SIMULATIONS OF RNA REPLICATOR SYSTEMS

COMPARING PROTOCELL AND SURFACE-BASED MODELS OF RNA REPLICATOR SYSTEMS AND DETERMINING FAVOURABLE CONDITIONS FOR LINKAGE OF FUNCTIONAL STRANDS

By VISMAY SHAH, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the Requirements for the Degree Master of Science

MASTER OF SCIENCE (2019) McMaster University

(Physics and Astronomy) Hamilton, Ontario

TITLE: Comparing Protocell and Surface-Based Models of RNA Replicator Systems and Determining Favourable Conditions for Linkage of Functional Strands

AUTHOR: Vismay Shah, B.Sc. (University of Toronto)

SUPERVISOR: Professor P. Higgs

NUMBER OF PAGES: XII, 101

Lay Abstract

Collections of RNA polymers are good candidates for the origin of life. RNA is able to store genetic information and act as polymerase ribozymes allowing RNA to replicate RNA. Polymerases have been experimentally developed in labs, however none are sufficiently general to work well in an origins of life setting. These polymerases are vulnerable to mistakes during copying, making survival of RNA systems difficult. Such systems have been studied by computer simulations, showing that the strands need to be kept together for survival, either on surfaces or in primitive cells. Differences in the details of the models has made comparing the surfaces to cells difficult. This work creates a unified model base allowing for comparison of these two environments. We find that the existence of primitive cells is very beneficial to systems of RNA polymers and thus it is likely such cells existed at the origin of life.

Abstract

In hypothesized RNA-World scenarios, replication of RNA strands is catalyzed by error-prone polymerase ribozymes. Incorrect replication leads to the creation of non-functional, parasitic strands which can invade systems of replicators and lead to their death. Studies have shown two solutions to this problem: spatial clustering of polymerases in models featuring elements to limit diffusion, and group selection in models featuring protocells. Making a quantitative comparison of the methods using results from the literature has proven difficult due to differences in model design. Here we develop computational models of replication of a system of polymerases, polymerase complements and parasites in both spatial models and protocell models with near identical dynamics to make meaningful comparison viable. We compare the models in terms of the maximum mutation rate survivable by the system (the error threshold) as well as the minimum replication rate constant required. We find that protocell models are capable of sustaining much higher maximum mutation rates, and survive under much lower minimum replication rates than equivalent surface models. We then consider cases where parasites are favoured in replication, and show that the advantage of protocell models is increased. Given that a system of RNA strands undergoing catalytic replication by a polymerase is fairly survivable in protocell models, we attempt to determine whether isolated strands can develop into genomes. We extend our protocell model to include additional functional strands varying in length (and thus replication rate) and allow for the linkage of strands to form proto-chromosomes. We determine that linkage is possible over a broad range of lengths, and is stable when considering the joining of short functional strands to the polymerase (and the

same for the complementary sequences). Moreover, linkage of short functional strands to the polymerase assures more cells remain viable post division by ensuing a good quantity of polymerase equivalents are present in the parent cell prior to splitting.

Acknowledgements

First and foremost, I would like to express my gratitude to my supervisor, Dr. Paul Higgs, for the guidance and assistance I received throughout the course of my degree. His office door was always open whenever I had questions or wanted to discuss new results or ideas

A special thanks to Andrew and Armin for the many insightful conversations we had leading to new directions of research for this thesis, and also for the many times you helped troubleshoot elusive bugs in my code.

A special thanks to the students and faculty of the Origins Institute and the Astrobiology program at McMaster University. It was a pleasure getting to know all of you and having the opportunity to work alongside you. The passion you show and the effort you put in to understand the various aspects of origins of life research is incredible.

Finally, I give a special thanks to my mom and dad, without your support this work would not be possible.

Table of Contents

Chapter 1: Introduction 1
1.1 Motivation 1
1.2 Structure of thesis 3
1.4 Ribozymes Catalyzing RNA Replication 7
1.5 Methods of Higher-level Selection 9
1.6 Error Threshold 10
1.7 Computational Models 14
1.8 Aims of Thesis 21
Chapter 2: Survival of RNA replicators in protocells and surface-based systems 22
2.1 Summary 22
2.2 Shah et al., 2019 24
Chapter 3: Additional content for Shah et al. (2019) 66
3.1 Summary of Additional Work 66
3.2 Additional Results 66
Chapter 4: Linkage of strands into a genome 72
4.1 Prior Work 77
4.2 Methods 80
4.3 Results 82
4.4 Discussion 91
Chapter 5: Conclusions 93
References: 95

List of Figures and Tables

Figures

Figure 1.1 Schematic overview of various prebiotic replicator models. Adapted and simplified from Ecology and evolution in the RNA World Dynamics and stability of prebiotic replicator systems by Szilágyi et al., 2017. 15

Figure 1.2 A 5-membered hypercycle model (A-E) featuring a parasite (X). Each member of the network catalyzes the replication of the next. B also catalyzes the formation of a parasitic species, X.....17

Figure 1.3 A 5-membered hypercycle model (A-E), featuring a shortcut where species A can catalyze the formation of species D. Over time this network will evolve to omit species B and C entirely, causing it to become a 3-membered hypercycle.....17

Figure 1.4 A schematic of a metabolically coupled replicator system with four species of replicators (A-D) and a parasitic species (X). Each member draws from the same population of metabolites (M) to catalyze their own replication. All members must be present for the network to work optimally.....20

Figure 2.1 A cartoon representation of the spatial model with local diffusion dynamics (left) and the protocell models (right). The red strands are polymerases (P), orange strands are complements to polymerases (C), and black strands are parasites (X). The blue arrows indicate the possibility of diffusion to and from the eight neighboring sites.....36

Figure 2.2 Average numbers of strands per cell in the PCP model. (**a**) $S_0 = 10$, (**b**) $S_0 = 20$. $k = 25$ in both cases. Points are from finite population simulations. Smooth lines are from deterministic theory. 43

Figure 2.3 Average numbers of strands per site in the SMF model. (**a**) $N = 100$, $k = 25$, $h = 0.4$ and (**b**) $N = 400$, $k = 20$, $h = 0.4$. Points are from finite population simulations. Smooth lines are from deterministic theory. 45

Figure 2.4 Comparison of the error threshold of the various models studied as a function of the polymerization rate k. $S_0 = 10$ in all models except the one per site model, and $h = 0.4$ in the lattice models. All results are from stochastic simulations except for SMF, which results are from the deterministic method. OSPS is the one strand per site model from [18]. Other models are defined in Table 2.1.....46

Figure 2.5 Comparison of the error threshold of the various models studied as a function of S_0. $k = 25$ in all models, and $h = 0.4$ in the lattice models. Results for SMF are obtained

from the deterministic method, except for the points with $S_0 > 150$, where the deterministic method becomes much slower than the stochastic simulation. Results for the other models are obtained from stochastic simulations. OSPS is the one strand per site model from [18]. Other models are defined in Table 2.1. 47

Figure 2.6 Comparison of the error thresholds of the spatial models with long distance diffusion and local diffusion as a function of the diffusion rate h. Made using $S_0 = 10$, k = 25. 52

Figure 2.7 Error thresholds versus k for PCP and SLD models in which parasites and polymerases have equal replication rates (same as Figure 2.4) compared with equivalent models where parasites have double the replication rate of polymerases (denoted PCP2X and SLD2X). Both error thresholds are reduced when the parasites multiply faster, but the SLD model is reduced more, meaning that the relative advantage of the protocells over the spatial model is increased. $S_0 = 10$ and h = 0.4. 55

Figure 2.8 Error thresholds versus S_0 for PCP and SLD models in which parasites and polymerases have equal replication rates (same as Figure 2.5) compared with equivalent models where parasites have double the replication rate of polymerases (denoted PCP2X and SLD2X). Both error thresholds are reduced when the parasites multiply faster, but the SLD model is reduced more, meaning that the relative advantage of the protocells over the spatial model is increased. Note that $M_{max} = 0$ for $S_0 > 20$ for SLD2X, because faster parasites kill the polymerases in the spatial model. k = 25 and h = 0.4. 56

Figure 3.1 Strand distributions for (**a**) protocell model with constant volume at a split size of 10; (**b**) protocell model with constant volume at a split size of 100. 68

Figure 3.2 Strand distributions for (**a**) spatial model – local diffusion with a maximum size of 10; (**b**) spatial model – local diffusion with a maximum size of 100 69

Figure 3.3 Snapshots from a SLD simulation performed with M=0.01, $S_0 = 15$, k = 25 and h=0.4. At such low mutation rates, effects of clustering are difficult to determine. 73

Figure 3.4 Snapshots from a SLD simulation performed with M=0.06, $S_0 = 15$, k = 25 and h=0.4. **b)** and **c)** clearly show the effects of clustering allowing groups of polymerases to survive. 75

Figure 3.5 Snapshots from a SMF simulation performed with M=0.06, $S_0 = 15$, k = 25 and h=0.4.**b)** and **c)** show a lack of the large, empty regions seen in figures 3.4b and 3.4c 77

Figure 4.1 An overview of the possible equilibrium states in the simulation. Green arrows represent successful invasions and black arrows failures to invade. 83

Figure 4.2 Phase plot of the invasion of a P state by Y+ and Y−. Red circles represent the cases where invasion was successful, blue squares where invasion was unsuccessful, purple diamonds where invasion was stochastic and black X's where invasion caused the death of the system. The purple dotted line represents a stochastic boundary between the two states. 84

Figure 4.3 Phase plot of the invasion of a P state by L_+ and L_-. Red circles represent the cases where invasion was successful, blue squares where invasion was unsuccessful, purple diamonds where invasion was stochastic. The purple dotted line represents a stochastic boundary between the two states. 85

Figure 4.4 Phase plot of the invasion of an intermediate system by linked or unlinked strands. Red triangles represent the cases where linked strands alone form the equilibrium state, blue triangles the cases where the equilibrium state is a mixture of P, C, Y_+ and Y_- and black circles where the equilibrium state is comprised of P and C alone. The purple dotted line represents a stochastic boundary between the states.....86

Figure 4.5 Phase plot combining information from Figures 4.2-4.4. (1) L state invades P state. PY state is not possible. (2) L state invades P and PY states. PY state invades P state. (3) PY state invades L and P states. L state invades P state. (4) PY state invades P state. L state is not possible. (5) Only P state is possible. 87

Figure 4.6 The average cell composition at equilibrium for simulations where a P, C state was invaded by unlinked strands for various l_y. Error bars are standard errors from a time average of the equilibrium state.....89

Figure 4.7 The average cell composition at equilibrium for simulations where a P, C state was invaded by linked strands for various l_y. Error bars are standard errors from a time average of the equilibrium state.....90

Figure 4.8 A comparison of the number of cell divisions per 100 δt for the cases studied in Figures 4.6 and 4.7.....90

Figure 4.9 A comparison of the number of viable cells in the population at equilibrium divisions for the cases studied in Figures 4.6 and 4.7.....91

Tables

Table 2.1 Overview of models in this study. 36

Table 3. 1 Mutation in Different Spatial Models 70

List of all Abbreviations and Symbols

RNA	Ribonucleic acid
DNA	Deoxyribonucleic acid
PCP	Protocell model with constant population
PCPCV	Protocell model with constant population and constant volume
PML	Monomer limited version of the Protocell model
SLD	Spatial model with local diffusion dynamics
SMF	Mean field approximation version of the spatial model
SML	Monomer limited version of the spatial model
P	Polymerase
C	Polymerase complement
X	Parasite
Y_+	Nucleotide synthase
Y_-	Nucleotide synthase complement
L_+	Linked functional strands
L_-	Linked complementary strands

Declaration of Academic Achievement

Chapter 2 includes a paper published in *Life* which is based on preliminary work from from Q. Pauli and features contributions from J. de Bouter, A. Tupper and P. Higgs. J. de Bouter developed the deterministic version of the spatial model and Q. Pauli and P. Higgs developed the deterministic version of the protocell model. All other research presented in this thesis was conducted by V. Shah between September 2017 and August 2019 at McMaster University.

Chapter 1: Introduction

1.1 Motivation

The question of how life arose on Earth and developed into the intricate living systems we see today is a complicated one. Modern life features elaborate reaction networks involving numerous biomolecules interacting with each other and the environment they are contained in. Biopolymers of deoxyribonucleic acid (DNA) make up the genome of all organisms, encoding the various proteins that each lifeform can synthesize (Phillips et al., 2012). Polymers of ribonucleic acid (RNA) have a variety of functions: messenger RNA is used in translating the genetic code into a form which directs the synthesis of proteins, transfer RNA carries amino acids to the site of protein synthesis and ribosomal RNA catalyzes the formation of peptide bonds, linking amino acids together to form proteins (Phillips et al., 2012). Proteins catalyze the vast array of chemical reactions living organisms use to survive, grow and replicate.

It is doubtful that such complexity could be present from life's inception, rather it must have developed over time starting from a collection of simpler prebiotic entities. The most popular theory for the origin of life is the RNA world theory, which was first posited by Alexander Rich and further developed by Gilbert in 1986 (Rich, 1962; Crick, 1968; Orgel, 1968; Gilbert, 1986; Higgs and Lehman, 2015). The theory states that the current system used by all modern cells, in which information flows from DNA to RNA to proteins, originated from a self-sustaining system composed entirely of RNA molecules with the ability to store genetic information and act as replicases, with replication dictated by

Watson-Crick base-pairing (Rich, 1962; Gilbert, 1986; Pace and Marsh, 1985; Joyce, 2002; Robertson and Joyce, 2012).

Akin to DNA, RNA has the ability to store information in polymers of itself (Joyce, 1989). While all known life uses DNA to store genetic information due to it being more stable than RNA, genomes made of RNA do exist, in fact most viruses have RNA genomes (Ahlquist, 2002). Moreover RNA is functional: self complementary parts of the single stranded polymer can undergo conformational changes in which parts fold together to form helices, stem loops, knots and other complicated structures (Tinoco and Bustamante, 1999). These tertiary structures can directly catalyze chemical reactions or associate with other RNA polymers and metal cofactors to form large scale complexes of RNA which are catalytic. These RNA enzymes are commonly referred to as ribozymes. Prebiotically, through this folding mechanism RNA is able to perform the role played by proteins in modern life.

The first naturally occurring ribozyme was a self-splicing intron discovered in 1982 by Cech (Kruger et al, 1982). The intron was found to catalyze the breaking of phosphodiester bonds between it and the exons, allowing the ribozyme to cut itself out of the overall RNA strand (Kruger et al, 1982). Since then, many other ribozymes have been discovered which can catalyse the splicing, ligation, recombination and even replication of RNA (Hayden et al., 2008; Lincoln and Joyce, 2009; Attwater et al., 2013). A subset of these ribozymes are capable of catalyzing the synthesis of other RNA strands from provided templates. If such polymerase ribozymes were autocatalytic – that is if they could catalyze their own replication and replication of their complementary templates – then systems of polymerases

would be good candidates for the first living systems. While we have yet to discover evidence that such general replicators existed in an RNA world, work has been done by many groups in developing polymerase ribozymes in the lab (Johnston et al, 2001; Lawrence and Bartel, 2005; Zaher et al, 2007; Cheng et al, 2010, Wochner et al., 2011; Attwater et al., 2013; David and Joyce, 2016; Attwater et al., 2018).

Replicators in these first living systems would be highly error prone as they lack the error proofing machinery employed by cells today. Under erroneous replication, the maximum length of RNA molecules is inversely proportional to the mutation rate, leading to lengths too small to comprise a polymerase, let alone a genome; the theory behind this will be covered in the following sections. In other words, for a non-zero mutation rate, non-functional products resulting from erroneous replication will overwhelm and kill the system of replicators. Much work has been done to address this issue and has led to the development of many models featuring interactions between different molecules and different higher-level selection mechanisms, primarily spatial clustering of RNA strands or compartmentalization of strands into protocellular structures. The absence of any organism relying on spatial clustering and the ubiquity of membranes in modern biology suggests that the first cell-like structures were a significant development for life. In this thesis we compare these two mechanisms to see their impact on a system of error-prone polymerases in an RNA world scenario.

1.2 Structure of thesis

The remainder of this chapter serves to provide background information for various topics developed in chapters 2, 3 and 4. Chapter 2 presents a paper submitted to *Life*

currently awaiting publication, focusing on comparing spatial models and protocellular models of a system of replicators in the RNA world. Chapter 3 provides additional material created during the research process on this topic that was deemed not pertinent for publication. Chapter 4 presents an investigation into the formation of a proto genome formed by the linkage of strands. Chapter 5 discusses the major conclusions of this thesis and suggests avenues for future investigation.

1.3 Prebiotic Replication

This section adopts the notation of Wu and Higgs (2012), Shay et al. (2015), Kim and Higgs (2016), Higgs (2017) and Higgs (2018) and summarizes work presented in Higgs (2017) and Higgs (2018).

There are three primary ways in which synthesis of RNA strands may occur in prebiotic systems. The first is through random chemical synthesis, which is referred to as *s*. This *s* reaction continuously produces random oligomers from monomers present in the environment, extends existing oligomers and can join existing oligomers together to form longer ones (Higgs, 2018). Mechanisms by which this can occur include wet-dry cycling (Damer and Deamer, 2015) or clays assisting phosphodiester bond formation (Himbert et al., 2016). The oligomers formed by this process are inherently random or intrinsically biased in composition depending on the nature of the monomer source. This synthesis method is required to produce the first oligomers as the latter two methods require a pre-existing template sequence to be present. Additionally, this method is capable of producing a diverse set of oligomers. It is thought that rarely this synthesis would create a functional sequence capable of acting as a ribozyme. Through the *s* reaction, there is no aspect of

selection for such functional or beneficial sequences; every sequence produced will be random so the likelihood of producing the same sequence repeatedly is infinitesimal. Even if a functional sequence were to arise, it could not create more of itself unless it was autocatalytic (thus not dependent on the *s* reaction). With no way to increase the concentrations of beneficial sequences, it is hard to imagine the *s* reaction sustaining a living system. It follows that once the first oligomers were formed, this would rapidly cease to be the dominant replication method, replaced by one of the two methods discussed below.

The second method to synthesize RNA strands is via non-enzymatic template directed replication of RNA, which is termed the *r* reaction (Higgs, 2018). An existing oligomer or RNA strand is used as a template onto which monomers in the environment can bind following the Watson-Crick base pairing rules, or the wobble base pairing rules (Varani and McClain, 2000). If these monomers are ligated together, a complementary sequence to the template is synthesized. Upon separation, the two strands can again be used as templates. Prebiotically, this may occur in a variety of ways as outlined by Szostak (2012). By its mechanism, the *r* reaction is able to increase the concentration of the template (and its complementary sequence) in a system; models by Chen and Nowak (2012) show that this process selects sequences with the highest fitness. Diversity can be generated from mutations – erroneous base pairing. Provided that monomers are plentiful, the higher the concentration of templates, the more often template directed replication will occur. Whether or not the *r* reaction can sustain a system depends on how fast it acts. All oligomers in solution are vulnerable to breakdown from hydrolysis. If the rate of the *r* reaction is fast

compared to the hydrolysis rate, then on average a strand will be able to create a copy of itself before it breaks down, allowing a system reliant on template directed replication to survive. Indeed, the work of Higgs (2018) shows using a simple spatial model that a rate $r \gtrsim 1.4$ times the breakdown rate is enough for survival.

The last method is by means of a catalyst, termed the *k* reaction (Higgs, 2018). In this method a ribozyme is used to catalyze the synthesis of a strand complementary to the template. The exact mechanism in which the catalyst acts is dependent on the catalyst – it can span from simple unitary primer extension, extension in groups of nucleotides, assisting in the formation of a phosphodiester bond, etc. (Higgs, 2018). In essence this reaction works in much the same way as the template directed one, however there are now two concentrations of importance, that of the catalyst and that of the template. Similar to the template directed case, a system can be sustained by the *k* reaction if it is fast enough.

In principle a living system can be sustained by the both the *r* reaction and the *k* reaction, or even a mixture of the two. Should the *r* reaction be faster than the *k* reaction, there is no need for catalyzed replication and vice versa. Based on the fact that all modern life undergoes catalyzed replication, it is fair to hypothesize at some point during the RNA world a transition was made to catalyzed replication. There is another important distinction: if a beneficial sequence arises with an intrinsic $r < r_{min}$ and replication occurs solely by the *r* reaction, then the sequence will become extinct, whereas if replication occurs by the *k* reaction, or both the *r* and *k*, then sequences with $r < r_{min}$ will still be maintained as it is the activity of the catalyst that matters (Higgs, 2018). This suggests that systems sustained by catalytic replication allow for a wider variety of strands. In their work, the groups of Ma et

al. and Wu and Higgs considered the case featuring all three reaction types (Ma et al., 2007; Ma et al., 2010; Wu and Higgs, 2012). In this thesis we consider models which involve only the *k* reaction under the premise that the *s* reaction is irrelevant on the length scales of polymers that we are interested in, and that the *r* reaction occurs at a much slower rate than the *k* reaction.

1.4 Ribozymes Catalyzing RNA Replication

Replication that is catalytic in nature is still a very broad topic. Is this catalytic replication mediated by a single ribozyme or is it the result of several ribozymes working as part of a network? Both methods are prebiotically viable. The following are examples of ligases and recombinases, different to the polymerases considered in this thesis. In 2002, Paul and Joyce developed the R3C ligase ribozyme which was capable of binding two substrates and catalyzing their ligation to generate a copy of itself. Further development by Lincoln and Joyce (2009) converted this ribozyme to a cross-catalytic form in which the enzyme catalyzes the ligation of two substrates creating a minus strand of the enzyme, which in turn can catalyze ligation of two different substrates forming the original ribozyme. A more complex network which is autocatalytic is that of fragments from the Azoarcus group I intron (Hayden and Lehman, 2006; Draper et al., 2007; Hayden et al., 2008). This network has 4 fragments, W, X, Y and Z which assemble into a complex that catalyzes the recombination of the fragments into a WXYZ ribozyme (Hayden and Lehman, 2006; Draper et al., 2007; Hayden et al., 2008). The WXYZ ribozyme in turn catalyzes the ligation of the fragments into various intermediaries that eventually lead to its production (Hayden and Lehman, 2006; Draper et al., 2007; Hayden et al., 2008).

While such autocatalytic systems do count as catalytic replication, they have the issue that they are sequence specific. Such networks can only catalyze the formation of themselves or other members of the network; should a novel beneficial strand come about in these systems it will not be replicated. Similarly, the substrates used in these networks tend to be relatively long oligomers that would have to be synthesized in some manner. The work of Lincoln and Joyce (2009) does show that these sets are capable of evolution, however it is limited to evolving variations of the existing members of the set, often focusing on refining existing domains of the strands to increase catalytic ability. As of yet no study has shown the ability of such networks to incorporate a significantly different strand.

An ideal ribozyme would be one able to catalyze the replication of any sequence provided to it as a template, a general polymerase ribozyme. Naturally, a general polymerase ribozyme is autocatalytic and non-sequence specific. The presence of such a ribozyme would allow a living system to sustain itself, incorporate any beneficial strands and develop complexity (Higgs, 2018). While a general polymerase has yet to be discovered, polymerase ribozymes have been developed in laboratory settings. In 2001, Johnston et al. developed a polymerase 189 bases long that could extend a primer by 14 nucleotides provided with a template. Lawrence and Bartel (2005) further derived 8 other polymerase ribozymes from that of Johnston et al. which could perform similar primer extension on any template provided. Zaher and Unrau (2007) improved on this with polymerase ribozyme B6.61 which extended a primer by at least 20 nucleotides, as did Wochner et al. (2011) who developed the tC19 and the tC19Z polymerase ribozymes which

were more general and could extend primers by up to 95 bases on favourable templates. In 2013, Attwater et al. developed the first polymerase (tC9Y) capable of synthesizing an RNA sequence longer than its own length of 202 nucleotides and in 2018, Attwater et al. discovered the $t5^{+1}$ ribozyme which uses RNA triplets as substrates. Horning and Joyce (2016) discovered the 24-3 polymerase which achieved primer extension at rates about 100 times faster than the tC19Z, and was capable of forming a full length tRNA, a first, from a 15-nucleotide template, although in low yields. These results are promising, supporting the notion that a general RNA polymerase could have been present during the RNA world. In the studies presented in chapters 2-4, we use a processive general polymerase as the replicator in our models. We propose that it can synthesize a complement to any template provided to it at a rate inversely proportional to the length of the template.

1.5 Methods of Higher-level Selection

Prebiotically, spatial clustering of RNA strands is possible due to the presence of physical environments capable of limiting the effects of diffusion such as in porous, connected mineral lattices (Branciamore et al., 2009) or RNA molecules adsorbed on clay (Franchi and Gallori, 2005). In these environments, RNA strands face limited movement and the localization provided can help with increasing encounters between strands, raising the chance for reactions to occur. Previous spatial models discussed in section 1.7 study the means by which spatial clustering helps the survival of the RNAs, often by simulating large lattices capable of housing one sequence at each site. Our models presented in Chapters 2 and 3 extend this by considering cases where multiple RNA strands can be present on a single site. In this manner new dynamics arise as the presence of a single parasite on a

lattice site is no longer enough to deem the site dead Moreover, replication cannot be halted by the neighbouring sites being occupied as the products of replication are placed in the same site as the parents. Site composition plays a role in determining the fate of the system, larger site volumes can even lead to the coexistence of parasites with polymerases.

In the modern world, all organisms make extensive use of membranes. Cellular membranes formed from lipids and sterols isolate cells from their environments, provide structural support and selective intake capabilities. Viruses can contain lipid envelopes and the even the most basic of them have a protein capsid enclosing their genome. While complicated membranes would not have existed prebiotically, lipids or lipid-like molecules (long chain hydrocarbons) could be synthesized (Segré et al., 2001). Some examples are fatty acids and fatty alcohols that could arise from Fischer-Troph synthesis (Segré et al., 2001). Analysis of the Murchison meteorite revealed amphiphilic compounds that have the propensity to self-aggregate could form membranes, films, and even vesicular structures (Deamer, 1985; Deamer and Pashley, 1988). Damer and Deamer (2015) propose that the presence of such molecules in a warm little pond environment could lead to RNA strands squished in layers of a multilamellar structure of the lipids when the pond is dry, and inside vesicles formed by the lipids when the pond is hydrated. Protocellular models outlined in section 1.7 study the case in which RNA systems grow and develop in such vesicles, drawing much similarity to modern cells.

