Simulation Model of Prebiotic Evolution of Genetic Coding

Sidney Markowitz[1], Alexei Drummond[1], Kay Nieselt[2] and Peter R Wills[3]
[1]Department of Computer Science, University of Auckland, PB 92019, Auckland, New Zealand
[2]Zentrum für Bioinformatik Tübingen, Universität Tübingen, D-72076 Tübingen, Germany
[3]37 Fairlands Avenue, Waterview, Auckland 1007, New Zealand
sidney@sidney.com

Abstract

Common to all life on Earth are the mechanisms of genetic encoding, in which specific trinucleotide sequences in DNA and RNA map to specific amino acids in synthesized proteins. This paper investigates a novel gene-replicase-translatase (GRT) system to determine whether emergence of genetic encoding from an initially random population of genes and proteins is feasible. The model incorporates gene replication with mutation, error-prone protein translation, and an arbitrary encoding from codons to amino acids. Simulations on the order of 10^9 event steps demonstrate self-organization to evolutionary stability with distinct phase transitions. The ranges of parameters that lead to an apparent attractor state are consistent with the notion of error threshold as a determinant of stability in error-prone autocatalytic systems.

Background

One of the processes in prebiotic evolution that is not yet very well understood is the mechanism of emergence of genetic coding: a regular mapping from the set of trinucleotide codons onto the 20 standard amino acids. The mapping is mediated by proteins, the aminoacyl-tRNA-synthetases (AARS), each of which catalyzes the assignment of a particular amino acid to its set of cognate codons. The proteins which catalyze the decoding that constitutes protein synthesis are themselves products of that synthesis. The extant autocatalytic system of coded protein synthesis poses the question of how it bootstrapped itself into existence and how it maintains stability.

Some indication of the path of evolution of the current system of genetic coding from simpler systems can be found in the structure of the AARSs. O'Donoghue and Luthey-Schulten (O'Donoghue and Luthey-Schulten, 2003) have created a structural phylogeny of AARSs that suggests an early mapping from classes of codons to classes of amino acids, evolving through steps of increasing specificity of mapping.

The focus of the current work is on modeling the characteristics of an autocatalytic system that can achieve stability and can self-organize to the type of informational complexity seen in modern genetic systems. An overview of the work that this paper extends can be found in Wills (Wills, 1993; Wills, 2004). Papers by Wills (Wills, 1993) and Nieselt-Struwe and Wills (Nieselt-Struwe and Wills, 1997) that develop a model of protein translation which allows the derivation of formal constraints on genetic coding systems are the immediate predecessors of the present work.

Eigen (Eigen, 1971) elucidated the notion of an error threshold as a solution to the so-called "error catastrophe" problem with error prone replication in autocatalytic systems. Hoffman (Hoffman, 1974) applied a similar threshold to the process of translation of codons to amino acids. Wills (Wills, 1994) showed that in a chemically homogeneous self-organized coding system mutation during gene replication inevitably leads to collapse.

Füchslin and McCaskill (Füchslin and McCaskill, 2001) added an error-prone replication process to a simplified version of the error-prone translation process model to show that reaction-diffusion coupling can allow the two processes to self-organize into a stable system. They described what they call a Gene Replicase Translatase (GRT) system that includes gene replication with mutation catalyzed by a "replicase" protein, protein synthesis with coding catalyzed by a "translatase" and two alternate translatase proteins that implement non-target codings.

The present work models a GRT system that is more complex and more realistic than that of Füchslin and McCaskill (Füchslin and McCaskill, 2001). Separate translatases for each codon-to-amino acid assignment are embedded as catalytic centers in the protein sequence. Our reaction employs a separate translatase for each amino acid (like the real process) whereas Füchslin and McCaskill employed a single translatase for the entire synthesis of a new protein. Their model was embedded on a three dimensional lattice with one molecule per node and diffusion between nodes. Our model also demonstrates stabilization through reaction-diffusion coupling, but uses a one-dimensional lattice of "well-mixed" compartments containing variable numbers of molecules, with diffusion between compartments.

Eigen (Eigen, 1971) formulated an abstract and general model that provides insights that are useful when attempting to design more realistic models of self-organizing processes. The subsequent cited work added features that bring the model closer to natural genetic systems, while elucidating constraints on self-organization. Our goal is to eventually model processes that can describe self-organization starting from a random initial state through simple autocatalytic systems all the way to the modern systems of genetic coding. To that effect, we have developed a model that is intermediate, in complexity and realism, with respect to the models of prior work and the systems of natural biology. By finding the ranges of parameters in which self-organization is demonstrated, we can then further refine the model in the direction of added complexity and realism. This paper has the modest goal of demonstrating that there are values of parameters of the model in which self-organization occurs. Some of the implementation decisions that are described in the Methods section were made to make the simulation computationally tractable. Others represent choices that were found to produce workable results. Future work will explore the range of parameter space and more precisely characterize the elements that are necessary for self-organization.

