Supporting Information

Moran Model Simulating a Primordial World

We construct a more concrete Moran model, a population simulation where the number of organisms is held constant throughout the simulation. As organisms reproduce, they randomly replace another organism in the population. The reproductive rate of the organisms is governed by their fitness (see below). The model we implemented begins with 100 organisms with random genome sequence and a genetic code of all stop codons. The genetic code of an organism is a circular array of length 9, where each location (codon sequence) in the array maps to either a red or green amino acid or stop codon. The genome is a linear array of up to 25 codons.

During each translation step, the genome is translated into a string of amino acids and stop codons via the genetic code. During this process, a codon may receive the amino acid encoded by one of its neighbors in code space with a probability of 0.05. At the end of the translation step, the polypeptide sequence is evaluated. The longest polypeptide of sequence {red-green-}^{N}-stop adds N points toward an organism's reproductive score. Other polypeptides are considered incidental and do not add to the reproductive score; however, during each translation step each organism receives 1 point toward its reproductive score because we assume its RNA has replicative function even in the absence of protein synthesis. Over multiple translation steps, each organism accumulates points toward its reproductive score. When it reaches a threshold of 100 points, the organism produces a daughter cell by replacing a random organism in the population. During replication, each codon in the genetic code encodes a different amino acid or stop codon with probability 0.01 per codon per replication, and each position in the genome mutates to a different codon with probability 0.01 per codon per replication. This process allows for the coevolution of the genetic code and the genome. Each trial of the simulation was run for 40,000 translation steps.

We note that another way in which our simulation differs from previous ones [2][3] is that the pressure to minimize genetic code entropy comes from competition between tRNAs during translation and not from mutation. This difference stems from a choice to consider error minimization from the perspective of robustness with regard to tRNA wobble and other thermodynamic effects rather than closeness in mutational space.

Results

We start from a simple pooling equilibrium, in which all codons are assigned to stop instructions, and genomes are random codon strings. During the course of the simulation, organisms are reproductively favored if they synthesize polypeptides of a certain pattern. We simplify the genetic code as a one-dimensional ring. During the translation process a signal for any particular codon in the ring may be misread as a signal for one of the neighboring

codons in code space. As in other studies [3], such a simplification allows us to view many of features of a real genetic code without specifying base pairs and positions, which may have varying substitution rates based on complex thermodynamic and copy number effects. During replication, a codon in the genome may mutate to any random codon at a rate of 0.01 per position per replication, and the genetic code may change one of its assignments to a random amino acid or stop codon with a probability of 0.01 per codon per replication. These rates are relatively high because we assume that prior to the advent of protein translation and replication machinery, the frequency of errors might have been higher.

We quantify the entropy of the genetic code based on the likelihood that an incorrect amino acid will be incorporated during the translation process (Figure S4A: Genetic Code Assignments). After 100 trials, the genetic codes of organisms at the end of the simulation have low entropy when compared to random codes (Figure S4B,C). For illustrative purposes, the dominant lineage from one simulation is shown in Figure S5A. In this example the genetic code goes through a diversifying step followed by a consolidation step, at the end of which neighbor codons tend to code similar amino acids, as expected [1][3][4][5]. The consolidation step is the result of a single common ancestor emerging as the dominant organism in the population, although its spread can be attributed to either drift or fitness advantage. Thus, during the diversification step, there is also a wider degree of heterogeneity in the codes of the population as a whole, in agreement with the stochastic simulations presented in the main text. In addition, there is a wider degree of genomic heterogeneity during the diversification step, as the variety in genetic codes allows for greater experimentation with different genomes.

Further reduction in entropy is afforded by codon usage bias in the genome, as expected from observations of real genomes [6]. In our simulations, the equal and concentrated presence of both red and green amino acids manifests as a large and equal representation of those amino acids in code space. By contrast, the stop codon, which is used only once during the translation of a long peptide, tends to assume a smaller portion of code space.

In this setting, proteome length increases when codons abutting an existing gene mutate so that they encode amino acids in a proper sequence, which lengthens the protein-encoding gene (See Figure S4A: Proteome). A plot of the length of the longest encoded protein as a function of the entropy of the genetic code is shown in Figure S4C. As expected, longer genes do not emerge if the genetic code is not accurate enough to translate them properly. As shown by the red points in Figure S5C, dominant species at the end of the simulation tend to have long genes and genetic codes that are error-minimizing when compared to random codes. The range of entropy values observed reflect the tension between the need to encode a diversity of amino acids and the need to reduce entropy; this finding accords with previous theoretical studies [4][5].