1.6 Error Threshold

The idea of an error threshold comes from the work of Manfred Eigen (1971). The error threshold in Eigen's study is the maximum amount of information a genome can store at a

given replication fidelity. We outline the essence of Eigen's model following Szilágyi et al. (2017). Eigen considered a virus-like first order system in free space: a population made of a wild type master sequence (x_W) competing against its own mutants (x_M) both of which are being replicated externally to the model (Eigen, 1971). Replication of wild type sequences is error-prone: for a sequence L nucleotides long with each base having a point mutation rate μ, the fidelity of replication is given by $Q = (1-\mu)^L \cong e^{-L\mu}$. Thus, replicating a wild type sequence produces another wild type sequence with a fidelity Q or a mutant otherwise. The mutant sequences may also replicate, in which case we do not care about the fidelity of replication (assuming that back mutations are negligible due to their rarity of occurrence). Assigning the replication rates A_W for the wild type sequence and A_M for the mutant sequence, and employing an outflow Φ to keep the total concentrations fixed, the equations:

$$\frac{dx_W}{dt} = x_W(QA_W - \Phi)$$

$$\frac{dx_M}{dt} = x_M(A_M - \Phi) + (1-Q)A_W x_W$$

describe the system (Szilágyi et al., 2017). In this case, if the wild type was less fit than its mutants ($A_W < A_M$) then it would go extinct, while coexistence of both the wild type and the mutants requires that $QA_W > A_M$. This gives $\frac{A_W}{A_M} > \frac{1}{Q}$, and solving the fidelity relation for sequence length, $L < \frac{-ln\, Q}{\mu}$, gives the error threshold inequality:

$$L < \frac{ln\left(\frac{A_W}{A_M}\right)}{\mu}$$

which shows the maximum amount of information that the system can maintain is limited by the superiority of the master sequence and the mutation rate. Eigen's paradox comes about when solving for the length with reasonable values: $ln\left(\frac{A_W}{A_M}\right) \sim 1$, and a nominal mutation rate of 1% gives that a sequence of length up to 100 nucleotides can be maintained. Past this value, the wild type sequence is lost and the equilibrium population is an amalgamation of the various mutants. This is far shorter than a genome, and shorter than the polymerase ribozymes that have been developed thus far (see section 1.4). Eigen's paradox states that without a large genome, there could be no accurate replication as the error proofing machinery cannot be encoded in a small genome, and without accurate replication large genomes cannot be maintained. This issue can be resolved to some degree by requiring cooperation and interdependence between molecules to avoid out-competition of the ensemble by a single species, as well as through higher level selection mechanisms acting upon the population.

Suggestions have been made that the limit on information by the error threshold is too strict. Studies studying RNA secondary structure folding revealed that there is a large amount of redundancy in going from sequence to structure in that many different sequences may fold to the same secondary structure, and that a subset of structures are found far more often than others (Schuster et al., 1994). The realization of such neutral networks in which sequences differing by a few nucleotides formed the same structure give rise to the possibility that the impact of mutations on a sequence may not be as severe as is thought (Huynen et al., 1996). Instead, a distinction must be made between the classical genotypic error threshold where sequence information cannot be maintained, and a new, more

important phenotypic error threshold where the folded structure cannot be maintained (Huynen et al., 1996). Takeuchi et al., 2005 show that even considering base substitutions not impacting each other, the actual increase in error threshold is limited as the average number of base substitutions rendering the phenotype unchanged is small. Work by Kun et al. (2005) builds on their results and demonstrates the tolerance provided by the phenotypic error threshold. They experimentally determined the number of neutral single nucleotide substitutions leaving two ribozymes unchanged, then assumed a replication accuracy of 0.999 as in the worst viral replicators and found the phenotypic error threshold to be about 7000 nucleotides, sufficient for a riboorganism to exist (Kun et al., 2005). The relaxation of the genotypic error threshold applies more for common sequences compared to rare ones and more work must be done to determine whether considering only the phenotypic error threshold is viable.

In the study presented in chapters 2 and 3 of this thesis, we use the idea of a mutational error threshold as a metric to test various spatial and protocellular models. This error threshold differs from that of Eigen. Instead of looking for the amount of information the system can maintain, we look for the maximum mutation rate that the system can tolerate. Moreover, the systems we test feature second-order catalysis, where the replicator itself is part of the system and reactions require the presence of a replicator and another molecule in order to get underway. Since strands are replicated externally in Eigen's model, going beyond the error threshold leads to a population of only mutants, whereas in our study, exceeding the maximum supported mutation rate (going beyond the mutational error

threshold) results in the loss of the replicators in the system and subsequently the rest of the strands.

1.7 Computational Models

Computational models have been used extensively to study the survival of prebiotic replicator systems across various parameters. The exact methodology adopted and the metrics tested vary with the study. Generally speaking, these models are dynamical, may consider spatial structure and feature local and global dynamics, and either model time discretely or continuously (Szilágyi et al., 2017). Here we present an overview of a few important model types, with figure 1.1 indicating the chronology of development.

The basis for all of these computational models stem from the study of exponential growth. In the field of population dynamics, a population in an environment with surplus resources and not otherwise limited by external factors will grow exponentially (Szilágyi et al., 2017). Mathematically, this idea is expressed in the simple differential equation $\frac{dN}{dt} = rN(t)$, which has the solution $N(t) = N(0)e^{rt}$, where N(t) is the size of the population and r the growth rate, which takes on some positive value.

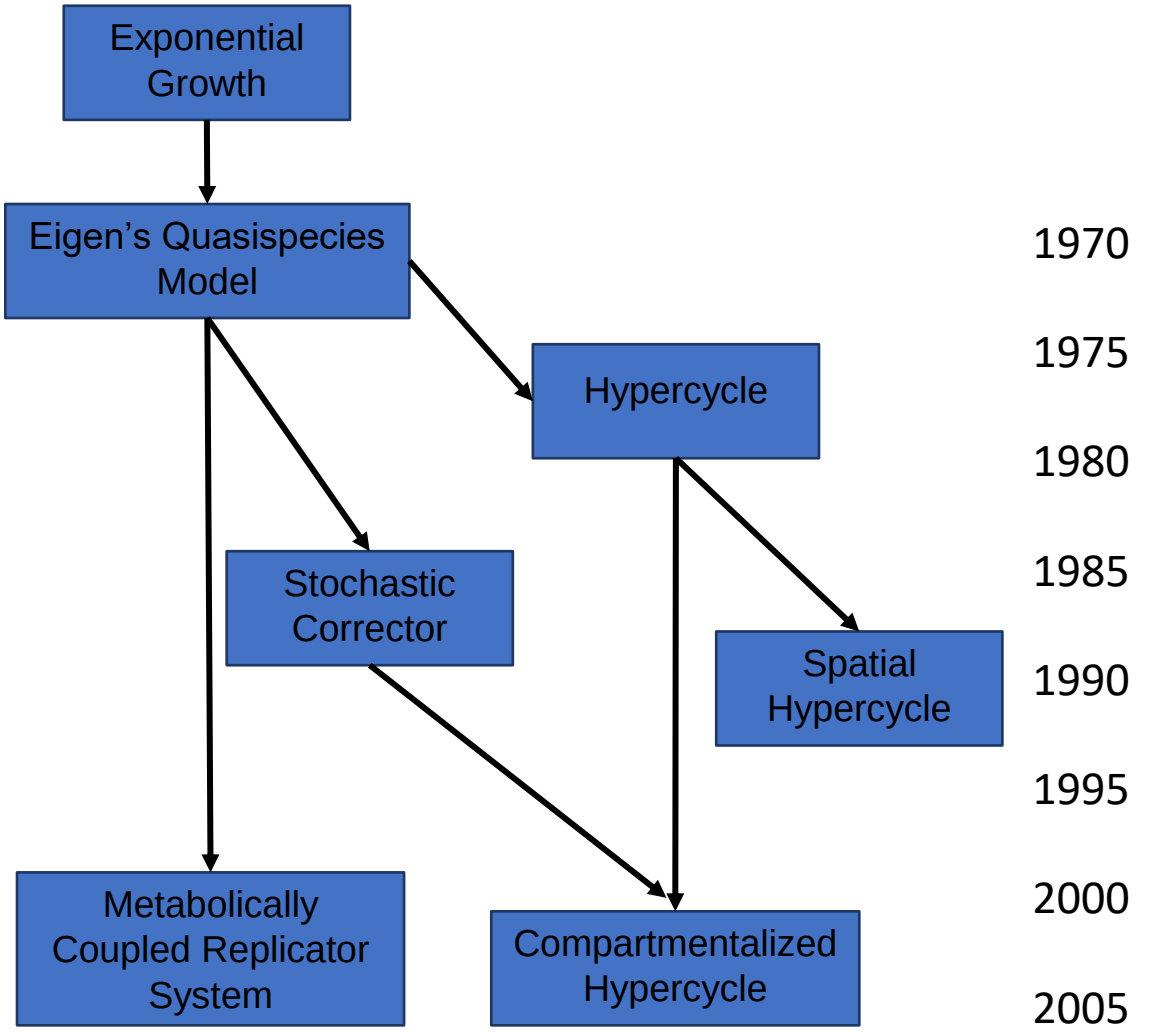


Figure 1.1 Schematic overview of various prebiotic replicator models. Adapted and simplified from Ecology and evolution in the RNA World Dynamics and stability of prebiotic replicator systems by Szilágyi et al., 2017.

Studying multiple competing populations, N_1 and N_2 with characteristic growth rates r_1 and r_2 reveals that the population with the higher growth rate will exponentially outcompete the other one. This is a problem of importance: if only the population with the highest growth rate can survive, how do we develop complexity in an RNA world scenario?

The solution to this is to look at regulating factors which affect the growth rates of one or more species. Such regulating factors include resources consumed for replication, spatial requirements for replication products, mixing, enforcing local interactions or limiting interactions in general.

Eigen's quasispecies model is described in detail in section 1.6. It is a free space model, meaning it does not have any defined spatial structure. Eigen's analysis of this model revealed that so long as the master sequence was more fit than its mutants, it could be maintained until an error threshold was reached. The inability to maintain long sequences can be circumvented if the information is contained in several shorter sequences. Eigen was one of the proponents of this, suggesting a set of short catalytic sequences forming a closed network such that each sequence in the network catalyzed the formation of the next, a hypercycle (see figures 1.2, 1.3) (Eigen, 1971). Despite the sequences competing for the same resources they can coexist via the forced cooperation as they require the presence of the previous member in the cycle to replicate. In this manner a set of sequences can store more information than a single, long sequence, and has the ability to undergo more rapid evolution. In theory there is no limit to the number of members in a hypercycle, however stable solutions only exist for up to 4-membered cycles in which all members are fully cooperative (Hofbauer et al., 1984; Schuster et al., 1979).

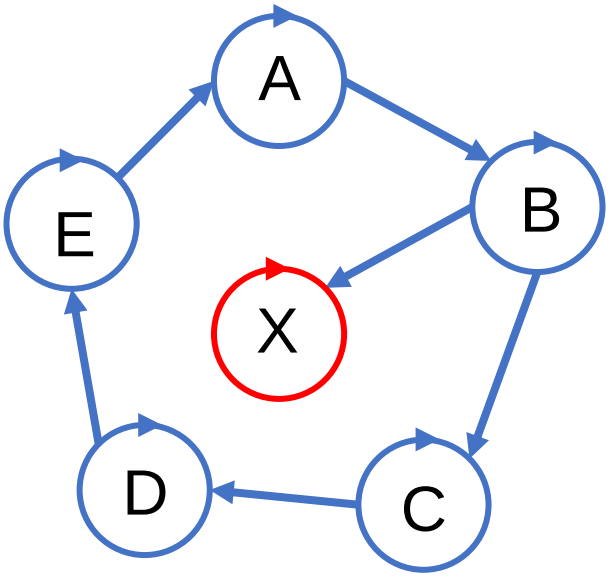


Figure 1.2 A 5-membered hypercycle model (A-E) featuring a parasite (X). Each member of the network catalyzes the replication of the next. B also catalyzes the formation of a parasitic species, X.

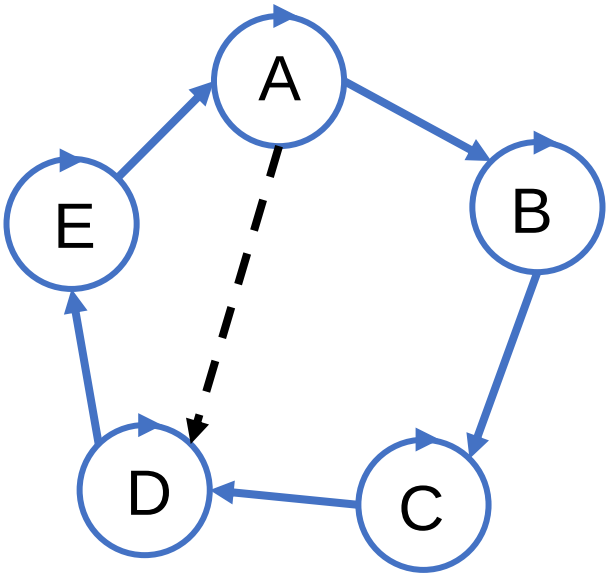


Figure 1.3 A 5-membered hypercycle model (A-E), featuring a shortcut where species A can catalyze the formation of species D. Over time this network will evolve to omit species B and C entirely, causing it to become a 3-membered hypercycle.

Further issues arise in cases where the members of the hypercycle are in competition, or the rates of formation and catalysis vary across the network, in which case only 3-membered or lower systems remain stable while higher membered cycles can exhibit oscillatory solutions (Hofbauer, 1984; Schuster et al., 1979). Other difficulties faced by hypercycles include the inability to incorporate new strands leading to limited evolvability, and there is also the issue of information loss should one of the members gain the ability to catalyze a member not directly following it, disrupting and shortening the network (see figure 1.3).

All of the work done on the free space models presented assume the superiority of the master sequence compared to its mutants. In reality this may not be the case, as was shown by Spiegleman's experiment involving replication of the Qβ genome *in vitro* (Mills et al., 1967). Here the wild type (the full genome) rapidly evolved to cut all unnecessary RNA away and form a mutant 17% the size of the wild type, capable of being replicated 15 times faster (Mills et al., 1967). When considering such prolific mutants, the free space models and hypercycle models fail to survive as the quickly replicating mutants outcompete the members of the network, starving them of the required resources.

Attempts were made by Boerlijst and Hogeweg (1991) to implement the hypercycle idea on a toroidally wrapped square lattice in which each site could be empty, occupied by a member of the hypercycle or occupied by a parasite. The spatial definition made it possible for hypercycle members to cluster together and be isolated away from parasites, increasing their survival. From their simulations, this clustering effect produced spiral wave patterns emerged which stabilized the system to a degree from parasitic invasion. Parasites

were still harmful when they emerged into the centre of the spiral formations. Zintzaras et al. (2002) instead simulated hypercycles in isolated compartments and found that the hypercycle system could sustain higher mutation rates, as in the spatial case. In either method, the inherent evolvability of the hypercycle remains limited and thus is a reason the hypercycle, spatial hypercycle and compartmentalized hypercycle models should be abandoned (Szilágyi et al., 2017).

A different class of models was proposed by Szathmáry and Demeter (1987) called the stochastic corrector model. This model aims to emulate the behaviour of prebiotic cellular compartments: it features different replicator types contained within compartments which grow as the number of replicators increases. There is competition amongst replicators as they use the same resources for replication, but this is offset by them contributing to a common metabolism affecting the entire compartment. At sufficiently high internal concentrations, the compartments split, randomly distributing the contained strands among the two daughters. The stochastic corrector model is able to tolerate higher mutation rates and thus is more successful than the hypercycle models mentioned before, while also retaining evolvability stemming from the variety produced by random assortment and the stochasticity of replication (Szilágyi et al., 2017). *In vitro* experiments show that even just transient compartmentalization in vesicles is sufficient to sustain an array of functional replicators (Matsumura et al., 2016).

The metabolic coupled replicator system model (MCRS) depicted in figure 1.4 is similar to the hypercycle models in that it suggests the coexistence of an array of replicator species in a network.

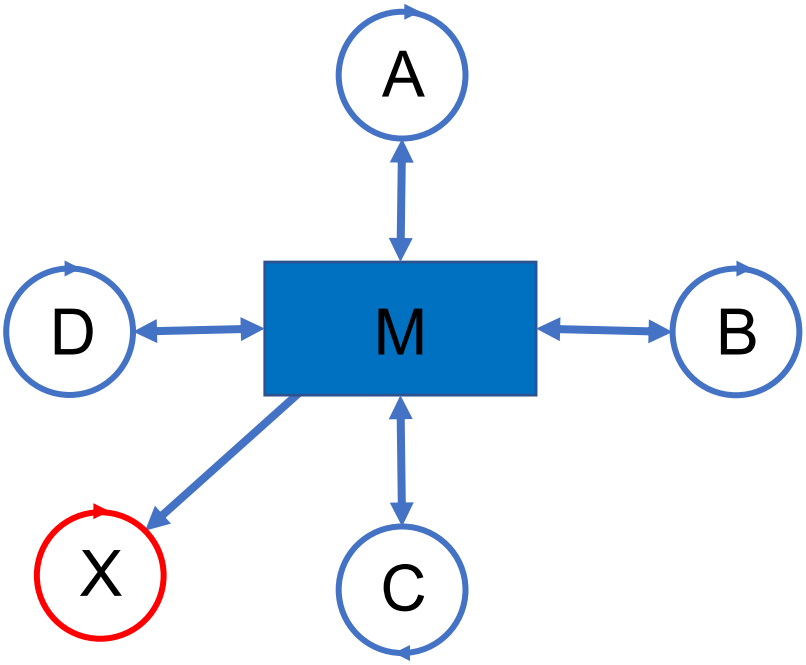


Figure 1.4 A schematic of a metabolically coupled replicator system with four species of replicators (A-D) and a parasitic species (X). Each member draws from the same population of metabolites (M) to catalyze their own replication. All members must be present for the network to work optimally.

The difference in the MCRS is that the various species do not catalyze the replication of the next member in the cycle – rather there is no cycle. All species draw from the same pool of metabolites and are able to template replicate themselves (Könnyű and Czárán, 2015). Additionally, the members are mutually dependent in the sense that all are required to replenish the pool of metabolites they draw from (Könnyű and Czárán, 2015). MCRS systems cannot survive in the well mixed case as the best replicator in the system outcompetes the others and then is starved for resources, however defining an MCRS model with spatial definition (by means of a lattice model) allows the replicators to coexist (Szilágyi et al., 2017). Furthermore, the MCRS is resistant to parasites as their appearance only locally drains the monomer supply, halting replication in a region and starving the

strands there without affecting the system at large. The MCRS has been widely studied in its ability to tolerate variances in ribozyme activity, system size, neighbourhood size, and phenotype-genotype distinction (Könnyű and Czárán, 2013; Könnyű and Czárán, 2014; Könnyű et al., 2015). Additionally, the replicator network is an open network capable of incorporating novel strands which can amplify the production of the metabolite, showing the MCRS demonstrates evolvability.

1.8 Aims of Thesis

The aims of this thesis are twofold. We wish to reconcile the various spatial and protocellular models discussed into a form lending to quantitative comparison of the benefits provided spatial clustering versus group selection on a system of replicators in an RNA world scenario. We also wish to investigate the conditions in which the linking of functional strands to form a proto-genome is favourable. A comparison of protocells with a spatial self-organization in a lattice based spatial model has previously been done by Takeuchi and Hogeweg (2009) by projecting protocells onto the same lattice structure as the spatial model. While their work demonstrated that the two model types display a similar response in stability, they note that the comparison made was qualitative due to the inherent differences in the models developed in the study (Takeuchi and Hogeweg, 2009).

To allow for proper quantitative comparison, this research uses simulations of protocellular and surface models developed to be dynamically identical with the exception of cell division and diffusion. We model replicators in the form of an error-prone general processive RNA polymerase capable of replicating any template provided to it, the complementary sequence to that polymerase as well as non-functional parasitic sequences

arising from mutation. Systems of these strands are placed into compartments or onto a toroidally wrapped square lattice capable of containing multiple strands per site. The success of the model is tested by determining the maximum mutation rate that can be sustained, as a function of both the size of the compartment or lattice site and the minimum polymerase activity constant.

To study the viability of linking functional strands together, a computational protocell model developed in the prior study is modified to allow additional replicators in the form of synthases and their complementary sequences, as well as linked versions of the functional strands and their complementary sequences. The synthases are assigned a parameter σ indicating their benefit to the system of replicators. Mixtures of the unlinked strands are invaded by the linked strands, and mixtures of the linked strands are invaded by the unlinked strands to determine the conditions under which the linkage of strands is more beneficial than having the strands remain unlinked.

Chapter 2: Survival of RNA replicators in protocells and surface-based systems

2.1 Summary

In this chapter, I present a paper published on August 7, 2019 in *Life*, *9*(3), 65. There is a plethora of studies on the survival of replicators in protocells or bound to a surface in a prebiotic RNA world scenario, and accordingly many computational models have previously been developed and analyzed in an attempt to determine which of the two methods works best. In the studies conducted, disparate dynamics in all the models developed and use of different metrics to judge success make a grounded, quantitative

comparison impossible. In this paper we address this by developing a suite of protocell and spatial lattice models with near identical dynamics, and use the error threshold (see section 1.6) as a metric for success to allow a quantitative comparison to be made. The models discussed are all Monte Carlo computational models, most similar to the stochastic corrector models described previously. The spatial models simulated feature multiple strands per lattice site, analogous to the multiple strands a protocell may contain. Thus, the difference between spatial models and protocell models is the presence of diffusion allowing for mixing between sites in a spatial model, and cellular division in protocell models. The models are compared across a variation in the size (number of strands allowed) of the lattice site/protocell as well as polymerase rate constant (representative of how good the polymerase is). It is found that protocellular models can sustain higher mutation rates in both cases. A further comparison is done with rapidly replicating parasites where again the protocellular model considered performs better than the spatial model.

This project was designed in collaboration with my supervisor, and was motivated by the issue outlined above as well as the partial work of an undergraduate summer student Q. Pauli. I developed the various models presented in the paper in C++ and performed the data analysis primarily using Python code that I wrote, except for the deterministic version of the spatial model which was coded and analyzed by J. de Bouter and the deterministic version of the protocell model which was developed in part by my supervisor and Q. Pauli. The paper makes use of data from an older work by A. Tupper and my supervisor. The paper was written by myself and my supervisor, with assistance during development and

revision from A. Tupper. Figure numbers and table numbers have been reformatted to match the style of this thesis.

2.2 Shah et al., 2019

Survival of RNA replicators is much easier in protocells than in surface-based, spatial systems

Vismay Shah [1], Jonathan de Bouter [1], Quinn Pauli [1], Andrew S Tupper [2], and Paul G Higgs [1,*]

[1] Origins Institute and Department of Physics and Astronomy, McMaster University, Hamilton, Ontario, Canada
[2] Origins Institute and Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario, Canada.
***** Correspondence: higgsp@mcmaster.ca; Tel.: +1 905 525 9140 ext 26870

Received: 11 April 2019; Accepted: 30 July 2019; Published: 7 August 2019

Abstract:

In RNA-World scenarios for the origin of life, replication is catalyzed by polymerase ribozymes. Replicating RNA systems are subject to invasion by non-functional parasitic strands. It is well-known that there are two ways to avoid the destruction of the system by parasites: spatial clustering in models with limited diffusion, or group selection in protocells. Here, we compare computational models of replication in spatial models and protocells as closely as possible in order to determine the relative importance of these mechanisms in the RNA World. For the survival of the polymerases, the replication rate must be greater than a minimum threshold value, k_{min}, and the mutation rate in replication must be less than a maximum value, M_{max}, which is known as the error threshold. For the protocell models, we find that k_{min} is substantially lower and M_{max} is substantially higher than for the equivalent spatial models; thus, the survival of polymerases is much easier in

protocells than on surfaces. The results depend on the maximum number of strands permitted in one protocell or one lattice site in the spatial model, and on whether replication is limited by the supply of monomers or the population size of protocells. The substantial advantages that are seen in the protocell models relative to the spatial models are robust to changing these details. Thus, cooperative polymerases with limited accuracy would have found it much easier to operate inside lipid compartments, and this suggests that protocells may have been a very early step in the development of life. We consider cases where parasites have an equal replication rate to polymerases, and cases where parasites multiply twice as fast as polymerases. The advantage of protocell models over spatial models is increased when the parasites multiply faster.

Keywords: RNA World; polymerase; error threshold; protocell; membranes; spatial lattice model; evolution of cooperation; parasites.

---

1. Introduction

The most widely studied theory for the origin of life—specifically, the transition from a mixture of prebiotic chemical components to a living system that is sustained by autocatalytic replication—is the RNA World theory [1–7]. Most versions of the RNA World propose the existence of RNA polymerase ribozymes that use a second strand as a template to synthesize the complementary strand to the template. RNA polymerase ribozymes have been experimentally developed while using in vitro evolution with maximum template lengths of 200 nucleotides and per-base error rates of a few percent [8–10]. Although none of the laboratory ribozymes is yet able to replicate its own sequence,

these experiments point to the likelihood that ribozymes working in this way could have supported life in its earliest stages.