METHODS

Genes and proteins are represented as bit sequences of codons and amino acids, respectively. Our model uses two bits to represent four types of codons and four types of amino acid. Genes and proteins are each 12 units long, represented by 24 bits. These numbers were chosen as a balance between having sufficient complexity to be likely to demonstrate interesting behavior in a random system and computational feasibility. The choice follows (Wills, 1993) who used a similar size of 4 for the codon and amino acid alphabets to represent the putative GNC codons and four amino acids thought to have existed in early prebiotic systems. Wills (Wills, 2004) discusses the evolution of a similar quaternary code from simpler binary codes, and the possibility of further stepwise evolution to the more complex codes of the biological genetic system.

At the start of a simulation, sixteen 24-bit sequences are randomly selected to be the "catalytic centers" (Wills, 1993) for protein translation, corresponding to the sixteen possible assignments of codon/amino acid pairs. The rate of protein translation and the selection of amino acids by codons are determined by the concentration in a compartment of proteins with sequences at or near the catalytic centers.

Error proneness is introduced into the system as in (Wills, 1993) and (Füchslin and McCaskill, 2001) by calculating a protein sequence's catalytic activity as a monotonically decreasing function (in our case Gaussian) of the Hamming distance from the catalytic center, measured as number of amino acid differences. Every protein sequence has some probability of catalyzing any coding, including erroneous ones. This also allows protein translation to occur with random coding at a minimal rate in the presence of random protein sequences, and at sharply higher rates when there are catalysts for specific assignments coding present.

Co→AA	**Translatase**	*** Target gene sequences**
00→W	WYZWZXWYYZYY	
00→X	WWXZWWWWYWYY	
00→Y	**ZYZZXZXZZWYX**	*** 100010101110111010010011**
00→Z	YZWYZWZYYYYY	
01→W	**ZYWYYXZXZWWZ**	*** 100001000011101110010110**
01→X	YYZZZXYWZYXY	
01→Y	WYXXXXWZZZYZ	
01→Z	XYXZWXYXZXZW	
10→W	ZZZWYWYXZZZX	
10→X	XWWXWZYYYZZY	
10→Y	XWYWZZYYYZZY	
10→Z	**XWYYWXYWWZYZ**	*** 110010100111000101100010**
11→W	XZWYZXZXYZWZ	
11→X	**WZXZWWWZWYXW**	*** 011011100101011001001101**
11→Y	WYYYZWYYXWYY	
11→Z	YZYWZXZZWWZW	
Replicase	**WYZWZZYXXWWY**	*** 010010011010001111010100**

Table 1: *An example coding. For each of the 16 possible translations from codons in the alphabet* {0,1} X {0,1} *to amino acids in* {W,X,Y,Z} *a random translatase sequence is chosen as a point in a 12-dimensional protein sequence space. Gene sequences are shown for one randomly selected coding.*

The "R" in GRT is introduced by randomly selecting a bit sequence to be the replicase catalytic center. Proteins catalyze gene replication with activity defined as a decreasing Gaussian function of Hamming distance from the replicase catalytic center. As with protein translation, this allows some replication to occur in the presence of random proteins, but does not introduce any inaccuracy. Replication error, i.e., mutation, is specified by a mutation rate parameter of the model expressed as average number of bit errors per genome replication. The introduction of gene mutation to the model is an extension to (Wills, 1993) that is found in (Füchslin and McCaskill, 2001).

A "coding" is a choice of translatase catalytic centers containing one mapping from each of the codons and one mapping to each of the amino acids. In our model there are 4! possible codings, each containing a set of four translatase catalytic centers. Each coding uniquely specifies the codon sequences of the four genes that translate to the four translatase catalytic centers, and that of the gene that translates to the replicase catalytic center. The five genes and five proteins form an autocatalytic system, in which the genes produce the proteins that catalyze the production of proteins and the replication of the genes. Table 1 shows an example of a randomly selected coding.

One coding is randomly selected to be the target coding for the simulation. The simulation model does not itself contain any bias towards a particular coding. We seed the initial state with one or more genomes at or near the target. Our goal is to determine ranges of parameters and initial conditions that cause the model to evolve to a stable operational system that expresses the selected target coding.

