Journal of Theoretical Biology 428 (2017) 34–42

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/jtbi

# Error thresholds for RNA replication in the presence of both point mutations and premature termination errors

Andrew S Tupper, Paul G Higgs*

*Department of Biochemistry and Department of Physics and Astronomy, McMaster University, Hamilton, Ontario, Canada*

ARTICLE INFO

*Article history:*
Received 15 February 2017
Revised 18 May 2017
Accepted 31 May 2017
Available online 9 June 2017

*Keywords:*
RNA world
Polymerase
Ribozyme
Error threshold
Termination error
Point mutation
Parasite

ABSTRACT

We consider a spatial model of replication in the RNA World in which polymerase ribozymes use neighbouring strands as templates. Point mutation errors create parasites that have the same replication rate as the polymerase. We have shown previously that spatial clustering allows survival of the polymerases as long as the error rate is below a critical error threshold. Here, we additionally consider errors where a polymerase prematurely terminates replication before reaching the end of the template, creating shorter parasites that are replicated faster than the functional polymerase. In well-known experiments where Q$\beta$ RNA is replicated by an RNA polymerase protein, the virus RNA is rapidly replaced by very short non-functional sequences. If the same thing were to occur when the polymerase is a ribozyme, this would mean that termination errors could potentially destroy the RNA World. In this paper, we show that this is not the case in the RNA replication model studied here. When there is continued generation of parasites of all lengths by termination errors, the system can survive up to a finite error threshold, due to the formation of travelling wave patterns; hence termination errors are important, but they do not lead to the inevitable destruction of the RNA World by short parasites. The simplest assumption is that parasite replication rate is inversely proportional to the strand length. In this worst-case scenario, the error threshold for termination errors is much lower than for point mutations. We also consider a more realistic model in which the time for replication of a strand is the sum of a time for binding of the polymerase, and a time for polymerization. When the binding step is considered, termination errors are less serious than in the worst case. In the limit where the binding time is dominant, replication rates are equal for all lengths, and the error threshold for termination is the same as for point mutations.



## 1. Introduction

The transition from chemistry to life on Earth may have occurred in an RNA World (Bartel and Unrau, 1999; Gilbert, 1986; Joyce, 2002; Higgs and Lehman 2015), with RNA taking the central role as both genetic storage and enzymatic function in the first organisms. Central to this idea is the existence of a polymerase ribozyme capable of synthesizing a complementary sequence from a template, thereby forming a self-replicating chemical system. Support for such a ribozyme has come from in-vitro evolution experiments which have made significant progress in recent years (Attwater et al., 2013; Johnston et al., 2001; Lawrence and Bartel, 2005; Wochner et al., 2011; Zaher and Unrau, 2007). In the most recent case, a polymerase ribozyme was created that is capable of synthesizing 206 nt extensions which are approximately the same length as itself (Attwater et al., 2013). This polymerase ribozyme however is not perfect and has a fidelity of 97.4% (accuracy of base additions) and processivity of 97.5% (probability of sequential nucleotide addition prior to dissociation with the template strand).

During polymerase-mediated replication, a point mutation error is the incorporation of an incorrect nucleotide into the growing product strand, whereas a termination error is the premature termination of replication before reaching the end of the template, which creates an incomplete sequence that is shorter than the template. Both kinds of errors create non-functional template strands that have the potential to overrun the replicating system if the error rates are too high. Following Eigen et al. (1988), we will use the term 'error threshold' to describe the maximum error rate for which the replicating system can survive without being overrun by mutations. The fidelity of replication in the point-mutation case is a well studied question, but the termination problem has received much less attention. Here we consider these two kinds of error in the same model. Spatial lattice models have been used extensively to study replicating RNA systems

* Corresponding author.
*E-mail address:* higgsp@mcmaster.ca (P.G. Higgs).

http://dx.doi.org/10.1016/j.jtbi.2017.05.037
0022-5193/

(Szabó et al., 2002; Könnyű et al., 2008, 2013; Ma and Hu, 2012; Ma et al., 2010a,b, Ma et al., 2007a,b; Takeuchi and Hogeweg, 2012; Walker et al., 2012; Wu and Higgs, 2012; Shay et al., 2015; Kim and Higgs, 2016; Colizzi and Hogeweg, 2016a,b). Here we consider a spatial model designed to more accurately represent the process of strand replication by allowing for termination errors as well as point mutations.

The standard error threshold theory (Eigen et al., 1988) deals with point mutations. This theory considers a well-adapted 'master sequence', *e.g.* a wild-type RNA virus, in competition with all the mutant sequences that surround it in sequence space. The mutant sequences have a lower replication rate than the master sequence, but they are generated by continual mutations from the master sequence. The concentration of master sequences in the mixture is found to go to zero at a critical value of the point mutation rate called the error threshold. The value of the error threshold depends on the ratio of replication rates of the master sequence and the mutant sequences, and this can be calculated fairly easily (Eigen et al., 1988).

We have previously studied point mutations for an RNA polymerase in the RNA World (Kim and Higgs, 2016; Shay et al., 2015). In this case, the polymerase is an RNA and the polymerase sequence is mutating, whereas in the standard theory, the polymerase is a protein that is not subject to mutation. The simplest assumption for the RNA World model is that sequences with point mutations are non-functional as catalysts but are equally good templates as the polymerase. This means that the mutant sequences are parasites of the polymerase. In the well-mixed version of this model, the polymerase is overrun by parasites for any non-zero error rate. Survival of the polymerase at finite error rate requires cooperating groups of polymerase sequences, either in a surface-based model with slow diffusion, or in a protocell model with group selection (see Higgs and Lehman, 2015, and references therein). We have shown by simulation (Kim and Higgs, 2016) that there is an error threshold in the two-dimensional surface-based problem, and that spatial clustering allows survival of the polymerase for error rates below this threshold. Calculation of the error threshold for the spatial lattice model is not easy because it depends on spatial correlations of the states of neighbouring sites, which cannot be determined exactly. In the appendix of this paper, we give a paired-site approximation to the lattice model that explains the error threshold behaviour of this model at least qualitatively.

The main aim of this paper is to compare termination errors with point mutations. The strands generated by premature termination are shorter than the template and therefore replicate faster. The simplest assumption is that the replication rate of a strand is inversely proportional to its length. Hence there will be selective pressure for parasites of shorter lengths. This selective pressure was shown in well-known experiments with Q$\beta$ RNA, which shrunk its genome by 83% after many rounds of selection, causing it to replicate 15 times faster (Mills et al., 1967). The Q$\beta$ example uses a protein polymerase that does not evolve in the experiment, whereas a polymerase ribozyme in the RNA World would have to compete with the short parasites generated by premature termination. The worry that motivates this paper is that parasites might evolve to shorter and shorter lengths until they inevitably destroy the polymerase. If this were true, this would essentially rule out the idea of an RNA World that depended on an RNA polymerase ribozyme. However, the central result that we show here is that, while short parasites are indeed lethal to the polymerase when alone, termination errors generate a mixture of parasites of different lengths, and this mixture is not always lethal. The polymerase survives in the presence of the mixed parasites up to a finite error threshold.

