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.