Prevolutionary dynamics and the origin of evolution

Martin A. Nowak[†] and Hisashi Ohtsuki

Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge, MA 02138

Communicated by Clifford H. Taubes, Harvard University, Cambridge, MA, July 14, 2008 (received for review May 31, 2008)

Life is that which replicates and evolves. The origin of life is also the origin of evolution. A fundamental question is when do chemical kinetics become evolutionary dynamics? Here, we formulate a general mathematical theory for the origin of evolution. All known life on earth is based on biological polymers, which act as information carriers and catalysts. Therefore, any theory for the origin of life must address the emergence of such a system. We describe prelife as an alphabet of active monomers that form random polymers. Prelife is a generative system that can produce information. Prevolutionary dynamics have selection and mutation, but no replication. Life marches in with the ability of replication: Polymers act as templates for their own reproduction. Prelife is a scaffold that builds life. Yet, there is competition between life and prelife. There is a phase transition: If the effective replication rate exceeds a critical value, then life outcompetes prelife. Replication is not a prerequisite for selection, but instead, there can be selection for replication. Mutation leads to an error threshold between life and prelife.

prelife | replication | selection | mutation | mathematical biology

The attempt to understand the origin of life has inspired much experimental and theoretical work over the years (1–10). Many of the basic building blocks of life can be produced by simple chemical reactions (11–15). RNA molecules can both store genetic information and act as enzymes (16–24). Fatty acids can self-assemble into vesicles that undergo spontaneous growth and division (25–28). The defining feature of biological systems is evolution. Biological organisms are products of evolutionary processes and capable of undergoing further evolution. Evolution needs a generative system that can produce unlimited information. Evolution needs populations of information carriers. Evolution needs mutation and selection. Normally, one thinks of these properties as being derivative of replication, but here, we formulate a generative chemistry ("prelife") that is capable of selection and mutation before replication. We call the resulting process "prevolutionary dynamics." Replication marks the transition from prevolutionary to evolutionary dynamics, from prelife to life.

Let us consider a prebiotic chemistry that produces activated monomers denoted by 0* and 1*. These chemicals can either become deactivated into 0 and 1 or attach to the end of binary strings. We assume, for simplicity, that all sequences grow in one direction. Thus, the following chemical reactions are possible:

$$i + 0^* \rightarrow i0$$
$$i + 1^* \rightarrow i1. \qquad [1]$$

Here i stands for any binary string (including the null element). These copolymerization reactions (29, 30) define a tree with infinitely many lineages. Each sequence is produced by a particular lineage that contains all of its precursors. In this way, we can define a prebiotic chemistry that can produce any binary string and thereby generate, in principle, unlimited information and diversity. We call such a system prelife and the associated dynamics prevolution (Fig. 1).

Each sequence, i, has one precursor, i', and two followers, $i0$ and $i1$. The parameter a_i denotes the rate constant of the chemical reaction from i' to i. At first, we assume that the active monomers are always at a steady state. Their concentrations are included in the rate constants, a_i. All sequences decay at rate, d. The following system of infinitely many differential equations describes the deterministic dynamics of prelife:

$$\dot{x}_i = a_i x_{i'} - (d + a_{i0} + a_{i1})x_i. \qquad [2]$$

The index, i, enumerates all binary strings of finite length, 0,1,00,. . . . The abundance of string i is given by x_i and its time derivative by $\dot{x}_i$. For the precursors of 0 and 1, we set $x_{0'} = x_{1'} = 1$. If all rate constants are positive, then the system converges to a unique steady state, where (typically) longer strings are exponentially less common than shorter ones. Introducing the parameter $b_i = a_i/(d + a_{i0} + a_{i1})$, we can write the equilibrium abundance of sequence i as $x_i = b_i b_{i'} b_{i''} \ldots b_\sigma$. The product is over the entire lineage leading from the monomer, σ (= 0 or 1), to sequence i. The total population size converges to $X = (a_0 + a_1)/d$. The rate constants, a_i, of the copolymerization process define the "prelife landscape." We will now discuss three different prelife landscapes.

For "supersymmetric" prelife, we assume that $a_0 = a_1 = \alpha/2$, and $a_i = a$ for all other i. Hence, all sequences grow at uniform rates. In this case, all sequences of length n have the same equilibrium abundance given by $x_n = [\alpha/2a][a/(2a + d)]^n$. Thus, longer sequences are exponentially less common. The total equilibrium abundance of all strings is $X = \alpha/d$. The average sequence length is $\bar{n} = 1 + 2a/d$.

Selection emerges in prelife, if different reactions occur at different rates. Consider a random prelife landscape, where a fraction p of reactions are fast ($a_i = 1 + s$), whereas the remaining reactions are slow ($a_i = 1$). Fig. 2*A* shows the equilibrium distribution of all sequences as a function of the selection intensity, s. For larger values of s, some sequences are selected (highly prevalent), whereas the others decline to very low abundance. The fraction of sequences that are selected out of all sequences of length n is given by $(1 - p)^2[1 - p(1 - p)]^{n-1}$. See supporting information (SI) for all detailed calculations.

Another example of an asymmetric prelife landscape contains a "master sequence" of length n (Fig. 2*B*). All reactions that lead to that sequence have an increased rate b, while all other rates are a. The master sequence is more abundant than all other sequences of the same length. But the master sequence attains a significant fraction of the population (= is selected) only if b is much larger than a. The required value of b grows as a linear function of n. In this prelife landscape, we can also discuss the effect of "mutation." The fast reactions leading to the master sequence might incorporate the wrong monomer with a certain probability, u, which then acts as a mutation rate in prelife. We find an error threshold: The master sequence can attain a significant fraction of the population, only if u is less than the inverse of the sequence length, $1/n$.