## 2. Methods

The model used is an extension of that in Kim and Higgs (2016). We use a square lattice in which each site can either be vacant or occupied by a single RNA strand. A strand is either a polymerase (P), the complement to a polymerase (C), or a parasite (X). Lengths of strands are measured in numbers of nucleotides. Polymerases and complements have a fixed length $L_{pol}$. Parasites may have any length up to and including $L_{pol}$. We assume that the surface environment limits diffusion, thus preventing strands from moving between lattice sites. The model allows only polymerase-catalysed replication and assumes that non-enzymatic template-directed replication is negligible. For a similar model that incorporates both kinds of replication, see Wu and Higgs (2012) and Shay et al. (2015). Polymerases replicate neighbouring template strands at rate $k(L)$, where $L$ is the length of the template. The simplest case, which was also used by Kim and Higgs (2016), is to assume that $k(L)$ is inversely proportional to $L$. It is convenient to write

$$k(L) = k_{pol}\frac{L_{pol}}{L}, \tag{1}$$

where $k_{pol}$ is the replication rate constant for a strand of length $L_{pol}$. We also introduce a cutoff length, $L_{cut}$, which is the minimum length of template that can be replicated. This is motivated by experiments with Q$\beta$ replicase, where the optimum RNA strands for replication have a length much shorter than the full virus RNA, but must be long enough to have a secondary structure that is recognized by the Q$\beta$ replicase protein (Biebricher and Luce 1992, 1993). For all the simulations in this paper, we have $L_{pol} = 100$ and $L_{cut} = 10$. For simplicity, we assume that one strand occupies one lattice site, irrespective of its length. However, if strands shorter than $L_{cut}$ are generated by termination errors, these strands are too short to be replicated again, and they are also assumed to be too short to take up a lattice site. These very short strands are simply ignored because they play no role in the model. The form for $k(L)$ in Eq. (1) is the worst-case scenario for survival of the polymerase, because it gives the short parasites the maximum advantage. Later in the paper we will consider more realistic forms for $k(L)$ in which the replication rate is less strongly length dependent.

The model also incorporates loss of strands. Strand sites are turned into vacancies at a rate that is assumed constant for all lengths, and is set to 1. This provides a scale for comparison of the other rates in the model. The loss rate represents either the escape of a strand from the surface or the breakdown of a strand back to individual monomers.

The simulations proceed in time steps of length $\delta t$. In one time step, we visit every strand in a random order and give it a chance to be a template. For each strand, we select one of the eight neighbouring sites at random from the Moore neighbourhood. If this site is occupied by a polymerase, the template is replicated with a probability $k(L)\delta t$. Only strands next to polymerases can be replicated. The new strand is the complement of the template: a P creates a C, and a C creates a P. Note that C strands are non-functional, but they are necessary for replication of the polymerases. If a point mutation occurs, a parasite of length $L_{pol}$ is generated, instead of a P or C. If a termination error occurs, a parasite of a length less than $L_{pol}$ is generated (further details below). If the template is a parasite of length $L$, accurate replication creates another parasite of the same length. Since all parasites of a given length are equivalent in this model we do not keep track of their plus and minus forms. Point mutations are not relevant to parasites for the same reason. Premature termination of replication of a parasite can generate shorter parasites.

When a new strand is created, we select a second random neighbour site of the template strand different from the site occupied by the polymerase. If the second neighbour site is a vacancy,

the new strand is placed on this site. If the second site is already occupied by a strand, the new strand is eliminated and no change occurs. After giving each strand a chance to replicate, we again go through each strand in a random order and give it a chance to be lost/broken down. This occurs with a probability $\delta t$, because the loss rate is defined as 1. This completes one time step $\delta t$.

We now discuss errors in more detail. For point mutations, we assume an error probability $m_{point}$ per nucleotide. The probability of at least one point mutation occurring during replication of a sequence of length $L$ is

$$M(L) = 1 - (1 - m_{\text{point}})^L. \tag{2}$$

P and C sequences have length $L_{pol}$. Hence, when a P or C is replicated, there is a probability $M(L_{pol})$ that the new strand is a parasite of length $L_{pol}$.

For termination errors, we assume that there is a probability $m_{term}$ of premature termination at each nucleotide. Let $p(l)$ be the probability of generation of a new strand of length $l$ from a template of length $L$.

$$\begin{aligned} p(l) &= (1 - m_{term})^l m_{term} \qquad (\text{for } 0 \le l \le \mathrm{L} - 1) \\ p(L) &= (1 - m_{term})^L \qquad (\textit{accurate replication}) \end{aligned} \tag{3}$$

This is normalized so that $\sum_{l=0}^{L} p(l) = 1$. There is a subtlety here regarding the replication rate. The rate of production of the new strand should depend on its own length, $l$, not on the length of the template, $L$. Hence, when a template is about to be replicated, we first determine the length of the product $l$ as a random value from the distribution $p(l)$. If $l < L$, the probability of the replication occurring is $k(l)\delta t$, rather than $k(L)\delta t$. In simulations that include both kinds of error, we first check for premature termination, then if the strand is accurately replicated according to Eq. (3), we check for point mutations according to Eq. (2).

All results reported here are from simulations using a square lattice of size $1024 \times 1024$ with periodic boundaries to limit edge effects. The time step is $\delta t = 0.001$ in all cases, except for runs with very high polymerization rates ($k_{pol} \ge 1000$), in which case it was necessary to decrease the time step to $\delta t = 0.0001$.

## 3. Polymerase survival with point mutation errors

We will first consider the simplest case with only point mutations and no termination errors. Due to the clustering of polymerases that arises in this spatial model, it is less likely for a parasite to be adjacent to a polymerase ribozyme than for a polymerase to be adjacent to another polymerase or complement. This means that parasites have a disadvantage, and they die out if there is no continued mutation ($m_{point} = 0$). For moderate $m_{point}$, there is coexistence of the parasites with the polymerase and complement. An example of this situation is shown in Fig. 1(a). For the corresponding animation see Video S1.

Fig. 2 shows the time averaged strand concentrations as a function of $m_{point}$ when all other variables are fixed. The error threshold is close to $m_{point} = 2.5 \times 10^{-3}$ for these parameters, which means the mutation probability per strand is $M = 0.22$ for a strand of length $L_{pol} = 100$. If the error threshold occurs at $M$ of order 1, then the maximum per-base error rate is $m_{point}$ of order $1/L_{pol}$. This is a similar conclusion to the usual error threshold in the master-sequence landscape. Eigen et al. (1988) showed that the minimum fidelity in that case is $Q = 1/\sigma_0$, where $\sigma_0$ is the relative replication rate (superiority parameter) of the master sequence to the mutants. In our case, however, the polymerase survives not because of a replication rate advantage, but due to an advantage arising from spatial clustering.

The example in Fig. 2 is calculated for $k_{pol} = 25$. Fig. 3 shows the way the error threshold in $m_{point}$ depends on $k_{pol}$ (with $L_{pol}$ fixed at 100). Firstly, we note that, even in absence of errors, there is a minimum value of the replication rate necessary for the survival of the polymerase. This is close to 8.6. Below this, the rate of loss/breakdown of strands is faster than the multiplication rate. The error threshold in Fig. 3 is therefore zero below $k_{pol} = 8.6$. Above this, the error threshold increases as $k_{pol}$ increases, because speed partially compensates for accuracy. If the replication rate is larger, then the number of accurate copies produced from one strand in its lifetime is larger, which helps the survival of the polymerase, even if mutant copies are also produced. However, Fig. 3 shows that the error threshold begins to decrease again very slowly at very high replication rates above $k_{pol} = 50$. This is because the spatial structure of the model breaks into very small clusters of polymerase and complement that are effectively "walled-in" by parasites. The parasites stop the polymerase clusters from growing, even though the parasites cannot spread. The concentration of the polymerases actually decreases with $k_{pol}$ over this range, and the error threshold also decreases slightly in consequence.

