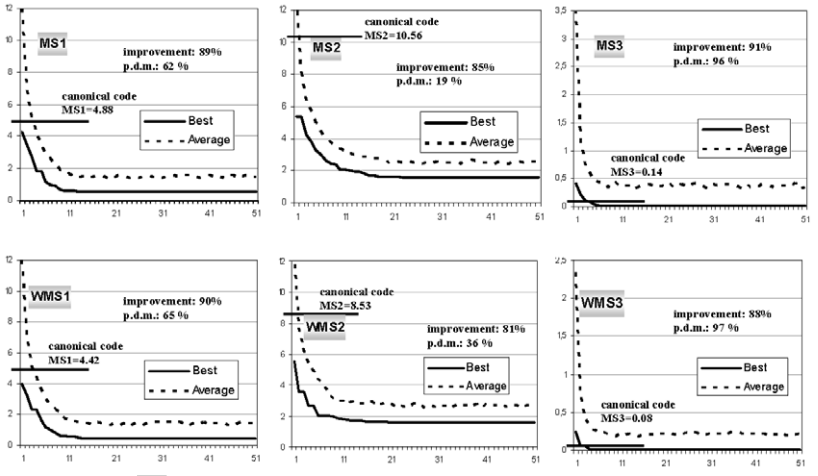

Fig. 5. Evolution of the individual components of MS with restrictive codes, with equal transition/transversion bias (top graphs) and with a bias of 3 (bottom graphs). In the graphs the y-axis indicates the fitness value of the best individual or the average quality of the genetic population, whereas the x-axis indicates the GA generation. MS1, MS2 and MS3 correspond to the mean squared errors of the individual bases without a bias in the transition/transversion mutations (W=1), and WMS1, WMS2 and WMS3 are the errors with a transition/transversion bias of 3 (W=3). The values indicated with a horizontal line are the values of the corresponding MS error of the canonical code.

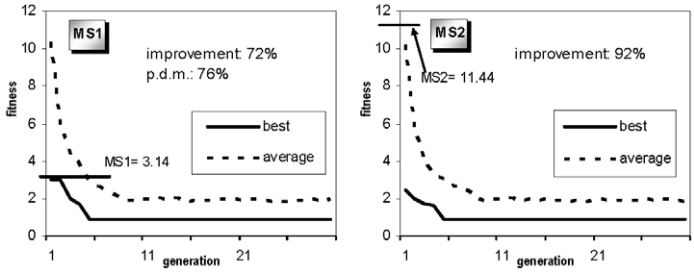

Fig. 6. Evolution of MS1 and MS2 in codes with two-base codons. MS1 and MS2 correspond to the mean squared errors of the two individual bases. The values indicated with a horizontal line are the values of the corresponding MS error of a supposed canonical code taken as reference (Yockey, 2005)

Codon set	Best code		Yockey's code
GG	Ala	7.0	Gly
CC	Arg	9.1	Pro
CU	Glu	12.5	Leu
CG	Gln	8.6	Arg
AC	Thr	6.6	Thr
GU	Asn	10.0	Val
GC	Ser	7.5	Ala
UC	Val	5.6	Ser
UU	Stop		Phe
UG	Ile	4.9	Cys
CA	Gly	7.9	Gln
AA	Leu	4.9	Asn
GA	Pro	6.6	Glu
AU	Stop		Ile
UA	Cys	4.8	Stop
AG	Phe	5.0	Stop

Fig. 7. Best evolved precursor code and first extension code proposed by Yockey (2005). From left to right, the columns indicate the 16 possible codons with two bases, the amino acids (plus the stop signal) codified with those codons in the best evolved code, the polar requirement value of those amino acids of the best code, and the amino acids codified in the code proposed by Yockey (2005)

Combined Weighting	Transitions	Transversions
First Base	1	0.5
Second Base	0.5	0.1
Third Base	1	1

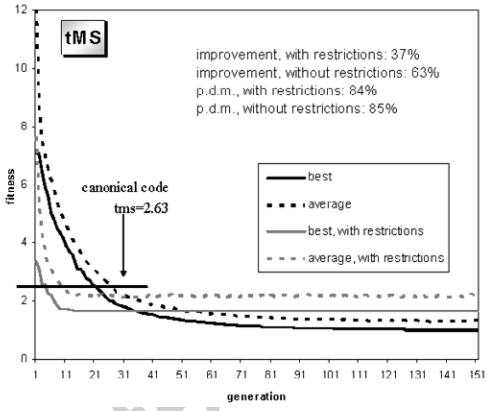

Fig. 8. Evolution of tMS (the error is a function of the base position in the codon), with and without restrictions in the evolved codes, together with the obtained quality improvements. The value indicated with a horizontal line is the value of the tMS error of the canonical code. The table at the left shows the quantification of mistranslation used to weight the relative efficiency of the three bases in the tMS calculation.