DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

Originator Dynamics

(Article begins on next page)

Originator dynamics

Michael Manapat [a,b], Hisashi Ohtsuki [c], Reinhard Bürger [d], Martin A. Nowak [a,e,f,*]

[a] Program for Evolutionary Dynamics, Harvard University, Cambridge, MA 02138, USA
[b] Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[c] Department of Value and Decision Science, Tokyo Institute of Technology, Tokyo 152-8552, Japan
[d] Department of Mathematics, University of Vienna, 1090 Vienna, Austria
[e] Department of Mathematics, Harvard University, Cambridge, MA 02138, USA
[f] Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

ABSTRACT

We study the origin of evolution. Evolution is based on replication, mutation, and selection. But how does evolution begin? When do chemical kinetics turn into evolutionary dynamics? We propose "prelife" and "prevolution" as the logical precursors of life and evolution. Prelife generates sequences of variable length. Prelife is a generative chemistry that proliferates information and produces diversity without replication. The resulting "prevolutionary dynamics" have mutation and selection. We propose an equation that allows us to investigate the origin of evolution. In one limit, this "originator equation" gives the classical selection equation. In the other limit, we obtain "prelife." There is competition between life and prelife and there can be selection for or against replication. Simple prelife equations with uniform rate constants have the property that longer sequences are exponentially less frequent than shorter ones. But replication can reverse such an ordering. As the replication rate increases, some longer sequences can become more frequent than shorter ones. Thus, replication can lead to "reversals" in the equilibrium portraits. We study these reversals, which mark the transition from prelife to life in our model. If the replication potential exceeds a critical value, then life replicates into existence.

© 2008 Elsevier Ltd. All rights reserved.

1. Introduction

The attempt to understand the origin of life has inspired much empirical and theoretical work over the years. Some basic building blocks of living systems can be produced under simple, prebiotic conditions (Miller, 1953; Szostak et al., 2001; Benner et al., 2002; Ricardo et al., 2004), although many questions are still unanswered (Shapiro, 2006). RNA can store genetic information and catalyze certain reactions (Ellington and Szostak, 1990; Cech, 1993; Sievers and von Kiedrowski, 1994; Ferris et al., 1996; Joyce, 1989, 2002; Johnston et al., 2001). This idea has led to the hypothesis of an RNA world, where both genetics and metabolism are governed by RNA alone. In one experiment, Bartel and Szostak (1993) isolated a ribozyme from a large number of randomly generated sequences. This ribozyme can ligate two RNA molecules that are aligned on a template. Subsequent experiments have resulted in the shortening of this ribozyme and the enhancement of its catalytic activity (Steitz and Moore, 2003).

Eigen (1971) and Eigen and Schuster (1977, 1979) have introduced a chemical theory for the origin of life. They study populations of binary sequences under mutation and selection. A central result of their celebrated "quasispecies theory" is the error threshold: adaptation is only possible if the mutation rate per bit is less than the inverse of the sequence length (Swetina and Schuster, 1982; McCaskill, 1984; Eigen et al., 1988; Nowak and Schuster, 1989; Nowak, 1992). They also propose the hypercycle as a concept for the evolution of further complexity.

Kauffman (1986, 1983) investigates catalytic protein networks in the context of the origin of life. Szathmáry and Demeter (1987) study replicating units within randomly dividing vesicles. Fontana and Buss (1994a, b) use the λ-calculus as a tool for investigating how a generative chemistry can lead to biological organization and evolution. For further theoretical approaches to questions concerning the origin of life, see Dyson (1982, 1999), Stein and Anderson (1984), Maynard Smith and Szathmáry (1995), and Segre et al. (1998, 2000).

Evolutionary dynamics need populations of individuals that are capable of reproduction and inheritance. Offspring inherit some information from their parents. If the term "replication" is used specifically to denote the reproduction of information, then biology becomes the study of replication.

Mutation arises if replication is not perfectly accurate but can also generate mistakes. Selection emerges if mutants differ in their replication rate, which is also called their "fitness." Thus, evolutionary dynamics are based on replication, mutation, and selection.

It is generally assumed that mutation and selection are consequences of replication, but here we want to challenge this perspective. We propose a generative chemistry (prelife) that is capable of mutation and selection prior to replication. We study how selection can favor (or oppose) replication.

Consider a binary soup of activated monomers, 0^* and 1^*. These monomers are produced and removed at certain rates. In addition, the following chemical reactions are possible:

$$i + 0^* \to i0,$$
$$i + 1^* \to i1. \tag{1}$$

Here i denotes any binary sequence (including the null sequence). We make several assumptions about this chemical system. First, we assume that some buffering mechanism exists so that the concentrations of the activated monomers are always at a fixed steady-state level. Second, we assume that elongation can occur in only one direction: sequence i can become $i0$ or $i1$. This is analogous to the polymerization of DNA where a new base can only be added to the $3'$ end. Under this assumption, each string has a unique chemical precursor and exactly two successors. For example, the precursor of sequence 001 is 00 and its two successors are 0010 and 0011. In particular, each string has a unique production lineage: the lineage of 0010 is $0 \to 00 \to 001 \to 0010$. Finally, we assume that strings of all lengths are removed at a fixed "death-rate" d. One can view sequence death as the degradation of the string into its constituent monomers, which are then absorbed into the buffered pool of activated monomers.

The dynamics within this binary soup are described by the following system of infinitely many differential equations, where the abundance of sequence i is given by

$$\dot{x}_i = a_i x_{i'} - (a_{i0} + a_{i1} + d)x_i, \quad i = 0, 1, 00, 01, \ldots. \tag{2}$$

The parameter a_i denotes the rate constant of the chemical reaction which produces sequence i from its precursor i'. For the abundances of the precursors of 0 and 1, we set $x_{0'} = x_{1'} = 1$. All sequences are removed at rate d. The above system converges to a unique equilibrium where, typically, longer sequences are exponentially less common than shorter ones. For a discussion of prevolutionary dynamics, see Nowak and Ohtsuki (2008).

Let us now assume that (some) sequences can reproduce. In the simplest scenario, they use activated monomers to make copies of themselves. This is called "direct replication" rather than replication via the complimentary bitstring (Eigen and Schuster, 1979). Suppose the relative replication rate (i.e., the fitness) of sequence i is given by f_i. We then have the following "originator equation":

$$\dot{x}_i = a_i x_{i'} - (a_{i0} + a_{i1} + d)x_i + r x_i (f_i - \phi), \quad i = 0, 1, 00, 01, \ldots. \tag{3}$$

The parameter r determines the relative magnitude of replication and prelife dynamics. It could depend on the supply of activated monomers or other chemical and physical properties of the system, such as the temperature. In the limit $r \to 0$, the originator equation describes prelife, (2). In the limit $r \to \infty$, we obtain the standard selection equation of evolutionary dynamics (Nowak, 2006). The parameter ϕ is chosen so that the total population size is constant. Without loss of generality we set $\sum_i x_i = 1$, so x_i denotes the frequency of sequence i. Since we must require

that $\sum \dot{x}_i = 0$, we obtain

$$\phi = \bar{f} + \frac{a_0 + a_1 - d}{r}. \tag{4}$$

The average fitness of the population is $\bar{f} = \sum_i f_i x_i$. If we set $a_0 + a_1 = d$, then $\phi = \bar{f}$.

In this paper, we will study some aspects of system (3), but we will mostly investigate a somewhat simpler equation which shares many properties with (3). This unary originator equation has the form

$$\dot{x}_i = a_{i-1} x_{i-1} - (a_i + d)x_i + r x_i (f_i - \phi), \quad i = 1, 2, 3, \ldots. \tag{5}$$

Here x_i is the abundance of the sequence of length i. As before, sequences grow on one side by the addition of activated monomers, but now there is only one type of monomer, 0^*. Thus we study the unary sequences $0, 00, 000, \ldots$. For the abundance of the precursor of 0, we set $x_0 = 1$. We can interpret the unary model as a binary model where all the properties of a sequence depend only on its length: the abundance x_i in the unary model corresponds to the total abundance of all strings of length i in the binary model. In addition, if we assume that all sequences of a given length and greater have the same fitness, the unary originator equation (5) can be written as a quasispecies equation with a very special mutation–selection matrix. The details of this reduction are described in Appendix A.