We do not have an exact calculation of strand concentrations or the error thresholds for this model. An approximate solution can be found using a pair approximation that considers correlations in the states of pairs of neighbouring sites but ignores correlations beyond two sites. This calculation is shown in the Appendix, and the strand concentrations as a function of $m_{point}$ are shown in Fig. A1. The result is qualitatively similar to Fig. 2, although the value of the error threshold found from the approximation is substantially higher than that found from simulation of the lattice model.

## 4. The existence of lethal parasites

As pointed out in the previous section, parasites of the same length as the polymerase have a disadvantage due to spatial clustering of the polymerases. Therefore, these parasites die out unless they are continually created by mutation. However, shorter parasites have an advantage due to faster replication which can overcome the disadvantage due to clustering. We already gave examples (Kim and Higgs, 2016) where short independent parasites coexist with the polymerase, and where very short parasites destroy the polymerase entirely. Here, we investigate how short the parasites need to be in order to be lethal.

A simulation of just P and C sequences was allowed to reach a steady state without mutations or parasites of any kind. A small number of parasites of a fixed length $L$ was then added in order to see if these parasites could invade the replicating system. The mutation rates were kept at zero, so the only parasites present are copies of the initial few that were added.

Three outcomes are possible, depending on $L$. For $L > L_{max}$, the parasites die out. For $L_{min} \le L \le L_{max}$, the parasites coexist with the polymerase and complement. For $L < L_{min}$, the parasites destroy the polymerases and everything dies out. Fig. 4 shows the time averaged concentrations of the strands as a function of $L$. For these parameters, where $k_{pol} = 25$ and $L_{pol} = 100$, we estimate $L_{min} = 14$ and $L_{max} = 76$. This problem can also be studied approximately using the pair approximation shown in the Appendix. The results shown in Fig. A2 are qualitatively similar to those in Fig. 4.

In contrast to the case of point mutations (Fig. 1a), the spatial dynamics in the case of coexisting parasites and polymerases without mutations gives rise to travelling waves (Fig 1b and 1c). Polymerases and complements form the leading edge of the waves while parasites survive on the trailing edge. Travelling waves in similar models to this have been observed previously (Takeuchi and Hogeweg, 2012; Colizzi and Hogeweg, 2016a,b). For the longest parasites in the range $L_{min} \le L \le L_{max}$, the traveling waves are small and constantly colliding (Fig. 1b) whereas shorter parasites result in larger traveling waves (Fig 1c). The large scale travelling waves for the shorter parasites are only possible if the lattice size is large enough, *i.e.* short parasites are more lethal if the lattice size

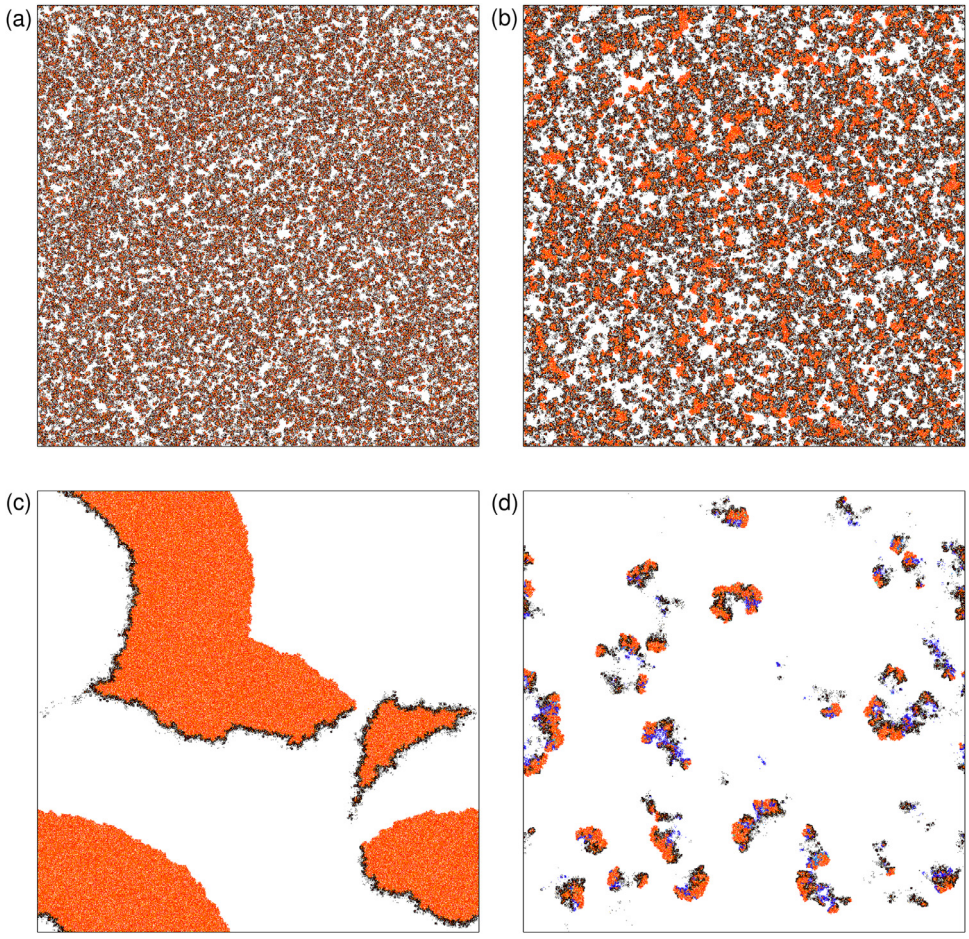


**Fig. 1.** Snapshots from simulations, where polymerases are shown in red, complements in orange, and parasites in black. In all cases $k_{pol} = 25$ and $\delta t = 0.001$. (a) Small clusters are seen in the case of point mutation rate $m_{point} = 1.6 \times 10^{-3}$ and no termination errors. (b) Small chaotic waves emerge when parasites of fixed length 45 coexist with polymerases and complements and no replication errors are allowed. (c) Large travelling waves are seen when parasites are shortened to length 15 under the same constraint of no replication errors. (d) When termination errors occur with $m_{term} = 1.0 \times 10^{-5}$, small waves emerge which constantly collide, split, and die. Parasites are coloured according to length, from black (short) to long (light blue). "(For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)".

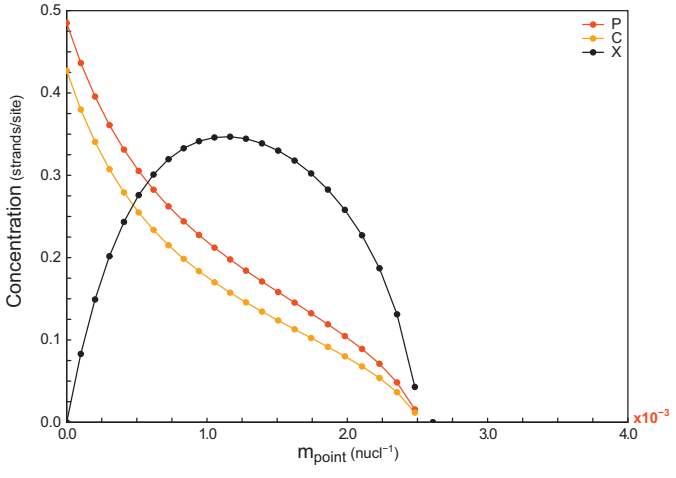


**Fig. 2.** Concentrations of polymerase, complements, and parasites are shown as a function of $m_{point}$, with $k_{pol} = 25$ and $\delta t = 0.001$. The point mutation error threshold occurs around $2.5 \times 10^{-3}$ where the average polymerase population goes to zero. All simulations were run until $t = 1000$ and repeated 100 times for each value of $m_{point}$ shown. Each data point represents time and simulation averaged strand concentration if at least 5% of the trials had a surviving polymerase population, and 0 otherwise.

