Replication of Genes in Rolling Circles

Replication of Genes in Rolling Circles

encoding functions in circular replicators at the origin of life

By Felipe Rivera-Madriñan, HBSc

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

McMaster University

Master of Science (2022)

Hamilton, Ontario (Department of Physics and Astronomy)

TITLE: Replication of Genes in Rolling Circles

AUTHOR: Felipe Rivera-Madriñan (McMaster University)

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: ix, 90

Lay Abstract

The origin of life is a topic that many people are inherently curious about. However, science is only just making progress towards an answer. The first organisms must have been able to replicate. Modern organisms use proteins, DNA, and RNA to do this, however it is unlikely these three molecules could have co-ordinated at the origin of life. A simpler model for replication uses only RNA, which can be both a gene and a catalyst. Here we propose some computational models which study RNA replication. These models simulates strands undergoing rolling circle replication, a method of replication some viruses use which has been suggested to be sustainable at the origin of life. We show that rolling circle replication can create long strands which can have new helpful sequences of RNA. This mechanism could have helped the first organisms achieve better replication and evolution, which is a key characteristic of life.

Abstract

The origin of life is one of the most captivating and difficult questions that science has yet to answer. Several different questions remain, including how genetic replication may have begun. Replication is a fundamental property of life that allows for evolution and the long-term survival of life. Non-enzymatic replication should have been present at the origin of life. The RNA world theory proposes that because it can act as both an enzyme and gene, RNA could have performed the function of a replicator at the origin of life. Abiotic chemistry for RNA nucleotides is known, as well as mechanisms for simple but random RNA sequence synthesis. However non-enzymatic replication of RNA sequences which might hold functions, has only achieved mild success. This is in no small part because of replication infidelity between RNA bases, and product inhibition during template directed replication. The rolling circle mechanism found in viroids and some RNA viruses, is a likely way to avoid these issues in the RNA World. Here we present a summary of key topics to origins of life and the RNA world, a deterministic model for rolling circle replication, followed by an original computational model for gene fixation in rolling circle replication. In these simulations we observe the dynamics of populations of protocells, each containing multiple copies of rolling circle RNAs that can replicate non-enzymatically. Selection for speed of replication tends to reduce circles to a minimum length. However, errors provide a natural doubling mechanism that creates strands multiple times the length of the minimal sequence. We show that if a beneficial gene appears in this new space, the longer sequence with the beneficial function can be selected, even though it replicates more slowly. This provides a route for the evolution of longer circles encoding multiple genes.

Acknowledgements

A special thanks to Dr. Paul Higgs without who this thesis would not have been possible. Your guidance these last two years has been instrumental both in this work and my career. Thank you for helping me make the best of these two very chaotic years, I look forward to continuing our work together in the coming years. I would also like to thank my lab mates and friends who collectively spent many hours listening to my various thoughts and providing support in many ways, your friendship makes the most challenging tasks just a little easier and much more rewarding. Finally thank you to my family and my partner for bringing so many things outside my work into my life. Your constant influence keeps me working towards my goals every day.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation and structure of thesis

Understanding the origin of life (OOL) would be an outstanding scientific achievement. The origin of life brings together our knowledge of the abiotic and biological worlds providing a connection between every living thing and the physical space they inhabit. Very few topics capture the imagination, our collective heritage, and a challenging appreciation for all physical sciences as well as the question of how life started. Befitting such a lofty goal, OOL is also a challenging, messy discipline with no clear-cut answers. There are many scenarios to consider and so few examples of life in the universe that OOL requires the use of ever major scientific principle in some way.

Bringing together all this disparate knowledge is perhaps the greatest challenge which faces OOL, though it can also be seen as an exciting opportunity for cooperation and innovation. Researchers of OOL often hope that discovering how life is created may uncover new ideas which can help us understand how life fits into

the world it inhabits. However for now, OOL is still trying to figure out how the characteristics of life could have come from a non-living world. The goal of this thesis is to contribute to our understanding of the origins of genetic replication, which is crucial to life. Specifically, we will argue that genetic replication may have started with a hammerhead ribozyme (HHRz), a small self cleaving catalyst that can improve its own reproduction and long-term stability while allowing for novel catalysts to emerge.