We will investigate the equilibrium structure of (5) as function of the replication potential, r. If r is less than a critical value, then the basic equilibrium structure of prelife prevails. Our focus will be on the "supersymmetric" case given by $a_i = a$ for $i > 0$ and $a_0 = d$, which implies that $\phi = \bar{f}$. Note that ϕ is a function of x_i and not a constant. For $r = 0$ (prelife), longer sequences are always less frequent than shorter ones. But as r increases there can be "reversals" in the equilibrium portrait after which some sequences become more frequent than their precursors.

The structure of this paper is as follows. In Section 2, we introduce the unary model and establish lower bounds for the values of the replication potential, r, at which reversals can occur. In Section 3, we compute an intersection point exactly for a certain fitness landscape. In Section 4, we study another class of fitness landscapes, show that the reversal points are given (approximately) by the roots of a simple polynomial, and then use this fact to derive a result on the replication potential required for the fittest sequence to dominate. In Section 5, we comment on the binary model, and in Section 6 we draw some conclusions.

2. The unary model

We study system (5) where $x_0 = 1$, $a_0 = d$, and $a_i = a$ for all $i > 0$. In this supersymmetric case, the system becomes

$$\dot{x}_1 = d - (a + d)x_1 + r x_1 (f_1 - \phi),$$
$$\dot{x}_i = a x_{i-1} - (a + d)x_i + r x_i (f_i - \phi), \quad i = 2, 3, \ldots,$$
$$\phi = \sum_{i=1}^{\infty} f_i x_i. \tag{6}$$

Here x_i denotes the frequency (i.e., the relative abundance) of the sequence of length i. We will assume that all sufficiently long sequences have the same fitness. This assumption allows us to perform exact numerical simulations and also guarantees that the system has a unique, globally stable equilibrium in the infinite simplex $\sum_i x_i = 1$, $x_i \geqslant 0$ for all i. Appendix A has a detailed discussion of how we simulate the system along with a discussion of its stability properties.

We will focus on the "reversal points" in the equilibrium portraits of these systems: for each value of r, we plot the equilibrium frequencies, $x_i^*(r)$. As r varies, these frequencies can
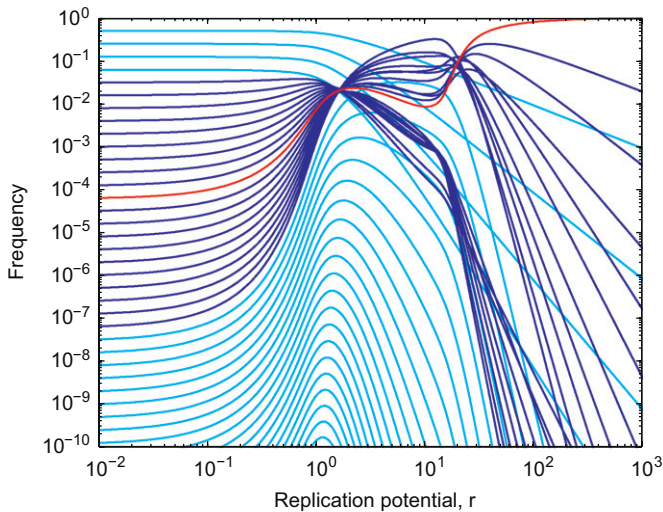
Fig. 1. The equilibrium portrait of the unary originator equation with $a = d = 1$. Sequences 5–24 (dark blue) have fitness values sampled from the uniform distribution on $[1, 1.1]$. All other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. The red curve is the graph of $x_{14}^*(r)$, the equilibrium frequency of sequence 14, which is the fittest sequence in this landscape. Surprisingly, the frequency of the fittest sequence does not increase monotonically: between $r \approx 1$ and ≈ 10, $x_{14}^*(r)$ actually decreases.

change order, and the values of r at which they do so are what we call reversal points. We are interested in these points because they mark the transition from prelife dynamics to evolutionary dynamics in the originator equation. Before a reversal, the chemically-determined structure of prelife is dominant; after a reversal, replication dominates.

Fig. 1 shows the equilibrium portrait of a random fitness landscape: sequences of length 5–24 have fitness values chosen from the uniform distribution on $[1, 1.1]$. The sequence of length 14 (in red) has the highest fitness. We will find lower bounds for the values of r at which reversals such as those near $r = 1$ can occur. This will allow us to make statements about necessary conditions on replication rates for life to dominate prelife.

The equilibrium structure in supersymmetric prelife is characterized by an inverse relationship between frequency and sequence length: longer sequences are less common than shorter ones. When some sequences have positive fitness, however, they may become more abundant than shorter sequences if r is large enough. In Fig. 1, the sequence of length 14 has the highest fitness and eventually becomes the most frequent sequence. However, its frequency does not increase monotonically as a function of r. From $r \approx 1$ to ≈ 10, the frequency of the fittest sequence actually decreases.

We start by establishing a lower bound for the values of r at which reversals can occur. In the following, we will speak of "sequence i," by which we mean the sequence of length i, the "frequency of i," by which we mean the equilibrium frequency of sequence i, and of "frequencies" generally, by which we mean the equilibrium frequencies of the sequences as functions of the parameter r. When we describe the behavior of sequences (e.g., in phrases such as "sequence i crosses sequence j at r_0"), we are referring to the equilibrium frequencies of the sequences as functions of r.

2.1. General lower bounds for reversals

We begin with the basic requirements that $f_i \geqslant 0$ for all i, at least one sequence has positive fitness, and all sufficiently long

sequences have the same fitness. These conditions ensure that (6) has a unique, globally stable equilibrium. Observe that the first intersection of sequences must involve the intersection of a sequence and its precursor. Therefore, it is enough to find a value of r below which a sequence and its precursor cannot intersect. From (6), we have at equilibrium

$$x_{i+1}^* = \frac{a}{(a+d) - r(f_{i+1} - \phi^*)} x_i^*, \tag{7}$$

where $\phi^* = \sum f_i x_i^*$. If $x_{i+1}^* = x_i^*$, we must have

$$\frac{a}{(a+d) - r(f_{i+1} - \phi^*)} = 1. \tag{8}$$

This equality holds if $r(f_{i+1} - \phi^*) = d$, and for this we must have $f_{i+1} > \phi^*$. These conditions become

$$r = \frac{d}{f_{i+1} - \phi^*} > \frac{d}{f_{i+1}}. \tag{9}$$

Thus, if the frequencies of i and $i+1$ intersect, they do so only if $r > d/f_{i+1}$.

We can justify this result intuitively as follows: each sequence has an input flow from its precursor (rate a), an output flow to its successor (rate a), and a death rate d. The absolute replication rate, rf_{i+1}, of sequence $i+1$ must exceed the death rate, d, to make net replication possible; the result is the bound $r > d/f_{i+1}$.

We note that (9) implies that a nonreplicating sequence, $i+1$, will never cross its precursor, i: if $f_{i+1} = 0$, the bound becomes infinity.

Taking into account all pairs of frequencies, we find that an intersection between any sequences can happen only if

$$r > \min_i \frac{d}{f_i}. \tag{10}$$

If r is less than the quantity on the right, then the equilibrium structure of prelife is maintained; as r increases beyond that value, we can have a transition after which the balance is tipped in favor of life (though nonreplicating sequences may still be more frequent than replicating ones). We collect the results above in the following theorem.