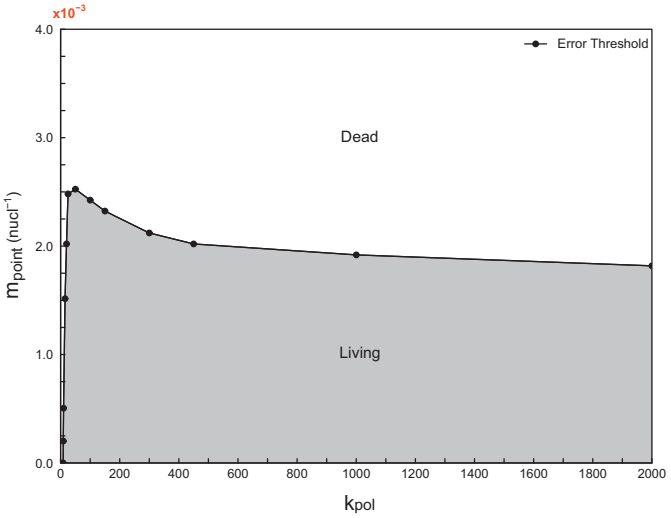


**Fig. 3.** The point mutation error threshold is shown as a function of $k_{pol}$. For each value of $k_{pol}$, simulations were run until $t = 1000$ and repeated 100 times for increasing values of $m_{point}$. Each data point represents the largest value of $m_{point}$ for which at least 5% of the trials had a surviving polymerase population.

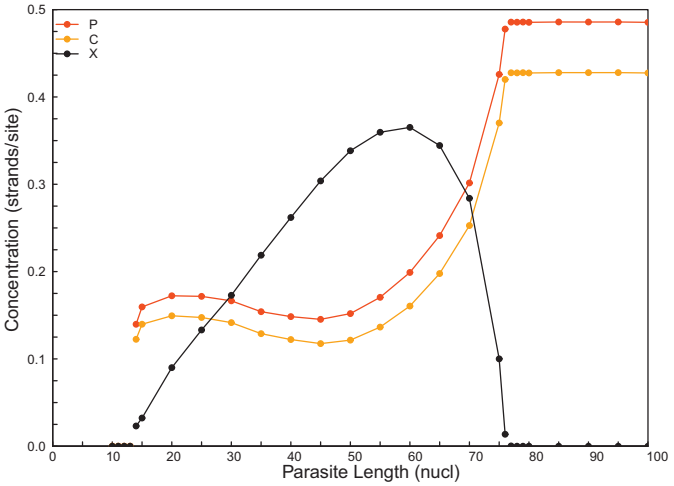


**Fig. 4.** Equilibrium concentrations of polymerase, complements, and parasites are shown as a function of parasite length. Coexistence of polymerases and parasites is possible when parasites have lengths in the range 14–76 nucleotides. All simulations were initialized with a small population of parasites with fixed length added to a polymerase population, and no point mutations or termination errors were allowed. Each simulation was run until $t = 1000$ and repeated 100 times for each length of parasite shown. Each data point represents time and simulation averaged strand concentration if at least 5% of the trials had a surviving polymerase population, and 0 otherwise. Common parameters used are $k_{pol} = 25$ and $\delta t = 0.001$.

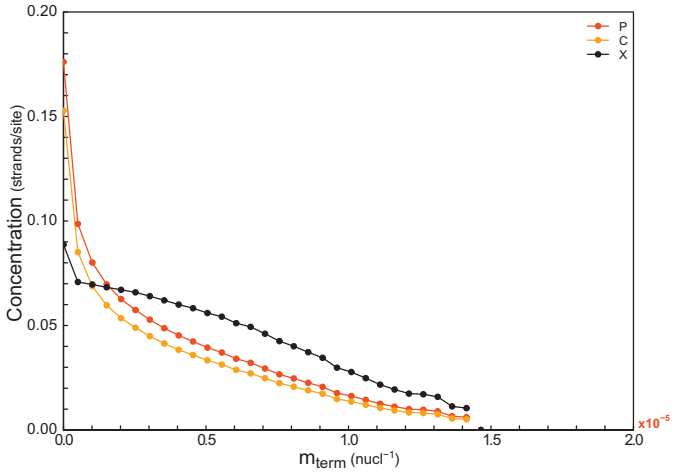


**Fig. 5.** Concentrations of polymerase, complements, and parasites are shown as a function of $m_{term}$, with $k_{pol} = 25$ and $\delta t = 0.001$. The termination error threshold occurs around $1.5 \times 10^{-5}$, a factor of 100 times smaller than the point mutation error threshold. All simulations were run until $t = 1000$ and repeated 100 times for each value of $m_{term}$ shown. Each data point represents time and simulation averaged strand concentration if at least 5% of the trials had a surviving polymerase population, and 0 otherwise.

is smaller. Animations of these cases are also available. Video S2 shows the case where $L < L_{min}$, where introduction of a few short parasites is lethal. Videos S3 and S4 show travelling waves.

The existence of short, lethal parasites is unsettling, as termination errors will inevitably generate parasites with $L < L_{min}$. One solution to this problem would be to evolve a polymerase that does not bind to templates that are too short. As we discussed in the introduction, this does seem to be the case for RNA replication by the Q$\beta$ replicase protein (Biebricher and Luce 1992, 1993). This feature is included in our model via the cutoff length $L_{cut}$, which is the minimum length template strand that can bind to the polymerase and be replicated. It is clear that if we set $L_{cut}$ larger than $L_{min}$, the parasites that are short enough to be lethal cannot be replicated, which eliminates the problem of lethal parasites immediately. However, the main aim of this paper is to look at the case of termination errors, which continually generate parasites of a mixture of lengths. We will now show that the short parasites are *not* lethal when present in this mixture, and that the polymerase can coexist with the mixture of parasites, even if $L_{min} > L_{cut}$.

## 5. Polymerase survival with termination errors

In this section, we consider simulations with termination errors but no point mutations. Parasites are generated of all lengths shorter than the template, with a probability distribution $p(l)$ as described in the Methods section. When parasites of all lengths are present, a new equilibrium emerges in which competition between parasites of varying lengths prevents the lethality of short strands. This stable state (Fig 1d) is visually distinct from the previous cases (Figs 1a–c), in that small traveling waves arise that are separated by large amounts of empty space. The wave structures are rather irregular and can split to generate new waves. Waves also collide sometimes, which tends to lead to death of the colliding waves because they are surrounded by parasites on all sides. For animations of polymerases surviving with termination errors see Videos S5-S7.

For the parameters in this example, $L_{min} = 14$, and $L_{cut} = 10$. This means that potentially lethal parasites in the range $L = 10$–$13$ can be replicated. Nevertheless, the system survives. Fig. 5 shows the concentration of polymerase and complement as well as the total concentration of parasite strands of all lengths as a function of $m_{term}$. These simulations were done in the following way. For each value of $m_{term}$, simulations were run until $t = 1000$ and repeated 100 times. Each data point represents the time and simulation averaged parasite concentration of all simulations which had a surviving polymerase population. Common parameters used are $k_{pol} = 25$ and $\delta t = 0.001$.

