

# Genetic Code Population Dynamics in a Simulated System of Protocells

Jerrisha Butler, Frances Uzowulu, David Digby and William Seffens

Biology Department Clark Atlanta University  
223 Brawley Dr, S.W. Atlanta, GA 30314  
[wseffens@cau.edu](mailto:wseffens@cau.edu)

**Keywords:** Genetic code, Evolutionary algorithm, and graph theory

## Abstract

An evolutionary algorithm was developed to investigate how the biological genetic code may have evolved. Graph theoretical "degree sequences" were derived from the simulated genetic codes of "organisms" produced by this algorithm. These graph theory metrics were used as a measure of similarity among evolving genetic codes, in order to classify the set of organisms into equivalence classes without requiring absolute identity. A detailed look at the population dynamics of these sequences revealed patterns in the rise and fall of specific sequences reminiscent of the succession of biological species. A new sequence first appears in small numbers, may become dominant for a time, but then fades away in the process of giving way to another, more "fit" sequence. The "fit" sequences are the remaining survivors after fidelity, tRNA, and stem loop rate selection. Except for a few scattered sequences that were accidentally reproduced in the first generation of entirely random genetic codes, these matching sequences frequently turned out to belong to offspring of a single "primordial" organism, as replicated during the genetic algorithm simulations.

## Introduction

Genetic Algorithms and Evolutionary Algorithms, although based on a biological theme, are computational methods used to solve various complex engineering problems unrelated to biology (Shapiro and Wu, 1996 & 1997). Here the problem is itself biological, involving both genetic and evolutionary concepts. Some of the terminology used here may therefore have two distinct meanings, but in most such cases, the biological meaning will take precedence.

The genetic code is the codification of the correspondence between RNA codons and protein amino acids. The origin and evolution of the genetic code is unknown, although many theories have been advanced (Maeshiro and Kimura, 1998; Di Giulio and Medugno, 1998; Di Giulio, 1995; Santos, et al, 1995; Wong, 1975). In the real world today, genetic codes are homogeneous within a given population, and in fact, the great majority of living organisms employ a common "universal" code (Seffens, 2002). We investigate the processes or factors

that influenced the different structures of the alternate genetic codes in this work. A model was developed in which a mixture of genetic codes was subjected to simulated evolution in an otherwise homogeneous population of competing organisms. We sought to determine if evolving random genetic codes settle into more optimum forms resembling the current genetic code.

Since there is a high probability that at least some element of the "frozen accident" theory is valid (Crick, 1966), it would not be reasonable to expect any simulation of the evolution of the genetic code to end up with the exact same real code. For this reason, abstractions of the real code were selected that model some of the mathematical symmetries of the code, while retaining structural features that may have some relevance to its viability. One such measure is presented here, the "degree sequence", which examines elements in the grouping of codons, independent of specific assignments. It consists of a set of numbers representing the counts of how many different codons are assigned to each amino acid.

## Codon redundancy

In attempting to understand the evolution of the genetic code, one must give some thought to the probability that a given characteristic of the real code might occur in a randomly chosen genetic code. Codon redundancy is one of these characteristics. In order to include the twenty amino acids, and to have at least one stop codon, the average level of redundancy must be more than three, but less than four. In order to leave at least one codon for each of the remaining entities, the largest set cannot have more than 44 codons. In like manner, not more than one set can have as many as 23 codons, there can be no more than 14 amino acids with a redundancy of four, etc. Under the assumption that all of the codons are utilized with the exception of the termination codons, a count of all such partitions gives us 59,755 combinations that could exist for codes selected completely at random. However, the set

of redundant codons assigned to each amino acid is linked to the mechanism required to translate mRNA into protein. Other properties in the structure of the real code are also the result of natural constraints, so that comparison to completely random alternatives such as the above is not justified.

In the real codes, there exists a disparity in the number of assignments for different amino acids. Every known real genetic code has two or three amino acids with more than four codons, but there are none with more than eight (Louis, et.al., 2005). These considerations led to a list of possible partitions restricted to the conditions above, drawn from the standard set of sixty-four codons. There are 21 fixed codon-groups, representing the 20 amino acids plus at least one stop codon. Some alternate codes also include a 22nd group of unused codons. Each of these was placed into a size subset according to its number of codons.