Replicating systems containing polymerase ribozymes are subject to invasion by non-functional sequences that act as parasites. Parasite strands are likely to frequently arise, either as a result of chemical synthesis of new random sequences, or as a result of errors in the replication of functional strands. It is known that parasites destroy the replicating system unless there is a mechanism that promotes cooperation among groups of functional polymerases. Two such mechanisms have been widely studied: spatial clustering of polymerases in two-dimensional lattice models representing RNA strands bound to a surface [11–18], and group selection of polymerases inside protocell compartments [19–21]. Spatial clustering is beneficial to polymerases, because a neighbour of a polymerase is more likely to be another polymerase and less likely to be a parasite than it would be if sequences were randomly mixed. Protocell compartments benefit polymerases, because compartments with more functional strands and fewer parasites grow and divide more frequently. There have also been several experimental studies of replication inside artificial protocells [22–25]. It is unknown whether the first replicating molecules arose outside of compartments, and they were later encapsulated in membranes, or whether they arose inside a pre-existing system of simple membrane compartments. The ubiquity of cellular life today and the absence of current surface- based, non-encapsulated life forms suggest that cells arose rather early in the history of life.

For polymerase ribozymes to survive, they must replicate faster than the rate at which they are destroyed by hydrolysis; hence, there is a minimum value of the replication rate

constant, k_{min}, which is required for survival. The polymerases must also replicate accurately enough to pass on their own sequence and avoid invasion by parasitic mutant sequences; hence, there is a maximum value of the mutation probability, M_{max}, required for survival. RNA replicating systems are most likely to survive in systems with low k_{min}, and high M_{max}. Here, we study several alternative versions of protocell and spatial models for RNA replication in order to compare the values of k_{min} and M_{max}.

The maximum tolerable error rate in models of sequence replication is usually called the error threshold. The original error threshold theory of Eigen et al. [26] dealt with first order replication, meaning that one strand has the ability to make a second. This is applicable if the catalyst that replicates the strand is not part of the evolutionary model, as is the case in experiments where the RNA sequence evolve in the presence of a supply of Qβ replicase protein that is provided by the experimenter. Here, we deal with second-order catalysis, meaning that two molecules have the ability to make a third. This is applicable in the RNA World, if a polymerase ribozyme acts on a template strand to make a third strand. In both the first- and second-order replication problem, selection favors functional molecules and mutation creates non-functional molecules (or slower replicators with reduced functionality). In the first-order problem, selection arises due to the functional molecule, or master sequence, replicates faster than the mutant sequences. There is no need for spatial structure or protocells in the first-order model, because the master sequence survives by selective advantage even in the well-mixed case. On the other hand, in the second-order problem that was studied here, mutant sequences are parasites that cannot replicate themselves, and we assume that the rate at which parasites are copied by the

polymerase is the same as the rate at which polymerases and complements are copied. There is no selective advantage of the polymerase in this case, and mutation favors the parasites. Therefore, the polymerases are destroyed by the parasites for any non-zero value of the mutation rate if the system is well-mixed. In the second-order case, either spatial structure or protocells are essential for the survival of the polymerases, as we have considered in several of our previous papers [6,16–18], because polymerases then have a selective advantage arising from clustering or group selection.

The simplest kind of deleterious mutation is one that destroys the function of the polymerase completely, without changing its ability to act as a template. We presume that such deleterious mutations will be frequent, because (i) a substantial fraction of point mutations in RNA sequences disrupt the secondary structure and (ii) a useful polymerase in the RNA World needs to be very insensitive to the sequence of the template, otherwise it cannot support further evolution and it cannot provide a means of replicating genes with other functions (for example, nucleotide synthetases [17]). Here, we focus on the case where the replication rate of the parasites is equal to the functional sequence, however the case of parasites whose replication rate is faster than that of the polymerases is also relevant, and it is considered at the end of this paper. In some cases, the advantage to the polymerase arising from clustering or group selection is sufficient to outweigh a large replication-rate advantage of the parasite (we consider a two-fold replication rate advantage in Section 3.4). One reason why parasites might replicate faster is if the mutation destroys the secondary structure of the polymerase, and the mutant sequence spends a larger fraction of its time in an unfolded state that is accessible as the template for another polymerase. A second reason

might be that parasites could be shorter in sequence, and hence more rapidly replicated. We previously considered the case where short parasites are created by incomplete replication that terminates before the end of the strand is reached [18].

Although catalytic replication in the RNA World is second-order, first-order replication could also be relevant if it is non-enzymatic. We refer to first- and second-order rates as r and k respectively. We have previously considered cases where both r and k are considered in the model [16,27]. These earlier papers focus on the transition to life/replication, rather than the maintenance of replication in system that is already living, as we do here. Some small r rate is necessary for creating the first catalysts, but the k rate is likely to be much higher than the r rate once well-adapted polymerases are present; hence, we assume that r can be neglected. The relative rates of r and k are important when we consider the question of which kind of ribozymes came first [28]. If r was small, the first biological catalysts must have been polymerases, whereas if r was high (because non-enzymatic replication was intrinsically fast), then other kinds of catalysts that contribute to the synthesis of RNA and its precursors could have preceded polymerases. These scenarios are qualitatively different, as we discuss in [28]. Functional ribozymes are likely to be quite long (maybe 100 nucleotides or more). We do not know whether the replication of sequences of this length could be non-enzymatically possible; however, non-enzymatic replication of short oligomers seems likely, and it may have preceded the origin of ribozymes. We have referred to non-enzymatic replication of oligomers as ″chemical evolution″ [29]. Such a system would really be evolving, because sequence information in the oligomers would be passed on during replication. However, it would be distinct from

the usual view of molecular evolution, which we have called ″biological evolution″, because sequences would be selected based on physicochemical properties rather than encoded function, and because oligomers would be short enough to be synthesized from scratch as well as by replication of existing oligomers. In the chemical evolution case, diversity can arise by chemical synthesis as well as by mutations of existing sequences, whereas, in biological evolution, diversity arises only by mutation, because functional gene sequences (including well-adapted ribozymes in the RNA world) would be too long to be synthesized by means other than copying an existing template. We have argued that chemical evolution is a significant step on the path to life [29], and that, if evolution is considered to be a defining feature of life, then it is the presence of biological evolution that defines life, not simply chemical evolution.

We now return the principle question in the present paper. Assuming that replication in the RNA World is maintained by second-order polymerases, and that first-order non-enzymatic replication can be neglected at this stage, then how can we quantitatively compare spatial clustering and encapsulation in protocells as mechanisms allowing for the survival of polymerases? This requires us to think carefully about what the spatial models actually represent. In spatial lattice models of RNA replicators [11–18], it is usually assumed that only one strand is allowed per lattice site, and that a polymerase on one site replicates a template on a neighboring site. One way to view this would be to say that the lattice represents a two-dimensional surface, on which the strands are fixed. There is evidence that synthesis of short RNA oligomers can occur in the presence of clay surfaces [30] and in alkaline hydrothermal vent systems [31], but it is less clear whether minerals

help in ribozyme function. It has been shown that clays can increase the rate of self-cleavage of the hammerhead ribozyme [32], but cleavage is not equivalent to replication. In a study of *in vitro* RNA evolution with and without the presence of clay [33], it was concluded that the effect of the clay was minimal, neither improving nor preventing the ability of RNA to evolve functional structures. Furthermore, ligases and polymerases are the most relevant laboratory ribozymes for replication in the RNA world [8–10], and these are not associated with mineral surfaces.

We suggest that the spatial models of replication are best viewed as representing the effects of confined geometry, slow diffusion, and spatial clustering of cooperating molecules, rather than literally representing molecules that are ″stuck″ on a surface. Spatial models require slow diffusion of molecules, so that clustering of polymerases arises. Nevertheless, some degree of motion is required so that replicating molecules can spread and encounters occur between polymerases and templates. One conceptual problem is that, if a molecule is stuck to a surface, it is difficult to see how it could slide along the surface without detaching. If it detached from the surface, then it would often diffuse away from the surface and be lost in open water. On the other hand, if the spatial lattice represents a confined geometry, such as pores in a rock [34], a mineral matrix [13], or cavities in which strands are trapped by thermophoresis [35], then diffusion will slowly occur and the molecules will remain within the restricted space.

In this paper, we consider the spatial models in which diffusion occurs slowly in a restricted space. When we think in this way, there is a small length scale (pore or cavity size) within which strands can quickly mix and interact, but motion of strands on large

scales is controlled by slow diffusion. We can think of one lattice site in the models as a pore size, and the volume of this pore will control the maximum number of strands that can be on one site, which we call S_0. One example of a model that is defined in this manner is featured in the study of Branciamore et al. [13]. In their paper, several types of autocatalytic replicators, each catalyzing one reaction in a metabolic network, could be present in a pore. Each pore was assigned a fitness that corresponded to the diversity of replicators it contained, with the requirement that at least one strand of each replicator type be present. Parasitic replicators were introduced through invasion and were also autocatalytic, competing with members of the network for resources without catalyzing any of the reactions [13]. In contrast, instead, we focus on a trans-acting polymerase. The parasitic sequences that we consider are fundamentally different: they cannot replicate without a polymerase present in the site. The polymerases may erroneously produce parasites from an improper replication. Hence, in our study, the central question is the maximum mutation rate the system can sustain rather than the number of different replicator species that can be sustained.

The minimum number of strands that must be allowed on a site if we want to allow for second order replication steps to occur on one site is three. The ″two′s company and three′s a crowd″ scheme that we previously studied [27,16], allows up to three strands per site for this reason. In the current paper, our object is to compare lattice models with protocell models; therefore, we want the underlying rules of the two types of models to be as similar as possible. In protocell models [19–21], it is typically assumed that the strands inside one protocell are well mixed and interact freely, but there is no interaction between strands in

different protocells. Here, we want to make a direct analogy between one lattice site in the spatial models and one protocell in the protocell models. We define S_0 in the protocell models as the number of strands at which cells divide. Thus, S_0 controls the maximum number of strands per cell/site in both types of model. The rules for replication of strands are identical in the two types of models when we define the models in this way. This enables us to focus on the *differences* between spatial models and protocells: in spatial models, replication occurs locally on one lattice site and the diffusion of strands occurs between neighbouring sites, whereas in protocell models, replication occurs locally in one protocell, new cells arise when cells divide, there is no diffusion of strands between cells, and there is no spatial structure of the cells.

An alternative way to compare spatial models and compartments is to use the Cellular Potts Model (CPM) [14], in which each compartment occupies multiple sites and only one strand is allowed per site. However, this model does not separate the effects of the compartments from the effects of spatial structure, because the compartments in the CPM are themselves on a lattice and they have spatial neighbours. Diffusion of strands can occur between neighboring compartments in the CPM, whereas in the protocell models that we use here, there is no diffusion between cells, simply growth, division, and death of cells. In some ways, the CPM model is closer to a spatial model in a restricted geometry than it is to a model of independent protocells. The study using the CPM [14] deliberately avoided making quantitative comparisons between the models with and without compartments, and concluded that it was impracticable to make a fair quantitative comparison. However, here, we have defined the rules of the models so that a fair quantitative comparison can be made,

and such that there is either slow diffusion and spatial clustering or group selection in compartments, and not both. We will show that the outcome of this comparison is that there is a substantial quantitative advantage to the protocell models over the equivalent spatial models, both in terms of the minimum catalytic rate that is required for survival and the maximum error rate that can be tolerated.

2. Materials and Methods

2.1. Overview of models

In this study, we compare protocell models and spatial lattice models in such a way that there is close analogy between one site in the lattice models and one protocell in the protocell models. Each lattice site (or each protocell) can hold multiple strands. Reactions that create and destroy strands occur locally on one site (or in one protocell) and are equivalently defined in the two kinds of models. Differences between the models are related to the dynamics of strands between sites (or protocells) and the factors that limit the replication of strands.

Each strand is one of three types: a polymerase (P), a complementary sequence to the polymerase (C), or a non-functional strand, which we refer to as a parasite (X). Accurate replication of a P produces a C, and vice versa. If a point mutation occurs during replication of a P or C, an X is produced. The replication of an X always produces another X. We do not allow back mutations that produce P or C strands from replicating an X. Replication only occurs via the action of a polymerase catalyst, and we ignore non-enzymatic template-directed replication. There is a rate of breakdown of strands back to monomers that is assumed to be equal for the three types of strands.

Figure 2.1 and Table 2.1 summarize the models that were studied in this paper. In the two protocell models, division occurs when the number of strands in a cell, *S*, reaches a specified value, S_0. This produces two daughter cells with the strands that were randomly divided between them. The number of protocells in the population, *N*, is a fixed parameter in the PCP model (Protocells with Constant Population), and is variable in the PML model (Protocells-Monomer Limited). In the PCP model, whenever a cell divides, another random cell is removed from the population to keep *N* fixed. This represents a situation where resources, such as lipids or available space, limit population growth. It is analogous to the standard Moran model that was used in population genetics [36]. In each model, there is a limiting factor *F* in the replication rates, which is required for preventing indefinite increase of either the population or the number of strands (details below in Section 2.2). In the PCP model, the population is already limited by fixing *N*, therefore no additional limiting factor is needed ($F = 1$). In the PML model, there is no limit to the number of cells, but the number of strands is limited by the availability of monomers (*i.e.* nucleotides). The limiting factor is $F = 1 - S_{tot}/S_{max}$, where S_{tot} is the total number of strands in the whole population , and S_{max} is the maximum allowed number of strands. We call this limit global, because it applies equally to all cells in the population. In the PML case, when a cell divides, it is not coupled to the removal of another cell. Instead, all empty cells with $S = 0$ are immediately removed in order to prevent the accumulation of empty cells. In the PCP model, we do not need to immediately remove empty cells because they are eventually removed at random due to the birth and death process of cells.

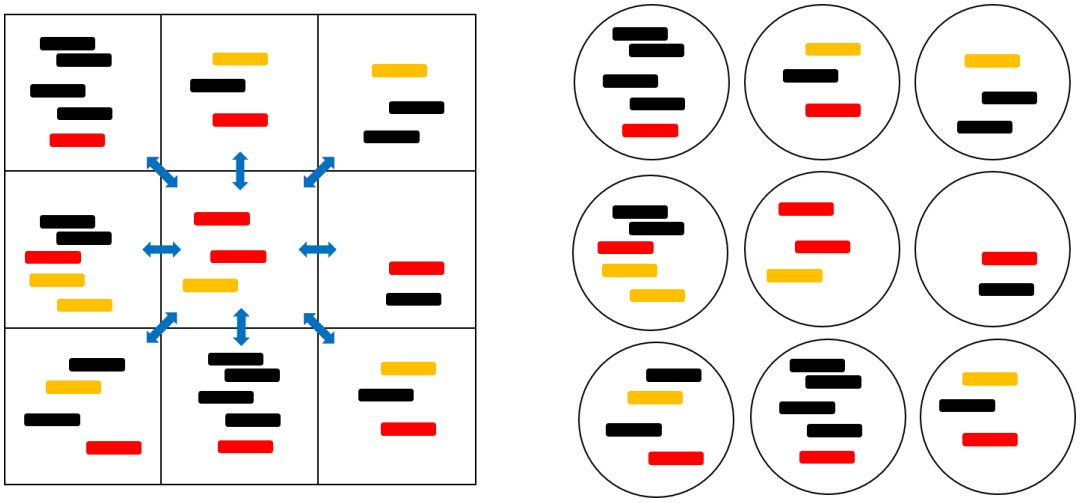


Figure 2.1 A cartoon representation of the spatial model with local diffusion dynamics (left) and the protocell models (right). The red strands are polymerases (P), orange strands are complements to polymerases (C), and black strands are parasites (X). The blue arrows indicate the possibility of diffusion to and from the eight neighboring sites.

Table 2.1 Overview of models in this study.

Model	**Dynamics**	**Limiting Factor**	**Volume**
PCP - Protocells with Constant Population	Division when $S \geq S_0$ N fixed	No limit, $F = 1$	Grows with cell $V = S$
PCPCV - Protocells with Constant Population and Constant Volume	Division when $S \geq S_0$ N fixed	No limit, $F = 1$	Constant $V = S_0$
PML - Protocells - Monomer Limited	Division when $S \geq S_0$ N variable	Global limit, $F = 1 - S_{tot}/S_{max}$	Grows with cell $V = S$
SLD - Spatial Model with Local Diffusion	Local diffusion rate h	Local limit, $F = 1 - S/S_0$	Constant $V = S_0$
SMF - Spatial Model with Mean Field dynamics	Mean field diffusion rate h	Local limit, $F = 1 - S/S_0$	Constant $V = S_0$
SML - Spatial Model – Monomer Limited	Local diffusion rate h	Global limit, $F = 1 - S_{tot}/S_{max}$	Constant $V = S_0$

The PCP and PML models correspond to different assumptions regarding the processes limiting protocell growth. However, we will show below that these two models are

surprisingly similar with regard to their error threshold behaviour. Hence, the differences that we observe between the protocells and spatial models do not depend on the process that limits protocell growth.

In spatial lattice models, there is no division process, but strands can diffuse from one site to another. The most natural spatial model, which we call SLD (Spatial model with Local Diffusion) has local diffusion of strands between each site and its nearest neighbours that is controlled by a hopping rate h. In this model, the parameter S_0 controls the number of strands on any one site. The limiting factor is $F = 1 - S/S_0$, which means that no further replication is possible on a site when $S \geq S_0$. The local motion of strands leads to a build up of correlations between the contents of one site and its neighbouring sites. This correlation causes the clustering of polymerases, which is part of the reason that the spatial model allows for the survival of polymerases and avoids destruction by parasites. Therefore, it is useful to consider the Spatial Model with Mean Field dynamics (SMF) model as a comparison to this. In mean field dynamics, whenever a strand hops to a different site, it is placed on any other site with equal probability, rather than on a neighbouring site. We have previously studied mean field models with small numbers of strands allowed per site [16, 27]. If only one strand is allowed per site, then the mean field model is the same as the well-mixed case, which is not useful, because polymerases are always destroyed by parasites. When up to three strands are permitted per site, the mean field model shows the correct qualitative behaviour, but is still quantitatively very different from the model with local dynamics. We will show here that when many strands are possible per site ($S_0 = 10$ or larger in the examples in this paper), there is very little difference between mean field and local

dynamics; hence, the mean field approximation is useful. An advantage of the SMF model is that it is possible to give a deterministic solution; whereas the SLD model requires stochastic simulations.

In both SMF and SLD, the limiting factor on strand growth is applied locally on each site. We also considered a third model, SML (Spatial - Monomer Limited), in which the global supply of monomers limits the strand growth, that is, the limiting factor is $F = 1 - S_{tot}/S_{max}$, as in the PML model. This corresponds to a case where monomers diffuse rapidly, hence the concentration is the same everywhere. In this way, we can compare protocell and spatial models when the monomer limitation is applied in the same way in the two cases. We did not study a protocell case where there is a local limit on growth, because we are assuming that there is no spatial structure in the protocell population above the level of the cells.

The last column in Table 2.1 gives the volume, V. This is fixed at $V = S_0$ in the spatial models, and it grows in proportion to the number of strands in the PCP and PML models, $V = S$. The volume determines the strand concentrations, and hence the reaction rates, as described in Section 2.2. Although it seems natural to keep V constant in the spatial models and to allow it to grow in the protocell models, it is useful for comparison to consider an additional model, PCPCV, in which the volume is kept constant. We will show below that there is a relatively small difference between the PCPCV and PCP models, so the question of whether the protocell volume grows or is fixed is a relatively minor one.

2.2. Model details

In all models, replication requires the encounter of a polymerase with another strand serving as a template, and produces a strand complementary to the template. Let *p*, *c* and *x* label the numbers of P, C and X strands in one site/protocell at a given moment in time, and let $K_P(p,c,x)$, $K_C(p,c,x)$, and $K_X(p,c,x)$ denote the rates of production of P, C and X strands in this site/protocell. For all models, we may write

$$K_P(p,c,x) = (1-M)\frac{pcF}{V}k.$$

$$K_C(p,c,x) = (1-M)\frac{(p-1)pF}{V}k.$$

$$K_X(p,c,x) = M\frac{pcF}{V}k + M\frac{(p-1)pF}{V}k + \frac{pxF}{V}k.$$

In the formula for K_P, *k* is the replication rate per polymerase, *c* is the number of C templates from which new P strands can be produced, and the concentration of polymerases is *p/V*, where *V* is the volume of the cell/lattice site. Note that the rate of increase in the *concentration* of product strands would be proportional to the concentration of the polymerases, *p/*V, times the concentration of the templates, *c/V*. However, K_p is the rate of increase in *number* of strands per cell, not the concentration, so there is an extra factor of *V*. Hence, K_p depends on *pc/V*, not pc/V^2. Equivalently, we may say that the rate of increase in the number of product strands is proportional to the concentration of polymerases, *p/V*, times the number of templates, *c*.

M is the probability that a mutation occurs from a P or C to an X during replication. *F* is the limiting factor that prevents the indefinite increase of strands, as discussed in section 2.1 and Table 2.1. The formula for K_C differs, in that the number of *P* templates is *p*, and

the concentration of *other* P strands that can act as polymerases is (p-1)/V. The formula for K_X includes the term for direct replication of X, plus the terms for creation of X strands by errors in replication of P and C.

The stochastic simulation of these models proceeds in time steps δt. In each time step, births and deaths of strands are considered separately on each site/protocell. The probabilities of adding one P, C or X strand are $K_P(p,c,x)\delta t$, $K_C(p,c,x)\delta t$, and $K_X(p,c,x)\delta t$. Strands break down at a constant rate, defined as $\nu = 1$. The probabilities of removal of one P, C, or X strand from a site/protocell are therefore $\nu p\delta t$, $\nu p\delta t$ and $\nu x\delta t$, respectively. After birth and death of strands, protocell division occurs in the protocell models and diffusion occurs in the spatial lattice models.

In the protocell models, $S = p + c + x$ is the current number of strands. Cells with $S \geq S_0$ undergo random division. Cell division is assumed to be rapid once the split size is reached, *i.e.* all cells with $S \geq S_0$ divide with probability 1 in one time step. The strands from the parent cell are assigned independently with equal probability to one of the two daughter cells. Even though cell division immediately occurs on reaching S_0 strands per cell, it is possible for a small number of cells with $S \geq S_0$ to remain in the population after cell division. Firstly, it is occasionally possible to create cells with more than S_0 strands, because replications of P, C and X strands are independently considered; hence, more than one replication can occur in the same cell in one time step. Secondly, it is possible for the random split to occasionally yield S_0 strands in one daughter and zero in the other; hence, there will sometimes still be S_0 strands after division.

In the PCP model, we begin with N cells, each having one P, one C and one X. The maximum number of strands that can arise in the PCP model is NS_0. We set S_{max} in PML to NS_0, where N is the fixed population size of the PCP model in order to compare PML with PCP. In the PML model, we begin with $S_{max}/2$ cells, each having one P, one C and one X.

In the spatial lattice models, the number of lattice sites is analogous to the population size. We consider a square lattice of $N = L \times L$ sites with periodic boundaries (edges connected in a torus). We begin with one P, one C and one X in each site. There is a probability $h\delta t$ per time step that a strand diffuses to another site. In the SLD and SML models, strands randomly move to one of the eight sites in their Moore neighbourhood (Figure 2.1). In the SMF model, strands move to any other site at random. The destruction rate of strands is $v = 1$ in the spatial models, in the same as for the protocell models.

All these models can be simulated by stochastic methods with finite population sizes and finite numbers of strands. However, in some cases, we can also consider deterministic versions of these models by solving the master equations for the probability distribution $P(p,c,x)$ that a site/protocell has p, c and x strands of types P, C and X. This is done in the Appendix for the protocell model with constant population size and the lattice model with long-distance diffusion.

3. Results and Discussion

3.1. Error Threshold Behaviour

Figures 2.2a and 2.2b show the concentrations of P, C and X strands as a function of mutation rate, M, for the PCP model with $S_0 = 10$ and 20. The smooth lines are obtained

from the deterministic theory in the Appendix, which applies for infinite populations. The points are measured by simulations with N = 1024. These show typical error-threshold behavior. The numbers of P and C strands per cell decrease steadily as the mutation rate is increased, while the number of X strands passes through a maximum. There are always slightly more P than C strands because of the (p-1) factor in $K_c(p, c, x)$, (*i.e.* a P cannot replicate itself, whereas a P can replicate all C's). All three strands die out at the error threshold, $M = M_{max}$. The deterministic theory predicts that the strand numbers smoothly decrease to zero as M approaches M_{max}. Close to this point, the expected number of viable cells in a finite population is very small; hence, the finite population simulations are vulnerable to stochastic fluctuations causing the death of the system. The average number of strands in the simulations in the long-time limit is then zero. This causes the simulated systems to die out at slightly smaller values of M than is predicted by deterministic models.