It can be seen in Fig. 5 that the system survives with non-zero $m_{term}$ up to a critical error threshold, in a similar way as for point mutations. The reason for this seems to be the fragmentary structure of the travelling waves. Since medium-length parasites coexist with polymerases in a traveling wave, the emergence of a short parasite is no longer lethal as it now has to compete for vacant sites with all non-lethal parasites present on the wave edge. We also observed occasions where enough short parasites accumulated to encapsulate and destroy a wave. However, since waves are widely separated from each other, the death of an individual wave does not result in the death of the entire polymerase population. Wave death is offset by the formation of new waves, which occurs when a wave splits due to the emergence of an internal parasite, or the escape of a polymerase from the trailing edge of the wave. This is an example of multi-level evolution acting at the higher level of the wave as well as the lower level of the single molecule (see also Takeuchi and Hogeweg, 2012; Colizzi and Hogeweg, 2016a,b).

The key point up to now is that the system with continued creation of parasites of all lengths by termination errors is stable up to a finite error threshold, even though the system with only very short parasites would be unstable, even with zero mutation rate. We can therefore be satisfied that the RNA World is not inevitably destroyed by the existence of termination errors. Nevertheless, comparison of Figs. 2 and 5 shows an important point. The error threshold for termination errors is two orders of magnitude smaller than for point mutations for the parameters we investigated ($m_{term} = 1.5 \times 10^{-5}$ in comparison to $m_{point} = 2.5 \times 10^{-3}$). This is not surprising, since termination errors result in parasites with a larger replication rate than those resulting from point errors. It does raise a substantial worry as to whether such low values of $m_{term}$ could be achieved by the earliest ribozymes.

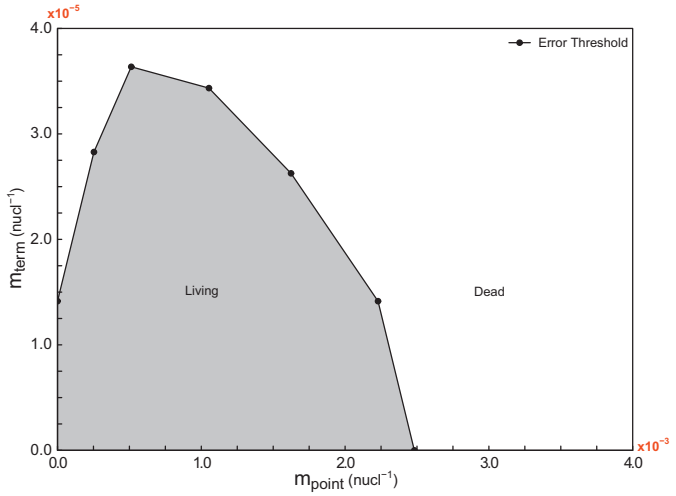


**Fig. 6.** The termination error threshold is shown as a function of point mutation rate. The shaded gray area corresponds to the combinations of point mutation and termination error rates for which the polymerase survives. For each value of $m_{point}$, simulations were run until $t=1000$ and repeated 100 times for increasing values of $m_{term}$. Each data point represents the largest value of $m_{term}$ for which at least 5% of the trials had a surviving polymerase population. Common parameters used are $k_{pol}=25$ and $\delta t=0.001$.

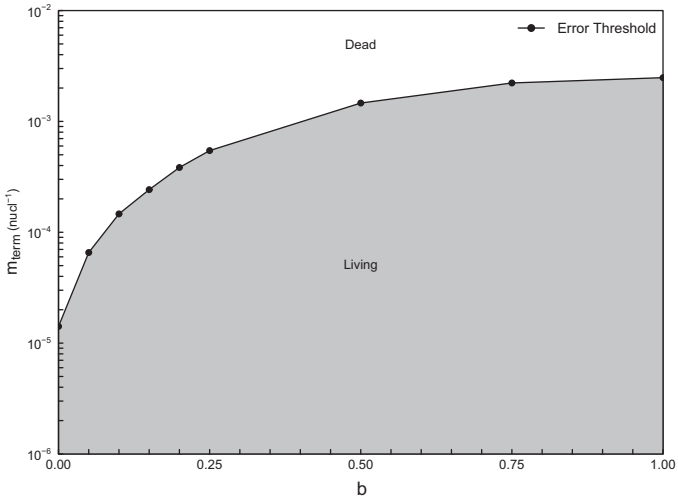


**Fig. 7.** The termination error threshold is shown as a function of the binding variable $b$. For each value of b, simulations were run until $t=1000$ and repeated 100 times for increasing values of $m_{term}$. Each data point represents the largest value of $m_{term}$ for which at least 5% of the trials had a surviving polymerase population. Common parameters used are $k_{pol}=25$ and $\delta t=0.001$.

Another difference between Figs. 2 and 5 is that, for the termination errors, the concentration of parasites approaches a non-zero constant as $m_{term}$ tends to zero, whereas for point mutations, the concentration of parasites tends to zero as $m_{point}$ tends to zero. This is because medium length parasites coexist with the polymerase when there is no mutation, whereas parasites of length $L_{pol}$ do not coexist with the polymerase in absence of continued mutation. All simulations reported here were initialized with 15% polymerases, 15% complementary sequences, and 15% parasites of random lengths. The point at $m_{term}=0$ in Fig. 5 appears at a non-zero parasite concentration because parasites were initially present and medium length parasites coexist with the polymerase indefinitely, even without mutation (as we showed in Fig. 4). Clearly, if we began with no parasites, the parasite concentration would remain zero when $m_{term}=0$. In contrast, in the point mutation case in Fig. 2, it doesn't matter whether a few parasites are present initially or not, because parasites of length $L_{pol}$ would disappear anyway when $m_{point}=0$.

## 6. Polymerase survival with point and termination errors

Now that we have shown polymerase ribozymes can survive in the presence of each type of error separately, we will consider the case in which both errors are possible. While it would seem that combining point mutations and termination errors would inevitability result in a new error threshold that is lower than either individually, this is not the case. The presence of a certain amount of point mutations in fact increases the termination error threshold by a factor of more than 2 in the best cases relative to the case with only termination errors (Fig. 6). Increasing the point error rate results in a new source of long parasites that limit the replicative advantage of short parasites that are competing for vacant sites on the trailing edge of a wave. This competition hinders short parasites, thereby allowing the polymerase population to survive for higher termination error rates.

## 7. Distinguishing binding and nucleotide addition steps

So far, we supposed that the replication rate was inversely proportional to the strand length. This is the worst-case scenario, because it gives maximum advantage to short parasites. Here we consider a slightly more realistic model of replication that distinguishes an initial step of binding of the polymerase to the template and a step of nucleotide addition. Let $T_B$ be the mean time for binding. We assume that this time is the same for templates of all lengths of at least $L_{cut}$, and that templates shorter than $L_{cut}$ cannot bind at all. Let $T_A$ be the mean time for adding one nucleotide. The mean time for copying a polymerase sequence of length $L_{pol}$ is $T_{pol}=T_B+T_AL_{pol}$. The fraction of time spent in the binding step is $b=T_B/T_{pol}$, while the fraction spent in the nucleotide addition steps is $(1-b)=T_AL_{pol}/T_{pol}$. We will keep the model simple by treating replication as a single effective step with a rate that is the inverse of the mean time. Hence for the polymerase, the rate is $k_{pol}=1/T_{pol}$, and for a strand of length $L$, the rate is

$$k(L)=\frac{1}{T_B+T_AL}=k_{pol}\frac{L_{pol}}{bL_{pol}+(1-b)L}. \tag{4}$$

If we set $b=0$, there is no time spent on binding, and we are in the worst case (same as Eq. (1)). If we set $b=1$, the time for nucleotide addition is negligible compared to the binding time. In this case, all strands replicate at rate $k_{pol}$ irrespective of their length. For an intermediate value of $b$, $k(L)$ increases for shorter parasites, but less drastically than in the worst case.