Genes undergo translation, replication, and diffusion in sets of five as “genomes”. The total number of genomes and the total number of proteins in the entire system are each kept constant. This simulates a regulated dilution flow that maintains “constant organization” as defined by Eigen (Eigen, 1971) and utilized by Wills (Wills, 1993). To prevent unconstrained growth in a single compartment from dominating the system growth, there is a volume limitation that sets a bound on the number of genomes and proteins in a compartment.

Genomes and proteins diffuse to adjacent compartments at independent rates proportional to the space available in the destination compartment. Available space in a compartment also modifies the rates of gene replication and protein translation.

There are four types of simulation events: translation (synthesis of a protein); diffusion of a protein to an adjacent compartment (with periodic boundary conditions); replication of a genome; and diffusion of a genome to an adjacent compartment.

The simulation clock is in units of events. The next event step is chosen from the four types using weighted probabilities that are proportional to the four rate parameters and the associated activity factors of each event type summed over all compartments.

The activity factors calculated per compartment are:

- Diffusion rate of genes or proteins is proportional to the ”osmotic pressure gradient” calculated as the product of the number of genes or proteins in a compartment and the number of spaces for them in adjacent compartments.

- Protein translation is proportional to the total catalytic activity of all proteins in a compartment relative to the translatase catalytic centers, and to the number of genes in the same compartment.

- Gene replication is proportional to the total catalytic activity of all proteins in a compartment relative to the replicase catalytic center, and to the number of genes in the same compartment.

Once an event type is chosen, a protein or gene is randomly selected using the weighted probabilities suitable for the component and the event, then the event is simulated.

The simulation model has 11 parameters that can be adjusted. The six parameters held constant in the simulation runs for this paper are number of compartments, number of genomes, number of proteins, maximum number of genomes in a compartment, maximum number of proteins in a compartment, and the exponential factor used to calculate catalytic activities as a function of Hamming distance. The five varying parameters are the four event rate factors and the mutation rate.

Results

Simulations were run with various values of parameters for 1 million generations to determine reasonable ranges of values. These runs were initialized by randomly placing genomes and proteins that exactly match the target sequences. With a mutation rate of one bit average per genome replication, the genome population deteriorated from the target into randomness. At an average of 0.1 bit mutation per genome replication there were indications of stabilization that might continue for more generations. The 0.1 mutation rate was used for all subsequent simulation runs.

As a result of the initial trials, the parameters selected for longer simulations of 1 billion generations were 300 genomes (1500 genes), 15000 proteins, 1500 compartments, a maximum of 2 genomes and 100 proteins in a compartment. The diffusion rates for both genes and proteins were set to a nominal value of 1, and gene replication rate to 0.3. Protein translation rates varied at 1, 3, and 10. The four rate factors have meaning only as values relative to each other.

The longer simulations were run using two different initial conditions. The first starting point consisted of randomly placed target genomes and proteins, as was done for the initial one-million-generation runs described above. These runs explore the characteristics of a mature population, to determine if long term stability appears to be possible for some values of parameters once the target coding has taken hold.

The second set of initial conditions used random sequences for genomes and proteins, except for a single seed genome from the target sequences placed in a randomly chosen compartment. Test runs with no seed genome produced only random results during the 10^{10} event steps that it was practical to run. This is consistent with the very low probability of a random 120 bit genome mutating into a genome for any coding at the mutation rate we were using of 0.1 bit mutation per genome per replication. In various runs a single seed genome placed in a random compartment was used. The seed was the target sequence with, 0, 16, 20, 24, or 32 bits inverted. Even with a 24 bit distance from the target genome, a single seed was usually enough to result in target genomes dominating after about 10^8 event steps. The results with 32 bit distance were not distinguishable from random within the scale of the simulations.

Fig. 1 shows the average Hamming distance of genes from their target sequence over time for a simulation run initialized with target genomes and proteins (“target”), a run initialized with random sequences and one target seed genome (“random”), and a run initialized with random sequences and a seed that is mutated from the target at 16 random bit locations (“random16”). Fig. 2 shows the total catalytic activity of proteins over time for the same three simulation runs. All simulations used the same set of parameters, with a relatively low gene replication rate, a relatively high protein translation rate, and medium diffusion rates. The gene