Author contributions: M.A.N. and H.O. wrote the paper.

The authors declare no conflict of interest.

[†]To whom correspondence should be addressed. E-mail: martin_nowak@harvard.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0806714105/DCSupplemental.

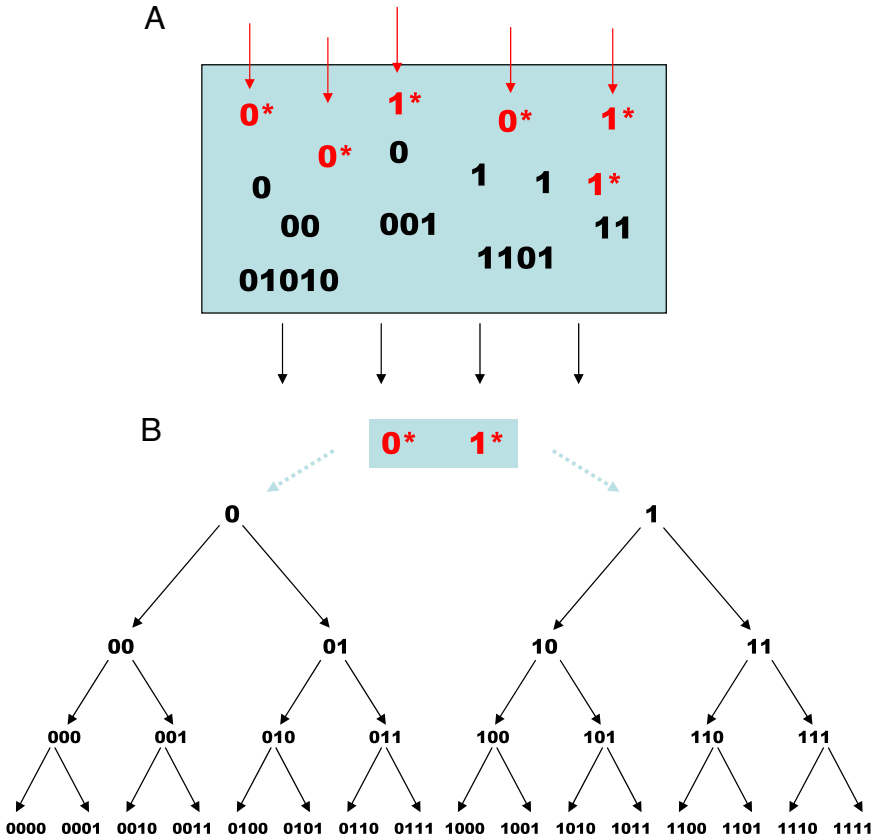

Fig. 1. A binary soup and the tree of prelife. (*A*) Prebiotic chemistry produces activated monomers, 0* and 1*, which form random polymers. Activated monomers can become deactivated, 0* $\rightarrow$ 0 and 1* $\rightarrow$ 1 or attach to the end of strings, for example, 00 + 1* $\rightarrow$ 001. We assume that all strings grow only in one direction. Therefore, each string has one immediate precursor and two immediate followers. (*B*) In the tree of prelife, each sequence has exactly one production lineage. The arrows indicate all of the chemical reactions of prelife up to length $n = 4$.

Let us now assume that some sequences can act as a templates for replication. These replicators are not only formed from their precursor sequences in prelife but also from active monomers at a rate that is proportional to their own abundance. We obtain the following differential equation

$$\dot{x}_i = a_i x_{i'} - (d + a_{i0} + a_{i1})x_i + r x_i(f_i - \phi) \qquad [3]$$

As before, the index i enumerates all binary strings of finite length. The first part of the equation describes prelife (exactly as in Eq. **2**). The second part represents the standard selection equation of evolutionary dynamics (28). The fitness of sequence i is given by f_i. All sequences have a frequency-dependent death rate, which represents the average fitness, $\phi = \Sigma_i f_i x_i / \Sigma_i x_i$ and ensures that the total population size remains at a constant value.