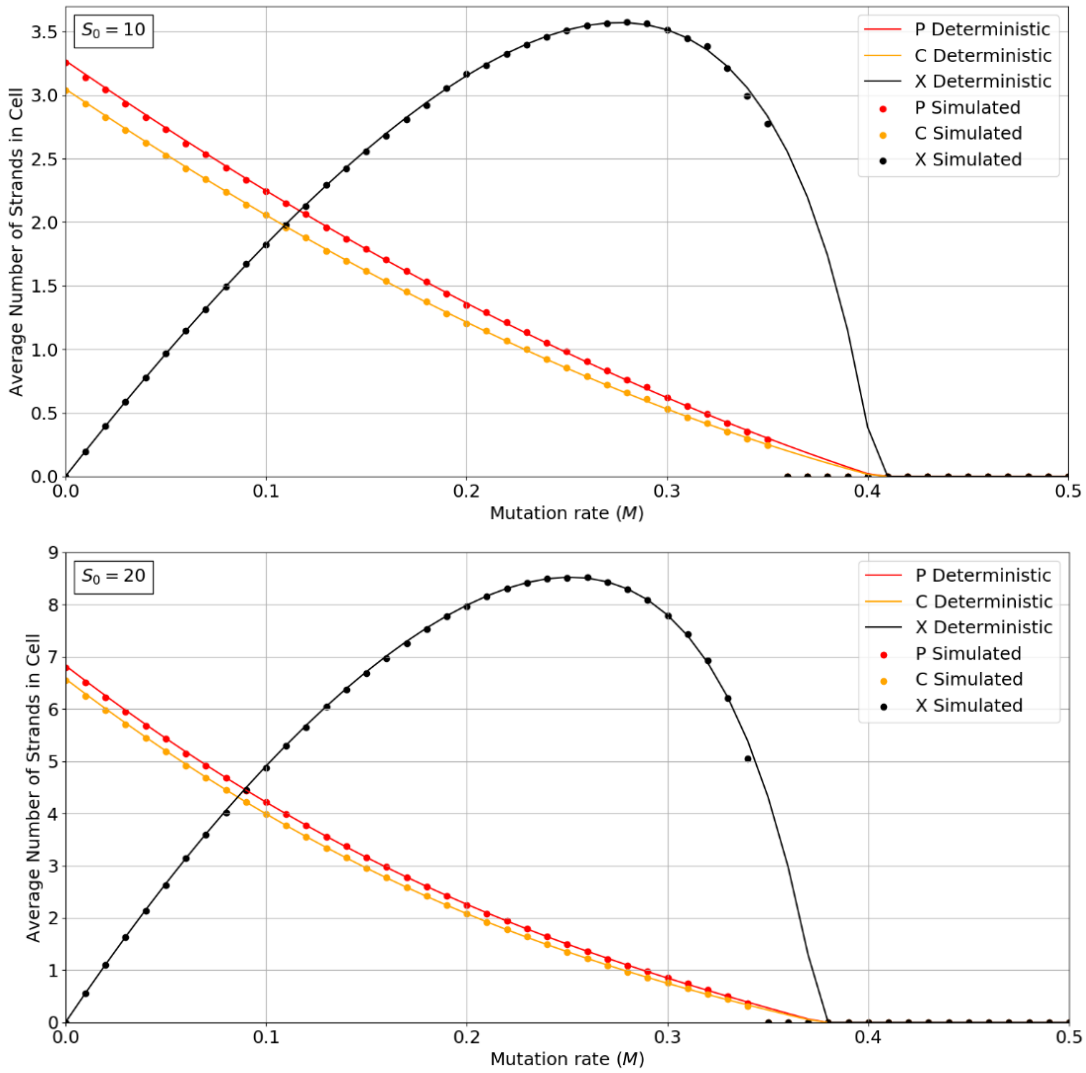
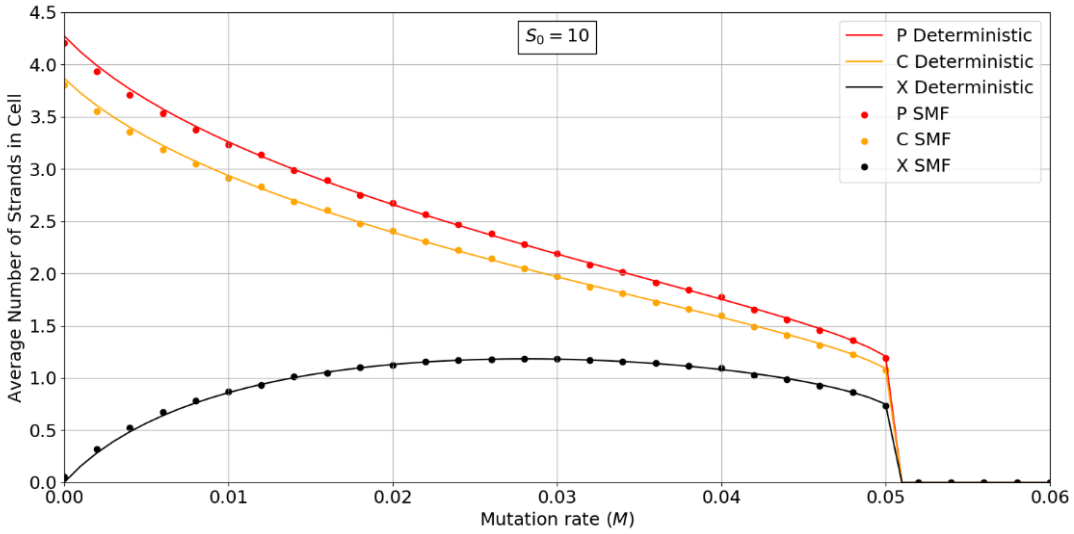


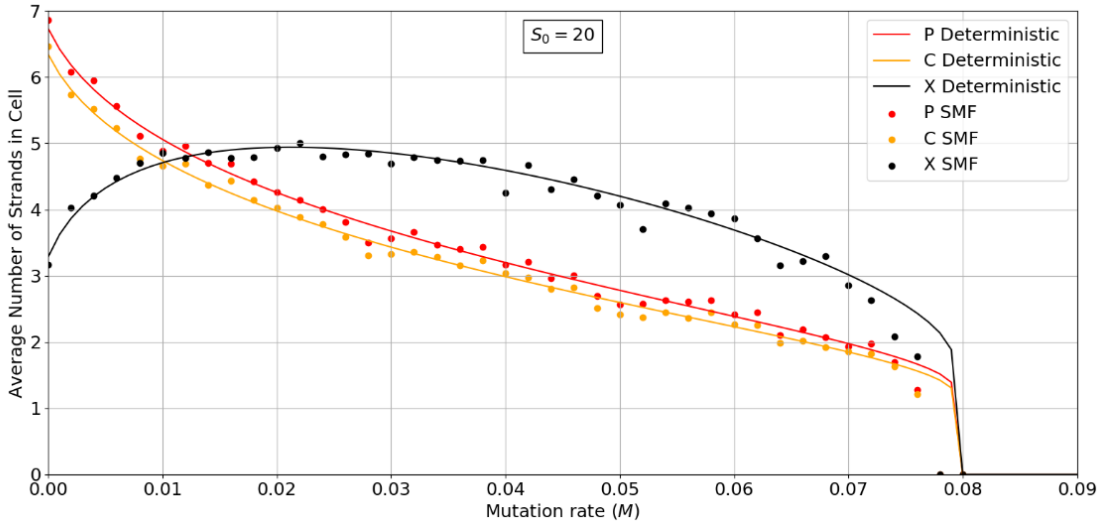
Figure 2.2 Average numbers of strands per cell in the PCP model. ***(a)*** $S_0 = 10$, ***(b)*** $S_0 = 20$. *k = 25 in both cases. Points are from finite population simulations. Smooth lines are from deterministic theory.*

Figures 2.3a and 2.3b show the error threshold behavior for the SMF model. In this case, the deterministic theory and the finite population simulation both show a discontinuous transition at the error threshold, i.e. the jump in the curve is not due to stochastic extinction in small populations, as it is in Figure 2.2. Comparison of Figures 2.2 and 2.3 shows that the error threshold is much larger in the protocell model than the lattice

model, as we discuss in more detail in section 3.2. Additionally of note is the fact that in both the deterministic and mean field versions of the spatial model, at higher S_0 values, there is a non-zero parasite population present even at zero mutation rates. In other words there is a coexistence of non-functional parasites with polymerases, even when the parasites are not replenished by mutations from the polymerases. This is a significant difference from the protocell models considered in Figure 2.2, where the parasites are always purged from the systems at zero mutation rates.



(**a**)



(b)

Figure 2.3 Average numbers of strands per site in the SMF model. ***(a)*** N = *100,* k = *25,* h = *0.4 and* ***(b)*** N = *400,* k = *20,* h = *0.4. Points are from finite population simulations. Smooth lines are from deterministic theory.*

3.2. Comparison of Error Thresholds in Different Models

The two key properties that we wish to compare between all of the models are the error threshold value, M_{max} (*i.e.* the maximum sustainable error probability per replication of the whole sequence) and the minimum catalytic rate, $k_{min,}$ required for survival of the polymerases. Figure 2.4 shows M_{max} measured from simulations as a function of k. The estimates of M_{max} were obtained by running a series of simulations at each value of k and gradually adjusting the mutation rate to zero in on the error threshold. A similar method was used to produce Figure 2.5, where S_0 was held fixed.

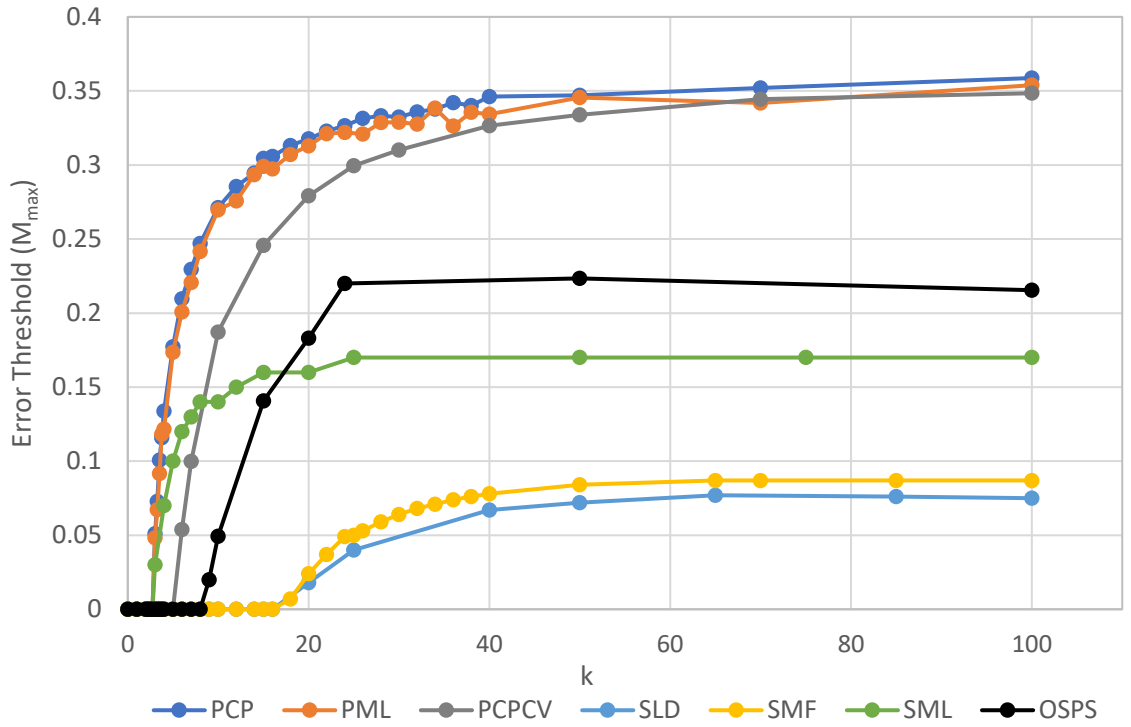


Figure 2.4 Comparison of the error threshold of the various models studied as a function of the polymerization rate k. S_0 = 10 in all models except the one per site model, and h *= 0.4 in the lattice models. All results are from stochastic simulations except for SMF, which results are from the deterministic method. OSPS is the one strand per site model from [18]. Other models are defined in Table 2.1.*

We will initially discuss the two principal protocell models, PCP and PML, in comparison to the two principal spatial models, SLD and SMF. The other models will be discussed later, because we consider them to be less realistic. The PCP and PML models show a higher error threshold than the SLD and SMF over the whole range of *k* studied, and require the lowest values of *k* to survive. The protocell models are thus "better" for the RNA World, in the sense that survival of the polymerases is substantially easier in the protocells than the lattice models.

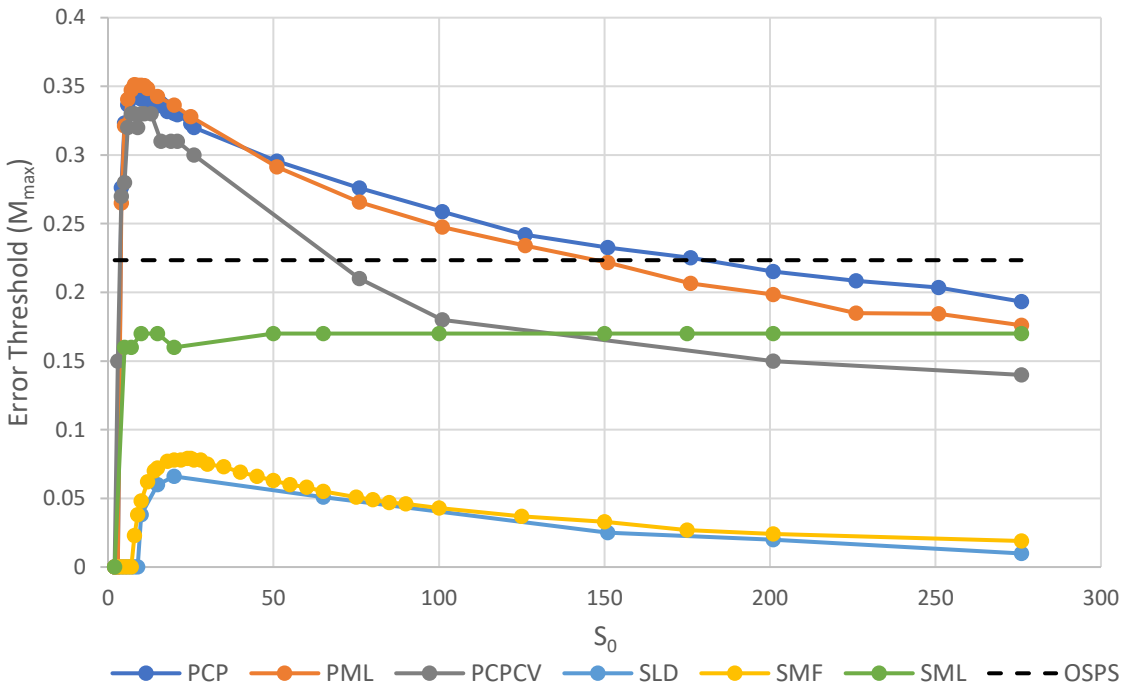


Figure 2.5 Comparison of the error threshold of the various models studied as a function of S_0. k = *25 in all models, and* h = *0.4 in the lattice models. Results for SMF are obtained from the deterministic method, except for the points with* $S_0 > 150$, *where the deterministic method becomes much slower than the stochastic simulation. Results for the other models are obtained from stochastic simulations. OSPS is the one strand per site model from [18]. Other models are defined in Table 2.1.*

It should be remembered that, even when there are no replication errors ($M = 0$), a minimum value of k is necessary for survival, because replication must be faster than the breakdown rate of the strands ($\nu = 1$). Thus k_{min} is the value of k at which M_{max} becomes zero. For the PCP and PML models, k_{min} is approximately 3, whereas it is approximately 18 for SLD and SMF. Thus there is a substantial range $3 \leq k \leq 18$ where replication is possible in protocells and not in spatial models. All these rates should be thought of as relative to the breakdown rate, because we have set $\nu = 1$.

For well-adapted ribozymes, where $k \geq k_{min,}$ we find M_{max} is around 0.36 for PCP, but only approx 0.075 for SLD and 0.09 for SMF. Thus the protocell models are four- to five-fold more tolerant of error. These figures are per-sequence. If they are converted to per-base error rates, this implies that there is a four- to five-fold greater limit in the maximum length of replicating sequences that can be maintained in protocells relative to spatial models.

Figure 2.4 also shows the PCPCV model. This model has the volume fixed to S_0 in the same way as it is in the spatial models, and therefore eliminates a minor difference in the definitions of protocell and spatial models. The error threshold of PCPCV is reduced slightly relative to PCP, but it is still much higher than the spatial models. Therefore the issue of whether the protocell volume is fixed or grows with the number of strands is only a minor effect. PCP seems more realistic because in reality a cell cannot keep constant volume when it divides.

We now turn to the SML model. This has M_{max} intermediate between the protocell models and the other spatial models, and has k_{min} almost equal to the protocell models. This comparison is interesting from a theoretical point of view, as it highlights the fact that the local limitation on growth that applies in the SLD and SMF models leads to much lower error thresholds than the global limitation in the SML model. However, there are problems with the SML model that mean that is not a biologically realistic model. Replication is fastest on sites with the largest number of polymerases. Strands tend to pile up with very large numbers of strands on a very small number of sites, and with many other sites being empty, as there is no local limit on the number of strands per site in the SML model. This

cannot be realistic, because sooner or later, local limits must take effect. Either the monomer limit becomes local, because the concentration of available monomers becomes depleted on sites when there is a lot of replication, or the local limit of space takes effect. Thus, we consider the SLD to be the most realistic of the spatial models, and the comparison between the SLD and the two protocell models as the most valid comparison of the differences between spatial models and protocells.

The final model on Figure 2.4 is the surface model with only one strand per site (OSPS), taken from Figure 3 of Tupper *et al.* [18]. In that model, a P strand replicates a strand on a neighbouring site (because there is no other strand on the same site). This model is also intermediate between the protocell models and the SLD model, but again, this seems less realistic than SLD, and it cannot be easily compared with the protocell models because there is no way of having a protocell with only one strand per compartment.

The parameter S_0, which controls the number of strands per site/cell has important effects on the error threshold, as shown in Figure 2.5. For a site/cell to be viable, there must be a minimum of either two P's or one P and one C. When S_0 is small, there are many sites that are not viable, and the whole system dies out. Once S_0 is above this minimum size for viability, M_{max} increases rapidly with S_o and then decreases slowly as S_0 becomes large. For very large S_0, each site is a well-mixed model, and there is no more clustering or group selection. Therefore, M_{max} must tend to zero for very large S_0. The SML model is an outlier here in that it is not affected by high S_0 values. In the SML model, the only effect of S_0 is to determine the total number of strands, because $S_{max} = S_oN$, and it does not limit the number of strands on one site, as it does in the other models. The one strand per site model

from Tupper *et al.* [18] is also shown as a comparison, but there is no equivalent of S_0 in this case. It should be remembered that polymerases replicate templates on neighbouring sites in ref. [18], but on the same site in this paper. Hence it is not possible to have $S_0 = 1$ in the spatial models in this paper. Once again, in Figure 2.5, we see that the PCP and PML models are very similar, and that the PCPCV is only slightly lower than the PCP model. The most useful comparison is between the PCP/PML models and the SLD model, and this shows a substantially larger error threshold for the protocell models, by a factor of 4 to 10.

An interesting observation in Figures 2.4 and 2.5 is that the PCP and PML models have almost equal error thresholds, even though the models differ in important respects. For example, all cells die with equal probability in the PCP model, but only empty cells die in the PML model. The average replication rate of strands is substantially faster than v in the PCP model, because replication has to balance the removal of strands occuring when cells die as well as when individual strands are removed. In contrast, the average replication rate of strands in the PML case is equal to v, because the limiting factor reduces this rate to balance the removal of individual strands. Nevertheless, we observe that these apparently large differences do not have a large effect on the error threshold. This is probably because the differences between the models disappear as the mutation rate approaches the error threshold. In the PCP model, the fraction of viable cells becomes very small when $M \rightarrow M_{max}$; therefore the cell division rate is very low, and the rate of removal of strands due to cell death becomes small relative to the rate of removal of individual strands. In the PML

model, there is a limiting factor $F = 1 - S_{tot}/S_{max}$ which is not present in the PCP model. However, when $M \to M_{max}$, the total number of cells is very small and the total number of strands is much less than S_{max}. This means that $F \to 1$, so this difference between the models also disappears close to the error threshold.

3.3 Effect of Diffusion Rate in the Spatial Models

It can also be seen in Figures 2.4 and 2.5 that SLD and SMF models give quite similar results. This means that the mean field approximation is quite a good one. With the parameters that are chosen in Figures 2.4 and 2.5, the error threshold is slightly lower with local diffusion than in the mean field case. However, this depends on the hopping rate *h*, which we have not yet considered. All of the above results were performed with a single value, *h* = 0.4, which was chosen at the beginning of this study because it gave fairly good survival of polymerases in the spatial models. If *h* is too large, the spatial model becomes well-mixed, and polymerases do not survive. If *h* is too small, there is no spread of strands between sites, and polymerases become extinct independently on each site. The effect of diffusion has also been studied by Branciamore et al. [13] in the case of metabolic replicators, rather than polymerases.

Therefore, we performed a further comparison of SLD and SMF models as a function of *h*, as shown in Figure 2.6. There is an optimum value of *h* close to 1, at which the error threshold is highest. The optimum *h* is slightly higher for the local diffusion model, but is of order $h \sim 1$ in both cases. Below the optimum *h*, the SMF has a slightly higher error threshold (as in Figs. 4 and 5), and above the optimum *h*, the SLD model has a slightly higher error threshold, but the difference is always small. The value *h* = 0.4 chosen initially

is slightly below the optimum value for both models. Hence, if h were tuned to the optimum, the results for SLD and SMF in Figs 4 and 5 would be slightly higher. Nevertheless, the error threshold in the two spatial models is only 0.07 at the optimum h, which is still very much less than the protocell models (M_{max} is approximately 0.32 for the protocell models with S_0 = 10 and k = 25, as shown in Fig 2a). Furthermore, there is no reason in nature why diffusion should be tuned to the optimum value. For most of the range of h, the error threshold for the spatial models would be even lower than those shown in Figures 2.4 and 2.5, and the system cannot survive at all (M_{max} = 0) if h is too high or too low. The problem of tuning diffusion does not arise in the protocell models, which is another advantage of protocells.

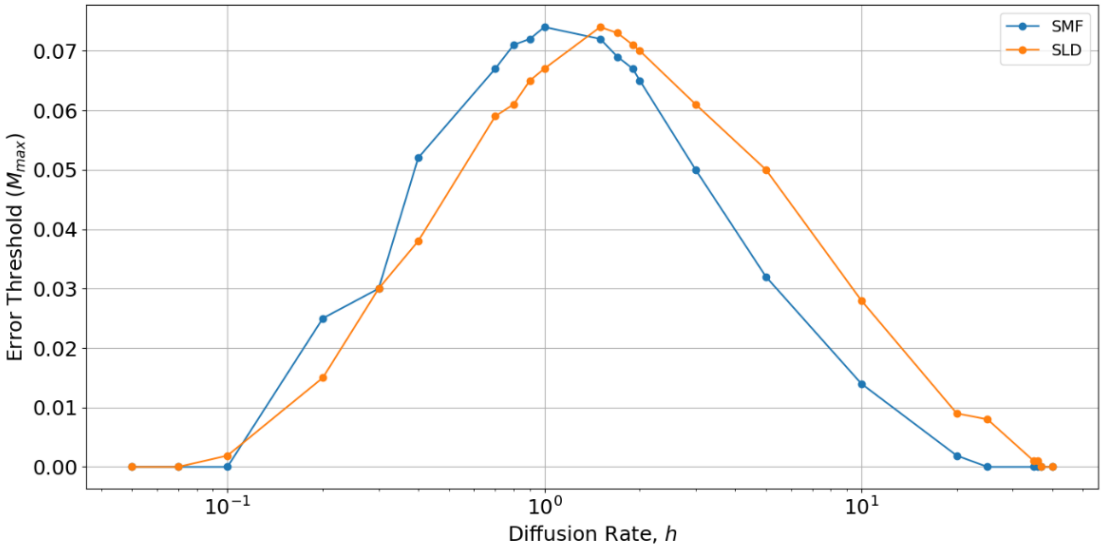


Figure 2.6 Comparison of the error thresholds of the spatial models with long distance diffusion and local diffusion as a function of the diffusion rate h. Made using S_0 = 10, k = 25.

The fact that the difference between local diffusion and mean field cases is small means that the effect of the dimension of space is small. Our spatial models are all studied on a

two-dimensional square lattice. However, we have argued that the spatial model is more usefully thought of as representing a confined geometry, such as pores in a rock rather than molecules stuck on a two- dimensional surface. There is no reason why the lattice needs to be two-dimensional. If we considered a three-dimensional lattice, the results would be between the two-dimensional (2D) and mean field cases, *i.e.* there would be little difference. However, it should be remembered that the spatial clustering mechanism only works if diffusion is very slow. Therefore, it would not apply to an open solution in three-dimensions (3D), in which diffusion and mixing would be rapid when compared to replication.

3.4 Rapidly-replicating Parasites

In all of the previous results, we have assumed that parasite templates are replicated at the same rate as functional strands. It seems likely that a substantial fraction of point mutations in the polymerase would disrupt the structure and prevent its function as a ribozyme, but have almost no effect on the ability of the sequence to be a template. However, it is also possible that some mutant sequences would be better templates than the original polymerase. Therefore, in this section we consider the case where the replication rate for parasites is $2k$, whereas it remains at k for the polymerase and complement. The parasite is thus favored by both mutation and speed of replication. Nevertheless, group selection and clustering effects mean that the polymerase can survive the presence of rapidly multiplying parasites for some parameter values.

Figure 2.7 shows the error thresholds of PCP and SLD models as a function of k in the case where parasites have the same replication rate as polymerases, together with the

equivalent models where parasites have double the replication rate of polymerases (denoted PCP2X and SLD2X). Both error thresholds are reduced when the parasites multiply faster, but the SLD model is reduced more. For well-adapted ribozymes with high *k*, the ratio of error thresholds for PCP and SLD is 0.36/0.075 = 4.8, whereas the ratio for PCP2X and SLD2X is 0.2/0.014 = 14.3. Thus, the addition of faster replicating parasites increases the advantage of protocells over spatial models.

Figure 2.8 shows the error thresholds for the same models as a function of S_0. The error threshold for PCP2X is reduced substantially relative to PCP for larger S_0. Nevertheless there is a non-zero error threshold in PCP2X up to at least $S_0 = 275$. On the other hand, there is only a very narrow range of S_0 (approximately 7 - 20) where the error threshold is non-zero for SLD2X, and even within this range, the error threshold is extremely low. For $S_0 > 20$ in the SLD2X model, fast replicating parasites multiply and lead to destruction of the polymerases (and themselves) even in the limit of zero mutation rate. We allowed the system to reach a steady state with only P and C strands present to test the limit of zero mutation rate. A very small number of parasites were then added, and replication continued with zero mutation rate. For $S_0 > 20$, the initial few parasites multiply and destroy the system even though there is no further production of parasites by mutation. In contrast, there is a finite error threshold for the PCP2X model at high S_0, as we just noted. Thus, once again, the advantage of the protocell over the spatial model is increased when we consider faster replicating parasites.

