Spatial Models of Persistence in RNA Worlds: Exploring the Origins of Life

William A. Maniatty[1], Thomas Caraco[2], Niles Lehman[3], and Boleslaw K. Szymanski[3]

[1] Department of Computer Science, University at Albany, Albany, NY 12222
maniatty@cs.albany.edu, http://www.cs.albany.edu/~maniatty/
[2] Department of Biological Sciences, University at Albany, Albany, NY 12222
caraco@albany.edu, http://www.albany.edu/faculty/caraco/index.html
[3] Department of Chemistry,, Portland State University, Portland, OR 97207-0751
niles@pdx.edu, http://chem.pdx.edu/faculty/lehman.htm
[4] Computer Science Dept., Rensselaer Polytechnic Institute, Troy, NY 12180
szymansk@cs.rpi.edu, http://www.cs.rpi.edu/~szymansk/

Abstract. Prevailing hypotheses concerning origins of life assume that rare configurations of prebiotic polymers allowed their accurate replication, the precursor of biological reproduction. In particular, the "RNA world" hypothesis equates the linear sequence of an RNA's nucleotides with its genotype. Individual RNA molecules may fold into different 3-dimensional structures, or conformers. Each such folding of a given genotype specifies a different phenotype. These phenotypes exhibit chemical properties that may result in replication of the genotype. Each phenotype has a fitness depending on replication rate, and so the phenotype distribution would be subject to natural selection. To model replication and extinction of prebiotic polymers, we combine computational and biological approaches. We consider a single genotype that can fold into two different phenotypes. Each phenotype's capacity to replicate its genotype depends on abiotic and biotic factors in the physical environment, which may change as time advances. We begin with a spatially detailed, individual based model, required for accurate modeling of small populations where the variability caused by random events among individual replicators dominates population dynamics. For efficient modeling of large populations where mean behavior tends to dominate, we derive a corresponding mean-field model that aggregates large, well mixed populations of common phenotypes to compare its behavior to the individual based version.

1 Introduction

The origins of life can be characterized as a process where the diversity of molecular replicators attains a quantitative permanence [2, 12]. Many biologists currently subscribe to the RNA world hypothesis, which (informally) states that sequences of simpler molecules were assembled over a sequence of steps, eventually leading to RNA replicator sequences, that in turn resulted in cellular organisms. Consider a system of catalytic polymers with heritable sequential structure.

System diversity should be bounded above by resource competition and the resulting selective evolutionary processes, and bounded below by replication error and functional variability of individual replicator sequences. Persistence of replicator diversity for an arbitrarily long period, a statistical permanence, should promote increases in biochemical complexity leading to cellular life. We hypothesize that rare events, i.e. the appearance of improbable genotypes or changes in abiotic conditions can influence whether such permanence can be achieved.

Evolutionary dynamics are governed by two biological processes: selection which promotes fit competitors over the unfit, and mutation which introduces diversity of genotypes. However, before evolution can take place, a minimum level of fitness must be met, that is the genotype must be able to avoid extinction in its environment in the absence of competition. At the molecular level, there is little prior art on modeling the one-to-many mapping from genotype to phenotype [1], and we are not aware of approaches that take the environment's spatial structure into account. We begin by considering the simplest case, where a single non-mutating genotype induces two phenotypes (by folding), and observe the population dynamics of the system as biotic and abiotic factors vary. Our approach uses multi-scale modeling, with two different spatially explicit models. The first model is a spatially explicit individual based model, using a generalized cellular automaton [7], which tends to be more accurate in modeling systems with a high level of spatial heterogeneity or small populations where variability induced by random localized interactions tends to dominate the population dynamics. The cellular automata (CA) model will let us explore (1) how spatially structured replicator interactions promote or reduce local and global replicator diversity, (2) how diversity patterns differ between 2 and 3-dimensional RNA worlds, and (3) how exogenous temporal variation (disturbance frequency) influences persistence of both replicator diversity and types of ecological interactions between replicators. The second model addresses a more extended scale. Our formulation assumes a series of spatially segregated replicator communities periodically linked by partial mixing between neighbors. This model of an RNA-world metapopulation will let us explore (1) how the degree of mixing governs within- and between-population components of replicator diversity, and (2) how temporal variation in the environment (e.g. cation concentration) influences within- and between-group selective evolution, and the resulting permanence of global diversity.