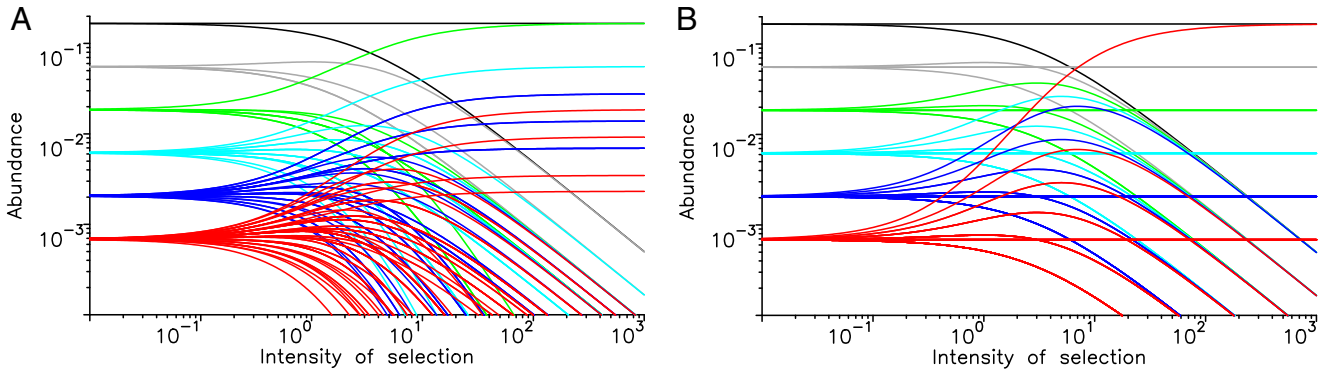

Fig. 2. Selection can occur in prelife without replication. The equilibrium abundances of all sequences of length 1 to 6 are shown as a function of the intensity of selection, s. There are 2^n sequences of length n. (*A*) In a random prelife landscape, half of all reactions occur at rate $1 + s$, the other half at rate 1. As s increases, a small subset of sequences is selected, whereas the others decline to very low abundance. (*B*) All reactions leading to the one ''master sequence'' of length 6 occur at rate $b = 1 + s$, all others at rate $a = 1$. As s increases, the master sequence is selected. Lineages that share sequences with the master sequence are suppressed, whereas other lineages are unaffected. Color code: black, gray, green, light blue, blue, and red for sequences of length 1 to 6, respectively. Other parameters: $a_0 = a_1 = 1/2$ and $d = 1$.

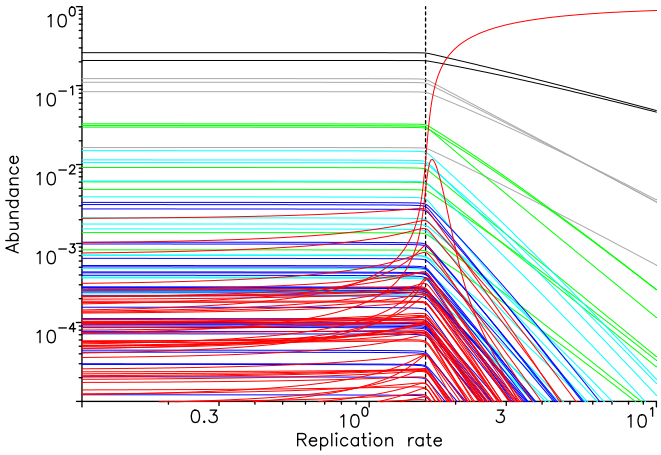

Fig. 3. The competition between life and prelife results in selection for (or against) replication. The equilibrium abundances of all sequences of length 1 to 6 are shown versus the relative replication rate, r. We assume a random prelife landscape, where the reaction rates a_i are taken from a uniform distribution on [0,1]. All sequences of length $n = 6$ can replicate. Their fitness values are also taken from a uniform distribution on [0,1]. For small values of r, prelife prevails. For large values of r, the fastest replicator dominates the population. As r increases, there is a phase transition at the critical value r_c. The fitness of the fastest replicator is given by $f_i = 0.999$, its extension rates are $a_{i0} = 0.4418$ $a_{i1} = 0.1284$. The death rate is $d = 1$. We have $r_c = (d + a_{i0} + a_{i1})/f_i = 1.572$, which is indicated by the broken vertical line and is in perfect agreement with the numerical simulation. The color code is the same as in Fig. 2.

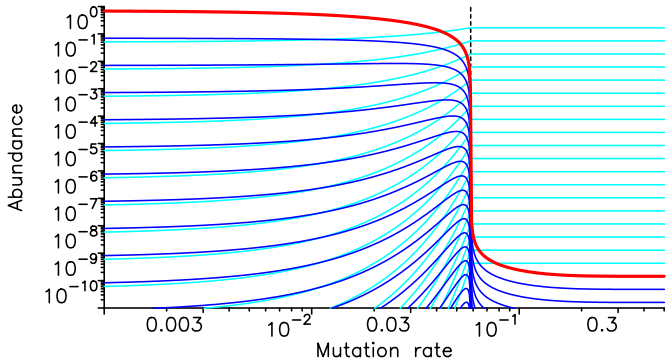

Fig. 4. There is an error threshold between life and prelife. We assume a "single-peak" fitness landscape, where one sequence of length $n = 20$ can replicate, but no other sequence replicates. Replication is subject to mutation. The mutation rate, u, denotes the error probability per base. Error-free replication of the entire sequence occurs with probability $q = (1 - u)^n$. We show all sequences that belong to the lineage of the replicator. The replicator is shown in red; shorter sequences are light blue, and longer ones dark blue. For small mutation rates, the replicator dominates the population, and the equilibrium structure is given by the mutation-selection balance of life. There is a critical error threshold. The theoretical prediction for this threshold, $u_c = 1 - [(d + 2a)/r]^{1/n} = 0.058$, is illustrated by the vertical broken line and is in perfect agreement with the numerical simulation. For larger mutation rates, we obtain the normal prelife equilibrium: Longer sequences (including the replicator) are exponentially less common than shorter ones. Parameter values: $a_0 = 1/2$, $a = 1$, $d = 1$; supersymmetric prelife; $r = 10$, $f_{20} = 1$.

The parameter r scales the relative rates of template-directed replication and template-independent sequence growth. These two processes are likely to have different kinetics. For example, their rates could depend differently on the availability of activated monomers. In this case, r could be an increasing function of the abundance of activated monomers. Template-directed replication requires double-strand separation. A common idea is that double-strand separation is caused by temperature oscillations, which means that r is affected by the frequency of those oscillations. The magnitude of r determines the relative importance of life versus prelife. For small r, the dynamics are dominated by prevolution. For large r, the dynamics are dominated by evolution.

Fig. 3 shows the competition between life (replication) and prelife. We assume a random prelife landscape where the a_i values are taken from a uniform distribution between 0 and 1. All sequences of length $n = 6$ have the ability to replicate. Their relative fitness values, f_i, are also taken from a uniform distribution on [0,1]. For small values of r, the equilibrium structure of prelife is unaffected by the presence of potential replicators; longer sequences are exponentially less frequent than shorter ones. There is a critical value of r, where a number of replicators increase in abundance. For large r, the fastest replicator dominates the population, whereas all other sequences converge to very low abundance. In this limit, we obtain the standard selection equation of evolutionary dynamics with competitive exclusion.