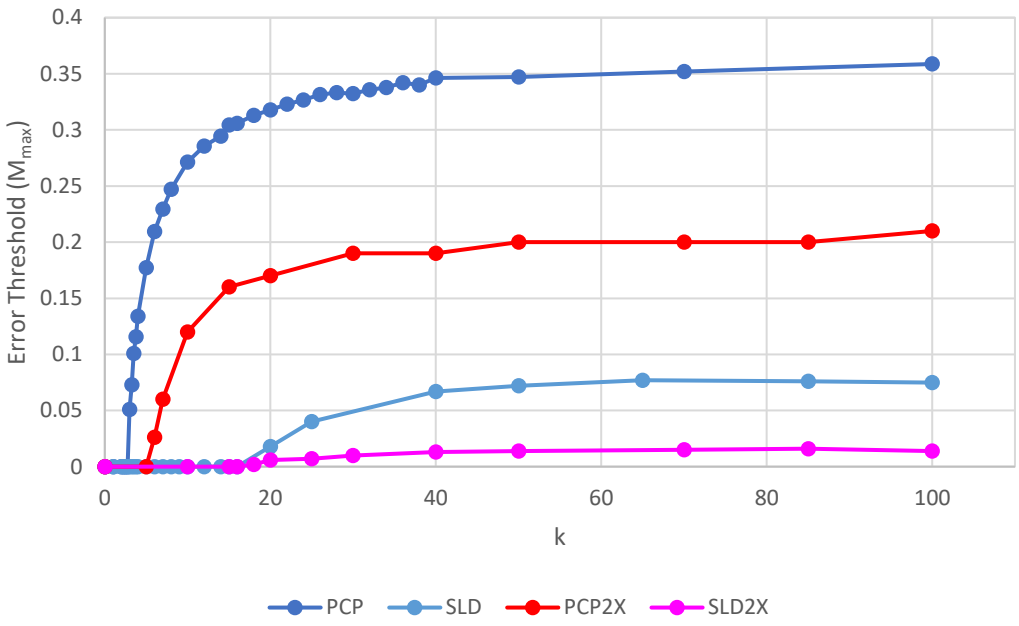


Figure 2.7 Error thresholds versus k *for PCP and SLD models in which parasites and polymerases have equal replication rates (same as Figure 2.4) compared with equivalent models where parasites have double the replication rate of polymerases (denoted PCP2X and SLD2X). Both error thresholds are reduced when the parasites multiply faster, but the SLD model is reduced more, meaning that the relative advantage of the protocells over the spatial model is increased.* $S_0 = 10$ *and* $h = 0.4$.

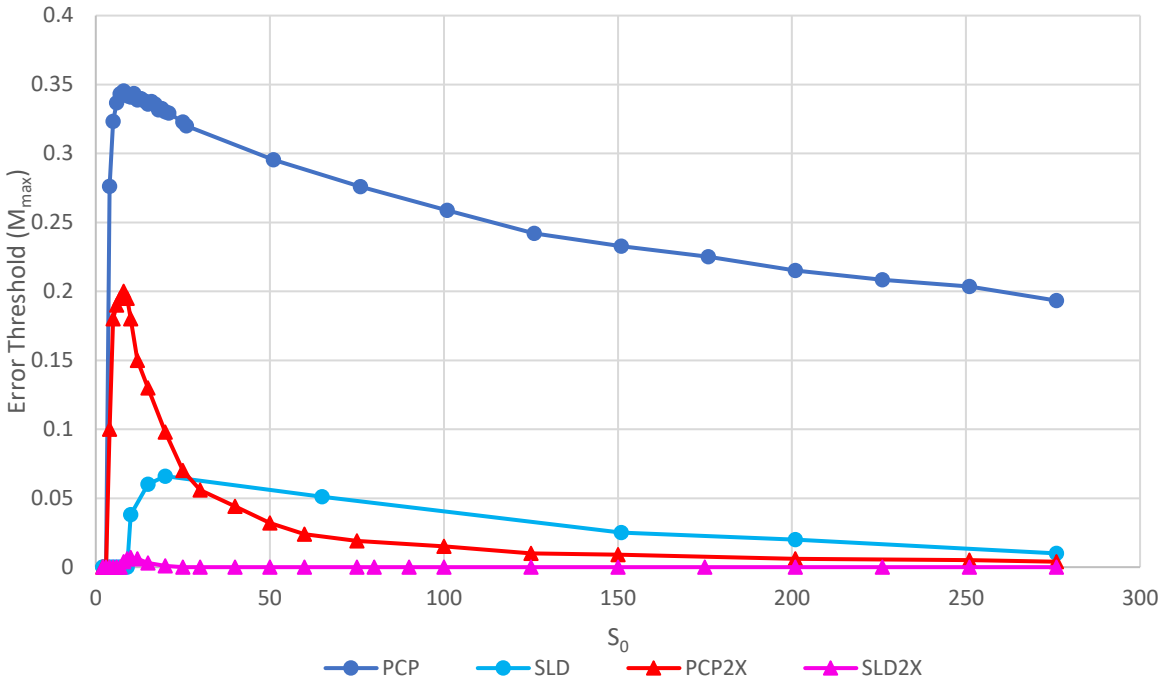


Figure 2.8 Error thresholds versus S_0 *for PCP and SLD models in which parasites and polymerases have equal replication rates (same as Figure 2.5) compared with equivalent models where parasites have double the replication rate of polymerases (denoted PCP2X and SLD2X). Both error thresholds are reduced when the parasites multiply faster, but the SLD model is reduced more, meaning that the relative advantage of the protocells over the spatial model is increased. Note that* $M_{max} = 0$ *for* $S_0 > 20$ *for SLD2X, because faster parasites kill the polymerases in the spatial model.* $k = 25$ *and* $h = 0.4$.

4. Conclusions

Although the mechanisms by which compartments and spatial clustering promote the survival of polymerases (or other kinds of cooperative replicators) have been understood for some time, there has not been much quantitative comparison of the two. It becomes apparent that it is necessary to clearly specify which factors limit the growth of strands when designing models to allow for this quantitative comparison. The limiting resource could simply be space, as is likely to be the case if there is a maximum number of strands

that can fit in the space represented by one lattice site. Alternatively, the limit may be the availability of monomers to synthesize the strands, or it may be the availability of another molecular resource, such as lipids, which limits the growth of protocells before the supply of monomers runs out.

It is also important to consider whether the limiting factor acts globally on the whole population, or locally on one protocell/site at a time. In the spatial models, it seems natural to apply the limitation locally, as in the SLD and SMF models, however we also considered the SML case where a global monomer limitation was applied. The survival of polymerases was somewhat easier when the limit was global, but this model does not seem realistic. The spatial models are intended to represent restricted geometries where diffusion will be slow, such as crevices in rocks. A lattice site would represent a region in which strands can interact with one another. Space is obviously limited in such environments. The diffusion of strands has to be slow in spatial models, otherwise the system becomes well-mixed and polymerases do not survive. Although the diffusion of monomers might be faster than that of strands, it is still finite. Thus the monomer limitation in spatial models has to be local. The locally-limited spatial models have a much lower error threshold and a much higher minimum catalytic rate that the protocell models. Hence, our principle conclusion that polymerase survival is much easier in protocells than in spatial models with restricted geometry, such as pores in a rock.

In protocell models, it seems natural to assume that the limiting factor is global, as will be the case if the protocells are free to move in the surrounding medium, and if supply of either lipids or monomers to the protocells is rapid and well-mixed. There is no restrictive

geometry or surface binding to slow things down in this case. A somewhat surprising finding of this paper is that the PML model, where strand growth is limited by monomers, and the PCP model, where protocell growth is limited by factors other than monomer supply, give such similar results in terms of the error threshold. Hence, our conclusion that the advantages of protocells found in this paper are robust to model variations.

Although both protocell models and spatial models have been widely studied, there has been little previous quantitative comparison. This paper enables us to make that comparison in a novel way and to distinguish carefully the different factors that limit replication. We will emphasize several detailed aspects of the models that emerge from this comparison. There are two different kinds of transitions that occur at the error threshold: either the strand concentrations go continuously to zero (as in Figure 2.2), or there is a discontinuous jump (as in Figure 2.3). Figure 2.3b also shows the unexpected result that parasites can coexist with polymerases in the limit of zero mutation rate where they are no longer being created. This occurs in the spatial models with large enough S_0 but not in protocell models. In the case where parasites multiply faster than polymerases, the advantage of the protocells over the spatial models is increased (as in Figures 2.7 and 2.8). In Figure 2.8, there is a qualitative difference between the protocells (PCP2X) and spatial model (SLD2X). For the protocells, there is a finite error threshold, even for the largest S_0 considered, whereas parasites destroy the spatial system even in the limit of zero mutation rate.

The results in this paper are somewhat different to those obtained from the Cellular Potts Model (CPM) [14] because they consider different cases. The CPM begins with a

spatial model of only one strand per site, and adds cellular compartments on top of this. It is found [14] that these surface models with and without compartments do not qualitatively differ from one another in respect to their stability against mutations. This is best thought of as representing a case where strands are stuck to surfaces. In contrast, we have referred to our lattice models as 'spatial' not 'surface', because they model a case where spatial clustering occurs due to slow diffusion in restricted geometries. In our case, mixing is fast locally, but slow on a global scale. Fast mixing on the local scale is equivalent to fast mixing inside a single protocell. This makes the protocell and spatial models equivalent on the local scale and allows direct comparison. Our models clearly separate the effects of diffusion between sites from the effects of group selection and cell division, and show that group selection is noticeably better as a means of limiting the growth of parasites.

The diversity and success of cellular life on Earth is evident, and there is no evidence for distributed living systems on surfaces. This paper goes some way to showing why this is. It also fits with our previous study [17] of interacting polymerase and nucleotide synthetase ribozymes, where we pointed out that survival of two complementary types of unlinked ribozymes is possible on a surface, but is difficult because it requires joint spatial patterns to form. There is no equivalent problem if the ribozymes are in compartments. Hence we expect evolution from single replicators to genetic systems involving multiple types of ribozymes to be easier in protocells.

We are very pleased to have the opportunity to contribute to this volume dedicated to Prof. David Deamer. The Origins Institute at McMaster has benefitted greatly from David′s help and advice over many years, and we are very grateful for his scientific input,

friendship, and enthusiasm. David has made many important contributions to the understanding of the origins of life. In particular, he has long been an advocate of the importance of membranes to early life [37–39]. His recent work has shown that lipid membranes can provide an environment in which RNA polymerization becomes possible [40]. The stability of membranes in fresh water conditions and the possibility of wetting and drying cycles occurring in shallow water has led to his current view that life began in freshwater pools associated with volcanic islands [41,42]. Under these conditions, it is likely that membranes were available to encapsulate the earliest replicating polymers at the time of the origin of life. This ties in with the conclusions of the current paper, in which we have shown that encapsulation greatly increases the ability of RNA replicators to survive when the replication accuracy is low. Thus, we conclude that the presence of protocells is beneficial to the development of early life, and it seems quite likely that the first replicating polymers may have functioned inside lipid vesicles from the outset.

Author Contributions: conceptualization, P.G.H., V.S. and A.S.T.; methodology, P.G.H., V.S. and A.S.T.; software, V.S., J.B., and Q.P.; investigation, V.S., J.B., and Q.P.; funding acquisition, P.G.H.; writing-original draft, V.S. and P.G.H.; writing-review and editing, V.S., A.S.T. and P.G.H.

Funding: This research was funded by the Natural Sciences and Engineering Research Council of Canada, Discover Grant number 2017-05911.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Appendix A

In the limit where the number of lattice sites or protocells becomes infinite, it is possible to calculate probability distribution $P(p, c, x)$ that a site/protocell will have p, c

and x molecules of types P, C and X. For the protocell model with constant population, we may write the rate of change of the probabilities due to birth and death of strands as

$$\frac{dP(p,c,x)}{dt} = K_P(p-1,c,x)P(p-1,c,x) + K_C(p,c-1,x)P(p,c-1,x)$$
$$+ K_X(p,c,x-1)P(p,c,x-1)$$
$$-\big(K_P(p,c,x) + K_C(p,c,x) + K_X(p,c,x)\big)P(p,c,x)$$
$$+v(p+1)P(p+1,c,x) + v(p+1)P(p,c+1,x) + v(x+1)P(p,c,x+1) -$$
$$v(p+c+x)P(p,c,x).$$

The rate of production of cells of type (p,c,x) by random division of cells of type (p_0,c_0,x_0) is

$$b(p,c,x|p_0,c_0,x_0) = 2\frac{p_0!}{2^{p_0}p!(p_0-p)!}\frac{c_0!}{2^{c_0}c!(c_0-c)!}\frac{x_0!}{2^{x_0}x!(x_0-x)!}.$$

All cells with at least S_0 strands divide with probability 1 in one time step. The total rate of production of cells of type (p,c,x) by division of all cells with at least S_0 strands is

$$B(p,c,x) =$$
$$\sum_{p_0\geq p;c_0\geq c;x_0\geq x;p_0+c_0+x_0\geq S_0} b(p,c,x|p_0,c_0,x_0)P(p_0,c_0,x_0).$$

As the population is fixed, the total rate of removal of cells when other cells divide is equal to the total division rate:

$$B_{tot} = \sum_{p_0+c_0+x_0\geq S_0} P(p_0,c_0,x_0)$$

The change in probability in one time step due to cell division is therefore $\Delta P(p,c,x) = B(p,c,x) - B_{tot}P(p,c,x)$, for cells with $p+c+x < S_0$, and $\Delta P(p,c,x) = B(p,c,x) - B_{tot}P(p,c,x) - P(p,c,x)$, for cells with $p+c+x \geq S_0$. Thus the probability

of each cell type after one time step, accounting for both strand birth and death and cell division is

$$P(p,c,x)|_{t+\delta t} = P(p,c,x)|_t + \delta t \frac{dP(p,c,x)}{dt} + \Delta P(p,c,x).$$

To solve the model by this deterministic method, we iterate forward in time till the stationary state is reached.

The spatial lattice model can also be solved deterministically if we make a mean field assumption, i.e. there is a single central site surrounded by a homogenous environment. The terms for birth and death of strands are the same as for the protocell case. Terms for diffusion into and out of the central site are added, as follows.

$$\frac{dP(p,c,x)}{dt} = birth\ and\ death\ terms$$
$$+h(p+1)P(p+1,c,x) + h(p+1)P(p,c+1,x) + h(x+1)P(p,c,x+1)$$
$$- h(p+c+x)P(p,c,x)$$
$$+h\overline{p}P(p-1,c,x) + h\overline{c}P(p,c-1,x) + h\overline{x}P(p,c,x-1) - h(\overline{p}+\overline{c}+\overline{x})P(p,c,x)$$

Here, $\overline{p}$, $\overline{c}$ and $\overline{x}$ are the current mean values of the numbers of strands per site across the whole lattice: $\overline{p} = \sum_{p,c,x} pP(p,c,x)$, $\overline{c} = \sum_{p,c,x} cP(p,c,x)$,and $\overline{x} = \sum_{p,c,x} xP(p,c,x)$. The mean field method has also been used in similar models by McCaskill et *al.* [43] and in our previous work [16, 27].

References

1. Pace, N.R.; Marsh, T.L. RNA catalysis and the origin of life. *Orig Life Evol Biosph* **1985**, 16, 97-116.

2. Gilbert, W. Origin of life: The RNA world. *Nat* **1986**, 319, DOI:10.1038/319618a0.

3. Bartel, D.P.; Unrau, P.J. Constructing an RNA wold. *Trends Cell Biol* **1999**, 9, M9-M13.

4. Joyce, G.F. The antiquity of RNA-based evolution. *Nat* **2002**, 418, 214-221, DOI:10.1038/418214a.

5. Robertson, D.L.; Joyce, G.F. The origins of the RNA world. *Cold Spring Harb Perspect Biol* **2012**, 1, DOI:10.1101/cshperspect.a003608.

6. Higgs, P.G.; Lehman, N. The RNA world: molecular cooperation at the origins of life. *Nat Rev Genet* **2015**, 16, 7-17, DOI:10.1038/nrg3841.

7. Pressman, A.; Blanco, C.; Chen, I.A. The RNA world as a model system to study the origin of life. *Curr Biol* **2015**, 25, R953-R963, DOI:10.1016/j.cub.2015.06.016.

8. Johnston, W.K.; Unrau, P.J.; Lawrence, M.S.; Glasner, M.E; Bartel, D.P. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science* **2001**, 292, 1319-1325, DOI:10.1126/science.1060786.

9. Wochner, A.; Attwater, J.; Coulson, A.; Holliger, P. Ribozyme-catalyzed transcription of an active ribozyme. *Science* **2011**, 332, 209-212, DOI:10.1126/science.1200752.

10. Attwater, J.; Wochner, A.; Holliger, P. In-ice evolution of RNA polymerase ribozyme activity. *Nat Chem* **2013**, 5, 1011-1018, DOI:10.1038/nchem.1781.

11. Szabó, P.; Scheuring, I.; Czárán, T.; Szathmary, E. In silico simulations reveal that replicators with limited dispersal evolve towards higher efficiency and fidelity. *Nat* **2002**, 420, 340-343, DOI:10.1038/nature01187.

12. Konnyu, B.; Czaran, T.; Szathmary, E. Prebiotic replicase evolution in a surface-bound metabolic system: parasites as a source of adaptive evolution. *BMC Evol Biol* **2008**, DOI:10.1186/1471-2148-8-267.

13. Branciamore, S.; Gallori, E.; Szathmary, E.; Czaran, T. The origin of life: chemical evolution of a metabolic system in a mineral honeycomb? J Mol Evol **2009**, 69, 458-469, DOI:10.1007/s00239-009-9278-6.

14. Takeuchi, N.; Hogeweg, P. Multilevel selection in models of prebiotic evolution ii: a direct comparison of compartmentalization and spatial self-organization. *PLoS Comput Biol* **2009**, 5, e1000542, DOI:10.1371/journal.pcbi.1000542.

15. Takeuchi, N.; Hogeweg, P. Evolutionary dynamics of RNA-like replicator systems: a bioinformatic approach to the origin of life. *Phys Life Rev* **2012**, 9, 219-263, DOI:10.1016/j.plrev.2012.06.001.

16. Shay, J.A.; Huynh, C.; Higgs, P.G. The origin and spread of a cooperative replicase in a prebiotic chemical system. *J Theor Bio* **2015**, 364, 249-259, DOI:10.1016/j.jtbi.2014.09.019.

17. Kim, Y.E.; Higgs, P.G. Co-operation between polymerases and nucleotide synthetases in the RNA world. *PLoS Comput Biol* **2016**, 12, e1005161, DOI:10.1371/journal.pcbi.1005161.

18. Tupper, A.S.; Higgs, P.G. Error thresholds for RNA replication in the presence of both point mutations and premature termination errors. *J Theor Biol* **2017**, 428, 34-42, DOI:10.1016/j.jtbi.2017.05.037.

19. Zintzaras, E.; Santos, M.; Szathmary, E. Selfishness versus functional cooperation in a stochastic protocell model. *J Theol Biol* **2010**, 267, 605-613, DOI:10.1016/jtbi.2010.09.011.

20. Ma, W.; Yu, C.; Zhang, W.; Zhou, P.; Hu, J. The emergence of ribozymes synthesizing membrane components in RNA-based protocells. *Biosystems* **2010**, 99, 201-209, DOI:10.1016/j.biosystems.2009.11.003.

21. Bianconi, G.; Zhao, K.; Chen, I.A.; Nowak, M.A. Selection for replicases in protocells. *PLoS Comput Biol* **2013**, 9, e1003051, DOI:10.1371/journal.pcbi.1003051.

22. Chen, I.A.; Roberts, R.W.; Szostak, J.W. The emergence of competition between model protocells. *Science* **2004**, 305, 1474-1476, DOI:10.1126/science.1100757.

23. Chen, I.A.; Salehi-Ashtiani, K.; Szostak, J.W. RNA catalysis in model protocell vesicles. *J Am Chem Soc* **2005**, 127, 13213-13219, DOI:10.1021/ja051784p.

24. Zhu, T.F.; Szostak, J.W. Coupled growth and division of model protocell membranes. *J Am Chem Soc* **2009**, 131, 5705-5713, DOI:10.1021/ja900919c.

25. Matsumura, S.; Kun, Á.; Ryckelynck, M.; Coldren, F.; Szilágyi, A.; Jossinet, F.; Rick, C.; Nghe, P.; Szathmary, E.; Griffiths, A.D. Transient compartmentalization of RNA replicators prevents extinction due to parasites. *Science* **2016**, 354, 1293-1296, DOI:10.1126/science.aag1582.

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

27. Wu, M.; Higgs, P.G. The origin of life is a spatially localized stochastic transition. Biol Direct **2012**, 7, 42, DOI: 10.1186/1745-6150-7-42.

28. Higgs, P.G. Three ways to make an RNA sequence: Steps from Chemistry to the RNA World. In *Handbook of Astrobiology*, **2019**, Chap 6.2. Ed. Kolb, V.M. CRC Press.

29. Higgs, P.G. Chemical Evolution and the Evolutionary Definition of Life. *J. Mol. Evol.* **2017**, 84, 225-235.

30. Ferris, J.P. Montmorillonite-catalysed formation of RNA oligomers: the possible role of catalysis in the origins of life. *Philos Trans R Soc Lond B Biol Sci* **2006,** 361:1777–1786.

31. Burcar, B.T.; Barge, L.M.; Trail, D.; Watson, E.B.; Russell, M.J.; McGown, L.B. RNA oligomerization in laboratory analogues of alkaline hydrothermal vent systems. *Astrobiology*, **2015**, 15, 509-522.

32. Biondi, E.; Branciamore, S; Fusi, L.; Gago, S.; Gallori, E. Catalytic activity of hammerhead ribozymes in a clay mineral environment: Implications for the RNA World. *Gene* **2007,** 389, 10-18.

33. Stephenson, J.D.; Popovic, M.; Bristow, T.F.; Ditzler, M.A. Evolution of ribozymes in the presence of a mineral surface. *RNA* **2016,** 22, 1893-1901.

34. Koonin, E.V.; Martin, W. On the origin of genomes and cells within inorganic compartments. *Trends in Genetics* **2005**, 21, 647-654.

35. Agerschou, E.D.; Mast, C.B.; Braun, D. Emergence of life from trapped nucleotides? Non-equilibrium behavior of oligonucleotides in thermal gradients. *Synlett* **2107**, 28, 56-63.

36. Moran, P.A.P. Random processes in genetics. *Math Proc Cambridge Phil Soc* **1958**, 54, 60-71.

37. Deamer, D.W. Boundary structures are formed by organic components of the Murchison carbonaceous chondrite. *Nat* **1985**, 317, 792-794, DOI:10.1038/317792a0.

38. Dworkin, J.P.; Deamer, D.W; Sandford, S.A.; Allamandola, L.J. Self-assembling amphiphilic molecules: synthesis in simulated interstellar/precometary ices. *Proc Natl Acad Sci USA* **2001**, 98, 815-819, DOI:10.1073/pnas.98.3.815.

39. Segré, D.; Ben-Eli, D.; Deamer, D.W.; Lancet, D. The lipid world. Origins Life Evol. Biosph. **2001,** 31, 119-145.

40. Rajamani, S.; Vlassov, A.; Benner, S.; Coombs, A.; Olasagasti, F.; Deamer, D. Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Orig Life Evol Biosph* **2008**, 38, 57-74.

41. Damer, B.; Deamer, D. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: a scenario to guide experimental approaches to the origin of cellular life. *Life* **2015**, 5, 872-887.

42. Deamer, D. The role of lipid membranes in life's origin. *Life* **2017**, 7, 5.

43. McCaskill, J.; Fuchslin, R.M.; Altmeyer, S. The stochastic evolution of catalysts in spatially resolved molecular systems. Biol Chem **2001**, 382, 1343-1363, DOI:10.1515/BC.2001.167.

Chapter 3: Additional content for Shah et al. (2019)

3.1 Summary of Additional Work

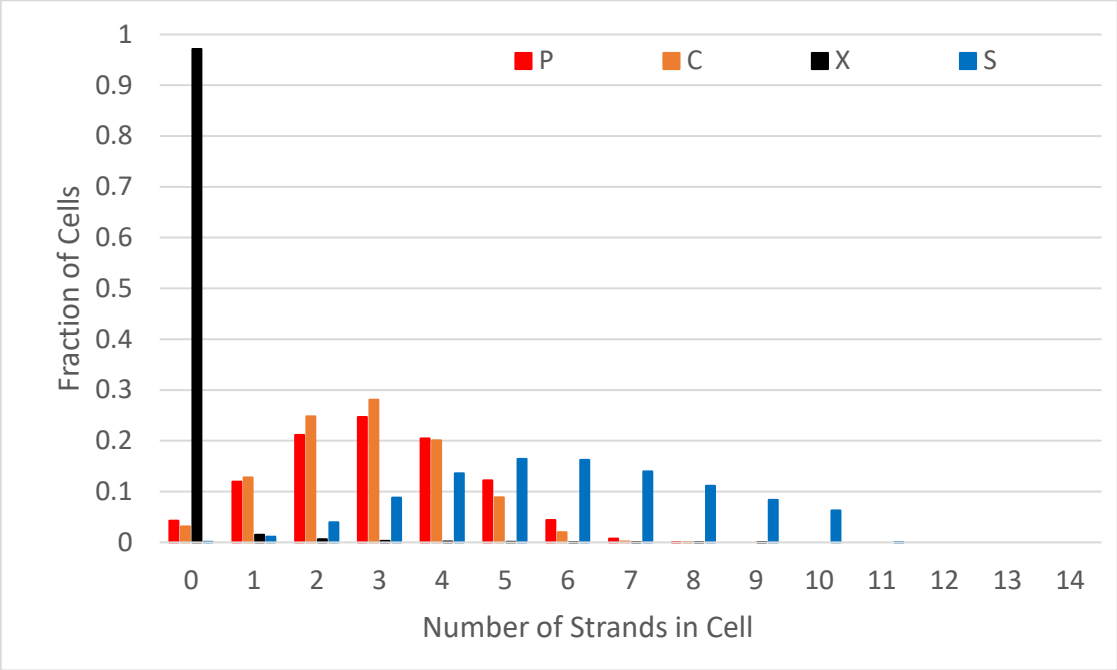
This section contains material that was not included in the version of the paper published in *Life*. The results presented compare how group selection and spatial clustering control the amount of parasitic strands present in a protocell and spatial model, respectively.

