Evolutionary Algorithm Analysis of the Biological Genetic Codes

David Digby

Department of Biological Sciences
Clark Atlanta University
Atlanta, GA 30134

William Seffens

Clark Atlanta University
Center for Theoretical Study of Physical Systems
Clark Atlanta University
Atlanta, GA 30134

Abstract

An evolutionary algorithm program has been used to investigate the evolution of the universal and several alternate genetic codes in biology. Various fitness functions were evaluated using a fixed set of genes and a genetic code that mutates every generation for up to 1000 organisms. The evolution of the population was found to be stable if starting with the universal genetic code.

1 INTRODUCTION & RESULTS

An evolutionary algorithm (EA) was implemented to investigate factors influencing the evolution of the universal and several alternate genetic codes in biology. The algorithm selects for fidelity in replication of protein sequences, without allowing the proteins themselves to mutate. Instead the genetic apparatus is mutated with every generation so that the interpretation of the genes can change. Up to one thousand organisms are started with randomly selected genetic codes, which are then mutated over many generations, with a continuous requirement to replicate a genome consisting of ten gene families of ten copies each. Representative real enzyme amino acid sequences were used, and the fidelity of replication is measured by reference to a standard table of amino acid similarities. An additional selection criterion demands minimization of a cost function roughly proportional to the number of tRNA genes required for the translation machinery. Several abstract characteristics of the most successful codes are compared to the same characteristics of the real genetic code and its several variants.

This EA is different from other implementations in that the genome is fixed while its interpretation by the genetic code can mutate. Each genetic code is defined as a 64x20 matrix relating each one of the 64 codons to each one of the 20 amino acids. The EA begins by constructing a random genetic code matrix for up to 1000 organisms, which are then allowed to mutate individually for each generation. Every organism is evaluated for fidelity of transcribing its genome, using its own genetic code (with weighted choices in case of degenerate assignments). The 100 fidelity values are summed to yield a total fitness parameter for each of the 1000 organisms, and these scores are sorted to establish a ranking of the organisms. Organisms with the lowest 50% of fitness values are killed, and the organisms with the highest 50% fitness values are further evaluated for the number of non-zero entries in their CODE matrix, each of which represents a "tRNA cost" to the organism. This additional fitness test was ranked in a similar manner as before. Those 25% organisms having the lowest numbers were replicated to produce three new copies, in order to replace those previously killed, whereas the remaining 25% just survived with a single copy. Once every fifty generations, a sample of the genetic codes for all 1000 organisms is stored, requiring about 2.5 Mb of space per sample. Analysis of these samples is performed at a later time. Abstract measures have been devised to measure structural or functional similarities to the real codes, without requiring exact duplication. Large populations were stable if started with the universal genetic code.

A typical study involves running five separate trials with each set of parameters, in order to assess the influence of the random starting set of genetic codes. For some EA parameters, the outcome after 500 generations depends upon the set of gene sequences used.

2 CONCLUSIONS

A modified EA has been used to investigate a possible scenario for the evolution of the universal and several alternate genetic codes. The set of genes in the genome appears to modify the outcome of the EA, implying that the genetic code is a result of the initial set of genes in early evolution.

Acknowledgments

This work was supported by NIH grant GM08247, by a Research Centers in Minority Institutions award, G12RR03062, from the Division of Research Resources, National Institutes of Health, and NSF CREST Center for Theoretical Studies of Physical Systems (CTSPS) Cooperative Agreement #HRD-9632844.