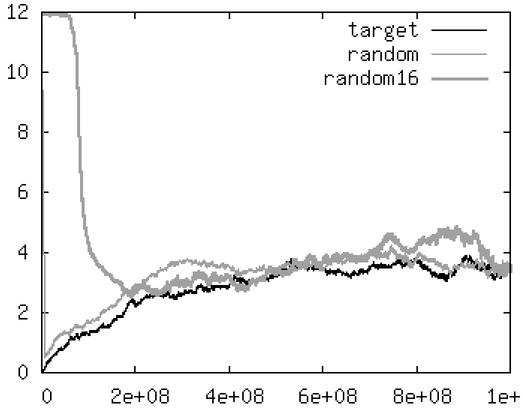

Figure 1: **Average mutation over time.** The average Hamming distance of genes from the target sequences, calculated based on bit differences, as they evolve over 10^9 event steps. Smaller numbers on the Y-axis indicate that genes are closer to the target sequence. The "target" simulation is initialized with all genes at the target, at Hamming distance 0, diverging over time due to mutation. The "random" simulation is initialized with random sequences, with an average Hamming distance of 12, except for one target seed. The "random16" run had one seed that is the target mutated in 16 bit positions. All simulations converge from their opposite directions to a similar apparent attractor state with an average Hamming distance near 3.5. The phase transition in the "random16" simulation after 500 million events is easier to see at this scale than the much earlier transition in the "random" simulation.

mutation curve for "target" starts at 0, increases as genes mutate during replication, then stabilizes after about 300 million event steps. The protein catalytic activity curves show corresponding behavior, starting high, decreasing as proteins are translated from mutated genes, then stabilizing. In the "random" simulation run, a sharp phase transition can be observed after $1.5 x 10^6$ events as the seed genome finally creates or drifts into proximity with enough of the target proteins to produce high local rates of replication and translation in the vicinity of some compartments. A second phase transition can be seen after $5 x 10^7$ events as the activity of the target sequences dominates the system, resulting in values for average mutation and protein activity that are similar to the apparent attractor state of simulation runs that are initialized with perfect target sequences. The "random16" simulation shows similar behavior after a longer initial period. The first phase transition is at almost 10^8 event steps, which is easier to see at the scale of the graph in Fig. 1. All three simulations settle to a similar apparent attractor state by the end of the 10^9 events of these runs. The average Hamming distances of between 3 and 4 are comfortably below the expected average for random sequences.

Stability has not been observed for parameter settings significantly different from those used in the simulations shown in Fig. 1 and Fig. 2. There may be other regions of the parameter space in which stability exists, but our aim was

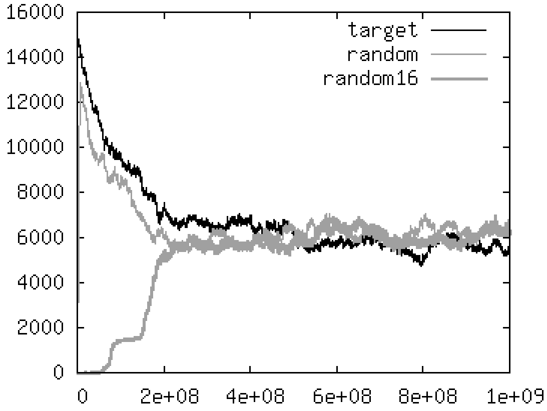

Figure 2: **Catalytic activity over time.** The average translatase catalytic activity in the same simulations used in Fig. 1. Larger numbers on the Y-axis indicate greater translatase activity from proteins that are closer to the target sequence. The curves follow closely those in Fig. 1.

purely to demonstrate the existence of stability in the system. However we did observe lack of stability for certain different parameter settings.

When the gene replication rate is high relative to protein translation, genes mutate faster than mutated genes produce proteins with low replicase catalytic activity. A relatively fast rate of protein translation allows there to be negative feedback from gene mutation to the rate of replication of the mutant genes. A high diffusion rate has the opposite effect, reducing feedback as the proteins created by a gene do not stay near enough to it to reinforce its effect.

Fig. 3 is a visualization of genome activity by compartment across generations for the "target" simulation run. The array of compartments is represented horizontally and time is represented vertically, increasing downward. The top row shows the initial random placement of genomes in compartments. Over time, genomes replicate in compartments that already contain suitable genomes and proteins.

A similar visualization of protein catalytic activity over time (not shown) looks almost identical, as proteins are concentrated in the compartments where they can be produced by translation from suitable genomes.

Fig. 4 shows the genome concentrations in compartments over generations for the seeded "random" simulation run. In the early generations, the genomes are scattered, giving the image the appearance of wisps of thread. The first phase transition occurs where a concentration of genomes first appears. In the second phase transition the narrow concentration of genomes diffuses out along with the proteins that they have produced, to look much more like the equivalent image in Fig. 3.