In 66% of 100 trials, 80% or more of the organisms at the end of the simulation used only one genetic code. In the vast majority of these cases, the organisms not using the dominant code were related to a common ancestor using the same code and varied by one mutation.

Universality of the genetic code was established at some point in time in virtually every trial; however, in many cases after universality was established other derivative genetic codes emerged, most likely a result of the fact that genome length in our simulations was bounded to be at most 25 codons. Thus a simpler theory without a need to enforce horizontal gene transfer may suffice to explain universality.

Availability

The code used to implement the mathematics described in the main text and simulation described here is freely available at: http://bioinformatics.nyu.edu/projects/genetic-code/

References

[1] Vetsigian, K., Woese, C., Goldenfeld, N. (2006) Collective evolution and the genetic code. *PNAS*. 103(28): 10696-10701.

[2] Ardell DH, Sella G (2002) No accident: genetic codes freeze in error-correcting patterns of the standard genetic code. *Phil Trans R Soc Lond B* 357:1625–1642

[3] Sella, G. and Ardell, D.H. (2006) The coevolution of genes and genetic codes: Crick's frozen accident revisited. *J. Mol. Evol*. 63(3):297-313.

[4] Tlusty, T. (2008) A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity, and cost. *Phys Biol*. 5 016001.

[5] Tlusty, T. (2010) A colorful origin for the genetic code: Information theory, statistical mechanics, and the emergence of molecular codes. *Phys Life Rev*. 100 048101.

[6] Moriyama, E.N., Powell, J.R. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Eschericihia coli*. (1998) *Nucl Acids Res* 26(13):3188-3193.

Figure S1. Simulations with two possible conventional signaling equilibria with different error tolerance. (A) Here, species may have one of two possible genetic codes, A and B, and possible proteome lengths 10, 20... 100. Code B (E=0.010) is error minimizing relative to Code A (E=0.015). It is possible to mutate from code B to code A (and vice versa) with a point mutation in a single tRNA. (B) The number of organisms of a given genetic code (A or B) is plotted against time, starting with one organism of A, proteome length=10. The plot titled "Normal" uses the same parameters as in Figure 2. Similarly, "Error(A) = Error(B)" depicts a simulation in which Code B has the same E value as Code A. "Low code mutation rate" depicts a simulation in which μ_0=0.1. "High code mutation rate" depicts a simulation in which μ_0 =4. "Low genome length mutation rate" depicts a simulation in which μ_{length} =0.01. "High genome length mutation rate" depicts a simulation in which μ_{length}=0.5.

Figure S2. A species map depicts three genetic codes. Code B and B* (E=0.010) are error minimizing relative to Code A (E=0.015). It is possible to mutate from code B to code A (and vice versa) and from B* to A (and vice versa) with a single tRNA mutation. Organisms with both codes may acquire genomes of longer length according to the dynamics previously described. In the six subplots shown, the number of organisms of a given genetic code (A, B or B*) is plotted against time, using the same parameters described in Figure S1.

Figure S3. Stochastic simulations run according to the same setup as the simulation in Figure 1, but with varying carrying capacities, K. In the top row, the genetic codes present in the population at a given time (as represented by their error value E) are shown across time for three different simulations, K=100, K=1000 as in Figure 1, and K=10000. In the bottom row, a snapshot of the population levels at the end of each simulation is shown according to the species map in Figure 1.

Figure S4. A) Schematic of a protocellular environment in which self-reproducing RNA vesicles containing both proto-mRNA (Genome) and proto-tRNA (Genetic Code) are immersed in a pool of highly concentrated amino acids (See SI: Primordial World Simulation for details). Genetic Code Assignments. The codon-amino acid assignments for one lineage are represented by a one-dimensional array. Any codon may substitute for its neighbors in code space with 5% probability. The genetic code assignments across generational time trend toward clustering similar amino acid assignments in blocks to reduce entropy, as well as reducing the representation of stop codons, which are used infrequently. To understand the evolution of proteome complexity, the longest protein encoded by the genome given the genetic code assignments above is recorded across generational time. B) The entropy for 10,000 randomly assigned genetic codes is shown for comparison with error-minimizing codes, which evolved in (C). C) Dominant species' genetic code entropy and encoded protein length from the end of 100 simulations are shown in red. Similar points from all ancestors of those final species are shown in blue.