To accomplish this the thesis is split into several chapters. Chapter 1 is used to introduce concepts in origins of life research which are important to our discussion. In this summary we mention the strengths of the RNA world theory which forms the basis for our hypothesis and aims to place RNA as the first functional and genetic molecule of life. We then discuss the difficulties that face the RNA world hypothesis (as well as OOL as a whole) and draw on prior work in the literature to show that a hammerhead ribozyme could alleviate some of these. We also discuss in some detail recent advances in computational models of the origins of life and give a brief description of how our model fits into the literature. Chapter 2 discusses a model for rolling circle replication which expands on work done by prior lab members. Chapter 3 contains a version of a paper currently under peer-review which makes the bulk of the work that has been done throughout 2 years of a Masters degree. The paper has been modified to fit into this thesis. Specifically, some of the introduction in chapter 1 contains portions of the introductory material from the paper, and an extra graph has been added to this thesis that was not present in the paper. Citations have been made to agree between the paper and introduction. Chapter 4 briefly describes how the model discussed in chapter 3

can be modified to explore the linking of different genes on a rolling circle, which is an important step towards the creation of life's first genome.

1.2 Basics of life

Finding a definition for life which encompasses the diversity we see, while excluding other complex natural phenomenon turns out to be a surprisingly difficult task [1]. Though not perfect, most widely accepted definitions of life tend to focus on what life can do, often defining life by its ability to self-replicate, and undergo Darwinian evolution [1–3]. OOL research is centered on the assumption that the first organisms shared these core functions with modern life, and often aims to understand how these behaviors could have begun in an abiotic world.

Today, life replicates using DNA and proteins. DNA acts as a blueprint from which catalytic proteins used for metabolism are made [4]. Metabolism is a broad term which often is undefined in the context of OOL, here we will use it to describe the chemical processes life use to maintain a steady state and grow. In any case, Individual metabolic proteins are made from individual segments of DNA called genes, which are translated into an RNA code by a protein, and read by an RNA catalyst called the ribosome that physically joins amino acids into proteins. Genes can be passed down to successive generations of closely related individuals which can share genetic material during reproduction. These organism form species, and the genetic variability within them forms the basis of evolution. Different genes create differences in proteins, and these differences impact the ability of individuals to reproduce in relation to each other. More successful organisms pass on their genes more often than others and over time this changes the overall genetic

pool of a species. Mutations and statistical fluctuations related to how genes are passed on can also influence the genetic pool of a species [2]. Many details and complexities surround the methods of protein synthesis and evolution which we have just described. The consensus is that for life to have started, some simpler method of replication, self-preservation and evolution should have been available [5].

1.2.1 The molecules of life and prebiotic chemistry

To understand how these processes could have started, we must first look at the molecules which facilitate them in modern life. As mentioned before, proteins composed of amino acids perform the majority of metabolic actions in cells [6], DNA and RNA nucleotides made from a common nucleobase alphabet mediate their creation [7,8]. Membranes composed of fatty acid chains also perform the important secondary task of keeping all the components of life together [9]. These three molecules are the basic materials from which all life is made, however it is not clear if they would have been available in a prebiotic earth. Considerable work has been done to find ways to make these molecules of life without prior biological cells or molecules. The famous Urey-Miller experiment is often considered the first modern publication in this topic, which is often called prebiotic chemistry [10]. The experiment aimed to create biomolecules by simulating lighting passing through the atmosphere of a prebiotic earth. To do this, a gaseous sample of water vapour, methane, ammonia, and hydrogen was shocked with electricity. The experiment succeeded in creating detectable amounts of 5 different amino acids [11] in less than a day. Though the experiment was monumental, it was not the first inorganic synthesis of amino acids. As far back as the 1800's chemist Adolph Strecker had

discovered a reaction which could make amino acids from aldehydes, ammonia, and hydrogen cyanide, in fact the same reaction is thought to be the source of amino acids in the Urey-Miller experiment [12].

The real impact of Urey-Miller was to conceptualise a procedure which applied concepts of inorganic chemistry to scenarios which might have been possible in a prebiotic earth. The concept invigorated discussion about the type of chemistry possible before life, and inspired experiments which up to today have proposed abiotic pathways for the creation of lipids [13], nucleobases [14–16] and sugars [17] (an important component of DNA, RNA and various other molecules), as well as some 20 amino acids used in life [18–21]. Progress has also been made in the creation of DNA and RNA nucleotides, which had previously been difficult due to the need to join a sugar, a phosphate and a nucleotide under the right conditions [22]. The field of prebiotic chemistry has even come so far as to leave behind its founding experiment, with most prebiotic-chemist now believing Urey-Miller experiment to be improbable in an early earth [23–26].