Between prelife and life, there is a phase transition. The critical replication rate, r_c, is given by the condition that the net reproductive rate of the replicators becomes positive. The net reproductive rate of replicator i can be defined as $g_i = r(f_i - \phi) - (d + a_{i0} + a_{i1})$. For $r < r_c$, the abundance of replicators is low, and therefore, ϕ is negligibly small. In Fig. 3, we have $d = 1$ and $a_{i0} + a_{i1} = 1$ on average. For the fastest replicator, we expect $f_i \approx 1$. Thus, the phase transition should occur around $r_c \approx 2$, which is the case. Using the actual rate constants of the fastest replicator in our system, we obtain the value $r_c = 1.572$, which is in perfect agreement with the exact numerical simulation (see broken vertical line in Fig. 3).

Replication can be subject to mistakes. With probability u, a wrong monomer is incorporated. In Fig. 4, we consider a "single-peak" fitness landscape: One seqence of length n can replicate. The probability of error-free replication is given by $q = (1 - u)^n$. The net reproductive rate of the replicator is now given by $g_i = r(f_i q - \phi) - (d + a_{i0} + a_{i1})$. The replicator is selected if the replication accuracy, q, is greater than a certain value, given by $q > (d + a_{i0} + a_{i1})/rf_i$. Thus, mutation leads to an error threshold for the emergence of life. Replication is selected only if the mutation rate, u, is less than a critical value that is proportional to the inverse of the sequence length, $1/n$. This finding is reminiscent of classical quasispecies theory (3, 4), but there, the error threshold arises when different replicators compete ("within life"). Here, we observe an error threshold between life and prelife.

Traditionally, one thinks of natural selection as choosing between different replicators. Natural selection arises if one type reproduces faster than another type, thereby changing the relative abundances of these two types in the population. Natural selection can lead to competitive exclusion or coexistence. In the present theory, however, we encounter natural selection before replication. Different information carriers compete for resources and thereby gain different abundances in the population. Natural selection occurs within prelife and between life and prelife. In our theory, natural selection is not a consequence of replication, but instead natural selection leads to replication. There is "selection for replication" if replicating sequences have a higher abundance than nonreplicating sequences of similar length. We observe that prelife selection is blunt: Typically small differences in growth rates result in small differences in abundance. Replication sharpens selection: Small differences in replication rates can lead to large differences in abundance.

We have proposed a mathematical theory for studying the origin of evolution. Our aim was to formulate the simplest possible population dynamics that can produce information and complexity. We began with a "binary soup" where activated

monomers form random polymers (binary strings) of any length (Fig. 1). Selection emerges in prelife, if some sequences grow faster than others (Fig. 2). Replication marks the transition from prelife to life, from prevolution to evolution. Prelife allows a continuous origin of life. There is also competition between life and prelife. Life is selected over prelife only if the replication rate is greater than a certain threshold (Fig. 3). Mutation during replication leads to an error threshold between life and prelife. Life can emerge only if the mutation rate is less than a critical value that is proportional to the inverse of the sequence length (Fig. 4). All fundamental equations of evolutionary and ecological dynamics assume replication (31–33), but here, we have explored the dynamical properties of a system before replication and the emergence of replication.

ACKNOWLEDGMENTS. This work was supported by the John Templeton Foundation, the Japan Society for the Promotion of Science (H.O.), the National Science Foundation/National Institutes of Health joint program in mathematical biology (NIH Grant R01GM078986), and J. Epstein.

