Information Sciences 425 (2018) 48–61

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Multi-objective genetic algorithms in the study of the genetic code's adaptability

Lariza Laura de Oliveira [a,*], Alex A. Freitas [c], Renato Tinós [b]

[a] Center of Health Informatics and Information, School of Medicine of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Brazil
[b] Department of Computing and Mathematics, University of São Paulo, Ribeirão Preto, Brazil
[c] School of Computing, University of Kent, Canterbury, Kent, UK

ARTICLE INFO

Article history:
Received 26 April 2017
Revised 12 August 2017
Accepted 8 October 2017
Available online 10 October 2017

Keywords:
Genetic code
Genetic algorithms
Lexicographic approach
Multi-objective genetic algorithm

ABSTRACT

Using a robustness measure based on values of the polar requirement of amino acids, Freeland and Hurst (1998) showed that less than one in one million random hypothetical codes are better than the standard genetic code. In this paper, instead of comparing the standard code with randomly generated codes, we use an optimisation algorithm to find the best hypothetical codes. This approach has been used before, but considering only one objective to be optimised. The robustness measure based on the polar requirement is considered the most effective objective to be optimised by the algorithm. We propose here that the polar requirement is not the only property to be considered when computing the robustness of the genetic code. We include the hydropathy index and molecular volume in the evaluation of the amino acids using three multi-objective approaches: the weighted formula, lexicographic and Pareto approaches. To our knowledge, this is the first work proposing multi-objective optimisation approaches with a non-restrictive encoding for studying the evolution of the genetic code. Our results indicate that multi-objective approaches considering the three amino acid properties obtain better results than those obtained by single objective approaches reported in the literature. The codes obtained by the multi-objective approach are more robust and structurally more similar to the standard code.

1. Introduction

The genetic code is responsible for mapping the four-letter DNA alphabet to the 20-letter protein alphabet. Almost all organisms use a unique standard code; non standard codes are very rare in nature [5,20,27,42]. However, the differences between these non-standard codes are considerably small and the similarities between the codes allow us to assume that all codes have a common origin [40].

The evolutionary context of the standard genetic code's origins has been an intriguing question [39]. Many approaches have been proposed in order to investigate the adaptation of the genetic code [24]. There are three main theories that are most accepted today. The first one is the stereochemical theory, which claims that the genetic code structure was determined by the physicochemical affinity between amino acids and codons or anti-codons [14,21,23,25].

The second one, adopted here, is the adaptive theory. This theory suggests that the genetic code acquired its standard form due to selective pressure to minimise the effects of errors introduced in the production of proteins [6,23,42]. In this theory, the genetic code evolved towards a frozen state or, in optimisation terminology, towards a local or global optimum.

* Corresponding author.
E-mail address: larizalaura@usp.br (L.L. de Oliveira).

https://doi.org/10.1016/j.ins.2017.10.022
0020-0255/

The third theory, called co-evolution [44], claims that the standard code evolved under the influence of the pathways of amino acid biosynthesis, together with the first species. The three theories, which are not mutually exclusive [10], can be used to explain the robustness of the standard Genetic Code (*SGC*).

One of the most evident features of the *SGC* is its robustness against errors or mutations. The robustness has been used as evidence to support the hypothesis that the genetic code has evolved [5]. Considering this hypothesis, two approaches have been used to investigate the relationship between the robustness and the evolution of the code [33]. The first one is the statistical approach, which estimates the number of random codes better than the standard genetic code by randomly generating many different codes [11,18]. The codes are evaluated by a robustness measure. A code A is better than a code B if the evaluation of the former, denoted $f(A)$, is better than the evaluation of the latter, $f(B)$. The other approach is the engineering approach, where the best genetic codes are obtained using an optimisation algorithm [33]. The problem with the statistical approach is that it is usually hard to find a significant number of hypothetical random codes better than the standard one by random sampling. On the other hand, when optimisation algorithms are used, it is generally easy to find hypothetical codes more robust than the *SGC*.