In recent years amino acids, nucleobases and fatty acids have also been observed to be present in several astronomical objects like meteorites and interplanetary dust [27,28]. These are now expected to have been an important source of organic molecules in a prebiotic earth. Topics in prebiotic chemistry continue to be debated and explored, but overal there is agreement that evidence points towards some possible synthesis of the building blocks of life on a prebiotic earth. However these components by themselves do not constitute a living system, and the question of how they could have come together to create life remains unanswered.

1.2.2 Theories for the origin of life

A mechanism for open-ended self-replication, which many agree would constitute life, would probably look different than what is seen in modern cells. DNA and proteins, the two most central molecules today, depend too heavily on each other to have spontaneously began working together in a prebiotic world. That being said, the genetic functions represented by DNA and the metabolic abilities shown by proteins each provide a foundation for the two most accepted theories for the origin of life.

The metabolism first theory proposes that self-replicating networks of small molecules provided a basis for metabolic activities from which life emerged. This theory tends to place the origin of life in hydrothermal vents where redox reactions like serpentinization can occur [29]. On the other hand, the RNA World theory proposes that life started from the catalytic and genetic capabilities of RNA [30,31]. This claim is bolstered by the plethora of catalytic RNA that has been found in recent years and the role of RNA in protein synthesis, these are considered evidence that an RNA world existed before the use of DNA or proteins by life.

The work performed for this thesis focuses mostly on the beginning of genetic inheritance and supports the RNA world theory. Though some questions still need to be answered by the RNA world hypothesis, several advancements have been achieved in recent years. On the other hand very little experimental evidence exists for the existence of small molecule replicating networks, or the ability of such molecules to evolve and become inheritable. Even if these did exist there is no hypothesis explaining how the modern apparatus for genes and protein synthesis could have emerged from these metabolic networks. Some recent work has

advocated for a mixed model [32], and though we agree that some components of metabolism first may have aided in the beginning of life (such as in the synthesis of nucleotides) these ideas have little impact on the issue of replication. In other words, we are confident that genetic inheritance is a key step to life that started in an RNA world. While metabolic-like mechanisms may have been present before or during the RNA world, once enough nucleotides were available (from either metabolism or direct chemistry) RNA replication could begin and become the key method through which life could appear. The source of nucleotides does not impact the issues facing RNA replication, therefore we assume in the model we will present that we start in an environment where nucleotides are available.

Before continuing, we should also note that there are several less studied theories of the origins of life which study the non-enzymatic replication of different molecules. For example, the lipid world hypothesis [33] argues that non-genetic, compositional information can be passed on by growing protocells. This is based off the observation that lipid membranes inherit some of the characteristics of the membranes they were created from. This shares the assumption with the RNA world, that life began with some of the molecules it uses today, however some believe that life can be created without much resemblance to life around us today, or even the processes that created life on earth to begin with [34]. In this thesis we will limit our discussion to the origin of life on earth as opposed theories of how any life may begin in general. Though these two ideas overlap, keeping our discussion to the context of what was possible in a prebiotic earth and what is most studied will ensure a more grounded approach to our arguments.

1.3 The RNA World

The RNA world theory suggests that instead of a complex system comprised of DNA and proteins, early life used ribonucleic acid (RNA) as the primary molecule of life [30,31]. As mentioned above, the capacity of RNA to act as both a catalyst and genetic carrier in modern life is most prominent during protein synthesis, where RNA acts as a messenger between DNA and the ribosome, a catalytic RNA complex which assembles proteins according to the RNA messenger. The existence of the ribosome lends significant validity to the RNA world hypothesis, specially given that the peptide bond between amino acids appears to be completely catalysed by ribosomal RNA [35]. This suggests that RNA preceded proteins in the development of life.