3.2 Additional Results

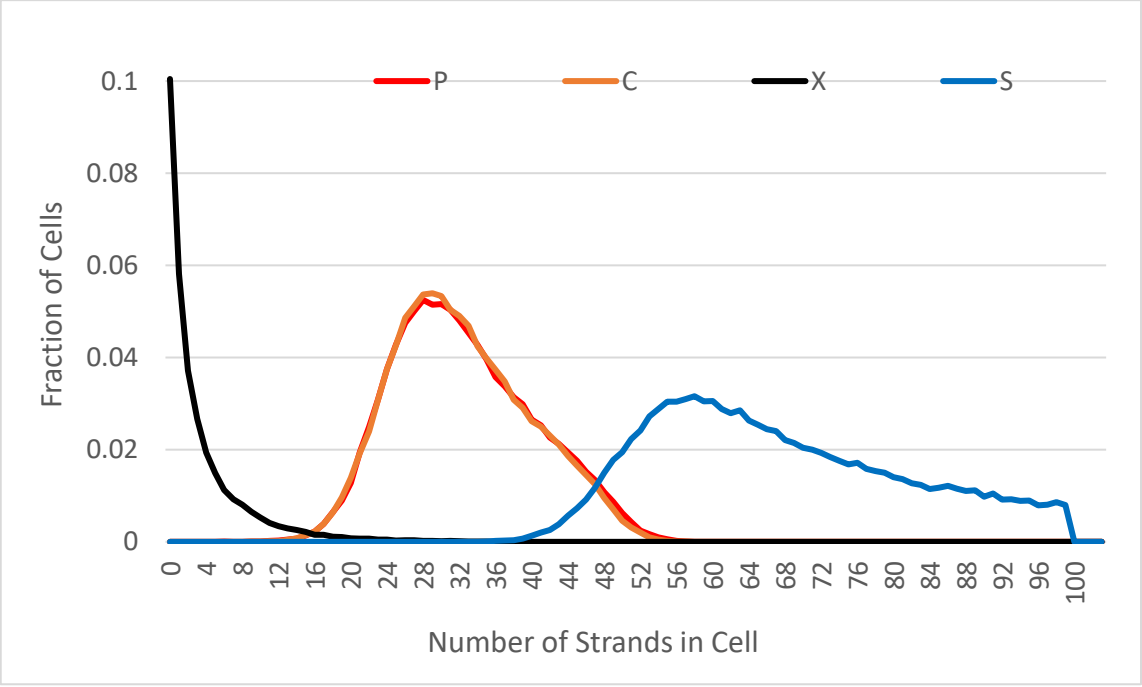
Figures 3.1 and 3.2 show the distribution of strands in a run of the PCP model SLD model respectively, along with the size distribution (blue bars/lines). The strand distribution is the fraction of the total population of cells/sites that contain a specific amount of strands of a given type, for example figure 3.2a shows just under 60% of all sites contain no parasites, just under 20% of all sites contain only 1 parasite, about 10% of all sites contain only 2 parasites, and so on. Similarly, the size distribution is the fraction of the total population of cells/sites containing a given amount of strands in total, again using figure 3.2a as an example shows 15% of all sites are empty and about 10% of all sites have 8 strands in them. The plots are done with different M values as our goal is to determine what happens to the distributions as the error threshold is approached. Comparing figures 3.1a and 3.2a shows the difference group selection has compared to spatial clustering: almost no protocells have any parasites remaining in them, whereas almost half of all lattice sites have some number of parasites. We can also see that for the protocells, there is a broad distribution in terms of sizes, with a peak at about half the split size, whereas for the spatial

model the size distribution is smeared and very slightly bimodal, with a large amount of sites which are close to empty and a large amount of sites which are close to being full.

Comparing 3.1b and 3.2b further shows that group selection prevents the accumulation of large numbers of parasites better than spatial clustering can. By allowing more strands per site/cell, we see that the behavior in the lattice model changes significantly. When only a small number of strands, say 10, are allowed per site, we see that the number of parasitic strands per site appears to decrease exponentially, and the number of polymerases and compliments tend to outnumber them. Increasing the number allowed per site to 100, we see that there is a much broader distribution of parasites, and that while their number still drops off, it occurs once they begin to outnumber the polymerases and compliments. Essentially, the majority of sites have a few polymerases and compliments – enough for them to continue to survive, but the majority of polymerase work is done to increase the number of parasites. Conversely, in the protocell model the number of parasites is always kept low by the effects of group selection favouring cells with more polymerases and compliments (as those with more P and C grow faster), regardless of how many strands are allowed per cell. This allows the majority of polymerase work to be focused on generating more copies compliments and themselves.

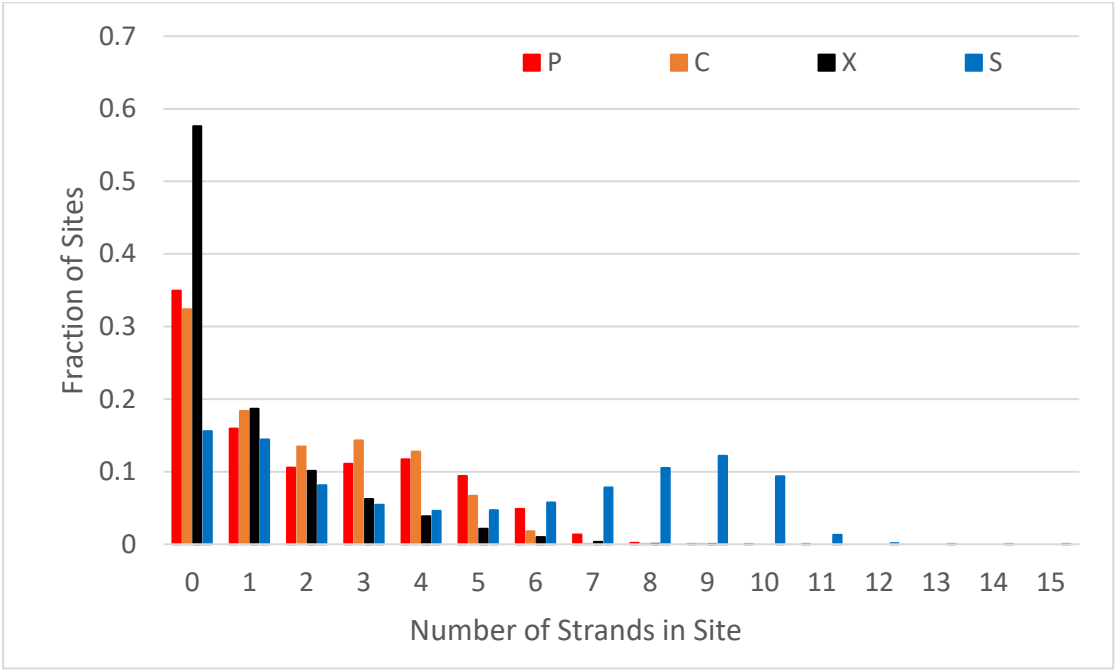


(**a**)

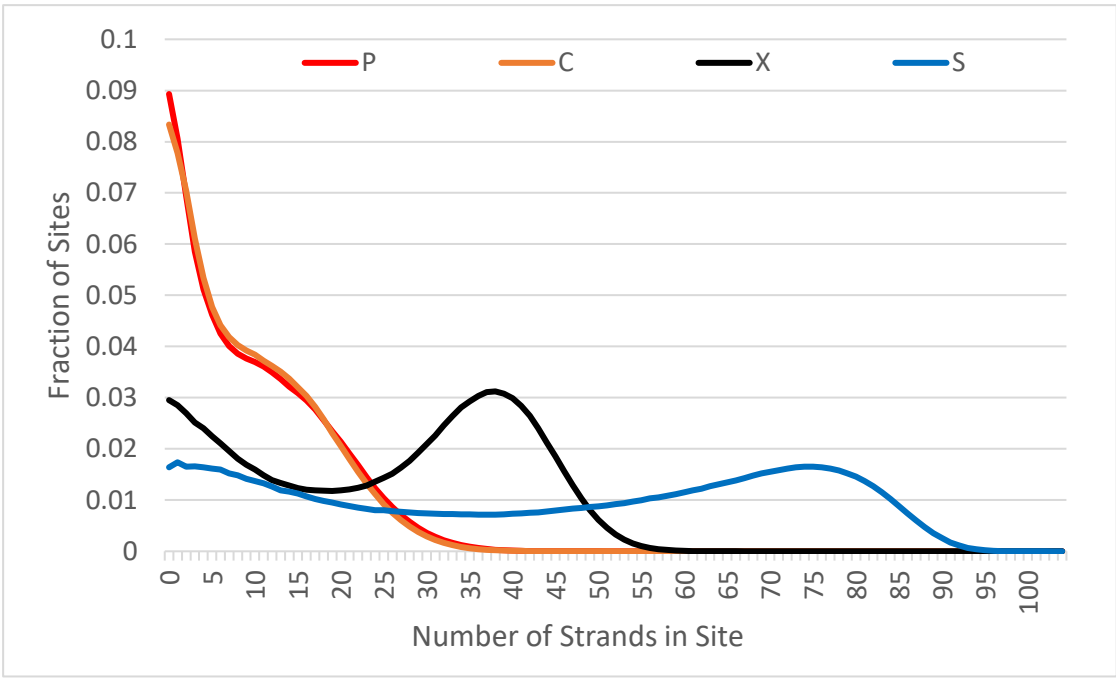


(**b**)

Figure 3.1 Strand distributions for ***(a)*** *protocell model with constant volume at a split size of 10;* ***(b)*** *protocell model with constant volume at a split size of 100*



(**a**)



(**b**)

*Figure 3.2 Strand distributions for **(a)** spatial model – local diffusion with a maximum size of 10; **(b)** spatial model – local diffusion with a maximum size of 100*

To visualize the effect of increasing mutation rate in spatial models, simulations were conducted using the SLD model for low (M = 0.01) and high (M = 0.06) mutation rates. These runs were performed with a lattice 100 × 100 sites in size, using the parameters S_0 = 15, k = 25 and h = 0.4. The sites depicted in Figures 3.3, 3.4 and 3.5 are coloured as follows: empty sites are coloured white, while non-empty sites are coloured according to

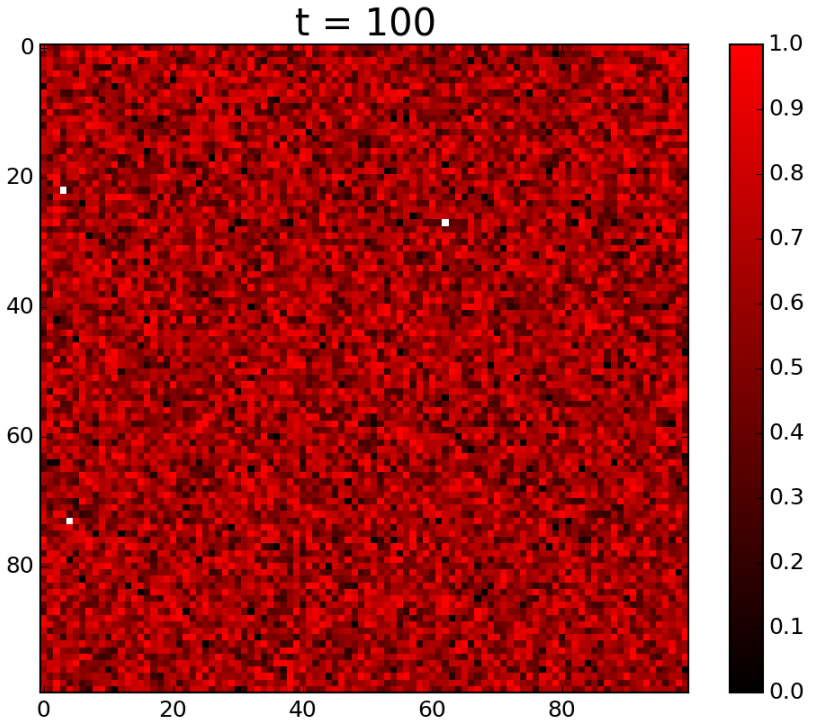
$$C(p,c,x) = \frac{p+c}{p+c+x}$$

The range of this function is mapped to a linear colour gradient spanning between black ($C(p,c,x) = 0$) and red ($C(p,c,x) = 1$). Thus, sites dominated by parasites are coloured darker and sites with relatively few parasites are coloured brighter. A summary of parameters and results is presented in Table 3.1.

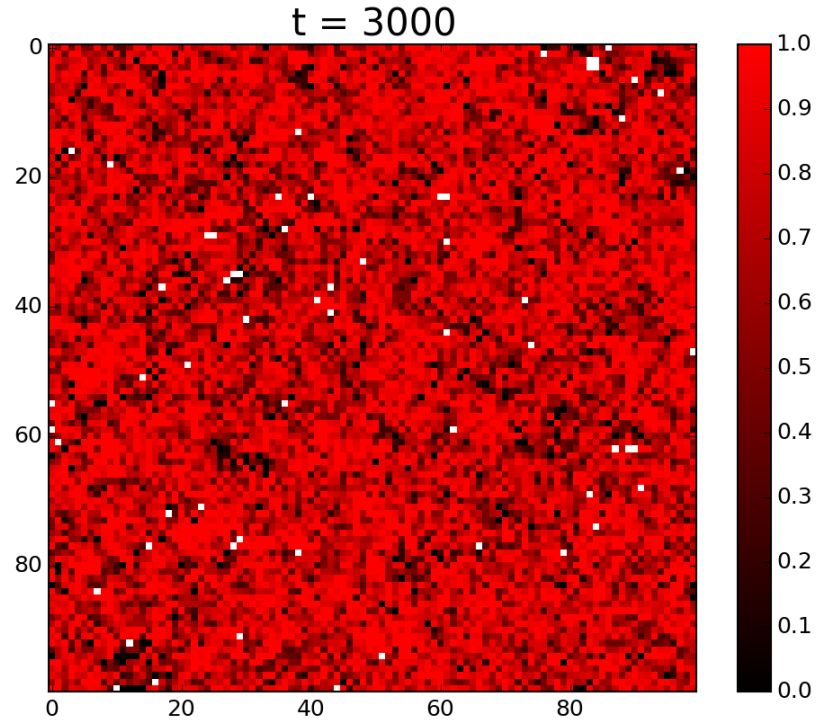
Table 3. 1 Mutation in Different Spatial Models

Figure	3.3	3.4	3.5
Model	SLD	SLD	SMF
M	0.01	0.06	0.06
S_0	15	15	15
k	25	25	25
h	0.4	0.4	0.4
Result	Full	Empty regions and clusters of living sites	Full

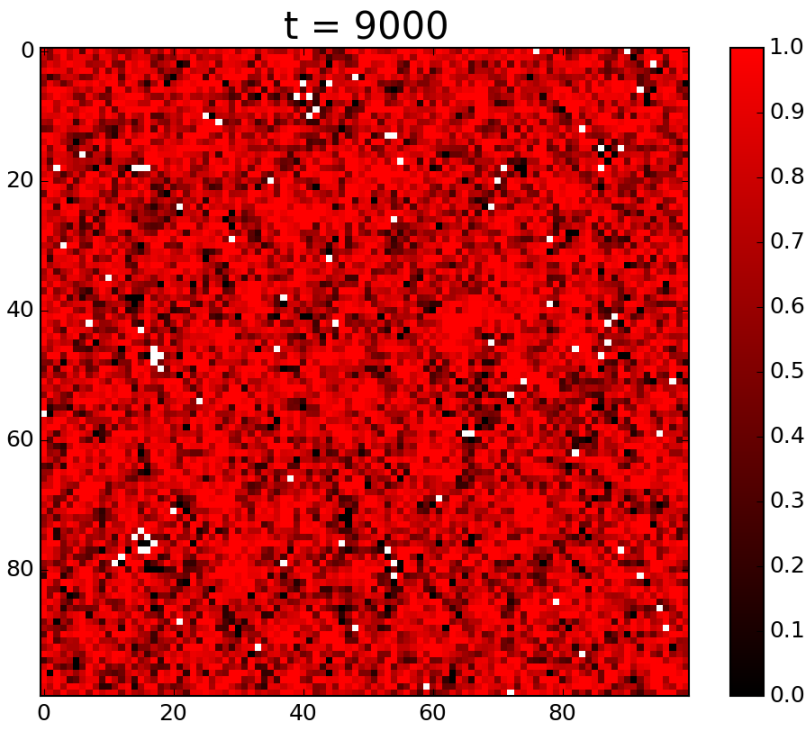
Figures 3.3 and 3.4 show three snapshots from each simulation, taken at t = 100 shortly after initialization, at t = 3000 partway through the simulation and at t = 9000 once the simulation had reached equilibrium. Comparing figure 3.3a and figure 3.4a shows little difference shortly after initialization, because only a relatively small number of replication events have occurred, limiting the effects of a higher mutation rate. Once more time has elapsed, the effects of higher mutation rates becomes obvious with many figures 3.4b and 3.4c showing an increasing number of empty lattice sites, while figures 3.3b and 3.3c show few in comparison. The effect of clustering is also readily visible in figures 3.4b and 3.4c, with populations of sites grouped together against the empty surroundings. Closer inspection also reveals sites with high relative parasite counts are largely found on the peripheries of the grouped sites, again showing spatial clustering in effect.


t = 100
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(a)

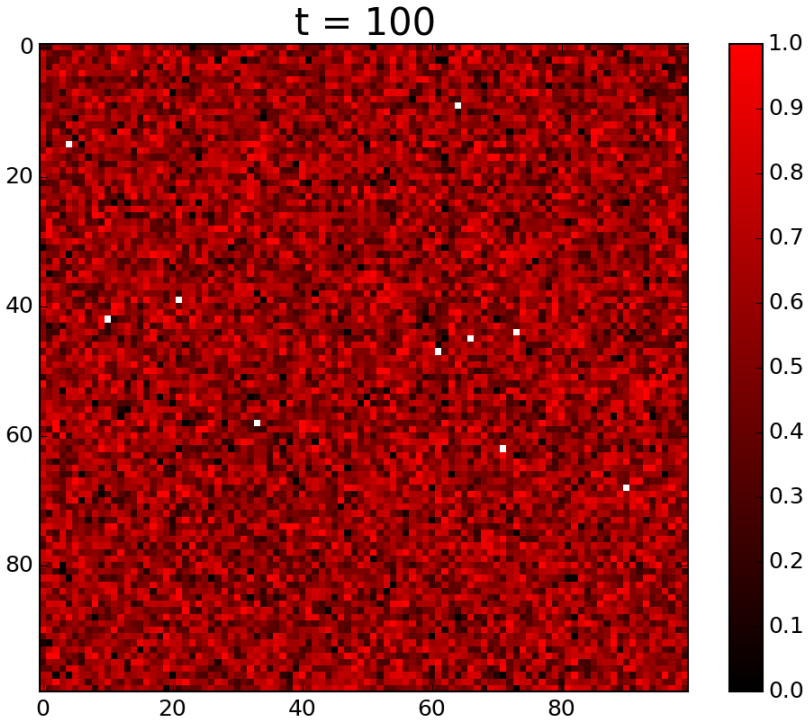

t = 3000
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(b)

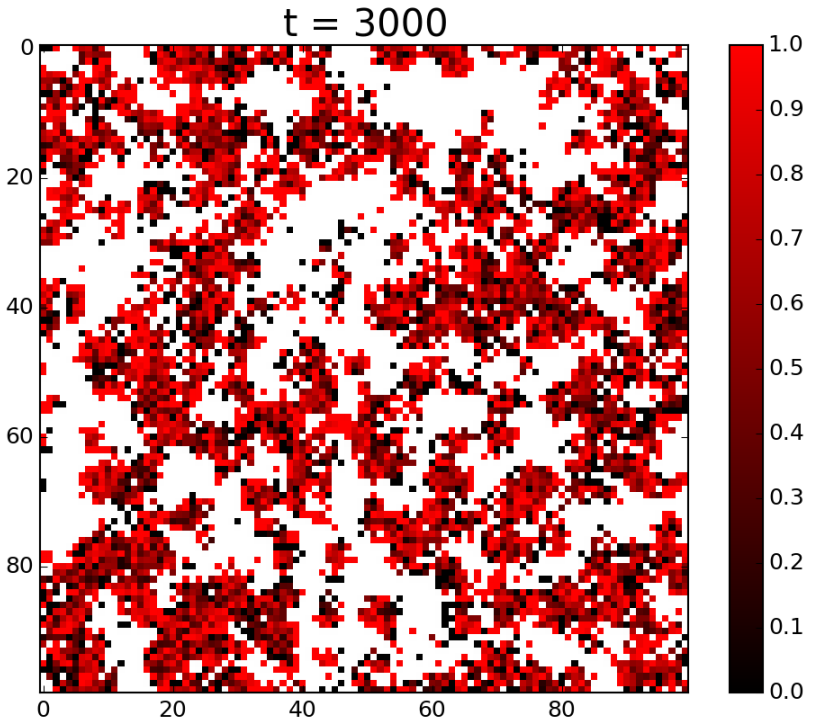


(c)

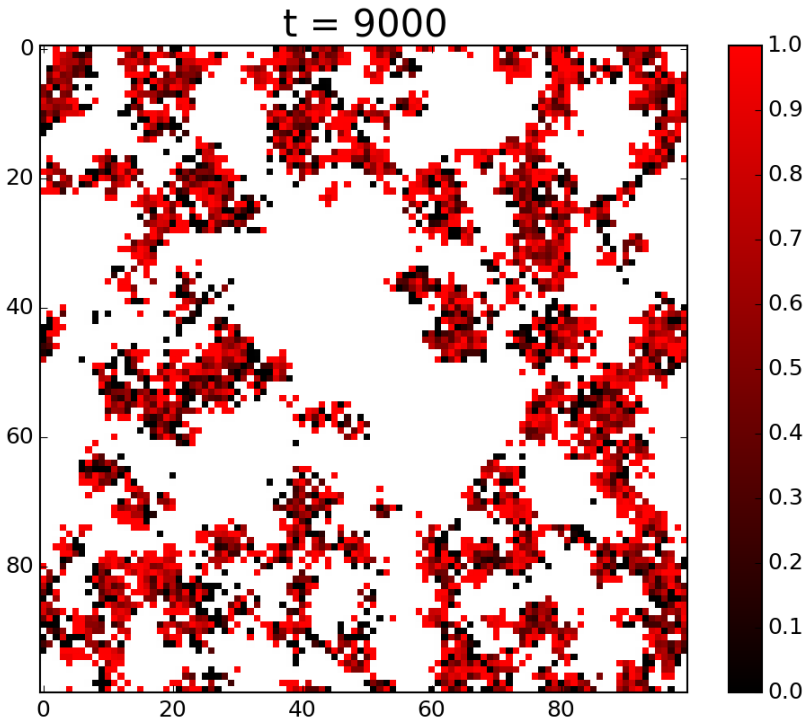
Figure 3.3 Snapshots from a SLD simulation performed with M=0.01, S_0 = 15, k = 25 and h=0.4. At such low mutation rates, effects of clustering are difficult to determine.


t = 100
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(a)


t = 3000
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

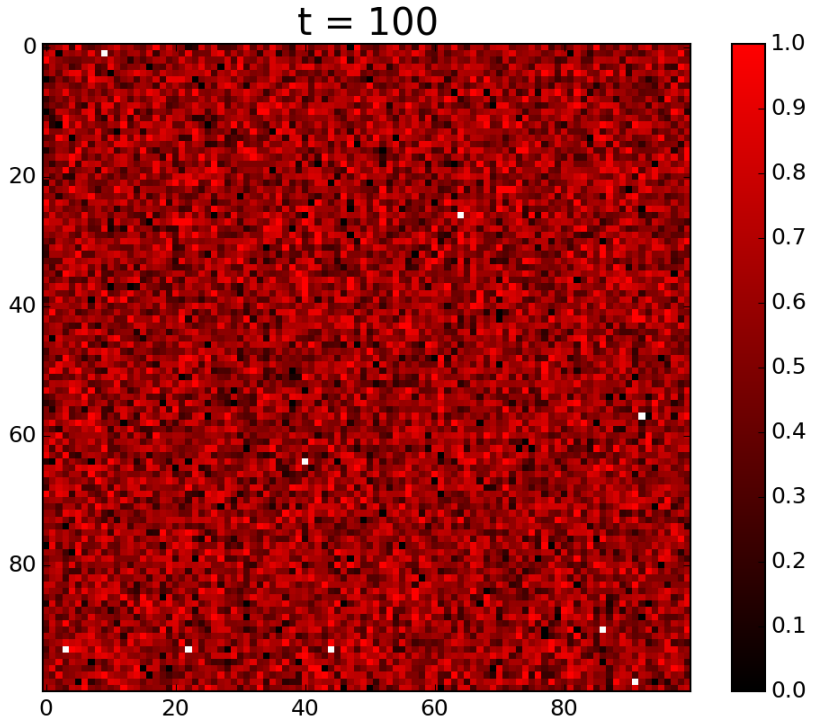
(b)