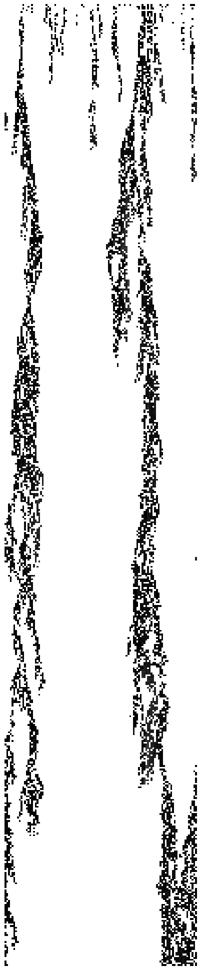

Figure 3: **Gene activity in compartments initialized using target sequences.** Each row of the graph shows which of the 1500 compartments of the simulation were occupied by genes at time *t*. The two sides of the graph are logically connected, as the surface of a cylinder. The top row shows the initial state (event 0). Each row going down the graph represents 0.5 million event steps. Only the first 750 million event steps of the full simulation are shown in this figure. At first many compartments are sites for gene replication and protein synthesis. Catalytic activity can be seen to spread due to diffusion. Eventually more active genomes are seen clumped together as they replicate and produce protein faster than mutated genomes in other regions. The apparent attractor state can be seen with the visual effect of a few narrow streams toward the bottom of the graph.

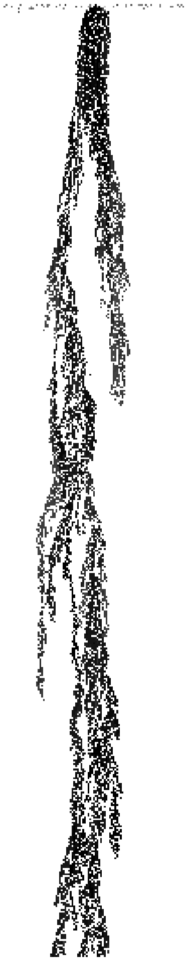

Figure 4: **Gene activity in compartments initialized using random sequences and one target seed.** This is the same visualization as Fig. 3 produced for a simulation initialized with random gene and protein sequences except for a single seed target genome in one compartment. By the end of the graph, the simulation reaches an apparent attractor state similar to the one shown in Fig. 3.

Discussion

As Wills (Wills, 2004) describes, it has been an open problem to demonstrate the establishment and stability of mutually self-sustaining coupled processes of replication and translation when they are error-prone. Our results show such stability starting from an ideal state, and show self-organization into such stability when starting from a state of minimal organization.

The model is simplified in a number of ways that indicate directions for future research. Biological genetic systems have a more complex system of coding. The two-bit codons and amino acids in our model can naturally be extended to more complex ones using stepwise evolution (Wills, 2004). The one-dimensional set of compartments can be extended to two and three dimensions. The diffusion flow model in which there is a fixed constant number of molecules is used to reduce the complexity of the simulation. That can be replaced with a decay rate and a model of resource supply and consumption.

We have demonstrated an autocatalytic system that achieves stability and that self-organizes a coding system that is a start towards that seen in modern genetic systems. It remains to determine more fully specific constraints on stability in such a system, and to demonstrate how such a system can bootstrap itself to greater, biologically plausible levels of complexity.

Supplemental materials

The software used for the simulations in this research will be made available as open source at a later date. Contact the corresponding author for details.

Acknowledgments

PRW is grateful to the University of Auckland for the use of office and library facilities. The authors thank the referees for extensive helpful criticism.

References

Eigen, M. (1971). Self-organization of matter and the evolution of biological macromolecules. *Naturwiss.*, 58:465–523.

Füchslin, R. M. and McCaskill, J. (2001). Evolutionary self-organization of cell-free genetic coding. *Proc. Natl. Acad. Sci. USA*, 98:9185–9190.

Hoffman, G. W. (1974). On the origin of the genetic code and the stability of the translation apparatus. *J. Mol. Biol.*, 86:349–362.

Nieselt-Struwe, K. and Wills, P. R. (1997). The emergence of genetic coding in physical systems. *J. Theor. Biol.*, 187:1–14.

O'Donoghue, P. and Luthey-Schulten, Z. (2003). On the evolution of structure in aminoacyl-trna synthetases. *Microbiol. Mol. Biol. Rev*, 67:550–573.

Wills, P. R. (1993). Self-organization of genetic coding. *J. Theor. Biol.*, 162:267–287.

Wills, P. R. (1994). Does information acquire meaning naturally? *Ber. Bunsenges. Phys. Chem.*, 98:1129–1134.

Wills, P. R. (2004). Stepwise evolution of molecular biological coding. In Pollack, J., Bedau, M., Husbands, P., Ikegami, T., and Watson, R. A., editors, *Artificial Life IX: Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*, pages 51–56. Cambridge, MA: MIT Press.