Fig. 7 shows results of simulations in which the replication rate follows Eq. (4) with different values of $b$. Termination errors occur at rate $m_{term}$ and no point mutations are included. The termination error threshold is shown as a function of the binding proportion $b$. When $b=0$, the error threshold is equivalent to our previous case of termination errors (Fig. 5). As the binding proportion increases, the termination error threshold increases by two orders of magnitude, *i.e.* when $b>0$, the system is much more tolerant of termination errors than in the worst possible case, and short parasites are much less dangerous. When $b=1$, all parasites have the same replication rate, and their length is not important. Hence termination errors are equivalent to point mutation errors, and the error threshold for $m_{term}$ when $b=1$ in Fig. 7 is the same as the error threshold for $m_{point}$ in Fig. 2. These results highlight the importance of the binding step, as even a 10% binding proportion is enough to increase the termination error threshold by a factor of 10.

## 8. Discussion

Parasites are likely to be important in the RNA world as it is easy for a sequence to be a template and difficult for it to be a polymerase. We have presumed that mutations that prevent the function of the polymerase will produce sequences that are still viable templates. Therefore, no special adaptation of the template sequence is required for it to act as a parasite. Nevertheless, selection on parasites will tend to increase their replication rate, and decreasing the template length is an easy way to do this without requiring any special adaptation. Hence, termination errors will provide a constant source of shorter and shorter parasites. Using our computational model we have shown that survival of a polymerase ribozyme is possible in a surface model when both point mutations and termination errors are considered. Hence, the tendency for selection of shorter parasites does not inevitably kill the polymerases.

In both the cluster patterns that arise in the point mutation case and the travelling wave patterns that arise in the termination error case, the polymerases are more likely to be next to other polymerases and complements and less likely to be next to parasites than they would be in the well-mixed case. This gives an advantage to the polymerases that allows them to survive up to a finite error threshold, whereas the polymerases are destroyed by parasites in the well mixed case for any non-zero error rate. There are several other models that show that spatial pattern formation can allow the survival of polymerases in the presence of parasites (Takeuchi and Hogeweg, 2012; Colizzi and Hogeweg, 2016a,b). These models also show travelling wave patterns similar to ours. These models allow the rate parameters of the parasites to evolve without explicitly considering the length of the template. Including the length as a parameter in our case makes the comparison possible between the termination errors and point mutations. We have also investigated the factors that affect the error threshold, which was not done previously.

We showed in Section 3 that the error threshold in the per-sequence error rate $M$ that is achievable by spatial clustering can be relatively large, but this still implies a per base error rate that must be very small and varies inversely with the length of the template. The value of the error threshold depends on many details, including whether the functional sequence is maintained by replication rate advantage or by clustering (as in this paper), the presence of neutral networks in the fitness landscape (Reidys et al., 2001; Wilke 2001; Takeuchi et al., 2005; Szilagyi et al., 2014) and the possibility of recombination (Santos et al., 2004). While these factors make quantitative differences, they do not really change the nature of the problem: replication must be accurate in order to sustain an RNA World. Szilagyi et al. (2014) calculate the phenotypic error threshold, and conclude that known ribozymes of lengths up to about 200 could be successfully replicated by currently known polymerases with per-base error rates of a few percent (Johnston et al., 2001; Wochner et al., 2011). However, they assume that there is a very large replication rate advantage to the sequences with the correct secondary structure, and that fitness only depends on the secondary structure. Furthermore, they assume that the polymerase is fixed, and is not itself subject to replication, whereas in the RNA World, the polymerase has to replicate other copies of itself, so a mechanism such as spatial clustering or compartmentalization is required to prevent the invasion of parasites. Hence, the conclusions of Szilagyi et al. (2014) seem somewhat optimistic to us. Nevertheless, we do not wish to argue against the RNA World hypothesis. In our view, there is a lot of evidence that supports the existence of an RNA World in the early stages of life on Earth (Higgs and Lehman, 2015), and theoretical treatments demonstrate that small per-base error rates are required for this to work. Therefore it becomes an experimental problem to demonstrate that sufficiently accurate polymerase ribozymes are possible. Work on error rates in non-enzymatic replication is also relevant here (Rajamani et al., 2010), which demonstrates that stalling of replication after an error slows down the replication of sequences with errors, giving an advantage to correctly copied sequences and an increase in the error threshold. This same effect could also occur in ribozyme catalysed replication.

We would like to summarize several important points that emerge from our studies in this paper. Short parasites are clearly lethal in this model when they are introduced into a connected system of polymerases and complements. Nevertheless, when parasites of all lengths are present, as is the case for termination errors, the lethality of short parasites is prevented because the system is broken up into separate fragments, and the lethal parasites cannot travel through the whole system. Our model allows us to compare the error thresholds from point mutations and termination errors. In general termination errors are more dangerous than point mutations because they create faster replicating parasites. Hence The error threshold in $m_{term}$ was found to be two orders of magnitude less than that for $m_{point}$ in the worst case, where replication rate varies inversely with template length. Competition between parasites of different lengths was further extended to the case in which we considered both errors simultaneously and showed that the addition of a point mutation rate in fact increased the termination error threshold above that which occurs with termination errors alone. This is again because the presence of non-lethal, long parasites created by point mutations changes the spatial pattern in which the shorter parasites evolve. An interesting factor that we did not yet consider is that point mutations are likely to impose a stalling effect on a polymerase similar to the stalling effect seen in non-enzymatic replication (Rajamani et al., 2010), and this might lead to an increase in the termination likelihood. Lastly, we showed that when a binding step was incorporated in the replication process, termination errors become much less dangerous than in the worst case because the relative advantage of the short parasites is reduced. We found that the termination error threshold approaches the point mutation threshold if the binding step is long compared to the polymerization step.

This paper therefore supports the idea that a polymerase ribozyme replicating on a surface may have supported the early stages of life. However, one limitation that should be borne in mind, is that in the present paper, there is no diffusion of strands across the surface. We already showed in the case with point mutations only (Kim and Higgs, 2016) that the incorporation of strand diffusion benefits parasites and reduces the error threshold. An alternative mechanism that prevents parasites from destroying polymerase systems is to place the replicators in protocells, in which case group selection at the cell level can overcome individual selection at the strand level (Takeuchi and Hogeweg, 2009; Higgs and Lehman (2015) and references therein). In an interesting recent experimental realization of the protocell case (Matsumura et al., 2016), it was also found that parasites of high replication rate do not inevitably destroy the system. Thus, both spatial surface models and protocell models agree that polymerase systems can survive, although it is still not clear which of these two factors was more important in actual evolutionary history, and whether there were surface-based replicating systems prior to the origin of cells. Another important question for early life is how a system with a single kind of functional ribozyme could evolve additional functions. Kim and Higgs (2016) showed that an unlinked strand functioning as a nucleotide synthetase can coexist and cooperate with the polymerase in some cases. We suggested that building up a metabolism controlled by many ribozymes might be easier in a proto-cell model than a surface based model, although this has not yet been fully tested. It will be also interesting to consider the relative sizes of error thresholds in proto-cell and surface based

models in future. Computational models provide a useful way to test ideas about the origin of life and early evolution. Here we attempted to solve just one piece of the puzzle by showing that short parasites are not lethal to a polymerase population even if there is a selective pressure for faster replication.

**Acknowledgement**

This work was supported by Discovery Grant 261831 from the Natural Sciences and Engineering Research Council of Canada.