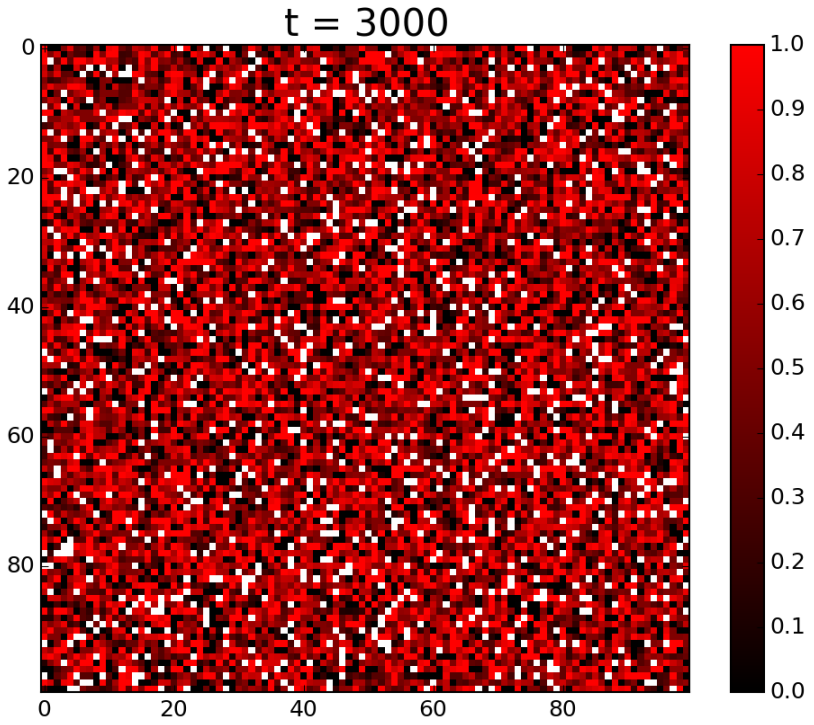
(c)

Figure 3.4 Snapshots from a SLD simulation performed with M=0.06, S_0 = 15, k = 25 and h=0.4. ***b)*** *and* ***c)*** *clearly show the effects of clustering allowing groups of polymerases to survive.*

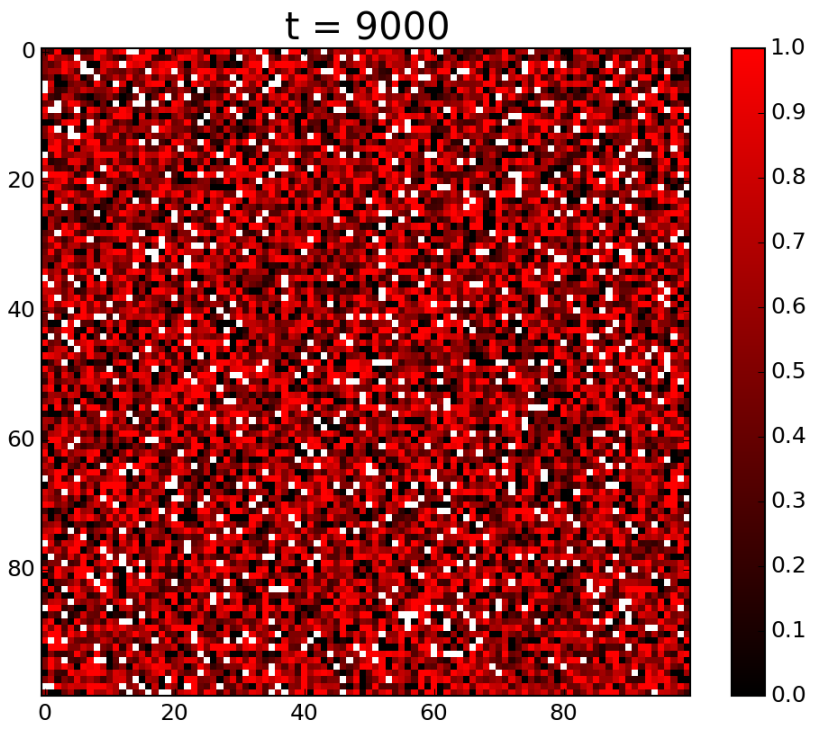
The SMF model features no correlation between sites, here diffusion permits strands to move from any one site on the lattice to another, as if they were all connected. Figure 3.5 shows snapshots taken at the same times for a run of the SMF model with identical parameters to those used to make figure 3.4; the colouring is also done in the same manner. The lack of correlation produces a much more random pattern, resembling static.


t = 100
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(a)


t = 3000
0
20
40
60
80
0
20
40
60
80
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0

(b)



(c)

*Figure 3.5 Snapshots from a SMF simulation performed with M=0.06, S_0 = 15, k = 25 and h=0.4.**b)** and **c)** show a lack of the large, empty regions seen in figures 3.4b and 3.4c*

Chapter 4: Linkage of strands into a genome

4.1 Prior Work

Even the simplest of biological entities, viruses, have genomes comprising of several individual genes linked together in a continuous strand. Evidently, at some point in the evolution of life, there was a transition from an assortment of loose functional and complementary strands in a compartment, to a defined genomic structure similar to the chromosomes seen in biology today. In modern life there are many asymmetries between the genome and the enzymes it codes for. The polymers used by the genome and enzymes differ (DNA vs proteins), as does the lifetimes of these polymers and their relative

importance in a cell. Additionally, cells feature at most a few copies of the genome, but many copies for enzymes. Determining how such differences came about is important to understanding the transition from an RNA world to more modern biology. While the origin of chromosomes is not known, the potential benefit of such structures existing in an RNA world scenario has been explored.

Szathmáry and Smith (1993) proposed that originally there were strands functioning as both a catalyst and template alongside strands functioning only as templates. The benefit derived from catalytic capability means evolution would increase the amount of catalysts compared to templates, leading to an asymmetry in the populations of the two strands (Szathmáry and Smith, 1993), in fact RNA viruses have shown to favour catalysts to templates (Winterberger and Winterberger, 1987). This asymmetry is thought to develop into the genome-catalyst split observed today (Takeuchi et al., 2017). The symmetry breaking of the catalyst from the genome has an additional benefit, by reducing the number of copies of the genome, there is less opportunity for variation via mutation to arise, protecting the protocell (Takeuchi et al., 2017).

Smith and Szathmáry (1993) proposed a model to study linkage of genes based on a stochastic corrector protocell model (Szathmáry and Demeter, 1987). They set up a system of protocells which contained a selfish gene, two cooperative genes as well as a linked version of the cooperative genes which was simply the two individual genes joined together (Smith and Szathmáry, 1993). Supposing the linked genes replicated slower due to their length, Smith and Szathmáry (1993) assigned fitness values to the various genes such that at the molecular level the selfish gene was favoured, and also assigned cell-level fitness

values in which cells with both the cooperative genes or linked gene were favoured. They found that cells with linked genes were readily established in the population so long as cells contained few total genes prior to division, and that some fraction of the population had linked genes at initialization (Smith and Szathmáry, 1993).

Linkage of genes and the benefit provided by chromosomes in a protocell model has also been studied by Szilágyi et al. (2012). Their focus was on comparing the survival of non-specialized enzymes and specialized enzymes capable of catalyzing different steps in a linear chain of reactions, with the non-specialized enzymes able to catalyze all steps inefficiently and specialized enzymes only able to catalyze a single step highly efficiently. Additionally, the evolution of generalized ribozymes to specialized ribozymes was permitted. They modeled the reaction chain to convert substrate to biomass accumulation (a proxy for fitness) which in turn drove cell division (Szilágyi et al., 2012). Models differed in the ability of the ribozymes to replicate only individually, replicate in both individual or linked states or only when linked in a chromosome. Szilágyi et al. (2012) found that initializing the model with all protocells featuring non-specific ribozymes, without chromosomes specific ribozymes never evolved. When linkage was allowed (but not required for replication), protocells with chromosomes did evolve specialized ribozymes and performed more efficiently, however the independent assortment upon division meant they did not dominate the system (Szilágyi et al., 2012). When chromosomes are required for replication, and they are assigned to daughter cells so that each gets a complete copy, cells evolve towards fully specialized enzymes (Szilágyi et al., 2012). This indicates that

without directed assortment of strands upon division, a system with linked strands may be unstable.

In our research, we seek to determine when linkage is favourable in more detail. Instead of explicitly assigning molecule-level and cell-level fitness values to genes with unspecified function, we will consider a system with multiple functional replicators, namely polymerases, nucleotide synthases and their complementary sequences contained inside protocells. Such a model has been considered previously in a spatial lattice by our group (Kim and Higgs, 2016). We allow for the polymerase and synthase to link together and allow the same behaviour for their complementary strands. We determine the cases in which incorporation of the individual genes and the linked genes is possible in a system of polymerases and polymerase complements. We also determine when the linkage of beneficial genes is favourable compared to the genes remaining unlinked.

4.2 Methods

The results from chapter 2 show the protocell models far outcompete the spatial models, in particular the protocell model with constant population (PCP) performs the best. Consequently, the investigation presented here considers only a modified version of the PCP model. The maximum size of the cells prior to division is now fixed, as is the polymerase rate constant so that $S_0 = 25$ and $k = 25$, corresponding to values that gave the best results in the previous study. The mutation rate has been set to zero and as such parasites are not considered. In addition to the polymerases (P) and complements to the polymerase (C), we now consider nucleotide synthases (Y_+) and their complementary sequences (Y_-) as well as linked versions of the functional strands (L_+) and linked versions

of the complementary strands (L_-). The linked functional strands are designated as doubly functional, capable of serving as a polymerase and synthase simultaneously. The length of the polymerases and their complementary sequences is nominally fixed at 100 base pairs as before, while the length of synthases and their complementary sequences, l_y, is allowed to vary. The function of the synthases is to amplify the availability of nucleotides in their cell, allowing the polymerase to operate at a higher rate. This benefit is hypothesized to be proportional to the number of synthases present, and is parameterized by σ. Following the nomenclature from (Shah et al., 2019, see chapter 2), the number of synthases, their complements, the linked functional strands and the linked complementary strands are given by y_+, y_-, l_+, and l_- respectively. We also adopt $N_{Syn} = y_+ + l_+$ for readability, modifying the rate equations to the following form:

$$K_P(p, c, y_+, y_-, l_+, l_-) = \left(1 + \sigma N_{Syn}\right) \frac{(p + l_+)c}{S} k.$$

$$K_C(p, c, y_+, y_-, l_+, l_-) = \left(1 + \sigma N_{Syn}\right) \frac{(p - 1 + l_+)p}{S} k.$$

$$K_{Y_+}(p, c, y_+, y_-, l_+, l_-) = \frac{l_y}{100}\left(1 + \sigma N_{Syn}\right) \frac{(p + l_+)y_-}{S} k.$$

$$K_{Y_-}(p, c, y_+, y_-, l_+, l_-) = \frac{l_y}{100}\left(1 + \sigma N_{Syn}\right) \frac{(p + l_+)y_+}{S} k.$$

$$K_{L_+}(p, c, y_+, y_-, l_+, l_-) = \frac{100}{100 + l_y}\left(1 + \sigma N_{Syn}\right) \frac{(p + l_+ - 1)l_-}{S} k.$$

$$K_{L_-}(p, c, y_+, y_-, l_+, l_-) = \frac{100}{100 + l_y}\left(1 + \sigma N_{Syn}\right) \frac{(p + l_+ - 1)(l_+ - 1)}{S} k.$$

The first term in the rate equations (absent in the K_P and K_C equations) represents a correction factor for the length of the strand being replicated. The polymerase considered in our model is a processive polymerase, so the replication rate will be inversely

proportional to the length of the strand being replicated. Synthase length was varied in the range $10 \leq l_y \leq 200$, corresponding to the synthases having an advantage in replication rate when under 100 base pairs in length and a disadvantage when over 100 base pairs. In contrast, the linked strands are always longer than the polymerase and will therefore always be disadvantaged in terms of replication rate. The second term in the equations represents the benefit of synthases present in the cell, either as part of a linked functional strand or present on their own. The third term represents the concentration of available replicators, whether polymerases or the polymerase parts of the linked strands, multiplied by the number of templates of the given strand type. The stochastic simulation, consideration of strand breakdown and cell division proceeds identically to the procedure outlined in the methods section of (Shah et al., 2019) in chapter 2.

4.3 Results

Here we explain the testing method as shown in figure 4.1. We initialize 1000 protocells to contain one P and one C strand each, and refer to this as the P state. The simulation is allowed to run to equilibrium, after which one of the following two things happen to test when incorporation of the second beneficial gene or the linked genes is possible:

1. The system is invaded with 10 cells containing 2 copies of P and C as well as 2 copies of Y_+ and Y_- and is allowed to re-establish equilibrium.
2. The system is invaded with 10 cells containing 2 copies of P and C as well as 2 copies of L_+ and L_- and is allowed to re-establish equilibrium.

Under scenario 1, successful invasion results in the creation of a state containing P, C, Y_+ and Y_-, termed the PY state, while unsuccessful invasion creates a state with the same

P, C equilibrium as before. Scenario 2 differs in that successful invasion by the linked strands removes all P and C from the system, creating a state with only linked strands – the L state – while an unsuccessful invasion again retains the P, C equilibrium.

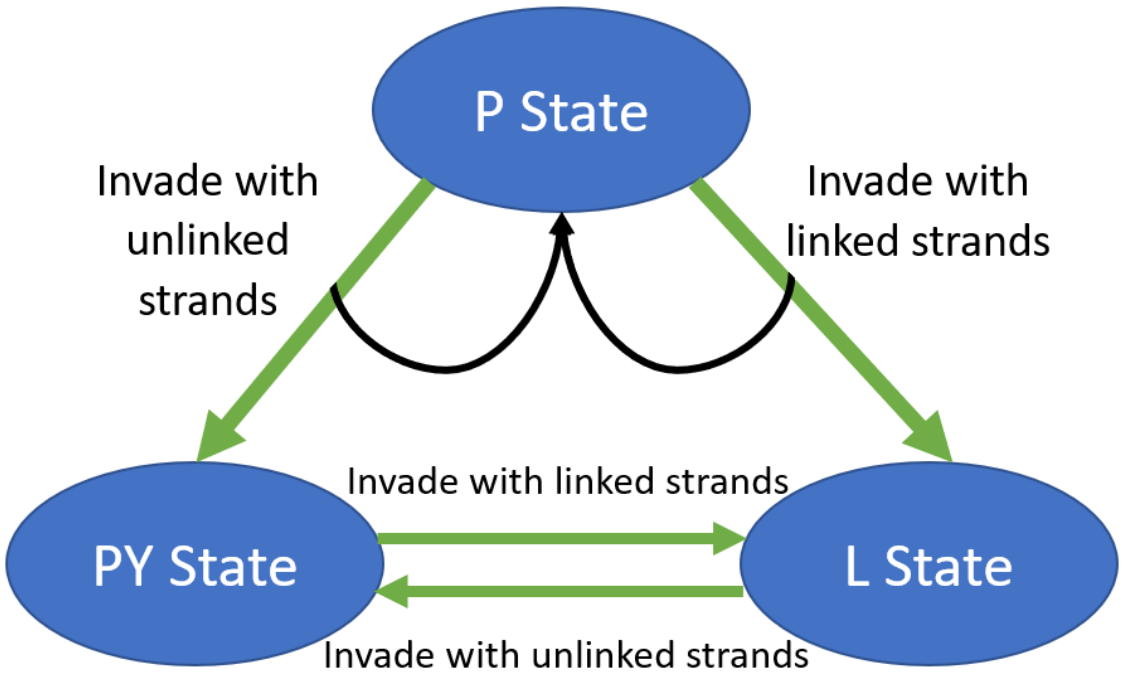


Figure 4.1 An overview of the possible equilibrium states in the simulation. Green arrows represent successful invasions and black arrows failures to invade.

The exact behaviour in either case is dependent on the l_y and σ chosen as is shown in the phase plots presented in figures 4.2 and 4.3. Looking at invasion by synthases shows four distinct regions. The region populated by blue squares represents simulations in which invasion was unsuccessful and the area with red circles shows simulations in which invasion was successful. The purple diamonds indicate a region where invasion was stochastic, and feature a purple dotted line indicating the stochastic boundary between the two states. Black x's show cases where the invasion was successful to such a degree that

the Y_+ and Y_- far outnumbered P and C, rapidly leading to no viable cells causing the death of the system. This dead region originally found from simulations with 1000 cells (not shown) was quite large compared to the region shown in figure 4.2 below. This area was reinvestigated using simulations performed with a larger population of 10000 cells for all points in the region $25 \leq l_y \leq 60$, and was found to shrink in size to the single column of x's depicted below. The higher population results were combined into figure 4.2 such that the points in $25 \leq l_y \leq 60$ are from simulations with 10000 cells and the other points are from simulations with 1000 cells. Larger sizes were not tested as it is computationally expensive to do so, however we predict that the dead region is merely a stochastic effect.

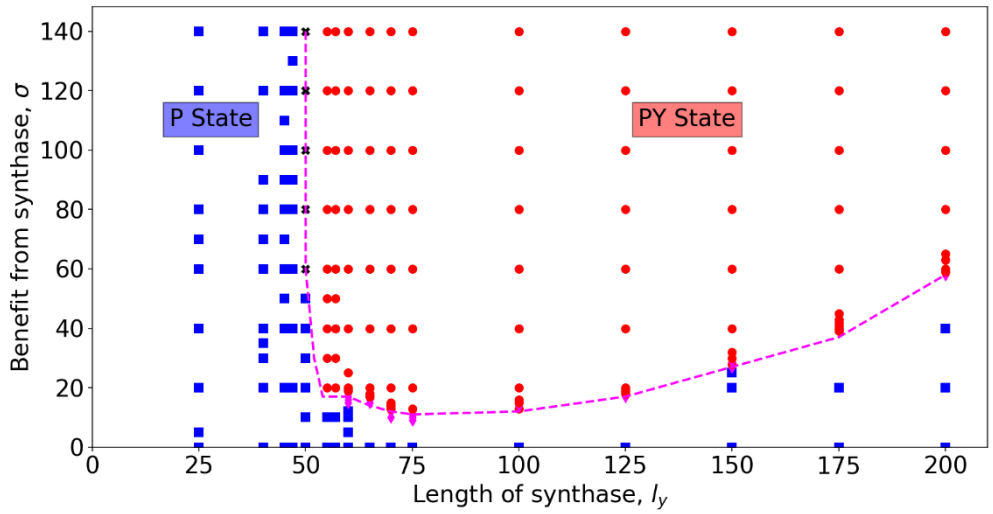


Figure 4.2 Phase plot of the invasion of a P state by Y+ and Y−. Red circles represent the cases where invasion was successful, blue squares where invasion was unsuccessful, purple diamonds where invasion was stochastic and black X's where invasion caused the death of the system. The purple dotted line represents a stochastic boundary between the two states.

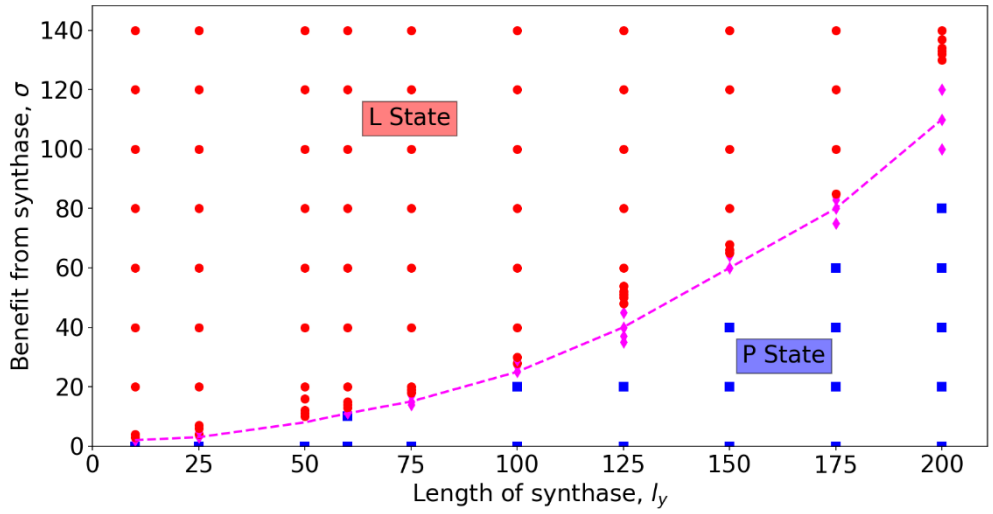


Figure 4.3 Phase plot of the invasion of a P state by L_+ and L_-. Red circles represent the cases where invasion was successful, blue squares where invasion was unsuccessful, purple diamonds where invasion was stochastic. The purple dotted line represents a stochastic boundary between the two states.

Looking at invasion by the linked strands (figure 4.3) shows similar behaviour, however there is no dead region. Here, the linked strands can replace only polymerases if the advantage σ is large enough and the length is not too large.

To determine whether the equilibrium state resulting from the invasion process is stable, we undertake another round of invasion. The stability of the PY state is tested by invading it with 10 cells containing 2 copies of P and C as well as 2 copies of L_+ and L_-. The result of this shows the conditions in which linked genes are favoured over unlinked genes. Similarly, the stability of the L state is tested by invading it with 10 cells containing 2 copies of P and C as well as 2 copies of Y_+ and Y_- where the result shows the conditions in which unlinked genes are favoured over linked genes.

The phase plot presented in figure 4.4 shows the results. The first phase, depicted by upwards facing red triangles, shows a region where L_+ and L_- successfully invade the PY state. The second phase, depicted by downwards facing blue triangles, shows a much larger region where P, C, Y_+ and Y_- invade the L state. The last phase, represented by black dots, is the region where the benefit provided by the synthase is too small for it to be incorporated, leaving the equilibrium as a P state. These regions are separated by purple dotted lines indicating stochastic boundaries between the phases.

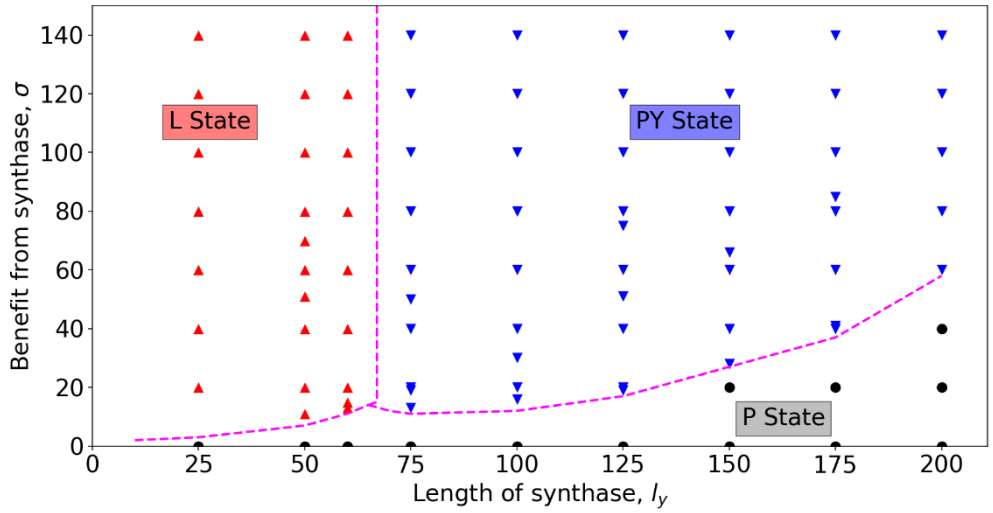


Figure 4.4 Phase plot of the invasion of an intermediate system by linked or unlinked strands. Red triangles represent the cases where linked strands alone form the equilibrium state, blue triangles the cases where the equilibrium state is a mixture of P, C, Y_+ and Y_- and black circles where the equilibrium state is comprised of P and C alone. The purple dotted line represents a stochastic boundary between the states.

The stochastic boundaries in figure 4.4 differ from those in figures 4.2 and 4.3 which separate the P state from the PY state or L state, respectively. Figure 4.5 plots the L state, PY state and P state as in figure 4.4 with the simulated points removed, retaining the phase

boundaries as purple dotted lines. Additionally, it plots the previous phase boundaries from figures 4.2 and 4.3 as black dotted lines. This shows that there are 5 distinct regions in the phase plot.

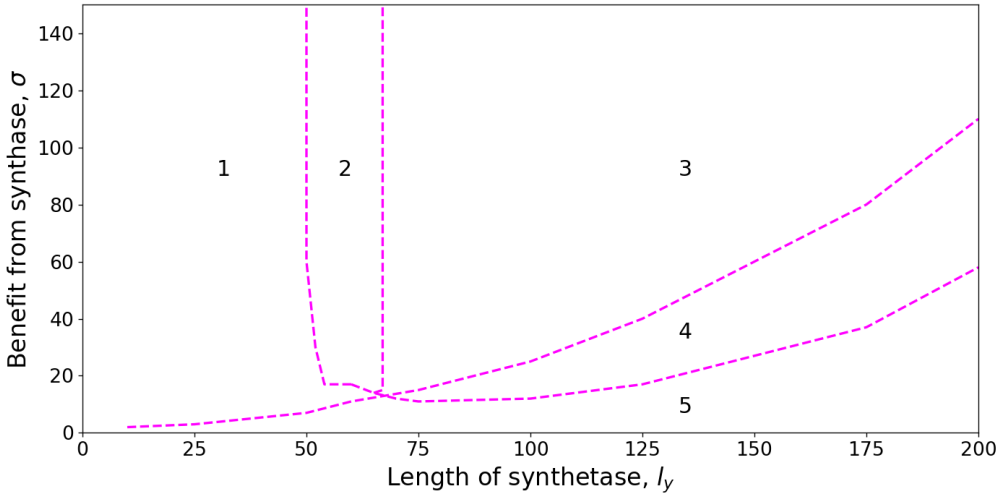


Figure 4.5 Phase plot combining information from Figures 4.2-4.4. (1) L state invades P state. PY state is not possible. (2) L state invades P and PY states. PY state invades P state. (3) PY state invades L and P states. L state invades P state. (4) PY state invades P state. L state is not possible. (5) Only P state is possible.