Theorem 1. *Let $\{f_i\}_{i=1}^{\infty}$ be a fitness landscape such that $f_i \geqslant 0$ for all i, $f_j > 0$ for some j, and $f_k = f$ for all sufficiently large k (f a fixed nonnegative number). Then*

1. *The equilibrium frequencies $x_i^*(r)$ and $x_{i+1}^*(r)$ do not intersect for $r \leqslant d/f_{i+1}$.*
2. *A nonreplicating sequence never intersects its precursor (the sequence of length one less).*
3. *There are no intersections between any sequences for $r \leqslant \min_i d/f_i$.*

Theorem 1 gives necessary but not sufficient conditions on r for reversals to occur. A reversal can fail to occur, even if r is sufficiently large, in one of two ways. First, the sequences can fail to intersect at all. In Section 2.3, we give an example of a fitness landscape in which many sequences can replicate but for which there are no intersections. Second, even if a group of sequences meets at a certain point, their order does not necessarily have to change after that point—the frequencies can be mutually tangent. Equality (8) is technically only a condition for intersection and not one for reversal. However, we have not observed, numerically, any situations in which two frequencies are tangent.

2.2. Sequences m to n replicate

Suppose sequences m to n have fitness $f > 0$ and all other sequences have fitness 0. Fig. 2 shows the equilibrium portrait for
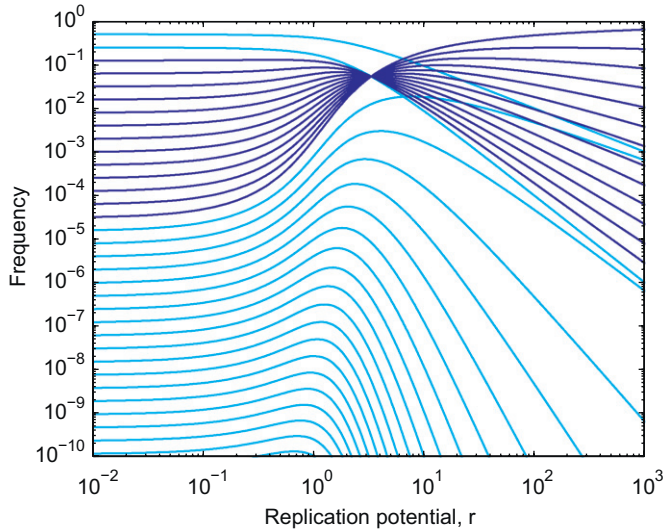
Fig. 2. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. Sequences 3–15 (dark blue) have fitness 1, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. The equilibrium frequencies of sequences 2–15 all intersect at $r \approx 3.31$. In general, if sequences m to n have the same fitness f and all other sequences have fitness 0, then the frequencies of sequences $m-1$ to n will intersect for the first time in a single point.

Fig. 3. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. Sequences 3–8 and 15–20 (dark blue) have fitness 1, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. The frequency of sequences 2–8 all intersect at $r \approx 2.27$ as do the frequencies of sequences 14–20 since all replicating sequences have the same fitness. In general, if sequence i has fitness f and $x_i^* = x_{i-1}^*$ at $r = r_0$, then $x_j^* = x_{j-1}^*$ at $r = r_0$ for all sequences j with fitness f.

$m = 3$, $n = 15$, and $f = 1$. Surprisingly, the frequencies of sequences $m - 1$ ($= 2$) to n ($= 15$) intersect in a single point at $r \approx 3.31$. To see this, observe that (8) implies that sequence n crosses sequence $n - 1$ at values of r satisfying

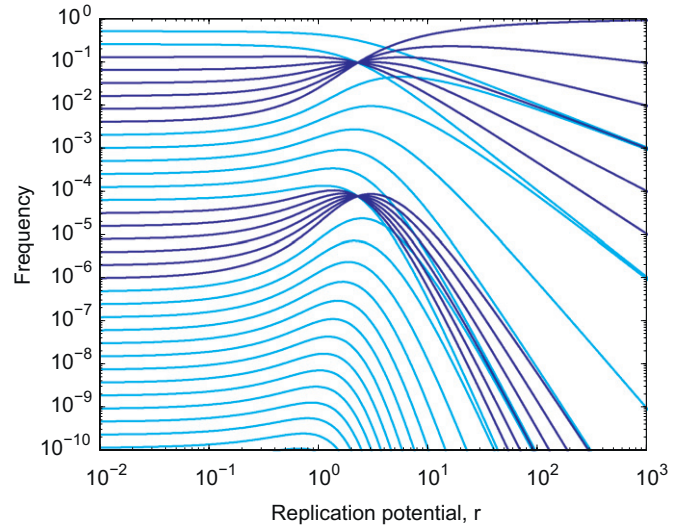$$r = \frac{d}{f} + \frac{r\phi^*}{f}. \tag{11}$$

The same is true for sequences $n - 1$ and $n - 2$, $n - 2$ and $n - 3$, etc., up to m and $m - 1$ since sequences m to n all have the same fitness. Thus at any r at which a replicating sequence crosses its precursor, all the replicating sequences cross their precursors and the frequencies of $m - 1$ to n are equal. Generalizing to arbitrary fitness landscapes, we observe that a sequence of fitness f intersects its precursor at values of r satisfying (11) and that this r is independent of the length of the sequence. We thus obtain the following theorem:

Theorem 2. *Under the assumptions of* Theorem 1, *if a sequence with fitness f crosses its precursor at $r = r_0$, then all sequences with fitness f cross their precursors at $r = r_0$. In particular, the value of r_0 does not depend on the fitness of the precursor.*

For $r = 0$, the prelife equilibrium structure is intact and sequences are ordered by sequence length. As r increases and selection dynamics become important, there is a value of r at which the frequencies of $m - 1$ to n are all equal. Thus if we have a series of consecutive sequences of the same fitness, they first intersect in a single point. Life is then selected over prelife only after all replicating sequences have achieved the same (equilibrium) frequency.

The equilibrium portrait of Fig. 3 shows two groups of replicating sequences: 3–8, which have fitness 1, and 15–20, which also have fitness 1. The frequencies of sequences 2–8 meet at $r \approx 2.27$ as do those of 14–20—the first intersections occur at the same r since all replicating sequences have the same fitness.

Now suppose sequences m to n have fitness 1, with the exception of sequences i to j ($m < i \leqslant j < n$), which have fitness $f > 1$. The fittest sequences cannot cross their precursors before $r = d/f$

and the sequences of intermediate fitness cannot cross their precursors before $r = d$. However, from our simulations, we know that there are not necessarily going to be intersections of both types.

In Fig. 4(a), sequences 4–18 have fitness 1, with the exception of sequence 10, which has fitness 1.5. Sequence 10 (the dashed line) crosses sequence 9 at $r \approx 0.91$ whereas the sequences with fitness 1 do not cross their precursors. In Fig. 4(b), sequences 4–18 have fitness 1, with the exception of sequence 10, which has fitness 1.1. Sequence 10 crosses sequence 9 at $r \approx 1.41$ whereas sequences 3–9 cross at $r \approx 2.19$, as do sequences 10–18.

2.3. Sequences 1 to n replicate

While life cannot dominate before $r = \min_i d/f_i$, it is possible for the prelife structure to remain intact in some cases regardless of how large r is. Here we give an example of such a landscape.