The size of the codon groups can be modeled as a number of lines incident to each codon group, or vertex. Each genetic code can be represented as a list of the number of lines for each vertex in a graph theoretical representation called a degree sequence. The complete degree sequence for the real standard, or "universal" code (in increasing order of size) is: 1,1,2,2,2,2,2,2,2,2,2,3,3,4,4,4,4,4,6,6,6. The numbers, in turn, represent the amino acid-codon assignments for this genetic code. In order to represent such partitions more compactly, they are compressed by listing only a single count to reserve a place for possible items of a size that are not present in this particular degree sequence. For instance, the compressed sequence for the standard, or "universal" code is : 2,9,2,5,0,3. Since each count in this compressed sequence represents either one of the amino acids or the set of stop codons, it is a partition of the number 21.

### Experimental Method

A series of computer algorithms were developed to simulate and analyze the involvement of mRNA structures (Digby and Seffens, 2004) in the evolution of the genetic code. The main program V05FEB constructed weighted random "genetic code" matrices for up to 1000 "organisms". Each such CODE matrix was used for translating the set of 64 codons into the set of 20 amino acids. Its structure allows for the possibility of degeneracy for amino acids, as well as for ambiguity of codons. Each entry in this 1280 element matrix represents a weight from 0 to 1, denoting the probability that codon

"i" will translate to amino acid "j". (The matrices for real genetic codes contain all zeros, except for a single "1" in each codon row.) Each genetic code was allowed to mutate individually, and its evolution was monitored over many generations.

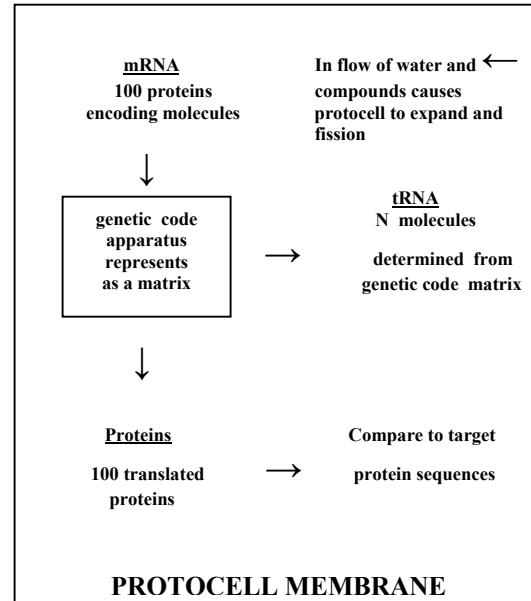


Figure 1. Model Protocell. The genetic code apparatus, the 100 mRNA sequences, and the number of tRNA molecules, N, and allowed to change in the simulation. Only the target protein sequences are fixed.

In order to exercise the genetic codes in a realistic, but strictly comparable way, all organisms were given the same identical "genome", consisting of multiple copies of from 10 up to 100 real protein sequences. Each of these was chosen to represent a fundamental enzyme such as ferridoxin, with a fixed, highly conserved sequence composed of no more than 100 amino acids. Therefore the complete genome in each organism has several gene families containing up to ten different versions of a gene for each enzyme, for a total of 100 genes representing up to 10,000 amino acids or 30,000 nucleotides. The genes are contained in a protocell package with a primitive fission function for reproduction (Figure 1).

At each generation, organisms were evaluated for fidelity of replicating these real enzymes, from each genome, using its own genetic code, as follows:

a) Representative mRNA nucleic acid sequences were predicted from each of the 100 proteins, based on codon assignments from the genetic code matrix weights. (This amounts to

"back- translation", which does not occur in nature).

b) These mRNAs were then (forward) translated back into proteins, again using each organism's own genetic code and weighted random choices.

c) The resulting protein sequences were compared to the original real genes by scoring mismatched amino acids based upon a standard similarity matrix (Maruyama et al, 1986) to yield a fidelity measure of fitness.

