On the Use of Fitness Sharing in Studying the Genetic Code Optimality

José Santos
Department of Computer Science,
University of A Coruña, A Coruña, Spain
Email: jose.santos@udc.es

Ángel Monteagudo
Department of Computer Science,
University of A Coruña, A Coruña, Spain
Email: angel.minsua@udc.es

***Abstract*—Since the canonical genetic code is not universal several theories arose to explain the evolution to its present form. Different computational methods were applied to analyze the optimality level of the canonical code organization, including our previous works using evolutionary computing in the problem. We discuss here the possibilities that the use of the classical fitness sharing technique provides for obtaining knowledge about the fitness landscape involved in the optimization of the genetic code.**

I. Introduction

The canonical genetic code (CGC), present in most superior organisms, establishes the association between codons and amino acids. In the canonical code there are 64 codons of three bases that encode 20 amino acids. Thus, the genetic code is redundant, since several codons codify the same amino acid. However, the canonical code is not universal, since there are other different associations between codons and amino acids. Mitochondrial DNA is an exception example. The exceptions show that the code could evolve in its origin, and it is an open question how the code evolved to the present form (the CGC).

Since other codes could appear, several models of hypothetical genetic codes were considered, with different associations between codons and amino acids. Considering only the present form of a genetic code with codons of three bases, if the codon set of the CGC is maintained, allowing only swaps of amino acids between the 20 codon sets, the possible codes are $20!$ $(2.43 \cdot 10^{18})$. Without that restriction, in the sense that every codon can codify every amino acid, the number of alternative codes is really huge, larger than $1.51 \cdot 10^{84}$.

In this research there are basically three theories about the reasons regarding the evolution of the genetic code: i) The stereochemical theory stablishes that codon assignments are dictated by physicochemical affinity between amino acids and the cognate codons, ii) the co-evolution theory states that a set of precursor amino acids passed part or all of their codon domain to the biosynthetically produced amino acids, and iii) the physicochemical or error minimization theory states that the main factor in the code's evolution was the minimization of the adverse effects of mutations.

In this last error minimization theory there are two different alternatives to assess the level of optimization of the CGC:

1. The statistical approach considers random alternative codes and compares the CGC against the average optimization quality of those codes [1]. The results of the works in this alternative tend to indicate that the CGC is quite better optimized with respect to random codes.

2. The engineering approach, which compares the CGC against the average random codes and with respect to better codes (typically obtained with local search algorithms) [2]. This alternative shows that the CGC has a certain level of optimization, but is still far from optimal.

In these alternatives, the "quality" of a hypothetical code is defined taking into account the consequences of all possible mutations in the codons. A mutation in a codon base can change the amino acid codified, which can take a serious consequence on the resultant protein. Usually, it is considered a property (or group of properties) of the codified amino acid, like the polar requirement (hydrophobicity), since it is the most important one in defining the folding of the protein. The quality is defined as the Mean Square (MS) of the difference of the property values of the corresponding amino acids before and after the mutation. This is averaged for all base mutations of all codons. The best code would be the one that minimizes MS, that is, the consequences of the mutations (with the lowest consequence on the phenotype, i.e., the proteins).

We introduced the possibility of evolutionary computation to discover better genetic codes [3], with a classical genetic algorithm (GA) (with *ad hoc* operators depending on the genetic code model of alternative codes), studying the level of optimality reached by the CGC in the huge landscapes considered, and taking into account aspects like the different mutation probability of transversion and transition mutations in the codon bases, or the biases in mutation probability in the three codon bases. In [4] we extended the analysis using a model of possible genetic codes that considers the known codon reassignments. Depending on the genetic code model and those commented aspects, the optimization level of the CGC varies, but in all cases is far from optimal, and in agreement with the engineering approach.

Since there is a discussion about whether the CGC is located in a local minimum (or close to a minimum) in the quality landscape and whether the landscape has a multimodal nature, in a recent work [5] (which is summarized here) we introduced the use of the classical Fitness Sharing (FS) technique [6] in the evolutionary algorithm to inspect that possible multimodal nature. FS allows the division of the population into different

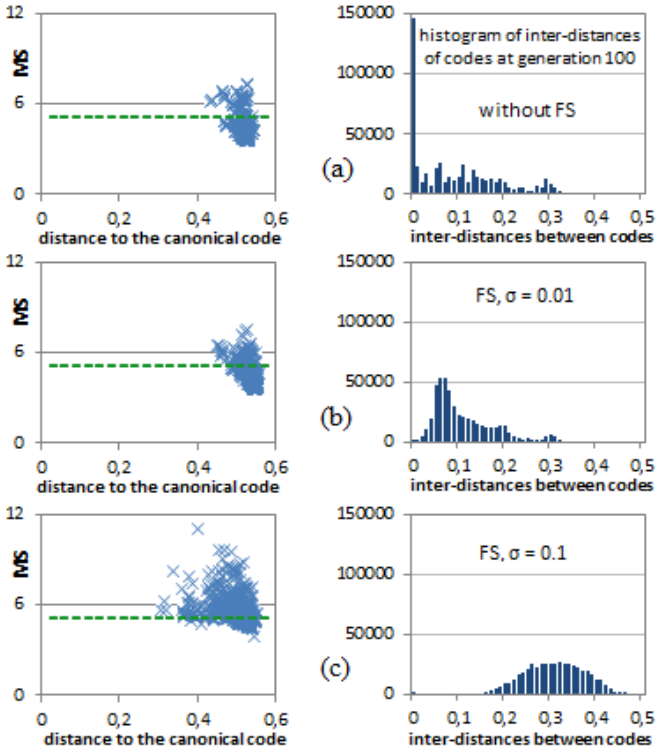

Fig. 1. Left: MS value of the hypothetical codes of the final genetic population vs. their distances to the canonical code and using restrictive codes. Right: Histograms of interdistances of final evolved codes.