Consider the situation in which sequences 1 to n have the same fitness f. Fig. 5 shows the case $n = 20$ and $f = 1$. We know that a nonreplicating sequence never intersects its precursor. If there are any intersections, they must involve two replicating sequences and thus must be preceded by the intersection of a replicating sequence and its precursor, say i and $i + 1$ ($1 \leqslant i < n$). From the analysis above, we know that sequences with the same replication rate intersect their precursors at the same values of r. Thus if sequences i and $i + 1$ first intersect at r_0, then sequences 1 to n must all intersect at r_0. To show that there are no intersections, then, it is enough to show that sequences 1 and 2 never meet. We have

$$x_1^* = \frac{d}{(a + d) - r(f - \phi^*)},$$
$$x_2^* = \frac{x_1^*}{(a + d) - r(f - \phi^*)}. \tag{12}$$

Sequences 1 and 2 have the same frequency only if $x_1^* = x_2^* = d$. If $d > 1/2$, this cannot happen.
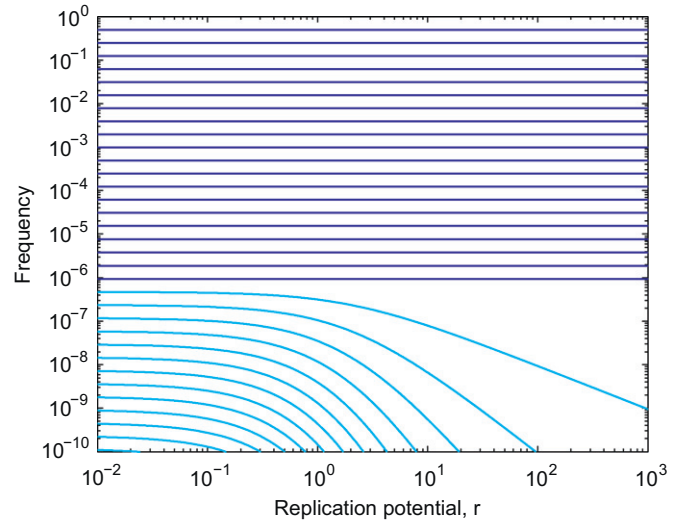
a

b
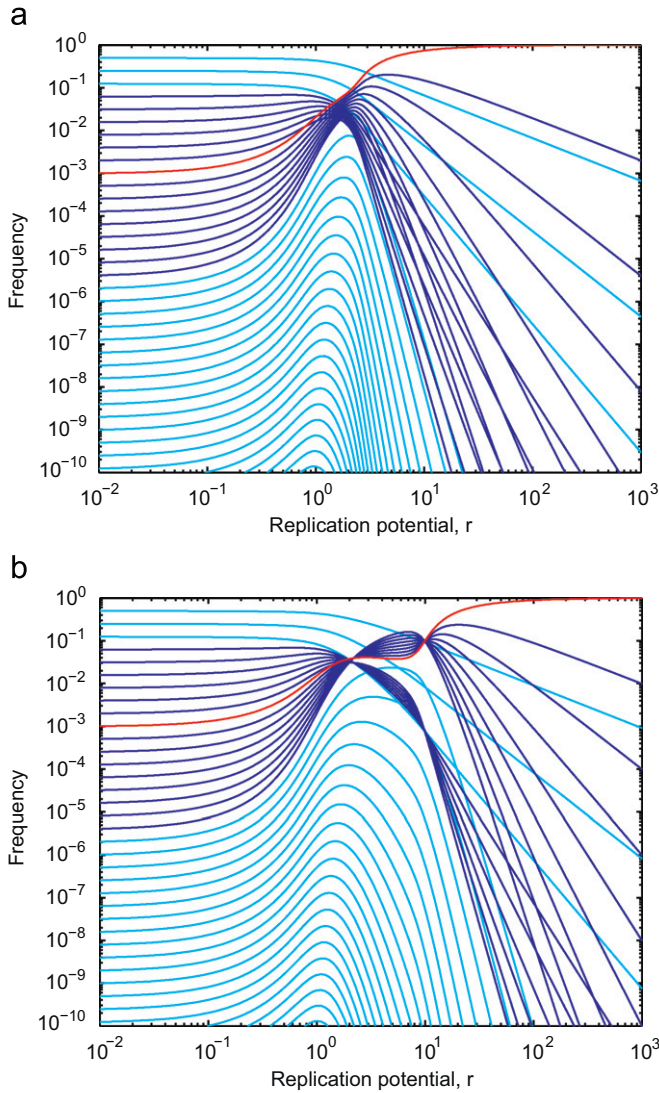
Fig. 4. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. (a) Sequences 4–18 (dark blue) have fitness 1, with exception of sequence 10 (red), which has fitness 1.5, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. According to (9), the frequency of sequence 10 can intersect the frequency of sequence 9 only when $r \geqslant 1/1.5 = 0.66$ and the frequencies of the sequences of fitness 1 can intersect their predecessors only when $r \geqslant 1$. In this case, however, there are no intersections of the latter type. (b) The same fitness landscape as in (a), but now sequence 10 has fitness 1.1. In this case there are five intersections: sequence 10 crosses its predecessor at $r \approx 1/1.1$, sequences 3–9 cross at $r \approx 1$ and then again at $r \approx 10$, and sequences 10–18 cross at $r \approx 1$ and $r \approx 10$ as well. The sequences within each of these two groups cross at the same values of r since sequences with the same fitness cross their predecessors at the same values of r.

3. Analytic calculation of an intersection in a special case

In the previous section, we found a lower bound for the values of r at which sequences can intersect, but there was no guarantee that there would be a reversal—or even just an intersection—if r exceeded that bound. In this section, we give an example in which we can guarantee—and compute the coordinates of—an intersection.

For simplicity, we set $f_2 = \cdots = f_n = 1$, all other $f_i = 0$, and $a = d = 1$, though what follows also applies for general a and d. Fig. 6 shows the $n = 20$ case. At equilibrium, system (6)

Fig. 5. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. Sequences 1–20 (dark blue) have fitness 1, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. Even if sequences 1–20 replicate, there are no reversals: the prelife structure of the population is maintained even in the limit $r \to \infty$.
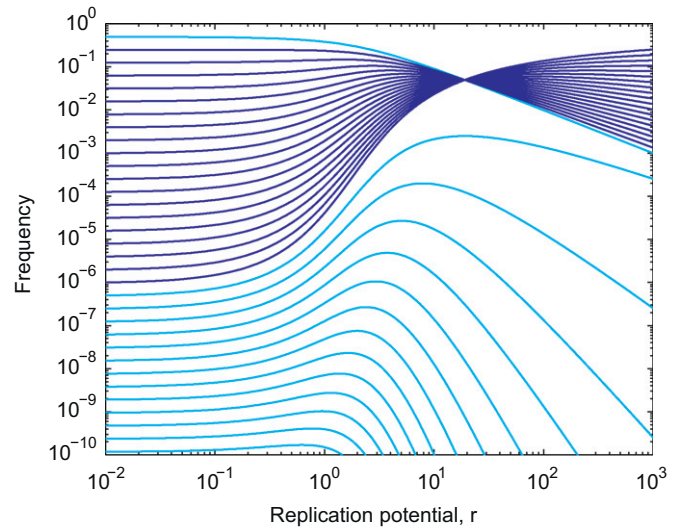
Fig. 6. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. Sequences 2–20 (dark blue) have fitness 1, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. In contrast to the fitness landscape of Fig. 5, here the frequencies of sequences 1–20 meet at $r = (19 + \sqrt{19^2 + 4})/2 \approx 20$.

becomes

$$1 - 2x_1^* - rx_1^*\phi^* = 0,$$
$$x_1^* - 2x_2^* + rx_2^*(1 - \phi^*) = 0,$$
$$\vdots$$
$$x_{n-1}^* - 2x_n^* + rx_n^*(1 - \phi^*) = 0,$$
$$x_n^* - 2x_{n+1}^* - rx_{n+1}^*\phi^* = 0,$$
$$\vdots \tag{13}$$

We know from Section 2.2 that if there are intersections involving any of the sequences $1, \ldots, n$, they must be preceded by the simultaneous intersection of all n sequences. We will thus attempt

to find a point at which such an n-way intersection occurs: we seek an intersection point of sequences 1 and 2 while assuming that sequences 1 to n meet there. This will give us an analytic condition on r. If we can find an r that satisfies it, we will have both confirmed the existence of the intersection point and determined the value of r at which it occurs. Now at this point, we have

$$r\phi^* = r\sum_{i=2}^{n} x_i^* = (n-1)rx^*. \tag{14}$$

Here x^* denotes the common equilibrium frequency of the first n sequences. Thus, the first two equations in (13) become

$$-(n-1)rx^{*2} - 2x^* + 1 = 0,$$
$$(n-1)rx^* - (r-1) = 0. \tag{15}$$

Solving (15) gives us

$$r = \frac{(n-1) + \sqrt{(n-1)^2 + 4}}{2} \tag{16}$$

and

$$x^* = \frac{(n-3) + \sqrt{(n-1)^2 + 4}}{(n-1)^2 + (n-1)\sqrt{(n-1)^2 + 4}}. \tag{17}$$

For $n = 20$, this gives $r \approx 19.0525$ and $x^* \approx 0.0499$. Note that there are two solutions for r in system (15), but one is negative and one is positive. Hence, there is precisely one positive value of r at which the frequencies of sequences 1 to n are equal. For large n, we obtain $r = n$ from (16) and $x^* = 1/n$ from (17).