2 Biological Background

One goal of our models is to understand the maintenance of replicator diversity, both genotype and phenotype, and the increase in complexity of interactions between replicator types in an RNA-world. We want to examine spatial structure at the scale of replicator interaction neighborhoods [3], and at the metapopulation scale. Spatial structure can affect extinction times (transients), and can influence the relative significance of selection at different levels of organization [9]. Finally, temporal variation in vitro (cation concentration [11]) can change or even reverse

fitnesses of different genotypes. We want to address these topics separately, and then in combination. The recently-described "continuous" methodology for in vitro RNA evolution [15] allows for dozens of iterated selection and amplification steps to be done within hours, rather than the days required by the earlier "stepwise" evolution protocols [4, 5, 10, 14] which needed laborious and slow purification steps between each round of RNA amplification to maintain sustained growth and evolution of an RNA population. Our modeling assumptions are designed to be consistent with this approach.

3 Modeling Persistence in RNA Worlds

Recall that we are beginning our investigations by considering a single genotype with multiple phenotypes, and in particular we consider the simplest form, analyzing extinction trends in a system governed by the interaction between two phenotypes. This is similar to our approaches to modeling selection in [6, 13], in the future we plan to extend our approach to model mutation, and selection much like in [8]. We consider two different models of selection in RNA worlds, first beginning with a two dimensional spatially explicit individual based model, and then exploring a spatially explicit aggregation based model.

3.1 An Individual Based Modeling Approach

The individual based model partitions a two-dimensional space into a lattice of $J \gg 1$ *sites* and synchronously advances time using uniformly sized time steps, like [6, 13]. It is assumed that RNA replicator molecules in this model have the same sequence (i.e. are of a common genotype G), and can belong to one of two phenotypes. Sites in this model are sufficiently small that they can hold at most one replicator. Uninhabited sites are partitioned into two groups, those that have sufficient resources (e.g. nucleotides and substrate as per [11, 15]) for replication (in state r) and those that lack the resource required for replication (in state ϕ). Sites with resources can become depleted (e.g. due to chemical decay) with probability μ_r and sites with insufficient resources can be replenished with probability ξ (due to chemical reactions or mixing).

Our initial model assumes that all replicators share the same genotype, g. Let p_h represent the phenotype of a replicator molecule with genotype g, so that the range of phenotypes assumed is $p_h \in \{p_0, p_1, \ldots, p_{n-1}\}$, for simplicity we begin with a model where $n = 2$, so there is a single genotype with two phenotypes. An abiotic factor for the entire environment (e.g. cation concentrations) is represented by the random variable $E(t) \in \{0, 1\}$. The value of $E(t)$ indicates which phenotype can reproduce at time t, and $E(t + 1)$ is computed using a stationary Markov chain, so that $E(t + 1) = 1 - E(t)$ with probability $\theta_{E(t)}$, otherwise $E(t+1) = E(t)$. About each site x is an interaction neighborhood, σ_x, the area of the interaction neighborhood is $\delta = |\sigma_x|$. The density of replicators of phenotype $p_h, h \in \{0, 1\}$ residing in σ_x is denoted d_h. Phenotype removal (death) is assumed to be local, with $\mu_h, h \in \{0, 1\}$ being the mortality rate of
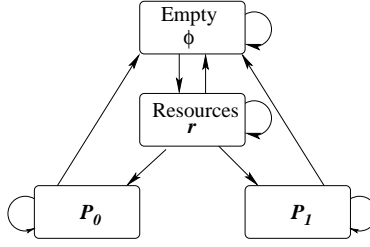
Fig. 1. Local State Transitions in the Individual Based Model

phenotype p_h. Thus the local transition probabilities from some state i to state j are denoted $t_{i,j}, i, j \in \{r, \phi, 0, 1\}$, correspond to the arcs in Figure 1. At each time step, each site does a Bernoulli trial to compute its future state, using the following local transition probabilities:

$$t_{\phi,r} = \xi$$
$$t_{\phi,\phi} = 1 - \xi$$
$$t_{r,\phi} = \mu_r$$
$$t_{r,0} = \pi_0(1 - \mu_r)[((1 - E)d_0) + (Ed_1)]$$
$$t_{r,1} = \pi_1(1 - \mu_r)[((1 - E)d_0) + (Ed_1)]$$
$$t_{r,r} = 1 - \mu_r - (1 - \mu_r)[((1 - E)d_0) + (Ed_1)]$$
$$t_{0,\phi} = \mu_0$$
$$t_{0,0} = 1 - \mu_0$$
$$t_{1,\phi} = \mu_1$$
$$t_{1,1} = 1 - \mu_1$$

3.2 An Aggregated Modeling Approach

The individual based approach captures dynamics at a microscopic level, however for computational efficiency in modeling larger populations aggregation is needed, leading to a spatially segregated replicator community model. The three-dimensional space is partitioned into regions, with homogeneous mixing assumed in each region. Flows exchange solution containing replicators and resources between adjacent regions. By taking the lower limit on the number of partitions, all the space could be put in the same partition, resulting in a spatially homogeneous model.