subgroups according to the similarity of the individuals, groups that perform a simultaneous search in the best found promising areas (global or local minima) of the search landscape.

The use of fitness sharing serves for two purposes: i) to have an indirect view of the huge search landscape and to locate the CGC in relation to the best possible codes that could be obtained. ii) Given the nature of the landscape obtained in i), to determine the difficulty of evolution of hypothetical genetic codes in such a landscape.

II. Results with a restrictive code model

An example of a run of the evolutionary algorithm is presented here. All the details about the GA implementation (genetic operators, encoding of hypothetical genetic codes) are in [5]. We present results of an example run with and without the use of fitness sharing and with a restrictive model of hypothetical codes. In this model, the pattern of synonymous coding found with the CGC is maintained, that is, the 21 CGC non-overlapping sets of codons are fixed (20 sets correspond to the amino acids and one set to the 3 stop codons). The 20 amino acids are randomly assigned to one of the 20 sets while the same 3 stop codons are fixed as in the CGC.

Figure 1 corresponds to a run of the GA without FS and with FS with two values of the parameter "sharing radius" ($\sigma_{share} = 0.01$ and $\sigma_{share} = 0.1$), which controls the extent of sharing [6]. In Figure 1, subfigures at the left, the x-axis corresponds to the distance of each encoded code to the CGC whereas the y-axis corresponds to the code MS value. These graphs correspond to the final populations at generation 100 in 3 cases: (a) without FS, (b) FS with $\sigma_{share} = 0.01$ and (c) FS with $\sigma_{share} = 0.1$. The figures at the right correspond to the histograms of interdistances between the encoded codes at that final generation. The distance d_{ij} between two codes i and j is measured taking into account the difference in polar requirement between the amino acids encoded in the same positions by both population codes, normalized in [0,1].

The distances of the population (1000 individuals) to the canonical code vary in a range between 0.3 and 0.6. Many of the evolved codes are better adapted than the canonical code (the dashed line represents the MS value of the CGC), showing that these better codes are far from the CGC. There is not any clustering of individuals, independently of the value of σ_{share}, which indicates that there are not deep local minima and that the CGC is not located in an area of a deep local minimum.

Moreover, the CGC was introduced in the initial population in the 3 cases. In all runs the CGC disappears from the population in few generations (as shown in [5]), even with the use of fitness sharing, which is another evidence that the CGC is not located in an area corresponding to a deep local minimum. The interdistances (Figure 1, right part) show that, without FS, the individuals are close to the best solution, whereas the solutions are more spread in the search landscape with larger values of σ_{share} when FS is used. Since there are not interdistance values close to 0 in the histogram (i.e., no clusters), both graphs (with FS) indicate that the landscape does not present clear localized areas of deep local peaks, as well as that the CGC is not located in one of such areas.

III. Conclusion

The main conclusion to be drawn from the results is that the fitness landscape, although is clearly rugged, does not have a multimodal nature with deep localized areas of low MS values and separated by barriers of high MS values. Therefore, the fitness landscape considered in the error minimization theory does not explain how the canonical code ended its evolution in an area that does not correspond to a deep local minimum.

Acknowledgment

This work was funded by Xunta de Galicia (project GPC ED431B 2016/035), Xunta de Galicia ("Centro singular de investigación de Galicia" accreditation 2016-2019 ED431G/01) and the European Regional Development Fund (ERDF).

References

[1] S. Freeland, R. Knight, and L. Landweber, "Measuring adaptation within the genetic code," *Trends in Biochemical Sci.*, vol. 25(2), pp. 44–45, 2000.
[2] M. Di Giulio, "The origin of the genetic code," *Trends in Biochemical Sciences*, vol. 25(2), p. 44, 2000.
[3] J. Santos and A. Monteagudo, "Study of the genetic code adaptability by means of a genetic algorithm," *Journal of Theoretical Biology*, vol. 264(3), pp. 854–865, 2010.
[4] J. Santos and A. Monteagudo, "Simulated evolution applied to study the genetic code optimality using a model of codon reassignments," *BMC Bioinformatics*, vol. 12:56, 2011.
[5] J. Santos and A. Monteagudo, "Inclusion of the fitness sharing technique in an evolutionary algorithm to analyze the fitness landscape of the genetic code adaptability," *BMC Bioinformatics*, vol. 18:195, 2017.
[6] D. Goldberg and J. Richarson, "Genetic algorithms with sharing for multimodal function optimization," *Proceedings 2nd International Conference on Genetic Algorithms*, pp. 41–49, 1987.