4. The fastest replicator is not always the most abundant

Fig. 7(a) is an equilibrium portrait of the fitness landscape $f_{10} = 1$, $f_{11} = 0.95$, and $f_i = 0$ for all $i \neq 10, 11$. Despite the facts that the initial (prelife) frequency of sequence 10 is greater than that of sequence 11 and that sequence 10 has a higher fitness than sequence 11, between $r \approx 1$ and ≈ 20, the longer, less fit sequence (11) is more frequent. In fact, for r between approximately 2 and 20, sequence 11 dominates the entire population. We now study this phenomenon.

We consider fitness landscapes with $f_i < f_{N+1} < f_N$ for all $i \neq N, N+1$. We would like to approximate the values of r for which the longer, less fit sequence can be more frequent than the fittest sequence. A perturbation theory argument allows us to estimate the average fitness ϕ^* at equilibrium for landscapes of this type (see Appendix C):

$$\phi^*(r) = f_N - \frac{a+d}{r} + \frac{1}{r^N}\frac{a^{N-1}d}{\prod_{i=1}^{N-1}(f_N - f_i)} + O\left(\frac{1}{r^{N+1}}\right). \tag{18}$$

For the fitness landscape of Fig. 7(a) and our choice of parameters ($a = d = 1$), the approximation becomes

$$\phi^*(r) = f_N - \frac{a+d}{r} + \frac{a^9 d}{r^{10}} + O\left(\frac{1}{r^{11}}\right)$$
$$= 1 - \frac{2}{r} + \frac{2}{r^{10}} + O\left(\frac{1}{r^{11}}\right). \tag{19}$$

Fig. 7(b) is a comparison of this estimate to the actual average fitness; the estimate essentially coincides with the actual average fitness for $r > 2$.

Now as we saw above, the frequency of sequence $N + 1$ will be greater than the frequency of sequence N precisely when

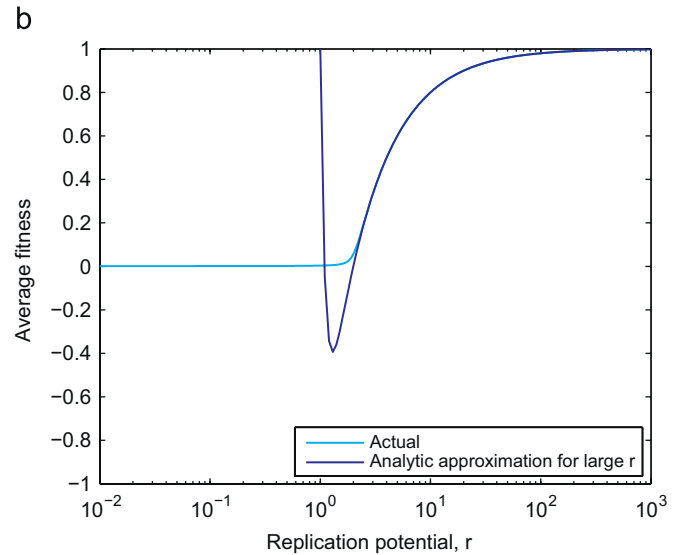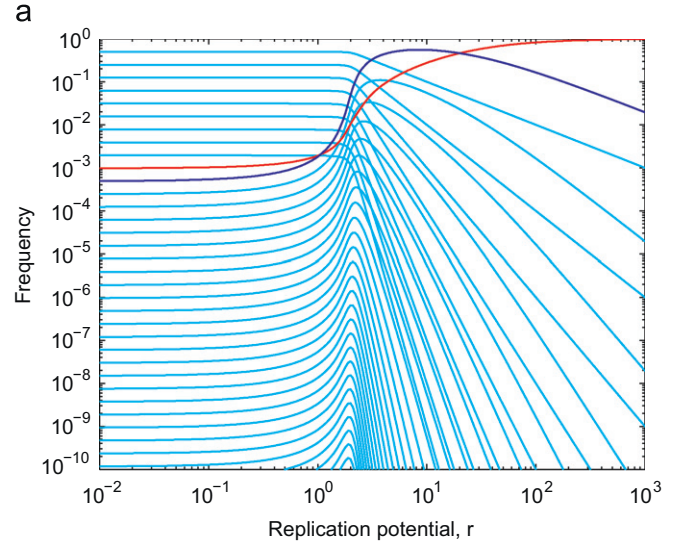$$\frac{a}{a + d - r(f_{N+1} - \phi^*)} > 0. \tag{20}$$

Fig. 7. The equilibrium portrait of the unary originator equation (6) with $a = d = 1$. (a) Sequence 10 (red) has fitness 1, sequence 11 (dark blue) has fitness 0.95, and all other sequences have fitness 0. Each curve is the graph of the equilibrium frequency $x_i^* = x_i^*(r)$ of a sequence of a particular length. Even if sequence 10 has an initial (when $r = 0$) frequency greater than that of sequence 11 and a higher fitness, between $r \approx 1$ and ≈ 20, $x_{11}^* > x_{10}^*$. The real positive roots of the polynomial $0.05x^{10} - x^9 + 1$ are the values of r at which $x_{10}^*(r) = x_{11}^*(r)$. (b) A comparison of the average fitness at equilibrium, $\phi^*(r)$, of the fitness landscape of (a) to the perturbation theory estimate (19). The actual average fitness and the estimate essentially coincide for $r \geqslant 2$.

Substituting (18) into (20) and rearranging terms, we see that sequence $N + 1$ will be more frequent than sequence N if

$$r(f_{N+1} - f_N) + a - \frac{a^{N-1}d}{r^{N-1}P} > 0, \tag{21}$$

where $P = \prod_{i=1}^{N-1}(f_N - f_i)$. Without loss of generality, we set $a = 1$, $f_N = 1$, and $f_{N+1} = 1 - s$, where $0 < s < 1$ (we are thus assuming that $f_i < 1 - s$ for $i \neq N, N+1$). Condition (21) then becomes

$$g(r) := sr^N - r^{N-1} + \frac{d}{P} < 0. \tag{22}$$

The real roots of this polynomial (if there are any) give approximations to the values of r at which the frequencies of sequences N and $N+1$ cross.

We study this polynomial in the special case $d = P = 1$ (the fitness landscape in Fig. 7(a) satisfies these conditions). Let

Table 1

A comparison of the actual and estimated values of r at which the equilibrium frequencies of sequences N and $N + 1$ cross (r_l is the r value of the first reversal and r_h the value of the second) when sequence N has fitness 1, sequence $N + 1$ has fitness $1 - s$, sequences of length $<N$ have a fixed fitness less than $1 - s$, and all other sequences have fitness 0.