For this model we will measure the concentration of each type of molecule in the solution of region i, denoted $0 \leq C_{r,i}, C_{0,i}, C_{1,i} \leq 1$. These concentrations are normalized to reflect the "carrying capacity" induced by environmental conditions. For consistency with the individual based model of Section 3.1, we impose $0 \leq C_{r,i} + C_{0,i} + C_{1,i} \leq 1$. The notation of the aggregated model is described in Table 1. Migration is modeled as flows of solution across region surfaces, edges

Symbol	Meaning
$C_{s,i}\, s \in \{0, 1, r\}, 0 \leq i < J$	Concentration of molecules in state s at site i
$E \in \{0, 1\}$	Environmental factor permitting $p_{0,E}$ to reproduce and excluding $p_{0,(1-E)}$
$F_{i,j}$	Flow of solution from aggregate site i to j
J	the number of aggregated sites in the lattice
$\mu_s,\, s \in \{0, 1, r\}$	Removal (mortality rate) of a type of molecule
$\pi_h,\, h \in \{0, 1\}$	Probability of a new RNA molecule assuming phenotype h
$\sigma_i, 0 < i < J$	Iteration neighborhood about site i.

Table 1. Notation Used in the Aggregate Model

and vertices to the nearest neighbors, carrying replicators and resources needed for replication. Solution is assumed to be conserved, thus, for each region, i, the inflow must match out-flow, that is: $\sum_{i \in \sigma_i} F_{i,j} = \sum_{j \in S | i \in \sigma_j} F_{j,i}$. For notational convenience, for region i, we denote the available capacity of the solution to hold more replicators or resources needed for their reproduction as A_i and the rate at which new replicators are generated in region i, N_i. For site i we derive:

$$A_i = 1 - C_{0,i} - C_{1,i} - C_{r,i}$$

$$N_i = C_{r,i} A_i [(1 - E)C_{0,i} + EC_{1,i}]$$

$$\frac{dC_{r,i}}{dt} = \xi A_i + \left(\sum_{j \in \{\sigma_i | F_{j,i} > 0\}} F_{j,i} C_{r,j} \right) - N_i - C_{r,i} \left(\mu_r + \sum_{j \in \{\sigma_i | F_{i,j} > 0\}} F_{i,j} \right)$$

$$\frac{dC_{0,i}}{dt} = \pi_0 N_i + \left(\sum_{j \in \{\sigma_i | F_{j,i} > 0\}} F_{j,i} C_{0,j} \right) - C_{0,i} \left(\mu_0 + \sum_{j \in \{\sigma_i | F_{i,j} > 0\}} F_{i,j} \right)$$

$$\frac{dC_1}{dt} = \pi_1 N_i + \left(\sum_{j \in \{\sigma_i | F_{j,i} > 0\}} F_{j,i} C_{1,j} \right) - C_{1,j} \left(\mu_1 + \sum_{j \in \{\sigma_i | F_{i,j} > 0\}} F_{i,j} \right)$$

4 Results of Simulations

In an experiment, we observed the behavior of our model according to the controls shown in Table 2. Our experiment focuses on exploring the impact of varying the per reproductive event probability, π_0, of having an offspring of phenotype p_0 on phenotype concentrations, C_0 and C_1. The per time step probability of the environment changing to favor the other genotype was held constant, that is $\theta_0 = \theta_1$. As a control, mortality was assumed to be identical between phenotypes, $\mu_0 = \mu_1$. Initially 1% of the nodes in both the individual based model and the cubic three-dimensional coupled map lattice models were populated with phenotype $h = 0$ and $h = 1$, positioned to be both equidistant. Since the laboratory systems we model have impermeable boundaries (the walls of a test tube and air), we did not allow flows across lattice boundary surfaces, edges or vertices.