1. Crick FH (1968) The origin of the genetic code. *J Mol Biol* 38:367–379.
2. Miller SL, Orgel LE (1974) *The Origins of Life on the Earth* (Prentice-Hall, Englewood Cliffs, NJ).
3. Eigen M, Schuster P (1977) The hyper cycle. A principle of natural self-organization. Part A: Emergence of the hyper cycle. *Naturwissenschaften* 64:541–565.
4. Eigen M, McCaskill J, Schuster P (1989) The molecular quasi-species. *Adv Chem Phys* 75:149–263.
5. Stein DL, Anderson PW (1984) A model for the origin of biological catalysis. *Proc Natl Acad Sci USA* 81:1751–1753.
6. Kauffman SA (1986) Autocatalytic sets of proteins. *J Theor Biol* 119:1–24.
7. Orgel LE (1992) Molecular replication. *Nature* 358:203–209.
8. Fontana W, Buss LW (1994) The arrival of the fittest: Toward a theory of biological organization. *B Math Biol* 56:1–64.
9. Fontana W, Buss LW (1994) What would be conserved if the tape were played twice? *Proc Natl Acad Sci USA* 91:757–761.
10. Dyson F (1999) *Origins of Life* (Cambridge Univ Press, Cambridge, UK/NY).
11. Miller SL (1953) A production of amino acids under possible primitive earth conditions. *Science* 117:528–529.
12. Szostak JW, Bartel DP, Luisi PL (2001) Synthesizing life. *Nature* 409:387–390.
13. Benner SA, Caraco MD, Thomson JM, Gaucher EA (2002) Planetary biology: Paleontological, geological, and molecular histories of life. *Science* 296:864–868.
14. Ricardo A, Carrigan MA, Olcott AN, Benner SA (2004) Borate minerals stabilize ribose. *Science* 303:196–196.
15. Benner SA, Ricardo A (2005) Planetary systems biology. *Mol Cell* 17:471–472.
16. Joyce GF (2005) Evolution in an RNA world. *Origins Life Evol B* 36:202–204.
17. Ellington AD, Szostak JW (1990) *In vitro* selection of RNA molecules that bind specific ligands. *Nature* 346:818–822.
18. Bartel DP, Szostak JW (1993) Isolation of new ribozymes from a large pool of random sequences. *Science* 261:1411–1418.
19. Cech TR (1993) The efficiency and versatility of catalytic RNA: Implications for an RNA world. *Gene* 135:33–36.
20. Sievers D, von Kiedrowski G (1994) Self-replication of complementary nucleotide-based oligomers. *Nature* 369:221–224.
21. Ferris JP, Hill AR, Liu R, Orgel LE (1996) Synthesis of long prebiotic oligomers on mineral surfaces. *Nature* 381:59–61.
22. Joyce GF (1989) RNA evolution and the origins of life. *Nature* 338:217–224.
23. Johnston WK, Unrau PJ, Lawrence MS, Glasner ME, Bartel DP (2001) RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science* 292:1319–1325.
24. Joyce GF (2002) The antiquity of RNA-based evolution. *Nature* 418:214–221.
25. Hargreaves WR, Mulvihill S, Deamer DW (1977) Synthesis of phospholipids and membranes in prebiotic conditions. *Nature* 266:78–80.
26. Hanczyc MN, Fujikawa SM, Szostak JW (2003) Experimental models of primitive cellular compartments: Encapsulation, growth, and division. *Science* 302:618–622.
27. Chen IA, Roberts RW, Szostak JW (2004) The emergence of competition between model protocells. *Science* 305:1474–1476.
28. Chen IA, Szostak JW (2004) A kinetic study of the growth of fatty acid vesicles. *Biophys J* 87:988–998.
29. Flory PJ (1953) *Principles of Polymer Chemistry* (Cornell Univ Press, Ithaca, NY).
30. Szwarc M, van Beylen M (1993) *Ionic Polymerization and Living Polymers* (Chapman and Hall, New York).
31. Nowak MA (2006) *Evolutionary Dynamics* (Harvard Univ Press, Cambridge, MA).
32. Hofbauer J, Sigmund K (1998) *Evolutionary Games and Population Dynamics* (Cambridge Univ Press, Cambridge, UK).
33. May RM (2001) *Stability and Complexity in Model Ecosystems* (Princeton Univ Press, Princeton).

Supporting Text for Prevolutionary Dynamics

Martin A. Nowak & Hisashi Ohtsuki

Program for Evolutionary Dynamics, Department of Organismic and Evolutionary Biology, Department of Mathematics, Harvard University, Cambridge MA 02138, USA

1 Prelife

Prelife dynamics are given by

$$\dot{x}_i = a_i x_{i'} - (d + a_{i0} + a_{i1}) x_i. \tag{1}$$

The index i represents all binary strings (sequences). Longer strings are produced from shorter ones by adding 0 or 1 on the right side. Each string, i, has one precursor, i', and two followers, $i0$ and $i1$. For example, the precursor of string 0101 is 010; the two followers are 01010 and 01011. For the precursors of strings 0 and 1 we set $x_{0'} = x_{1'} = 1$. The constants a_i denote the rate at which string i arises from i' by addition of an activated monomer (which is either 0^* or 1^*). Eq.(1) assumes that the concentration of activated monomers is constant. All strings are removed (die) at rate d.

Prelife dynamics define a tree with the activated monomers at the root. The tree of prelife has infinitely many lineages. A lineage is a sequence of strings that follow each other. For example, one such lineage is $0, 00, 000, ...$

At equilibrium, the right hand side of Eq.(1) is zero, so we obtain

$$x_i = b_i x_{i'}, \tag{2}$$

where b_i is given by

$$b_i = \frac{a_i}{d + a_{i0} + a_{i1}}. \tag{3}$$

Using Eq.(2) recursively gives us

$$x_i = b_i b_{i'} b_{i''} \cdots b_\sigma, \tag{4}$$

where σ is the ancestral monomer (0 or 1) of sequence i.

Let us consider super-symmetric prelife with $a_0 = a_1 = \alpha/2$ and $a_i = a$ for all other sequences, i. From Eq.(4), we obtain the following results.

The abundance of a sequence of length n is

$$x_n = \frac{\alpha}{2a}\Big(\frac{a}{2a+d}\Big)^n. \tag{5a}$$

The total abundance of all sequences of length n is

$$X_n = 2^n x_n = \frac{\alpha}{2a}\Big(\frac{2a}{2a+d}\Big)^n. \tag{5b}$$

The total abundance of all sequences is

$$X = \sum_{n=1}^{\infty} X_n = \frac{\alpha}{d}. \tag{5c}$$

The total abundance of all sequences in one lineage is

$$\tilde{X} = \sum_{n=1}^{\infty} x_n = \frac{\alpha}{2(a+d)}. \tag{5d}$$

The average sequence length is

$$\bar{n} = \frac{\sum_{n=1}^{\infty} n X_n}{X} = 1 + \frac{2a}{d}. \tag{5e}$$

Although there are infinitely many lineages, the abundance of any one lineage is a considerable fraction of the entire population. The reason is that short sequences belong to many lineages and they are much more abundant than long sequences.

2 Prelife landscape

Let us consider a random prelife landscape where reaction rates of sequences of length more than two are randomly given by

$$a_i = \begin{cases} a+s & (\text{with prob. } p) \\ a & (\text{with prob. } 1-p). \end{cases} \tag{6}$$

The other parameters are the same as before: $a_0 = a_1 = \alpha/2$.