Fidelity values were sorted to rank the organisms by fitness, and the highest, lowest and average scores were recorded. A tRNA cost fitness parameter was then used, after which the lowest scoring organisms were replaced by copies of the highest scoring organisms. Each element matrix that does not contain 0 will require a tRNA or other enzymes to implement. The greater the number of tRNAs, the greater the load and energy demanded by that particular cell.

After the selection above, the genetic codes were subjected to random "mutation", at a selectable rate, by altering a few entries in their genetic code matrices. Each of the 1280 entries in the CODE matrix was given some probability to be altered for each simulation run. A control run, made at the lowest rate of mutation, but with all selection and replacement disabled, showed a declining level of fitness over 1000 generations, from 60% to about 10% of the maximum possible. Control simulations were also seeded with the real universal genetic code to investigate if V05FEB recognizes and retains highly fit organisms.

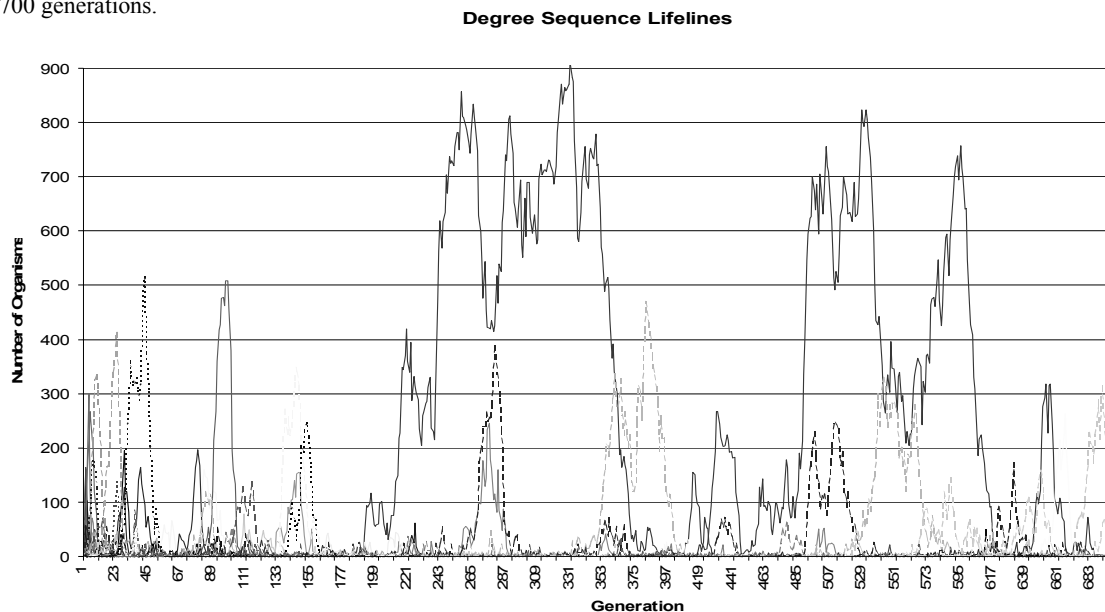
## Results

Various fitness functions and parameters were evaluated in the evolutionary algorithm V05FEB. Various strengths of the fitness test for fidelity of protein translation were evaluated to find parameter sets that exhibited behavior on the edge of chaos. We also examined parameters for "tRNA cost" to the organism, as measured by the number of corresponding genes required. Each simulated the evolution of randomly created genetic codes by duplicating codes with good fitness to replace codes with poor fitness, randomly mutating the surviving codes, creating a log of fitness and other values, and storing samples of the complete set of genetic codes.

A program (SHOWPART) was used to analyze the generation-interval samples taken of all the evolving genetic codes, by identifying a degree sequence represented by each of these sampled codes. For each sample, a list of all distinct degree sequences was recorded, and a count of the number of organisms sharing that partition. A wide variety of partitions was always found in the first generation, before any selection had taken place, which provided a verification that all organisms were started off with random genetic codes.