Parameter	Values Used
$\pi_0 = 1 - \pi_1$	$0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$
δ	$3^2, 11^2, 33^2, 101^2$ (Individual Based Model only)
$\theta_0 = \theta_1$	$0.001, 0.01, 0.1, 0.25, 0.5, 1.0$
$\mu_0 = \mu_1$	$0.001, 0.005, 0.01, 0.05, 0.1$
μ_r	0.1
ξ	0.1

Table 2. Parameters Used in Experimentation

One particular case that we are interested in is when a rare event triggers a change in the outcome of the simulation. One example of a rare event is a change in the abiotic conditions (e.g. when the value of E changes). This sort of rare event was observed under the condition of $\theta_0 = \theta_1 = 0.001$, which is expected behavior given the formulation of the models. The aggregate based model had $E = 0$ for early time steps, $0 \leq t < 92$, after which E was set to 1 for the remainder of the simulation, $92 \leq t \leq 2000$. As a control, we coerced the individual based model to have same timing for the change of E as the aggregation based model. For individual based models with large interaction neighborhoods, we expected to have a close approximation to both aspatial models and spatially segregated replicator community models. As shown in Figure 2(a) and (b), the population trends are similar. Interestingly, the rare event described above triggered an extinction of the replicator when $\pi_0 \in \{0.8, 0.9\}$, which seems to occur when the environment was saturated with phenotype $h = 0$ prior to the change in E and denying phenotype $h = 1$ locations suitable for establishment. After the change in E, phenotype $h = 1$ is thus unable to spread and dies out, while the now sterile phenotype $h = 0$ has used the majority of the systems resources needed for reproduction. For the individual based model, when a large interaction neighborhood is employed, the system approximates homogeneous mixing, and has similar qualitative behavior (but slightly different steady state populations), as seen in Figure 2(e) and (f). Smaller interaction neighborhoods tended to reduce stable state density, and promote extinction as seen in Figures 2(c) and (d).

5   Summary and Conclusion

We present a novel approach combining theory, simulation and experimentation to provide a preliminary model of RNA worlds. Following the principle of step-wise refinement, we begin with the fundamental phenomena of persistence. The novel aspects of this approach include addressing spatial effects and many-to-one mapping from genotypes to phenotypes at the molecular level. These models may accelerate experimental breakthroughs and interpretation of results, by avoiding blind experimentation. First, we plan to extend the individual based model to three dimensions. We then want to extend the individual based and aggregate
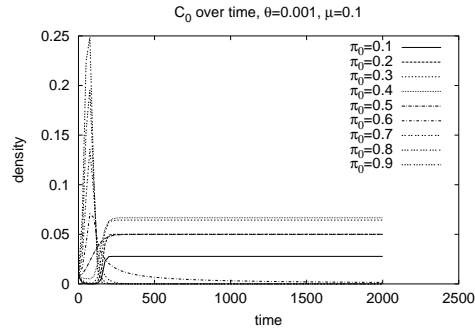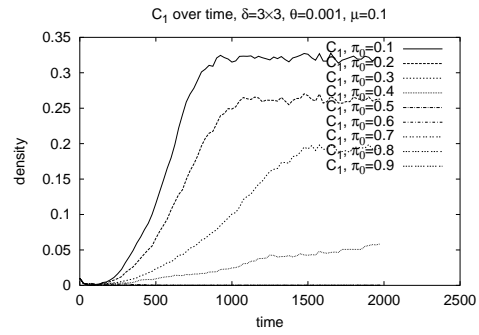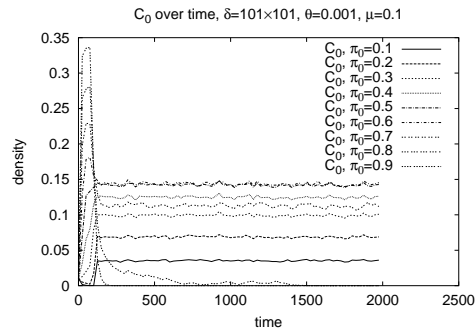
(a) C_0 as a function of time
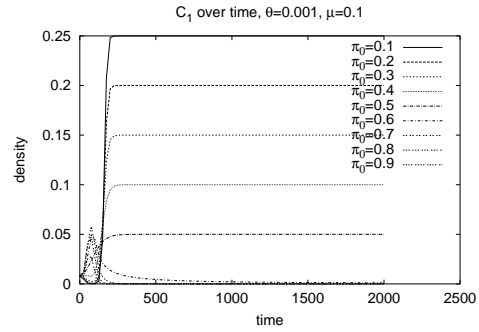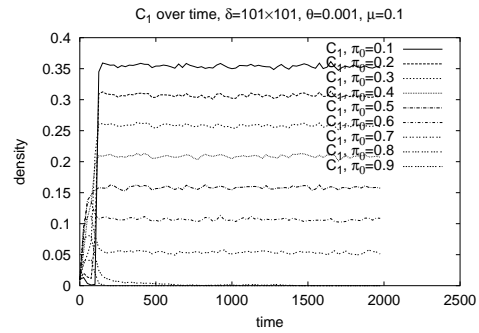