The L state from figure 4.4 is comprised of two subregions, regions 1 and 2. Region 1 is the parameter range in which the synthases as unlinked genes cannot successfully invade the P state, but the linked genes can. Region 2 is a small area in which both the unlinked and linked strands can invade the P state, however in this range the linked strands are more beneficial, resulting in the L state dominating the PY state. This is due to an overabundance of Y_+ and Y_- relative to P and C in each cell, leading to the random assortment of strands not producing two viable daughter cells a large portion of the time (see discussion for more).

Similarly, the PY state from figure 4.4 also has two subregions, regions 3 and 4. Much like region 2, region 3 is a parameter range where both the unlinked and linked strands can invade the P state, however here the opposite occurs as the PY state is favoured over the L state. At such large synthase lengths, there is no overabundance of the synthase and synthase complements relative to the polymerases and polymerase complements, thus the assortment issue disappears. In such a case, the strands staying unlinked is favourable due to quicker replication. Region 4 defines the area where the synthases as unlinked genes can successfully invade, but the linked genes cannot. This is because at the large length scales, the benefit provided by the synthase must be high to offset slower replication rates. The unlinked genes have a higher replication rate than the linked genes, thus require a lower σ to invade successfully. Lastly, region 5 is simply where the benefit provided by the synthase is not high enough for invasion of any kind to be successful.

The l_y = 60, 100 and 175 columns were selected, and simulations were done to determine the average equilibrium cell composition, shown in figures 4.6 and 4.7. We also determined the cell division rates and the number of viable cells for both the linked and unlinked cases for the same columns as a function of σ, shown in figures 4.8 and 4.9 below. In all of these plots, the error bars represent standard error. When considering a mixture featuring unlinked strands, the number of viable cells is determined by going through the population of cells and determining which ones have at least one P and one C strand or those with at least two P strands in them. In the case of the linked strand equilibrium state, we count the number of cells with at least one L_+ and L_- or at least two L_+ strands.

Figure 4.6 reveals that once invasion is successful, a particular ratio of Y_+ and Y_- to P and C is established, the value of which depends on l_y, and that Y_- outnumbers Y_+. When the length of the synthetase is short (see $l_y = 60$ column in figure 4.6), Y_+ and Y_- far outnumber P and C, to the point that the random assortment of strands upon cell division produces large amounts of non-viable daughter cells (see $l_y = 60$ column in figure 4.9). In contrast, figure 4.7 shows that L_+ and L_- always exist in equal amounts (a 1:1 ratio) post invasion.

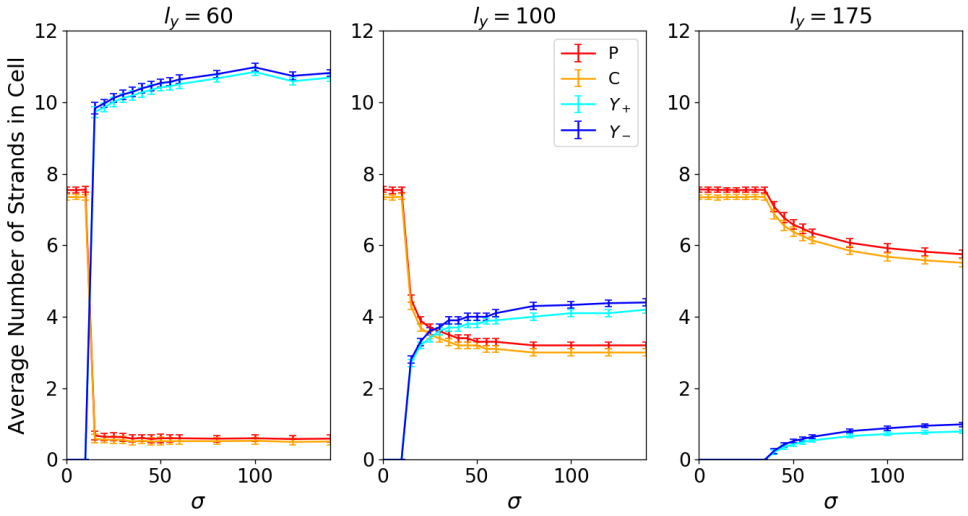


Figure 4.6 The average cell composition at equilibrium for simulations where a P, C state was invaded by unlinked strands for various l_y. Error bars are standard errors from a time average of the equilibrium state.

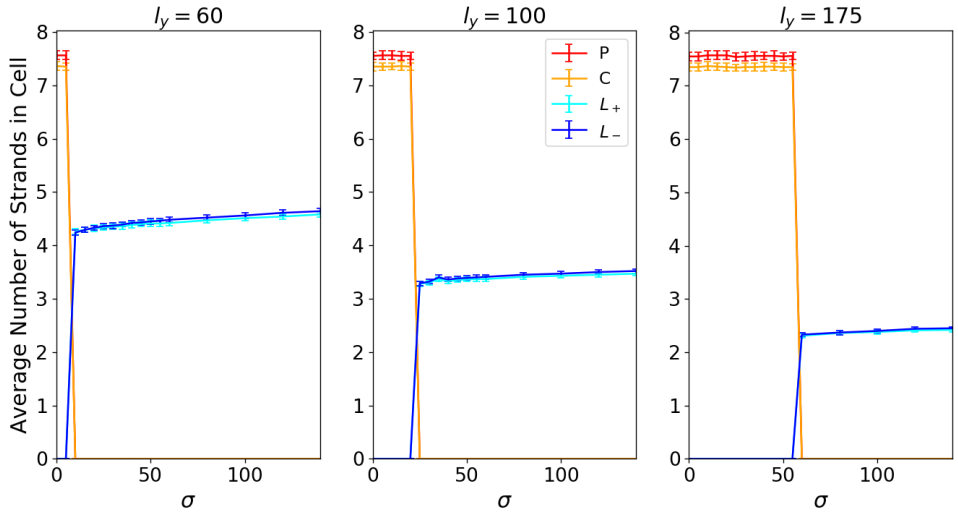


Figure 4.7 The average cell composition at equilibrium for simulations where a P, C state was invaded by linked strands for various l_y. Error bars are standard errors from a time average of the equilibrium state.

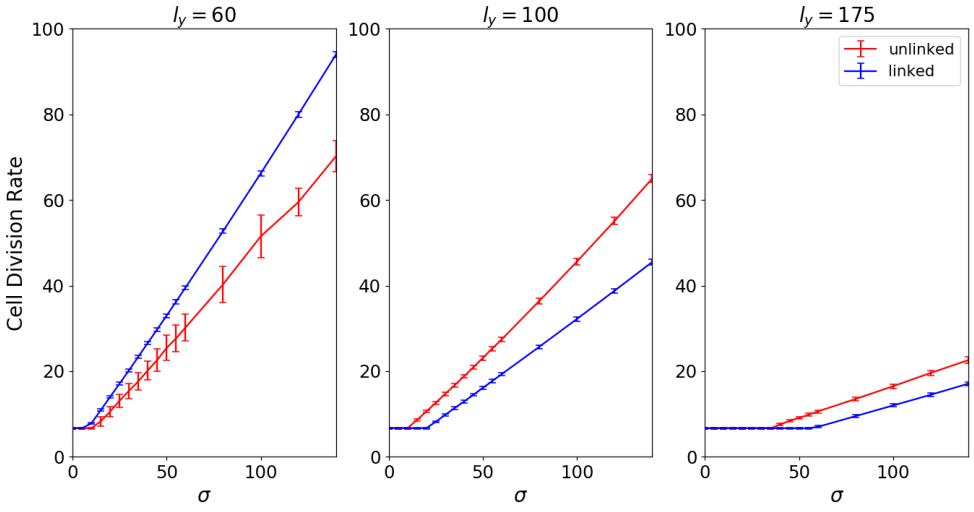


Figure 4.8 A comparison of the number of cell divisions per 100 δt for the cases studied in Figures 4.6 and 4.7.

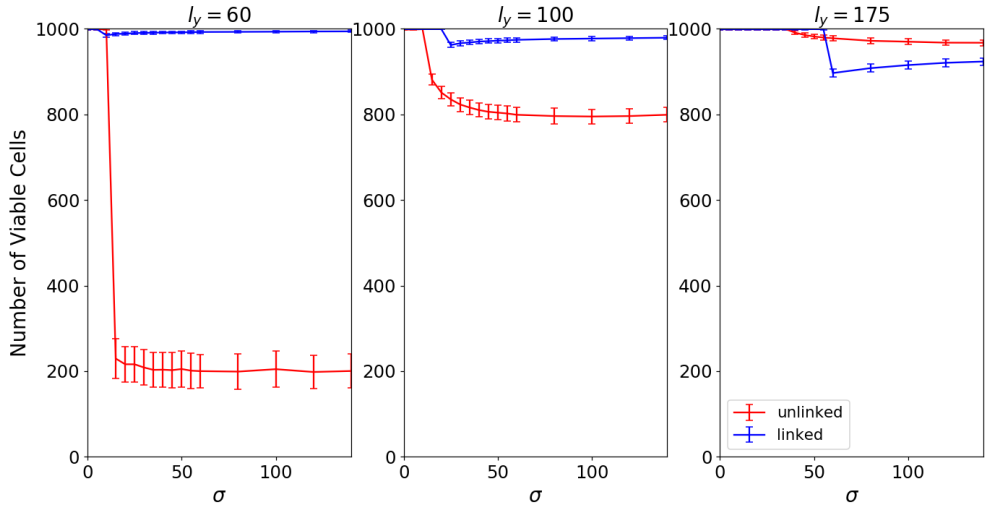


Figure 4.9 A comparison of the number of viable cells in the population at equilibrium divisions for the cases studied in Figures 4.6 and 4.7.

4.4 Discussion

The reason for successful invasion by either the linked or unlinked strands is made clear by figure 4.8: beyond a certain σ value, the incorporation of the unlinked strands and the replacement of P and C by the linked strands increases the replication rate due to the benefit provided by the synthase, which in turn increases the cell division rate. The linked strands tend to require a higher σ (better synthase) than their unlinked counterparts for this to occur due to their larger length, however this is not the case at short synthase lengths ($l_y \lesssim 60$). While incorporation of the synthase in is beneficial, figure 4.9 shows that the number of viable cells in the population has decreased when compared to the P and C equilibrium state. This is explained by looking at the strand distributions; when considering a system with only P and C strands, post random assortment of strands at cell division almost every daughter cell is viable. For the case with the unlinked strands at short synthase

lengths, Y_+ and Y_- outnumber the P and C strands, so when compared to a cell containing only P and C there will be far fewer P and C present prior to division. With a dearth of P and C to randomly assign to the daughter cells at division, the chances of a daughter cell not having a sufficient number of of P and C increases, leading to fewer viable cells. In the case with only linked strands the ratio of Y_+ and Y_- to P and C is fixed to be ~ 1:1, however each individual strand is longer, meaning fewer of them fit inside a cell. Thus, at large synthase lengths we run into a similar problem as above: upon division there are fewer polymerase equivalents to distribute leading to fewer viable cells.

At low values of l_y, the unlinked sequences cannot invade. The invading cells have rapidly replicating unlinked synthase strands which start taking over the system, but reach states where the number of synthases and synthase complements far exceeds the number of polymerases and polymerase complements (see $l_y = 60$ column in figure 4.9). This means that a large portion of the time the daughter cells produced are not viable, containing only synthases and synthase complements. The population rapidly fills up with cells which are unable to replicate, thus unable to divide which halts the takeover of the system. Simultaneously, the small fraction of the system featuring cells with only P and C continue to make viable daughter cells, leading to the viable daughter cells overwriting the stagnant ones with unlinked strands, eliminating the unlinked strands from the population. Linkage is beneficial for low values of l_y: the increased replication rate provided by adding a small portion to the existing polymerase sequence exceeds the length correction factor introduced due to the increased length. Linking the Y_+ to P and Y_- to C ensures a moderate replication rate enhancement and the ability to keep the number of polymerase equivalents in a cell

high. Since both genes are functional in the linked strand, the issue of non-viable daughter cells is lessened greatly and invasion is successful.

For intermediate to large values of l_y, upon invasion the synthases and synthase complements do not surge up in numbers on account of their longer lengths allowing fewer to fit in a given cell, slowing their replication. This means the system can maintain a good number of polymerases and polymerase complements in cells with the unlinked strands, avoiding the problem described above. In such cases they can outcompete the longer linked strands, as there is no need to moderate the ratio of synthases to polymerases the benefit of linkage disappears. At the longest lengths, the systems with linked strands produce fewer viable daughter cells than the systems with unlinked strands (see $l_y = 175$ column in figure 4.9), as their long lengths prevent many strands from existing in a single cell. This leads to the same issue that the unlinked synthases had at short synthase lengths, there are not enough polymerase equivalents prior to division to ensure the daughter cells remain viable post random assortment.

Chapter 5: Conclusions

In this thesis, we address the issue of incompatibility of dynamically different computational models (see section 1.7 for an overview) of a prebiotic system of replicators leading to difficulties in the comparison of higher-level selection methods. Specifically, comparison of spatial clustering in lattice models and group selection in protocellular models has been attempted before (Takeuchi and Hogeweg, 2009), however the comparison made was qualitative at best due to how the model was constructed, making it difficult to separate out the effects of diffusion. In the research presented in chapter 2, we

develop a variety of computational spatial and protocellular models with near identical dynamics, differing only in details of diffusion for spatial models and cell division in protocellular models. Our protocell models are most comparable to stochastic corrector models (Szathmáry and Demeter, 1987), however we avoid explicitly assigning fitness values to cells. Our spatial models resemble cellular automaton models, consisting of sites connected globally or locally depending on the model, allowing a large number of strands per site to make them analogous to protocells. We simulate polymerases, polymerase complements and parasites, and consider the polymerases to be general, error prone catalytic replicators. Two kinds of diffusion are studied in the spatial models, allowing strands to move to other sites in a Moore neighbourhood or to any other site on the lattice. With only the minor differences left between the spatial and protocellular models, quantitative comparison is done using the metric of the error threshold, the maximum mutation rate survivable by our systems. We are able to show that group selection in protocell models can sustain error rates is four- to five-fold higher than spatial clustering can when considering normal parasites, or five- to fourteen-fold higher when considering fast replicating parasites. Analysis of the models in chapter 3 reveals this is due to spatial clustering being unable to purge the parasites from the spatial models unlike how group selection can in protocell models. Having established that group selection does a far better job allowing a system of replicators to survive, we adopt the protocell model and conduct a study to determine the cases under which the linkage of strands to form a proto-genome is possible and stable. The results of chapter 4 show that linkage is beneficial and can be achieved for a wide range of parameters, however it is only stable for a small subset of this

range. Specifically, linkage of functional strands is stable for short sequences attaching onto the polymerase and is even advantageous in ensuring that daughter cells remain viable by controlling the ratio of polymerases to shorter functional strands.

This research supports origins of life theories involving RNA replicators in porous media, on mineral lattices, or in protocellular systems by showing that spatial clustering and group selection can provide stability to mutational pressure from erroneous replication. The large amount of stability provided by group selection supports the idea that protocell or protocell like structures were present at the onset of life, or arose quickly after. The work on linkage further shows the ability of protocellular systems to develop proto-genome like structures, a path to the chromosomes in modern biology.

Further investigations along the lines of this research could do similar tests comparing clustering and group selection using the phenotypic error threshold. This can be computationally intensive, and involves modelling full sequences and testing the impact of point mutations by finding the new folded structure of the strand. While a general processive polymerase has not yet been discovered (hence its sequence is unknown), it may be substituted by a proxy RNA strand with sufficiently complicated folded structure, or perhaps by using the sequence of one of the experimentally developed polymerases discussed in section 1.4.

References:

Ahlquist, P., 2002. RNA-Dependent RNA polymerases, viruses, and RNA silencing. *Science*, 296(5571), pp.1270-1273.

Attwater, J., Wochner, A., and Holliger, P., 2013. In-ice evolution of RNA polymerase ribozyme activity. *Nature Chemistry*, 5(12), p.1011.

Attwater, J., Raguram, A., Morgunov, A., Gianni, E. and Holliger, P., 2018. Ribozyme-catalysed RNA synthesis using triplet building blocks. *Elife*, 7, p.e35255.

Boerlijst, M., and Hogeweg, P., 1991. Spiral wave structure in pre-biotic evolution: hypercycles stable against parasites. *Physica D: Nonlinear Phenomena*, 48(1), pp.17-28.

Branciamore, S., Gallori, E., Szathmáry, E., and Czárán, T., 2009. The origin of life: chemical evolution of a metabolic system in a mineral honeycomb?. *Journal of molecular evolution*, 69(5), 458.

Chen, I. and Nowak, M., 2012. From prelife to life: How chemical kinetics become evolutionary dynamics. *Accounts of chemical research*, 45(12), pp.2088-2096.

Cheng, L. and Unrau, P., 2010. Closing the circle: replicating RNA with RNA. *Cold Harbor Perspectives in Biology*, 2(10), p.a002204.

Crick, F., 1968. The origin of the genetic code. *Journal of molecular biology*, 38(3), pp.367-379.

Damer, B. and Deamer, D., 2015. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: A scenario to guide experimental approaches to the origin of cellular life. *Life*, 5(1), pp.872-887.

Deamer, D., 1985. Boundary structures are formed by organic components of the Murchison carbonaceous chondrite. *Nature*, 317(6040), p.792.

Deamer, D. and Pashley, R., 1989. Amphiphilic components of the Murchison carbonaceous chondrite: surface properties and membrane formation. *Origins of Life and Evolution of the Biosphere*, 19(1), pp.21-38.

Draper, W., Hayden, E. and Lehman, N., 2007. Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. *Nucleic Acids Research*, 36(2), pp.520-531.

Eigen, M., 1971. Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften*, 58(10), pp.465-523.

Franchi, M., and Gallori, E., 2005. A surface-mediated origin of the RNA world: biogenic activities of clay-adsorbed RNA molecules. *Gene*, 346(14), pp.205-214.

Gilbert, W., 1986. Origin of life: The RNA world. *Nature*, 319(6055), pp.618.

Hayden, E. and Lehman, N., 2006. Self-assembly of a group I intron from inactive oligonucleotide fragments. *Chemistry & Biology*, 13(8), pp.909-918.

Hayden, E., von Kiedrowski, G. and Lehman, N., 2008. Systems chemistry on ribozyme self-construction: evidence for anabolic autocatalysis in a recombination network." *Angewandte Chemie International Edition*, 47(44), pp.8424-8428.

Higgs, P. and Lehman, N., 2015. The RNA world: molecular cooperation at the origins of life. *Nature Reviews Genetics*, 16(1), pp.7-17.

Higgs, P., 2017. Chemical Evolution and the Evolutionary Definition of Life. *Journal of Molecular Evolution*, 84(5-6), pp.225–235.

Higgs, P., 2018. Three Ways to Make an RNA Sequence: Steps from Chemistry to The RNA World, in Kolb, V. (ed.) *Handbook of Astrobiology*. CRC Press: Boca Raton, FL, USA.

Hofbauer, J., 1984. A difference equation model for the hypercycle. *SIAM Journal on Applied Mathematics*, 44(4), pp.762-772.

Horning, D. and Joyce, G., 2016. Amplification of RNA by an RNA polymerase ribozyme. *Proceedings of the National Academy of Sciences*, 113(35), pp.9786-9791.

Huynen, M., Stadler, P. and Fontana, W., 1996. Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences*, 93(1), pp.397-401.

Johnston, W., Unrau, P., Lawrence, M., Glasner, M. and Bartel, D., 2001. RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science*, 292(5520), pp.1319-1325.

Joyce, G., 1989. RNA evolution and the origins of life. *Nature*, 338(6212), pp.217-224.

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

Kim, Y. and Higgs, P., 2016. Co-operation between Polymerases and Nucleotide Synthetases in the RNA World. *PLOS Computational Biology*, 12(11), p.e1005161.

Könnyű, B. and Czárán, T., 2013. Spatial aspects of prebiotic replicator coexistence and community stability in a surface-bound RNA world model. *BMC evolutionary biology*, 13(1), p.204.

Könnyű, B. and Czárán, T., 2014. Phenotype/genotype sequence complementarity and prebiotic replicator coexistence in the metabolically coupled replicator system. *BMC evolutionary biology*, 14(1), p.234.

Könnyű, B. and Czárán, T., 2015. Template directed replication supports the maintenance of the metabolically coupled replicator system. *Origins of Life and Evolution of Biospheres*, 45(1-2), pp.105-112.

Könnyű, B., Szilágyi, A. and Czárán, T., 2015. In silico ribozyme evolution in a metabolically coupled RNA population. *Biology direct*, 10(1), p.30.

Kruger, K., Grabowski, P., Zaug, A., Sands, J., Gottschling, D. and Cech, T., 1982. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1), pp.147-157.

Kun, A., Santos, M. and Szathmáry, E., 2005. Real ribozymes suggest a relaxed error threshold. *Nature genetics*, 37(9), p.1008.

Lawrence, M., and Bartel, D., 2005. New ligase-derived RNA polymerase ribozymes. *RNA*, 11(8), pp.1173-1180.

Lincoln, T., and Joyce, G., 2009. Self-Sustained Replication of an RNA Enzyme. *Science*, 323(5918), pp.1229-1232.

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

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

Matsumura, S., Kun, Á., Ryckelynck, M., Coldren, F., Szilágyi, A., Jossinet, F., Rick, C., Nghe, P., Szathmáry, E. and Griffiths, A., 2016. Transient compartmentalization of RNA replicators prevents extinction due to parasites. *Science*, 354(6317), pp.1293-1296.

Mills, D., Peterson, R. and Spiegelman, S., 1967. An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proceedings of the National Academy of Sciences of the United States of America*, 58(1), p.217.

Orgel, L., 1968. Evolution of the genetic apparatus. *Journal of molecular biology*, 38(3), pp.381-393.

Pace, N. and Marsh, T., 1985. RNA catalysis and the origin of life. *Origins of Life and Evolution of the Biosphere*, 16(2), pp.97-116.

Paul, N. and Joyce, G., 2002. A self-replicating ligase ribozyme. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20), pp.12733-12740.

Phillips, R., Theriot, J., Kondev, J. and Garcia, H., 2012. *Physical biology of the cell*. Garland Science.

Rich A., 1962, "On the problems of evolution and biochemical information transfer," in: Kasha M. and Pullman B. (eds), *Horizons in Biochemistry*, New York: Academic Press, 103-126.

Robertson, M. and Joyce, G., 2012. The origins of the RNA world. *Cold Spring Harbor Perspectives in Biology*, 4(5), p.a003608.

Schuster, P., Sigmund, K. and Wolff, R., 1979. Dynamical systems under constant organization. III. Cooperative and competitive behavior of hypercycles. *Journal of Differential Equations*, 32(3), pp.357-368.

Schuster, P., Fontana, W., Stadler, P. and Hofacker, I., 1994. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 255(1344), pp.279-284.

Segré, D., Ben-Eli, D., Deamer, D. and Lancet, D., 2001. The lipid world. *Origins of Life and Evolution of the Biosphere*, 31(1-2), pp.119-145.

Shah, V., de Bouter, J., Pauli, Q., Tupper, A. and Higgs, P., 2019. Survival of RNA Replicators is much Easier in Protocells than in Surface-Based, Spatial Systems. *Life*, 9(3), p.65.

Shay, J., Huynh, C. and Higgs, P., 2015. The origin and spread of a cooperative replicase in a prebiotic chemical system. *Journal of Theoretical Biology*, 364(1), pp.249-259.

Smith, J. and Száthmary, E., 1993. The origin of chromosomes I. Selection for linkage. *Journal of Theoretical Biology*, 164(4), pp.437-446.

Szathmáry, E. and Demeter, L., 1987. Group selection of early replicators and the origin of life. *Journal of theoretical biology*, 128(4), pp.463-486.

Szathmáry, E. and Smith, J., 1993. The evolution of chromosomes II. Molecular mechanisms. *Journal of theoretical biology*, 164(4), pp.447-454.

Szilágyi, A., Kun, Á. and Szathmáry, E., 2012. Early evolution of efficient enzymes and genome organization. *Biology direct*, 7(1), p.38.

Szilágyi, A., Zachar, I., Scheuring, I., Kun, Á., Könnyű, B., and Czárán, T., 2017. Ecology and Evolution in the RNA World Dynamics and Stability of Prebiotic Replicator Systems. *Life*, 7(4), p.48.

Szostak, J., 2012. The eightfold path to non-enzymatic RNA replication. *Journal of Systems Chemistry*, 3(1), p.2.

Takeuchi, N., Poorthuis, P. and Hogeweg, P., 2005. Phenotypic error threshold; additivity and epistasis in RNA evolution. *BMC Evolutionary Biology*, 5(1), p.9.

Takeuchi, N. and Hogeweg, P., 2009. Multilevel selection in models of prebiotic evolution II: a direct comparison of compartmentalization and spatial self-organization. *PLoS computational biology*, 5(10), p.e1000542.

Takeuchi, N., Hogeweg, P. and Kaneko, K., 2017. The origin of a primordial genome through spontaneous symmetry breaking. *Nature communications*, 8(1), p.250.

Tinoco, I. and Bustamante, C., 1999. How RNA folds. *Journal of Molecular Biology*, 293(2), pp.271-281.

Varani, G., and McClain, W., 2000. The G·U wobble base pair: A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Reports*, 1(1), pp.18-23.

Wintersberger, U. and Wintersberger, E., 1987. RNA makes DNA: a speculative view of the evolution of DNA replication mechanisms. *Trends in Genetics*, *3*, pp.198-202.

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

Wu, M. and Higgs, P., 2012. The origin of life is a spatially localized stochastic transition. *Biology Direct*, 7(1), p.42.

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

Zintzaras, E., Santos, M. and Szathmáry, E., 2002. "Living" under the challenge of information decay: the stochastic corrector model vs. hypercycles. *Journal of theoretical biology*, 217(2), pp.167-181.