Before any selection had taken place, there were always sequences with some rather large codon groups. After several generations of selection, the size of the largest codon group in

Figure 2 Illustrates patterns of "fit" organisms over 700 generations.



most sequences seldom exceeded the range of four to six, much closer to those of real genetic codes. After just 50 generations, the number of different partitions was dramatically reduced, and the number of organisms sharing one "favorite" partition had grown accordingly.

Utilizing data from SHOWPART, Figure 2 was produced to illustrate a typical progression of the number of distinct partitions, revealing patterns in the rise and fall of specific partitions reminiscent of the fate of biological species in succession. A new partition first appears in small numbers, may become dominant for a time, but then fades away in the process of giving way to another that is more "fit". The identity number of each "founding organism" retained in each surviving organism indicated that, except for a few scattered sequences duplicated in the first generation of entirely random genetic codes, the large number of organisms sharing a common sequence were derived from reproduced copies of a single "primordial" organism.

Most simulations with a variety of parameter values show a three-part behavior. For the first 10 generations average fitness steadily increases as the unfit genetic codes are rapidly killed. Then a decrease in average fitness for the 1000 organisms dominates for the next 2000-10,000 generations. During this period the mutation rate parameter has a large influence and generally overwhelms the organisms. Eventually, for smaller mutation rates, the organisms begin to recover and the average fitness slowly increases.

This work will be used to examine the influences that gave rise to the alternate genetic codes in nature. Here we examined the effect of protein product fidelity and tRNA load on the pattern of degree sequences observed in the evolving genetic codes for 1000 organisms. The program V05FEB has options to include mRNA folding fitness functions to examine genome-wide biases in folding free energies of mRNAs (Digby and Seffens, 1999)

#### **Acknowledgements**

This work was supported (or partially supported) by NIH/NIGMS/MBRS/SCORE/RISE (SCORE grant #S06GM08247) and G12RR03062 from the National Center for Research Resources, NIH.

#### **References**

Crick, F., 1966, Codon-anticodon pairing: The wobble hypothesis. *J. Mol. Biol.* 19:548-555.

Digby, D. and W. Seffens, 2005, Step-Wise Mutations of mRNA Sequences Lead to Progressive Changes in Calculated Folding Free Energies. *Series in Mathematical Biology and Medicine* Vol. 8. p 341-350.

Giulio, M. Di., 1995, The phylogeny of tRNAs seems to confirm the predictions of the coevolution theory of the origin of the genetic code. *Orig. Life Evol. Biosph.* 25(6):549-564.

Giulio, M. Di and Medugno, M., 1998, The historical factor: the biosynthetic relationships between amino acids and their physicochemical properties in the origin of the genetic code. *J. Mol. Evol.* 26(6):615-621.

Louis, V., Digby, D and Seffens, W., 2005, Constructing Taxonomy Dendrograms of the Alternate Genetic Codes. *Proceedings of Annual Meeting Atlanta, GA, 21-25 May 2005.* Poster R-017.

Maeshiro, T and Kimura, M., 1998, The role of robustness and changeability on the origin and evolution of the genetic code. *Proc. Natl. Acad. Sci. USA* 26(3):5088-5093.

Maruyama, T., Gojobori, S. Aota, and T. Ikemura, 1986, Codon usage tabulated from the GenBank genetic sequence data. *Nuc. Acids Res.* 14:r151-r189.

Santos, M.A., Useda, T., Watanabe, K. and Tuite, M.F. On the origin and evolution of the genetic code. *J. Mol. Evol.* 26(3):712-716. 1995

Seffens, W. and D. Digby, 1999. mRNAs Have Greater Calculated Folding Free Energies Than Shuffled Or Codon Choice Randomized Sequences. *Nuc. Acids Res.* 27, 1578-1584.

Seffens, W., 2002, Graph Theory Patterns in the Genetic Codes. *Forma*, 17:309-320.

Shapiro, B.A. and Wu, J.C. Predicting RNA-H type pseudoknots with the massively parallel genetic algorithm. 13(4):459:471. 1997

Watson, J.D. and Crick, F.H.C. *Nature (London)* 171, 737-738. 1953. J.T. Wong. *Proc. Natl. Acad. Sci USA* 72: 1909-1912. 1975.