(b) C_1 as a function of time
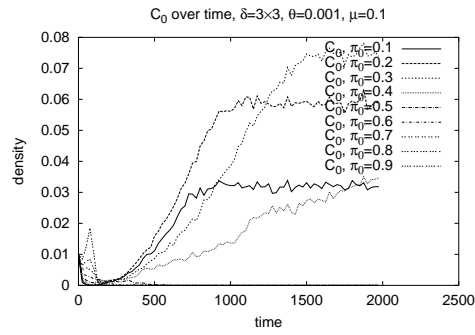
(c) C_0 as a function of time, $|\delta| = 3^2 - 1$

(d) C_1 as a function of time, $|\delta| = 3^2 - 1$

(e) C_0 as a function of time, $|\delta| = 101^2 - 1$

(f) C_1 as a function of time, $|\delta| = 101^2 - 1$

Fig. 2. Individual Based Model: Phenotype Concentration as a function of time

models to explore: (i) the impact on persistence if folding is sensitive to abiotic parameters (i.e. make π_0 and π_1 functions of E) and (ii) continuous abiotic factors (i.e. let $0 \leq E \leq 1$). Some open issues include modeling (i) selection between competing genotypes, (ii) relaxed synchronization of the models, (iii) model integration across different spatial scales and (iv) enabling mutation of genotypes.

References

1. L. W. Ancel and Fontana W. Plasticity, evolvability, and modularity in RNA. *Journal of Experimental Zoology*, 288(3):242–283, 2000.
2. C. Blomberg. On the appearance of function and organisation in the origin of life. *J. Theoretical Biology*, 187:541–554, 1997.
3. M. C. Boerlijst and P. Hogeweg. Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. *Physica D*, 48:17–48, 1991.
4. B.M. Chowrira and A. Berzal-Herranz. Novel guanosine requirement for catalysis by the hairpin ribozyme. *Nature*, 354:320–322, 1991.
5. R. Green, A.E. Ellington, and J.W. Szostak. In vitro selection of RNA molecules that bind specific ligands. *Nature*, 346:818–822, 1990.
6. W. A. Maniatty, B. K. Szymanski, and T. Caraco. High-performance simulation of evolutionary aspects of epidemics. In *Proceedings of the PARA98 Workshop on Applied Parallel Computing in Large Scale Industrial and Scientific Problems*, volume 1541 of *Lecture Notes in Computer Science*, pages 322–331, Umeå University, Umeå Sweden, June 1998. Springer-Verlag.
7. W. A. Maniatty, B. K. Szymanski, and T. Caraco. Parallel computing with generalized cellular automata. *Parallel and Distributed Programming Practices*, 1(1):85–104, January 1998. Also Technical Report 97-3 Department of Computer Science, Rensselaer Polytenchic Institute, Troy, NY 12180.
8. W. A. Maniatty, B. K. Szymanski, and T. Caraco. High-performance computing tools for modeling evolution in epidemics. In *Proc. 32nd IEEE Annual Hawaii International Conference on System Sciences, HICSS'99*. IEEE, January 1999. CD ROM Disk, 10 Pages (Abstracts pp. 318-319).
9. R. E. Michod. *Darwinian Dynamics: Evolutionary Transitions in Fitness and Individuality*. Princeton Univ Press, Princeton, NJ, USA, 1999.
10. D.R. Mills, R.L. Peterson, and S. Spiegelman. An extracellular darwinian experiment with a self-duplicating nucleic acid molecule. *Proc. Natl. Acad. Sci. USA*, 58:217–220, 1969.
11. T. Schmitt and N. Lehman. Non-unity molecular heritability demonstrated by continuous evolution *in vitro*. *Chemistry and Biology*, 6(12):857–869, December 1999.
12. E. Szathmary and J. Maynard Smith. From replicators to reproducers: the first major transitions leading to life. *J. Theoretical Biology*, 187, 555-571.
13. B. K. Szymanski and T. Caraco. Spatial analysis of vector-borne disease: A four species model. *Evolutionary Ecology*, 8:299–314, 1994.
14. C. Tuerk and L. Gold. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249:505–510, 1990.
15. M.C. Wright and G.F. Joyce. Continuous in vitro evolution of catalytic function. *Science*, 276:614, 1997.