From Eq.(4), at equilibrium we obtain the following results. The average abundance of a sequence of length n is

$$\bar{x}_n = \frac{\alpha}{2} A B^n, \tag{7}$$

where

$$A = \frac{(2a+d)^2 + (2a+d)(3-2p)s + 2(1-p)^2 s^2}{a(2a+d)^2 + (2a+d)(3a+pd)s + \{2a+p(2-p)d\}s^2} \tag{8}$$

and

$$B = \frac{a(2a+d)^2 + (2a+d)(3a+pd)s + \{2a+p(2-p)d\}s^2}{(2a+d)(2a+d+s)(2a+d+2s)}. \tag{9}$$

A sequence is selected if its equilibrium abundance is not vanishing as $s \to \infty$. For sequence i of length n, rewriting Eq.(4) yields

$$x_i = \frac{1}{d+a_{i0}+a_{i1}} \cdot \underbrace{\frac{a_i}{d+a_{i'0}+a_{i'1}} \cdot \frac{a_{i'}}{d+a_{i''0}+a_{i''1}} \cdots \frac{a_{\sigma\rho}}{d+a_{\sigma 0}+a_{\sigma 1}}}_{n-1 \text{ terms}} \cdot \frac{\alpha}{2}, \tag{10}$$

where $\sigma\rho$ represents the first two digits of sequence i. The first term in the right hand side of Eq.(10) is

$$\begin{cases} \frac{1}{(a+s)+(a+s)+d} \xrightarrow{s\to\infty} 0 & (\text{with prob. } p^2) \\ \frac{1}{(a+s)+a+d} \xrightarrow{s\to\infty} 0 & (\text{with prob. } 2p(1-p)) \\ \frac{1}{a+a+d} \xrightarrow{s\to\infty} \frac{1}{a+a+d} & (\text{with prob. } (1-p)^2). \end{cases} \tag{11}$$

The first term does not vanish with probability $(1-p)^2$.

For each of the next $n-1$ terms on the right hand side of Eq.(10) we have

$$\begin{cases} \frac{a+s}{(a+s)+(a+s)+d} \xrightarrow{s\to\infty} \frac{1}{2} & (\text{with prob. } p^2) \\ \frac{a+s}{(a+s)+a+d} \xrightarrow{s\to\infty} 1 & (\text{with prob. } p(1-p)) \\ \frac{a}{(a+s)+a+d} \xrightarrow{s\to\infty} 0 & (\text{with prob. } p(1-p)) \\ \frac{a}{a+a+d} \xrightarrow{s\to\infty} \frac{1}{a+a+d} & (\text{with prob. } (1-p)^2). \end{cases} \tag{12}$$

Each term does not vanish with probability $1-p(1-p)$. Therefore, the probability that a sequence of length n is selected (does not vanish) is given by

$$(1-p)^2[1-p(1-p)]^{n-1}. \tag{13}$$

The expected number of sequences of length n that are selected is

$$2^n(1-p)^2[1-p(1-p)]^{n-1}. \tag{14}$$

For example, if $a=1, d=1, \alpha=1$ and $p=1/2$ as in Figure 2, we obtain from Eq.(7) for the average abundance of sequences of length n

$$\bar{x}_n = \frac{18+12s+s^2}{36+42s+11s^2}\left(\frac{36+42s+11s^2}{12(3+s)(3+2s)}\right)^n. \tag{15}$$

Note that $\bar{x}_n(s)$ a monotonically decreasing function (of s) for $n \leq 3$, a one-humped function for $3 < n < 12$, and a monotonically increasing function for $n \geq 12$. From Eq.(14), the expected number of sequences of length n that survive for large s is given by $(1/3)(3/2)^n$.

3 Master sequence

In this section, we study the case where all reactions leading to one particular sequence (the master sequence) occur at the increased rate b, while all other reactions occur at rate a.

Suppose $0^n = \underbrace{00\cdots0}_{n}$ is the master sequence. The reaction rates are given by

$$\begin{aligned} a_0 = a_1 &= \alpha/2 \\ a_i &= b \qquad \text{for } \; i = 00, \cdots, 0^n \\ a_i &= a \qquad \text{for other } \; i. \end{aligned} \tag{16}$$

From the general formula, Eq.(4), the abundances of sequences $i = \underbrace{0\cdots0}_{\ell}\underbrace{1*\cdots*}_{m}$ at equilibrium are given by

$$x_i = \begin{cases} \dfrac{\alpha}{2a}\Big(\dfrac{a}{2a+d}\Big)^m & \text{if } \; \ell = 0 \\ \dfrac{\alpha}{2b}\Big(\dfrac{b}{a+b+d}\Big)^\ell\Big(\dfrac{a}{2a+d}\Big)^m & \text{if } \; 1 \le \ell \le n-1 \\ \dfrac{\alpha}{2a}\Big(\dfrac{b}{a+b+d}\Big)^{n-1}\Big(\dfrac{a}{2a+d}\Big)^{\ell+m+1-n} & \text{if } \; \ell \ge n. \end{cases} \tag{17}$$

In particular, we are interested in the abundances of all sequences that have the same length as the master sequence. Let x_i denote the abundance of a sequence of the form $\underbrace{0\cdots0}_{i}\underbrace{1*\cdots*}_{n-i}$. In this notation, x_n represents the abundance of the master sequence. From eq.(17), we obtain

$$x_i = \begin{cases} \dfrac{\alpha}{2a}\Big(\dfrac{a}{2a+d}\Big)^n & \text{if } \; i = 0 \\ \dfrac{\alpha}{2b}\Big(\dfrac{b}{a+b+d}\Big)^i\Big(\dfrac{a}{2a+d}\Big)^{n-i} & \text{if } \; 1 \le i \le n-1 \\ \dfrac{\alpha}{2a}\Big(\dfrac{b}{a+b+d}\Big)^{n-1}\Big(\dfrac{a}{2a+d}\Big) & \text{if } \; i = n. \end{cases} \tag{18}$$

Since $b > a$, we find

$$x_0 > x_1 < x_2 < \cdots < x_{n-1} < x_n \qquad \text{and} \qquad x_0 < x_n. \tag{19}$$

The master sequence is most abundant among all sequences of length n.