N	s	f_i $(i<N)$	r_l (actual)	r_l (estimate)	r_h (actual)	r_h (estimate)
10	0.01	0	1.014	1.001	100.000	100.000
10	0.05	0	1.057	1.006	20.000	20.000
10	0.10	0	1.116	1.012	10.000	10.000
10	0.01	0.33	1.520	1.495	100.000	100.000
10	0.05	0.33	1.619	1.506	20.000	20.000
10	0.05	0.33	1.762	1.520	10.000	10.000
10	0.01	0.5	2.047	2.005	100.000	100.000
10	0.05	0.5	2.230	2.024	20.000	20.000
10	0.10	0.5	2.511	2.052	10.000	10.000
20	0.01	0	1.010	1.001	100.000	100.000
20	0.05	0	1.053	1.003	20.000	20.000
20	0.10	0	1.111	1.006	10.000	10.000
20	0.01	0.33	1.515	1.494	100.000	100.000
20	0.05	0.33	1.613	1.499	20.000	20.000
20	0.10	0.33	1.754	1.505	10.000	10.000
20	0.01	0.5	2.041	2.002	100.000	100.000
20	0.05	0.5	2.222	2.011	20.000	20.000
20	0.10	0.5	2.500	2.024	10.000	10.000

We set $a = d = 1$. The estimates are obtained by finding the real positive roots of polynomial (22).

$k = 1/s$. Finding the roots of g is equivalent to finding the roots of $h(r):=r^N - kr^{N-1} + k$. Taking derivatives, we have

$$h'(r) = r^{N-2}(Nr - k(N-1)),$$
$$h''(r) = (N-1)r^{N-3}(Nr - k(N-2)). \qquad (23)$$

There is precisely one positive root $r_0 = k(N-1)/N$ of h' (and this root is not a root of h'') and the condition $h(r_0)<0$ is equivalent to

$$\frac{k^{N-1}}{N-1}\left(1 - \frac{1}{N}\right)^N > 1. \qquad (24)$$

When k and N satisfy this condition, $h(r)$ will have exactly two positive real roots. The smaller one is the value of r at which the frequency of sequence $N + 1$ exceeds that of sequence N. The larger one is the value of r at which sequence N starts to dominate the population. For large N, (24) can be rewritten as $k^N > eN$.

When inequality (24) holds and h has two real roots, the larger one is approximately equal to k. To see this, observe that $h(k) = k > 0$ and $h'(r) > 0$ for $r \geq k$, so h can have no roots larger than k. Since $h(r_0) < 0$, it follows that h has a root between $r_0 = k - k/N$ and k. In fact it can have only one root in that interval since h' never vanishes for $r > r_0$. We can interpret this in the following way: if the fittest sequence has advantage s over the next fittest sequence, then the fittest sequence can dominate the population only if

$$r \geq \frac{1}{s}. \qquad (25)$$

For $r<1/s$, sequences of intermediate fitness can be most frequent. Table 1 gives the values for the real and estimated crossing points of sequences N and $N + 1$ for various values of N and s. In all cases, the estimate $1/s$ for the second point, after which sequence N can dominate, coincides with the numerical determination of the last crossing point.

5. The binary model

The results above apply to the case of the binary originator equation with slight modifications. Here we carry out one explicit

computation. The analogue of (6) in the binary case is

$$\dot{x}_i = a_i x_{i'} - (a_{i0} + a_{i1} + d)x_i + rx_i(f_i - \phi), \quad i = 0, 1, 00, 01, \ldots. \qquad (26)$$

As in (7), we can write

$$x_i^* = \frac{a_i}{(a_{i0} + a_{i1} + d) - r(f_i - \phi^*)}x_{i'}^*. \qquad (27)$$

For the frequencies of i and i' to cross, the fractional quantity must be 1. This condition becomes

$$r = \frac{a_{i0} + a_{i1} - a_i + d}{f_i} + \frac{r\phi^*}{f_i}. \qquad (28)$$

For the supersymmetric case ($a_{i0} = a_{i1} = a_i = a$) we have

$$r = \frac{a+d}{f_i} + \frac{r\phi^*}{f_i} \geq \frac{a+d}{f_i}. \qquad (29)$$

In supersymmetric binary prelife, all sequences of the same length have the same equilibrium frequency; for the frequency of a sequence to increase above that of its precursor, we need $r>(a+d)/f_i$. In contrast to the unary case, the net flow to successors due to prelife occurs at rate $a + d$ (since there are now two downstream flows of rate a to a sequence's successors). Net replication then becomes possible when the absolute replication rate rf_i exceeds $a + d$, giving (29).

6. Conclusion

We have introduced the originator equation, a model that allows us to study the transition from the generative chemistry of prelife—a system that proliferates information without replication—and its prevolutionary dynamics to life and its evolutionary dynamics. The equilibrium portraits associated to this model show how the frequencies of sequences vary as a function of the replication potential, r. When these frequencies intersect, we have a transition after which life dominates prelife.

We derived a "local" lower bound (9) for r below which a sequence and its precursor cannot cross and a "global" lower bound (10),

$$r > \min_i \frac{d}{f_i},$$

below which no sequences can cross. We showed that r exceeding this bound is necessary but not sufficient for a reversal to occur: when sequences 1 to n replicate with the same fitness, replication cannot break the symmetry of the prelife structure.

For a special fitness landscape—when sequences 2 to n replicate with the same fitness—the first and only n-way intersection of sequences 1 to n occurs at

$$r = \frac{(n-1) + \sqrt{(n-1)^2 + 4}}{2}.$$

For large n, this is intersection point occurs at $r \approx n$.

For the fitness landscape in which sequence N has fitness 1, sequence $N + 1$ has fitness $1 - s$ ($0<s<1$), and all other sequences have fitness 0, we showed that the real positive roots of polynomial

$$h(r) = sr^N - r^{N-1} + 1$$

approximate the values of r at which the equilibrium frequencies of sequences N and $N + 1$ intersect. We have condition (24) under which the polynomial will in fact have two real positive roots and showed that the larger of these roots is (to good approximation)

equal to $1/s$. Thus, if we have two replicating sequences, N and $N + 1$, and sequence N has a fitness advantage of s, then the replication potential must be at least $1/s$ for the fittest sequence to dominate—before that, sequence $N + 1$ can be more abundant than sequence N.

In supersymmetric prelife, longer sequences are exponentially less frequent than shorter sequences. When sequences replicate, however, the symmetry of the prelife ordering can be disrupted, and the equilibrium distribution can be far from exponential: as r increases, selection becomes increasingly efficient in removing all but the fittest sequences.

Acknowledgements

Appendix A

In this appendix, we describe how we transform (6) to make it a finite system, and we deduce the uniqueness and stability of its equilibrium point in the simplex by reduction to the quasispecies equation. We will assume that all sequences of length $>N$ have the same fitness f.

We start with system (6),

$$\dot{x}_1 = d - (a + d)x_1 + rx_1(f_1 - \phi),$$
$$\dot{x}_i = ax_{i-1} - (a + d)x_i + rx_i(f_i - \phi), \quad i = 2, 3, \ldots,$$
$$\phi = \sum f_i x_i,$$

where $a, d > 0$, but now we assume that $f_i = f$ for all $i > N$. We introduce a new variable y representing the aggregate frequency of all strings of length $>N$:

$$y = \sum_{i=N+1}^{\infty} x_i. \tag{A.1}$$

We then have

$$\dot{y} = \sum_{i=N+1}^{\infty} \dot{x}_i$$
$$= a \sum_{i=N}^{\infty} x_i - (a + d) \sum_{i=N+1}^{\infty} x_i + r(f - \phi) \sum_{i=N+1}^{\infty} x_i$$
$$= ax_N - dy + ry(f - \phi). \tag{A.2}$$

We thus obtain a finite system of differential equations:

$$\dot{x}_1 = d - (a + d)x_1 + rx_1(f_1 - \phi),$$
$$\dot{x}_i = ax_{i-1} - (a + d)x_i + rx_i(f_i - \phi), \quad i = 2, 3, \ldots, N,$$
$$\dot{y} = ax_N - dy + ry(f - \phi), \tag{A.3}$$

where $\phi = f_1 x_1 + \cdots + f_N x_N + fy$. We can simulate the finite system (A.3) numerically without introducing any approximations: the trajectories and the equilibrium frequencies of sequences $1, \ldots, N$ determined by the simulation will be the same as the equilibrium frequencies of those sequences in the infinite system.