The engineering approach also needs a measure to evaluate the codes. Usually, like in the statistical approach, a robustness function based on one amino acid property is employed. Using the properties: polar requirement [43], hydropathy index [26], molecular volume [15] or isoeletric point [1], Haig and Hurst [18] showed that the standard genetic code is more robust than most random codes for the first three properties, with better results for the first one. In fact, Santos and Monteagudo, using the engineering approach, also concluded that the isoeletric point is the only property that is not good to compute the robustness of the genetic codes [33], and that polar requirement is the best measure. It is important to highlight that the results presented in both papers were obtained by using the amino acid properties individually, i.e, using a single objective approach.

Santos and Monteagudo [33] employed a Genetic Algorithm (GA) to optimise the robustness function based on the amino acids' polar requirement [34,42].

Other works used other objectives in the engineering approach. In this sense, the code is also optimised for the kinetic energy in polypeptide chains [17], compensation between codon-anticodon mismatches and tRNA misacylation [38], and secondary structure formation by mRNAs [19]. In [41], some intriguing questions about the genome structure are raised and discussed in the context of gene expression error minimisation.

The polar requirement was shown to be important to determine the organisation of the genetic code [11,42,43]. However, probably it was not the only factor considered during the evolutionary process. In this context, here we propose a multi-objective approach to investigate the robustness of the genetic code. We use a genetic algorithm (GA) as an optimisation algorithm to obtain hypothetical genetic codes and compare them to the standard genetic code. It is important to highlight that other optimisation algorithms could be used, but GAs, due to their intrinsic characteristics, e.g., the use of a population of candidate solutions, are natural approaches to deal with multi-objective problems [3].

In [31], a multi-objective Pareto approach was used to investigate the *SGC*'s robustness, but with a restrictive encoding. In the restrictive encoding, each amino acid is associated to a set of codons and the sets are the same found in the *SGC*. Hence, this encoding significantly reduces the search space and uses *a priori* information about the *SGC*.

In a more recent paper, Santos and Monteagudo [35] included the fitness sharing technique to explore the fitness landscape of the problem, considering a robustness function based only on the amino acids' polar requirement. They concluded that the *SGC* is not a deep local minimum in the fitness landscape. Also, their findings show that robustness based only on the polar requirement cannot explain the *SGC*'s structure by itself.

According to Freitas [13], when dealing with multi-objective problems, we can use three main approaches: (a) the weighted formula approach, which transforms the multi-objective problem into a single objective one; (b) the lexicographical approach, where the objectives are ranked in a priority order; and (c) the Pareto approach, which considers a set of non-dominated solutions (details will be given in Section 2). In this context, the main objective of this article is to investigate the hypothesis that a multi-objective optimisation approach is useful to study the genetic code's adaptability, since intuitively it is more biologically plausible to consider evolution as a multi-objective optimisation process than a mono-objective one. We compared the three proposed multi-objective approaches, considering their pros and cons. We also used a non-restrictive encoding with three amino acid properties which seem to be relevant to the computation of robustness. Regarding implementation, we used the well-known NSGA-II algorithm as the Pareto-based genetic algorithm, and implemented the weighted formula and lexicopraphic approaches using a standard genetic algorithm [7].

When comparing our results with previous ones [33], we found better values of fitness, which means that the best hypothetical solutions evolved by the GA are closer to the *SGC* in terms of the used evaluation function. In addition, the solutions found by the multi-objective approach have frequencies of codons associated with amino acids more similar to the *SGC* than those found by the single-objective approach. This result also indicates that it is not necessary to use a restrictive encoding to reduce the search space of the problem – a restrictive encoding is frequently used in the literature [31]. Also, it is important to highlight that the multi-objective approach seems to be more realistic, because it does not seem plausible that the robustness of the standard genetic code was optimised considering only polar requirement.