**Appendix. Paired-site approximation**

Spatial correlations in the states of neighbouring sites are essential in this model. An exact mathematical treatment of the spatial model is not possible, but it is possible to qualitatively explain some of the results in the simulations by using a paired site approximation. Let $X_1$ and $X_2$ be the concentration of the polymerase and complement and $X_3$ be the concentration of one kind of parasite. The concentration of vacancies is $X_0 = 1 - X_1 - X_2 - X_3$. In absence of spatial correlation, the following mean field equations apply:

$$\frac{dX_1}{dt} = k_{pol} X_0 X_1 X_2 (1 - M_{pol}) - X_1 \tag{A.1}$$

$$\frac{dX_2}{dt} = k_{pol} X_0 X_1^2 (1 - M_{pol}) - X_2 \tag{A.2}$$

$$\frac{dX_3}{dt} = k_3 X_0 X_1 X_3 + M_{pol} k_{pol} X_0 X_1 (X_1 + X_2) - X_3 \tag{A.3}$$

Here, $k_3$ is the rate of replication of the parasite. For the case in Section 3, the parasite is a point mutation, so $k_3 = k_{pol}$, and $M_{pol} = 1 - (1 - m_{\text{point}})^{L_{pol}}$. For the case in Section 4, the parasite is an independent shorter parasite with $k_3 = k(L)$ and there is no mutation ($M_{pol} = 0$). The mean field equations do not explain the behaviour seen in the lattice simulations in either case, because the point-mutation parasite always destroys the system if $M_{pol} > 0$, and the independent parasite can only coexist with the polymerase if it has exactly the same replication rate ($k_3 = k_{pol}$).

The simplest approximation that accounts for correlations in the states of neighbouring sites is to consider pairs of sites in the following way. Let $C_{ij}$ be the frequency with which the first site is in state $i$ and a random neighbour site is in site $j$. We are working with the Moore neighbourhood where there are 8 neighbouring sites. In this approximation there is no distinction between a horizontal/vertical neighbour and a diagonal neighbour. In the equations below, we will distinguish the order of the indices, although it is clear by symmetry that $C_{ij} = C_{ji}$. The concentrations of the single sites can be obtained from the pair frequencies: $X_i = \sum_j C_{ij}$, where the sum goes over states 0 to 3. The deterministic differential equations for the pair frequencies are as follows.

$$\frac{dC_{ij}}{dt} = C_{i0} R_{i0,ij} + C_{0j} R_{0j,ij} - 2C_{ij} \quad (\text{when i} \neq 0 \text{and j} \neq 0) \tag{A.4}$$

$$\frac{dC_{i0}}{dt} = C_{00} R_{00,i0} - C_{i0} \sum_{j \neq 0} R_{i0,ij} - C_{i0} + \sum_{j \neq 0} C_{ij} \quad (\text{when i} \neq 0) \tag{A.5}$$

$$\frac{dC_{00}}{dt} = -\sum_{i \neq 0} C_{00} R_{00,i0} - \sum_{j \neq 0} C_{00} R_{00,0j} + \sum_{i \neq 0} C_{i0} + \sum_{j \neq 0} C_{0j} \tag{A.6}$$

$R_{00,\,i0}$ is the rate at which an $i$ is synthesized in a 00 pair, and $R_{0j,\,ij}$ is the rate at which an $i$ is synthesized in a 0j pair. We will first consider these replication terms in absence of mutation. We will use a "prime", $R'$, to denote that the rate is in absence of mutation. Then we will write the full rates, $R$, in terms of the $R'$. As an example, consider $R'_{00,10}$. As a type 1 strand is being formed, the template must be a type 2 strand. The type 2 strand must be on one of the 7 neighbours of the vacancy other than the second site in the pair. The expected concentration of 2s on the neighbours using the paired-site approximation is $C_{02}/X_0$. From the definition of the lattice model, each template chooses two random neighbours, the first of which must be a polymerase, and the second of which must be a vacancy. There is a probability 1/8 that the second site chosen is the vacancy in the pair under consideration. The first site, on which the polymerase must be found, is a different neighbour of the template. The expected frequency of polymerases (type 1) on neighbours of type 2 sites is $C_{21}/X_2$. Putting these factors together, we obtain

$$R'_{00,10} = \frac{7k_{pol}}{8} \frac{C_{02}}{X_0} \frac{C_{21}}{X_2}.$$

It follows that $R'_{0j,1j} = R'_{00,10}$, as long as $j$ is not a template of 1. If $j = 2$ (the template of 1), then there is one neighbour where we know there is a template, in addition to the 7 neighbours where there might be a template. In this case

$$R'_{02,12} = \frac{k_{pol}}{8} \left(1 + 7\frac{C_{02}}{X_0}\right) \frac{C_{21}}{X_2}.$$

As another example, consider $R'_{00,30}$. This case the template of a 3 must be a 3. This template must be a neighbour of the 0 site. There must also be a polymerase (type 1) as a neighbour of the template 3. Hence

$$R'_{00,30} = \frac{7k_{pol}}{8} \frac{C_{03}}{X_0} \frac{C_{31}}{X_3}.$$

By similar logic,

$$R'_{03,33} = \frac{k_{pol}}{8} \left(1 + 7\frac{C_{03}}{X_0}\right) \frac{C_{31}}{X_3}.$$

In order to account for mutations, we note that replication rates involving ribozymes and their complements are reduced by factors $1 - M_{pol}$ and rates involving mutant sequences are increased by the corresponding amount. For example:

$$R_{00,10} = R'_{00,10}(1 - M_{pol})$$

$$R_{00,20} = R'_{00,20}(1 - M_{pol})$$

$$R_{00,30} = R'_{00,30} + M_{pol}(R'_{00,10} + R'_{00,20})$$

All the other $R$ functions can be obtained using the same method.

We found the stable values of $C_{ij}$ by numerical simulation of the paired-site Eqs. (A.4–A.6), and hence determined the concentrations $X_i$. Fig. A1 shows the point-mutation case as a function of the point mutation rate for the same parameters as Fig. 2. The approximation qualitatively predicts the shape of this curve, although the predicted error threshold is considerably larger than in the lattice simulation.

We also used the approximation to calculate the frequencies of the three types of strand when an independent parasite is coexisting with the polymerase and complement in absence of mutation. Fig. A2 is similar to Fig. 4. The approximation correctly predicts that there is a length $L_{max}$ above which the parasite dies, and a length $L_{min}$ below which the parasite destroys the system. The predicted values are not particularly close to the ones obtained in the lattice simulation, and the approximation considerably overestimates the parasite concentration in the region where coexistence occurs.

Better approximations could be obtained by accounting for correlations over more than two sites, but the paired site approximation already does a reasonable job at explaining the qualitative behaviour of the model.

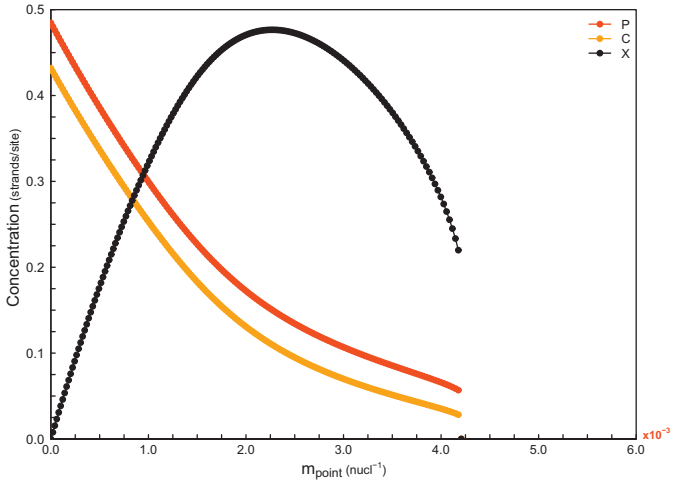


**Fig A1.** Error threshold as a function of mutation according to the paired-site approximation. Polymerase P$=X_1$, Complement C$=X_2$, Parasite $X=X_3$.