If $b \to \infty$, then the abundance of the master sequence converges to

$$x_{n,max} = \lim_{b\to\infty} x_n = \frac{\alpha}{2(2a+d)}. \tag{20}$$

Let us now calculate the condition for the abundance of the master sequence, x_n, to exceed a fraction, $1/k$, of the maximum value, $x_{n,max}$. From Eqs.(18) and (20), we have

$$\frac{\alpha}{2a}\Big(\frac{b}{a+b+d}\Big)^{n-1}\Big(\frac{a}{2a+d}\Big) > \frac{1}{k}\cdot\frac{\alpha}{2(2a+d)}. \tag{21}$$

This condition is rewritten as

$$b > \frac{a+d}{k^{\frac{1}{n-1}}-1} \approx \frac{a+d}{\ln k}\, n \qquad (n \gg 1). \tag{22}$$

Hence, for a master sequence of length n to make up a significant fraction of the population, the rate constant b must grow as a linear function of n.

4 Master sequence with mutation

As before, we assume that all reactions leading to the master sequence occur at an increased rate, b, but there is a probability u of incorporating the wrong monomer. The rate of those reactions that stay within the lineage leading to the master sequence is given by $b(1-u)$, while the reactions that come off the lineage occur at rate $a + bu$. We have

$$\begin{aligned} a_0 = a_1 &= \alpha/2 \\ a_i &= b(1-u) \qquad \text{for } \; i = 00, \cdots, 0^n \\ a_i &= a + bu \qquad \text{for } \; i = 01, \cdots, 0^{n-1}1 \\ a_i &= a \qquad \text{for all other } \; i. \end{aligned} \tag{23}$$

Consider sequences of the form $i = \underbrace{0\cdots0}_{\ell}\underbrace{1*\cdots*}_{m}$. As always the asterisks represent either 0 or 1. From the general formula, Eq.(4), the equilibrium abundance of

sequence i is given by

$$x_i = \begin{cases} \dfrac{\alpha}{2a}\left(\dfrac{a}{2a+d}\right)^m & \text{if } \ \ell = 0 \\ \dfrac{\alpha}{2b(1-u)}\left(\dfrac{b(1-u)}{a+b+d}\right)^\ell & \text{if } \ 1 \le \ell \le n-1, \ m = 0 \\ \dfrac{\alpha}{2b(1-u)} \cdot \dfrac{a+bu}{a}\left(\dfrac{b(1-u)}{a+b+d}\right)^\ell \left(\dfrac{a}{2a+d}\right)^m & \text{if } \ 1 \le \ell \le n-1, \ m \ge 1 \\ \dfrac{\alpha}{2a}\left(\dfrac{b(1-u)}{a+b+d}\right)^{n-1}\left(\dfrac{a}{2a+d}\right)^{\ell+m+1-n} & \text{if } \ \ell \ge n. \end{cases} \tag{24}$$

Let us now compare the abundances of all sequences of length n. Let x_i denote the abundances of sequences of the form $\underbrace{0\cdots0}_{i}\underbrace{1*\cdots*}_{n-i}$. In this notation, the abundance of the master sequence is given by x_n. From eq.(24), we obtain

$$x_i = \begin{cases} \dfrac{\alpha}{2a}\left(\dfrac{a}{2a+d}\right)^n & \text{if } \ i = 0 \\ \dfrac{\alpha}{2a} \cdot \dfrac{a+bu}{b(1-u)}\left(\dfrac{b(1-u)}{a+b+d}\right)^i \left(\dfrac{a}{2a+d}\right)^{n-i} & \text{if } \ 1 \le i \le n-1 \\ \dfrac{\alpha}{2a}\left(\dfrac{b(1-u)}{a+b+d}\right)^{n-1}\left(\dfrac{a}{2a+d}\right) & \text{if } \ i = n. \end{cases} \tag{25}$$

In order to understand the relative ranking of the equilibrium abundances of all sequences of length n, we must distinguish three cases.

Case (i) $u < \frac{a}{2a+d}$**:**

(i-a) If $b < \frac{a(a+d)}{(a+d)-u(2a+d)}$ then $x_0 < x_1 > x_2 > \cdots > x_{n-1} > x_n$.

(i-b) If $\frac{a(a+d)}{(a+d)-u(2a+d)} < b < \frac{a}{1-2u}$ then $x_0 < x_1 < x_2 < \cdots < x_{n-1} > x_n$.

(i-c) If $\frac{a}{1-2u} < b < \frac{a^2}{a-u(2a+d)}$ then $x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n$.

(i-d) If $b > \frac{a^2}{a-u(2a+d)}$ then $x_0 > x_1 < x_2 < \cdots < x_{n-1} < x_n$ and $x_0 < x_n$.

Case (ii) $\frac{a}{2a+d} \leq u < \frac{a+d}{2a+d}$**:**

(ii-a) If $b < \frac{a(a+d)}{(a+d)-u(2a+d)}$ then $x_0 < x_1 > x_2 > \cdots > x_{n-1} > x_n$.