We now investigate the equilibrium properties of (6). Let

$$W = \begin{pmatrix} f_1 r + d & d & d & \cdots & d & d \\ a & f_2 r & 0 & \cdots & 0 & 0 \\ 0 & a & f_3 r & \cdots & 0 & 0 \\ \vdots & & & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & a & fr + a \end{pmatrix}, \quad \vec{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \\ y \end{pmatrix}. \tag{A.4}$$

W is nonnegative and always irreducible. We will assume that $r > 0$ and that at least one $f_i > 0$ so that W is also aperiodic. When $r = 0$, originator dynamics reduce to prelife dynamics, for which the uniqueness and stability of the equilibrium is well-known (Nowak and Ohtsuki, 2008).

On the simplex $x_1 + \cdots + x_N + y = 1$, (A.3) is equivalent to the matrix equation

$$\dot{\vec{x}} = W\vec{x} - r\phi\vec{x} - (a + d)\vec{x}$$
$$= W\vec{x} - (r\phi + a + d)\vec{x}. \tag{A.5}$$

Note that, formally, (A.5) is the quasispecies equation (Eigen, 1971; Eigen and Schuster, 1977; Nowak, 2006). The equilibrium of originator dynamics is given by

$$W\vec{x} = (r\phi + a + d)\vec{x}. \tag{A.6}$$

This is, formally, exactly the equation for the equilibrium of quasispecies dynamics—W can be viewed as an ergodic (i.e., irreducible and aperiodic) mutation–selection matrix—so there is a unique, globally stable equilibrium in the $(N + 1)$-dimensional simplex (Bürger, 2000). This implies that originator dynamics (of the infinite system) converge to a "quasiequilibrium" where the frequencies of sequences 1 to N and the aggregate frequency of strings of length $>N$ are given by the equilibrium frequencies of the finite system (A.3). This does not immediately show, however, that the frequencies $x_{N+1}, x_{N+2}, \ldots$ have unique and stable equilibrium values.

We can now proceed by induction for $M \geqslant N$ given our assumption that $f_i = f$ for $i \geqslant N$. For M, we obtain an equilibrium solution of the finite-dimensional system that has the property that the equilibrium frequencies x_i^* of the system of dimension $M + 1$ coincide with those of dimension M if $i \leqslant M$. This implies that the infinite-dimensional system has an equilibrium solution that can be constructed in this way (clearly it sums to 1) and it is unique. Since we obtain global convergence to the unique equilibrium solution for every transformed system (of dimension M), we obtain convergence of all trajectories in the infinite system.

Appendix B

The originator equation (A.3) is a system of (deterministic) differential equations that is meant to describe what, in reality, is a stochastic system with a finite—sometimes small—number of molecules. To check that stochastic effects do not dominate, we performed two sets of simulations (with several values of r) for a representative fitness landscape—sequences of length 7 and 8 have fitness 1 and all other sequences have fitness 0. Our methodology was based on the Gillespie algorithm (Gillespie, 1977). In the first set of simulations, we started with 100,000 (unactivated) monomers and allowed one billion random reactions to occur (this was enough for the distribution of frequencies to be stationary). Each reaction was a sequence elongation, sequence death, or sequence replication. We then computed the frequency of each sequence. We repeated this procedure five times and computed the average frequency of each sequence over all the trials. In the second set of simulations, we started the system with an equal number of sequences of length 1–10, but all other aspects were the same. In both cases, we found that the

Table 2
A comparison of the equilibrium frequencies computed by a deterministic simulation of (A.3) to the stationary frequencies computed by stochastic simulations.

r	Sequence length	Deterministic	Stochastic (M)	Stochastic (E)
0	1	0.500000	0.500072	0.501320
	2	0.250000	0.249553	0.250157
	3	0.125000	0.125287	0.124887
	4	0.062500	0.062594	0.061808
	5	0.031250	0.031095	0.031161
	6	0.0156250	0.015618	0.015249
	7	0.007813	0.007959	0.007679
	8	0.003906	0.004069	0.003885
	9	0.001953	0.001820	0.001935
	10	0.000977	0.001027	0.000921
	>10	0.000977	0.000906	0.000998
1	1	0.493187	0.493156	0.493147
	2	0.243233	0.243205	0.243206
	3	0.119959	0.119959	0.120018
	4	0.059162	0.059153	0.059162
	5	0.029178	0.029193	0.029187
	6	0.014390	0.014400	0.014390
	7	0.014003	0.013996	0.014010
	8	0.013627	0.013665	0.013624
	9	0.006721	0.006722	0.006720
	10	0.003314	0.003319	0.003315
	>10	0.003225	0.003231	0.003221
2	1	0.441400	0.441449	0.441402
	2	0.194834	0.194908	0.194810
	3	0.086000	0.086001	0.860050
	4	0.037960	0.037973	0.037961
	5	0.016756	0.016763	0.016750
	6	0.007396	0.007403	0.007397
	7	0.027855	0.027812	0.027800
	8	0.104905	0.104837	0.104907
	9	0.046305	0.046274	0.046306
	10	0.020439	0.020426	0.020437
	>10	0.016151	0.016154	0.016145

We simulated the fitness landscape in which sequences 7 and 8 have fitness 1 and all other sequences have fitness 0.
The "Deterministic" column gives the equilibrium frequencies computed by the deterministic simulation, "Stochastic (M)" gives the (average) stationary frequencies when the system was started with 100,000 unactivated monomers and nothing else, and "Stochastic (E)" gives the (average) stationary frequencies when the system was started with an equal number of sequences of length 1–10 (10,000 each).

stochastic simulations agreed with our deterministic ones. Table 2 presents the detailed results.

Appendix C

Here we derive the asymptotic expansion (18) of the equilibrium average fitness ϕ^*. For the unary model (6), we obtain the equilibrium solution

$$x_1^* = \frac{d}{(a+d) - r(f_1 - \phi^*)}, \tag{C.1a}$$

$$x_n^* = \frac{ax_{n-1}}{(a+d) - r(f_n - \phi^*)} = \frac{a^{n-1}d}{\prod_{i=1}^{n}[(a+d) - r(f_i - \phi^*)]}. \tag{C.1b}$$

Since we are seeking an asymptotic expansion in the limit $r \to \infty$, we set $\varepsilon = 1/r$ and let $\varepsilon \to 0$. From C.1, we obtain straightforwardly

$$x_1^* = \frac{\varepsilon d}{\varepsilon(a+d) + (\phi^* - f_1)} = \varepsilon \frac{d}{\phi^* - f_1} + O(\varepsilon^2), \tag{C.2a}$$

$$x_n^* = \frac{\varepsilon a x_{n-1}}{\varepsilon(a+d) + (\phi^* - f_n)} = \varepsilon \frac{a x_{n-1}}{\phi^* - f_n} + O(\varepsilon^2 x_{n-1}), \tag{C.2b}$$

where in both cases the second equality requires $\phi^* - f_i = O(1)$ for $i = 1$ or $i = n$, respectively. In particular, if $\phi^* - f_i = O(1)$ for every $i = 1, \dots, n$, we obtain from (C.2) by iteration

$$x_n^* = \varepsilon^n \frac{a^{n-1}d}{\prod_{i=1}^{n}(\phi^* - f_i)} + O(\varepsilon^{n+1}). \tag{C.3}$$