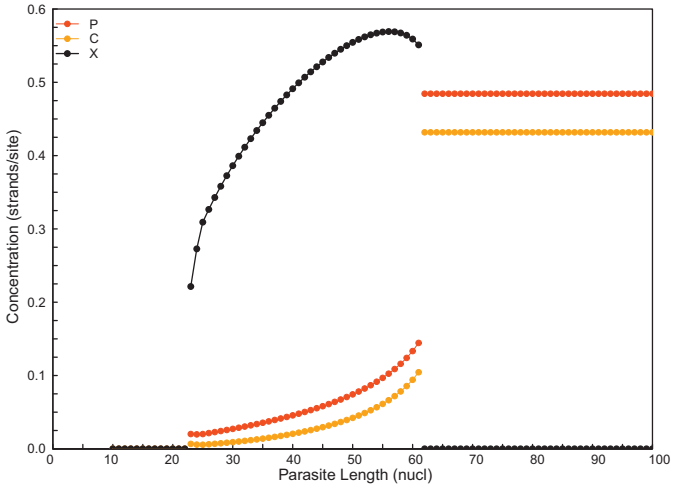


**Fig A2.** Range of L for coexistence of an independent parasite with a polymerase. Polymerase P$=X_1$, Complement C$=X_2$, Parasite $X=X_3$.

## Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jtbi.2017.05.037.

## References


Attwater, J., Wochner, A., Holliger, P., 2013. In-ice evolution of RNA polymerase ribozyme activity. Nat. Chem. 5 (12), 1011–1018.

Bartel, D.P., Unrau, P.J., 1999. Constructing an RNA world. Trends Biochem. Sci. 24 (12), M9–M13.

Biebricher, C.K., Luce, R., 1992. *In vitro* recombination and terminal elongation of RNA by Q$\beta$ replicase. EMBO J. 11, 5129–5135.

Biebricher, C.K., Luce, R., 1993. Sequence analysis of RNA species synthesized by Q$\beta$ replicase without template. Biochemistry 32, 4848–4854.

Colizzi, E.S., Hogeweg, P., 2016a. Parasites sustain and enhance RNA-like replicators through spatial self-organisation. PLoS Comput. Biol. 12 (4), e1004902.

Colizzi, E.S., Hogeweg, P., 2016b. High cost enhances cooperation through the interplay between evolution and self-organization. BMC Evol. Biol. 16, 31.

Eigen, M., McCaskill, J., Schuster, P., 1988. Molecular quasi-species. J. Phys. Chem. 92 (24), 6881–6891.

Gilbert, W., 1986. Origin of life: the RNA world. Nature 319 (6055).

Higgs, P.G., Lehman, N., 2015. The RNA World: molecular co-operation at the origin of life. Nat. Rev. Genet. 16, 7–17.

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 (5520), 1319–1325.

Joyce, G.F., 2002. The antiquity of RNA-based evolution. Nature 418 (6894), 214–221.

Kim, Y.E., Higgs, P.G., 2016. Co-operation between polymerases and nucleotide synthetases in the RNA world. PLoS Comput. Biol. 12 (11), e1005161.

Könnyű, B., Czárán, T., Szathmáry, E., 2008. Prebiotic replicase evolution in a surface-bound metabolic system: parasites as a source of adaptive evolution. BMC Evol. Biol. 8, 267.

Könnyű, B., Czárán, T., 2013. Spatial aspects of prebiotic replicator coexistence and community stability in a surface-bound RNA world model. BMC Evol. Biol. 13, 204.

Lawrence, M.S., Bartel, D.P., 2005. New ligase-derived RNA polymerase ribozymes. RNA 11 (8), 1173–1180.

Ma, W., Hu, J., 2012. Computer simulation on the cooperation of functional molecules during the early stages of evolution. PLoS One 7 (4), e35454.

Ma, W., Yu, C., Zhang, W., Hu, J., 2010. A simple template-dependent ligase ribozyme as the RNA replicase emerging first in the RNA world. Astrobiology 10 (4), 437–447.

Ma, W., Yu, C., Zhang, W., Zhou, P., Hu, J., 2010. The emergence of ribozymes synthesizing membrane components in RNA-based protocells. Biosystems 99 (3), 201–209.

Ma, W., Yu, C., Zhang, W., 2007. Monte Carlo simulation of early molecular evolution in the RNA world. Biosystems 90 (1), 28–39.

Ma, W., Yu, C., Zhang, W., Hu, J., 2007. Nucleotide synthetase ribozymes may have emerged first in the RNA world. RNA. 13 (11), 2012–2019.

Matsumura, S., Kun, A., Ryckelynck, M., Coldren, F., Szilagyi, A., Jossinet, F., Rick, C., Nghe, P., Szathmary, E., Griffiths, A.D., 2016. Transient compartmentalization of RNA replicators prevents extinction due to parasites. Science 354, 1293–1296.

Mills, D.R., Peterson, R.L., Spiegelman, S., 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. Proc. Nat. Acad. Sci. 58 (1), 217–224.

Reidys, C., Forst, C.V., Schuster, P., 2001. Replication and mutation on neutral networks. Bull. Math. Biol. 63, 57–94.

Rajamani, S., Ichida, J.K., Antal, T., Treco, D.A., Leu, K, Nowak, M.A., Szostak, J.W., Chen, I.A., 2010. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. J. Am. Chem. Soc. 132, 5880–5885.

Santos, M., Zintzaras, E., Szathmáry, E., 2004. Recombination in primeval genomes: a step forward but still a long leap from maintaining a sizeable genome.. J. Mol. Evol. 59, 507–519.

Shay, J.A., Huynh, C., Higgs, P.G., 2015. The origin and spread of a cooperative replicase in a prebiotic chemical system. J. Theor. Biol. 364, 249–259.

Szabó, P., Scheuring, I., Czárán, T., Szathmáry, E., 2002. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. Nature 420 (6913), 340–343.

Szilagyi, A., Kun, A., Szathmary, E., 2014. Local neutral networks help maintain inaccurately replicating ribozymes. PLoS One 9 (10), e109987.

Takeuchi, N., Poorthuis, P.H., Hogeweg, P., 2005. Phenotypic error threshold; additivity and epistasis in RNA evolution. BMC Evol. Biol. 5, 9.

Takeuchi, N., Hogeweg, P., 2009. Multilevel selection in models of prebiotic evolution II: A direct comparison of compartmentalization and spatial self-organization. PLoS Comp. Biol. 5 (10), e1000542.

Takeuchi, N., Hogeweg, P., 2012. Evolutionary dynamics of RNA-like replicator systems: a bioinformatic approach to the origin of life. Phys. Life Rev. 9, 219–263.

Walker, S.I., Grover, M.A., Hud, N.V., 2012. Universal sequence replication, reversible polymerization and early functional biopolymers: a model for the initiation of prebiotic sequence evolution. PLoS One 7 (4), e34166.

Wilke, C.O., 2001. Selection for fitness *versus* selection for robustness in RNA secondary structure folding. Evolution 55, 2412–2420.

Wochner, A., Attwater, J., Coulson, A., Holliger, P., 2011. Ribozyme-catalyzed transcription of an active ribozyme. Science 332 (6026), 209–212.

Wu, M., Higgs, P.G., 2012. The origin of life is a spatially localized stochastic transition. Biol. Direct 7, 42.

Zaher, H.S., Unrau, P.J., 2007. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. RNA 13 (7), 1017–1026.