(ii-b) If $\frac{a(a+d)}{(a+d)-u(2a+d)} < b < \frac{a}{1-2u}$ then $x_0 < x_1 < x_2 < \cdots < x_{n-1} > x_n$.

(ii-c) If $b > \frac{a}{1-2u}$ then $x_0 < x_1 < x_2 < \cdots < x_{n-1} < x_n$.

Case (iii) $u \geq \frac{a+d}{2a+d}$**:**

(iii-a) If $b < \frac{a}{1-2u}$, then $x_0 < x_1 > x_2 > \cdots > x_{n-1} > x_n$.

(iii-b) If $b > \frac{a}{1-2u}$, then $x_0 < x_1 > x_2 > \cdots > x_{n-1} < x_n$ and $x_1 > x_n$.

In summary, the equilibrium abundance of the master sequence is

$$x_n = \frac{\alpha}{2(2a+d)} \left(\frac{b(1-u)}{a+b+d} \right)^{n-1}. \tag{26}$$

The master sequence is most abundant among all sequences of length n if

$$u < \frac{a+d}{2a+d} \qquad \text{and} \qquad b > \frac{a}{1-2u}. \tag{27}$$

If $b \to \infty$, then the abundance of the master sequence converges to

$$x_{n,max} = \lim_{b \to \infty} x_n = \frac{\alpha}{2(2a+d)} (1-u)^{n-1}. \tag{28}$$

For x_n to exceed a fraction, $1/k$, of this maximum value, $x_{n,max}$, we need

$$\frac{\alpha}{2(2a+d)} \left(\frac{b(1-u)}{a+b+d} \right)^{n-1} > \frac{1}{k} \cdot \frac{\alpha}{2(2a+d)} (1-u)^{n-1}, \tag{29}$$

which is simplified to

$$b > \frac{a+d}{k^{\frac{1}{n-1}}-1} \approx \frac{a+d}{\ln k}\, n. \qquad (n \gg 1). \tag{30}$$

If $b \to \infty$ and $u \to 0$, then the abundance of the master sequence converges to

$$\hat{x}_{n,max} = \lim_{\substack{b\to\infty \\ u\to 0}} x_n = \frac{\alpha}{2(2a+d)}. \tag{31}$$

For x_n to exceed a fraction, $1/k$, of this maximum value, $\hat{x}_{n,max}$, we need

$$\frac{\alpha}{2(2a+d)} \left(\frac{b(1-u)}{a+b+d} \right)^{n-1} > \frac{1}{k} \cdot \frac{\alpha}{2(2a+d)}, \tag{32}$$

which is rewritten as

$$\left(\frac{a+b+d}{b(1-u)} \right)^{n-1} < k. \tag{33}$$

When $b \gg a+d$, $u \ll 1$ and $n \gg 1$, the left hand side of Eq.(33) is approximated by

$$\left[\left(1 + \frac{a+d}{b} \right) (1+u) \right]^n \approx \left(1 + \frac{a+d}{b} + u \right)^n \approx \exp\left[n \left(\frac{a+d}{b} + u \right) \right]. \tag{34}$$

Therefore condition (33) is simplified to

$$\frac{a+d}{b} + u < \frac{\ln k}{n}. \tag{35}$$

For $u = 0$ we obtain the previous condition on b. For $b \to \infty$ we obtain the error-threshold

$$u < \frac{\ln k}{n}. \tag{36}$$

The mutation rate of prelife must be less than the inverse of the sequence length, for the master sequence to reach a significant abundance in the population.

5 Replication

Let us assume that some sequences have the ability to replicate. Incorporating replication into prelife dynamics leads to the following differential equation:

$$\dot{x}_i = a_i x_{i'} - (d + a_{i0} + a_{i1})x_i + r x_i(f_i - \phi). \tag{37}$$

The first part of this equation describes prelife as before. The second part represents the standard selection equation. The coefficient, r, measures the relative contribution of selection dynamics in Eq.(37). The fitness of sequence i is given by f_i. The quantity, ϕ, is an additional death rate, which cancels out the additional production of sequences by replication. From

$$\sum_i r x_i(f_i - \phi) = 0, \tag{38}$$

we have

$$\phi = \frac{\sum_i f_i x_i}{\sum_i x_i}. \tag{39}$$

In other words, ϕ represents the average fitness of the population.

For $r = 0$, replication is absent and we recover prelife dynamics, Eq.(1). For $r \to \infty$, replication dominates and we obtain the standard selection dynamics.

We define the net reproductive rate of sequence i as

$$g_i \equiv r(f_i - \phi) - (d + a_{i0} + a_{i1}). \tag{40}$$

As in the main text, the sign of the net reproductive rate predicts a phase transition between prelife and life.

6 Replication with mutation

Imagine that sequence i of length n is the unique replicator, but its replication is susceptible to errors. In each elongation step a wrong monomer is attached with

probability u.

Let f_i be the fitness of the replicator in the absence of errors. As replication is error-free with probability $(1-u)^n$, the realized fitness of the replicator becomes

$$(1-u)^n f_i. \tag{41}$$

For the replicator to be selected, the net reproductive rate must be positive:

$$g_i = r\{(1-u)^n f_i - \phi\} - (d + a_{i0} + a_{i1}) > 0. \tag{42}$$

By using

$$(1-u)^n \approx \exp(-un) \quad (u \ll 1 \text{ and } n \gg 1), \tag{43}$$

and by neglecting ϕ (which is very small at the error threshold), condition (42) can be rewritten as

$$u < \frac{1}{n} \log \left[\frac{r f_i}{d + a_{i0} + a_{i1}} \right]. \tag{44}$$

Therefore, the replicator is selected if the mutation rate is less than the inverse of the sequence length.