Now assume that sequence N has the highest fitness, $N+1$ has the second highest fitness, and all other sequences are less fit, i.e., $f_i < f_{N+1} < f_N$ for all $i \neq N, N+1$. If $\varepsilon = 0$, then $\phi^* = f_N$ and $x_N = 1$. Hence, for small ε, we have $x_N^* = 1 + O(\varepsilon)$. From (C.2b), we infer that this requires

$$1 + O(\varepsilon) = \frac{\varepsilon a x_{N-1}^*}{\varepsilon(a+d) + (\phi^* - f_N)}. \tag{C.4}$$

Since we must also have $\phi^* = f_N + O(\varepsilon)$ and because $f_i < f_N$ for every i, (C.3) yields

$$x_{N-1}^* = \varepsilon^{N-1} \frac{a^{N-2}d}{\prod_{i=1}^{N-1}(f_N - f_i)} + O(\varepsilon^N). \tag{C.5}$$

By rearranging (C.4) and substituting (C.5), we obtain

$$\phi^* = f_N - \varepsilon(a+d) + \frac{\varepsilon a x_{N-1}^*}{1 + O(\varepsilon)}$$

$$= f_N - \varepsilon(a+d) + \varepsilon^N \frac{a^{N-1}d}{\prod_{i=1}^{N-1}(f_N - f_i)} + O(\varepsilon^{N+1}) \text{ as } \varepsilon \to 0. \tag{C.6}$$

Because $\varepsilon = 1/r$, this is equivalent to (18).

Now, it is straightforward to derive from (C.2) and (C.6) asymptotic expansions for the equilibrium frequencies, x_n^*. If $N > 1$, one obtains

$$x_n^*(\varepsilon) = \varepsilon^{n-1} \frac{a^{n-2}d}{\prod_{i=1}^{n-1}(f_n - f_i)} + O(\varepsilon^{n+N}) \quad (n < N),$$

$$x_N^*(\varepsilon) = 1 - \sum_{i \neq N} x_i,$$

$$x_{N+1}^*(\varepsilon) = \varepsilon \frac{a}{f_N - f_{N+1}} - \varepsilon^2 \left[\frac{a^2}{(f_N - f_{N+1})^2} + \frac{ad}{(f_N - f_1)(f_N - f_{N+1})} \right] + O(\varepsilon^3),$$

$$x_{N+k}^*(\varepsilon) = \varepsilon^k \frac{a^k}{\prod_{i=1}^{k}(f_N - f_{N+i})} + O(\varepsilon^{k+1}) \quad (k > 1).$$

References

Bartel, D.P., Szostak, J.W., 1993. Isolation of new ribozymes from a large pool of random sequences. Science 261, 1411–1418.

Benner, S.A., Caraco, M.D., Thomson, J.M., Gaucher, E.A., 2002. Planetary biology-palaeontological, geological, and molecular histories of life. Science 296, 864–868.

Bürger, R., 2000. The Mathematical Theory of Selection, Mutation, and Recombination. Wiley, Chichester.

Cech, T.R., 1993. The efficiency and versatility of catalytic RNA: implications for an RNA world. Gene 135, 33–36.

Dyson, F.J., 1982. A model for the origin of life. J. Mol. Evol. 18, 344–350.

Dyson, F., 1999. Origins of Life. Cambridge University Press, Cambridge and New York.

Eigen, M., 1971. Molecular self-organization and the early stages of evolution. Q. Rev. Biophys. 4, 149–212.

Eigen, M., Schuster, P., 1977. The hypercycle, a principle of natural self-organization. Part A: emergence of the hypercycle. Naturwissenschaften 64, 541–565.

Eigen, M., Schuster, P., 1979. The Hypercycle, A Principle of Natural Self-Organization. Springer, Berlin and New York.

Eigen, M., McCaskill, J., Schuster, P., 1988. Molecular quasi-species. J. Phys. Chem. 92, 6881–6891.

Ellington, A.D., Szostak, J.W., 1990. In vitro selection of RNA molecules that bind specific ligands. Nature 346, 818–822.

Gillespie, D.T., 1977. Exact stochastic simulation of coupled chemical reactions. J. Phys. Chem. 81, 2340–2361.

Ferris, J.P., Hill, A.R., Liu, R., Orgel, L.E., 1996. Synthesis of long prebiotic oligomers on mineral surfaces. Nature 381, 59–61.

Fontana, W., Buss, L.W., 1994a. What would be conserved if "the tape were played twice?". Proc. Natl. Acad. Sci. USA 91, 757–761.

Fontana, W., Buss, L.W., 1994b. "The arrival of the fittest": toward a theory of biological organization. Bull. Math. Biol. 56, 1–64.

Johnston, W.K., Unrau, P.J., Lawrence, M.S., Glasner, M.E., Bartel, D.P., 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. Science 292, 1319–1325.

Joyce, G.F., 1989. RNA evolution and the origins of life. Nature 338, 217–224.

Joyce, G.F., 2002. The antiquity of RNA-based evolution. Nature 418, 214–221.

Kauffman, S.A., 1986. Autocatalytic sets of proteins. J. Theor. Biol. 119, 1–24.

Kauffman, S.A., 1993. The Origins of Order: Self Organization and Selection in Evolution. Oxford University Press, Oxford and New York.

Maynard Smith, J., Szathmáry, E., 1995. The Major Transitions in Evolution. W.H. Freeman, Oxford.

McCaskill, J., 1984. A localisation threshold for macromolecular quasispecies from continually distributed replication rates. J. Chem. Phys. 80, 5194–5202.

Miller, S.L., 1953. A production of amino acids under possible primitive earth conditions. Science 117, 528–529.

Nowak, M.A., 1992. What is a quasispecies? Trends Ecol. Evol. 7, 118–121.

Nowak, M.A., 2006. Evolutionary Dynamics. Harvard University Press, Cambridge and London.

Nowak, M.A., Schuster, P., 1989. Error thresholds of replication in finite populations: mutation frequencies and the onset of Muller's ratchet. J. Theor. Biol. 137, 375–395.

Nowak, M.A., Ohtsuki, H., 2008. Prevolutionary dynamics and the origin of evolution. Proc. Natl. Acad. Sci. USA 105, 14924–14927.

Ricardo, A., Carrigan, M.A., Olcott, A.N., Benner, S.A., 2004. Borate minerals stabilize ribose. Science 303, 196.

Segre, D., Lancet, D., Kedem, O., Pilpel, Y., 1998. Graded autocatalysis replication domain (GARD): kinetic analysis of self-replication in mutually catalytic sets. Origins Life Evol. Biosphere 28, 501–514.

Segre, D., Ben-Eli, D., Lancet, D., 2000. Compositional genomes: prebiotic information transfer in mutually catalytic noncovalent assemblies. Proc. Natl. Acad. Sci. USA 97, 4112–4117.

Shapiro, R., 2006. Small molecule interactions were central to the origin of life. Q. Rev. Biol. 81, 105–125.

Sievers, D., von Kiedrowski, G., 1994. Self-replication of complementary nucleotide-based oligomers. Nature 369, 221–224.

Stein, D.L., Anderson, P.W., 1984. A model for the origin of biological catalysis. Proc. Natl. Acad. Sci. USA 81, 1751–1753.

Steitz, T.A., Moore, P.B., 2003. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. Trends Biochem. Sci. 28, 411–418.

Swetina, J., Schuster, P., 1982. Self-replication with errors: a model for polynucleotide replication. Biophys. Chem. 16, 329–345.

Szathmáry, E., Demeter, L., 1987. Group selection of early replicators and the origin of life. J. Theor. Biol. 128, 463–486.

Szostak, J.W., Bartel, D.P., Luisi, P.L., 2001. Synthesizing life. Nature 409, 387–390.