Though RNA is quite similar to DNA, the use of an RNA primer in DNA synthesis of some species [36], and the ubiquitous use of the RNA sugar ribose as a precursor to the deoxyribose sugar in DNA, suggests RNA may be more ancient [30,37]. RNA also maintains the polymeric structure of DNA, using three of the same nucleobases as monomers. This allows RNA to function as a genetic code, with some viral genomes even being entirely composed of RNA [38]. The similarity of the two polymers also allows for hybridization and translation to occur between them, providing a possible route from the RNA world to the modern DNA genome [39]. The presence of secondary and tertiary structures in single stranded RNA also allows some sequences to be catalytic, these sequences are know as ribozymes[40]. Though catalytic DNA has also been observed [41], it has been hypothesised that the presence of the hydroxyl group which differentiates RNA from DNA may make it a better catalyst [40,42]. However, little is actually known about the mechanism

of nucleic acid catalysts and some work suggests that the perceived edge of RNA over DNA as a catalyst is unfounded [43].

Other ribozymes have also been discovered. The first ribozyme ever identified was a self-cleaving sequence discovered from a protozoan genome in 1982 by Thomas Cech [44]. This discovery, as well as a supporting paper on RNA capable of cleaving in trans the next year [45], received a Nobel prize for its implications to health sciences and the origin of life [46]. Since then, several other classes of cleaving ribozymes like the hammerhead ribozyme [47,48] have been identified. Various ribozymes capable of synthesizing a host of functions like nucleotide synthesis [49], amide bond formation [50], aminoacylation [51], Acetyl-CoA synthesis [52], RNA capping (used in the creation of mRNA) [53], as well as a handful of ribozymes with dual functions [54,55] have also been found, or made in labs. Notable amongst advancements in synthetic ribozymes is the considerable progress that has been made in the development of an RNA polymerase [56–59], which would be key in an RNA world.

Though the large number of different ribozymes found suggests many functions were possible in the RNA world, there is still no ribozyme capable of universal or self replication. Such a universally replicating polymerase would be able to replicate itself along with a whole repertoire of ribozymes held in an RNA genome. Proponents of the RNA world theorem hope that if possible, such a cycle of self-replication would fulfill all the requirements for life and allow an organism to freely replicate and evolve. However before the advent of a polymerase, strands from which ribozymes can emerge need some way of being made and it seems likely that some form of non-enzymatic polymerization and replication would be

required. Simple random polymerization of RNA nucleotide has already been demonstrated in a plethora of environments [60–62]. RNA polymerization occurs through a condensation reaction which forms phosphodiester bonds between a phosphoric acid and a hydroxyl group of two separate nucleotides. Because of this requirement non-enzymatic RNA polymerization in aqueous environments can only occur favourably with activated RNA. These are RNA oligomers which are modified to provide or lower the activation energy required for polymerization. Different modifications exists, including the addition of a circular phosphate between the 2'and 3' carbons in ribose [63] and the addition of Imidazole to the phosphate group attached to the 5' carbon in ribose [64].

However polymerization of non-activated RNA has been shown to occur within lipid-nucleotide solutions undergoing dehydration [61]. Under these conditions, lipids form multilamellar matrices which restrict RNA to a two dimensional space [65]. In this confirmation entropic effects are decreased, and the phosphodiester bond can be created spontaneously. Other forms of RNA trapping have also been reported in laboratory, including within clay and salts [62]. These experiments demonstrate that polymerization of RNA is possible in the prebiotic world. Some strands created in this fashion even exceed the length of some ribozymes, opening the possibility that catalytic RNAs could appear under prebiotic conditions.

However strands made through free end-to-end polymerization are random. The process of polymerization does lead to replication of sequences, just the creation of a random pool of strands. For life to occur, ribozymes create through random polymerization must be inheritable and copiable, allowing a beneficial sequence to be passed on. Watson-Crick pairing between a template and monomeric, or

oligomeric nucleotide strands (which we call nmers) has been proposed as a potential way to achieve inheritable replication in a non-enzymatic world [66]. If several nmers attach to a template through Watson-Crick pairing, polymerization can occur between nmers which are next to each other on the template. The result is a complimentary copy of the template, which may be fully or partially complete depending on how many of the Watson-Crick pairs on the template were filled. Though several advances in this template directed non-enzymatic replication have been achieved [39,67–71] many issues remain. Polymerization rates between annealed nmers remain low and even the use of activated nucleotides has only achieved replication of less than 10 base pairs [66,72]. Replication fidelity is also problematic, mismatching in Watson-Crick pairing commonly causes errors in replication [73].

Non-enzymatic synthesis of a complementary strand on a template also leads to formation of a double strand. Double strands are stable under conditions in which the templating reaction occurs. This suggests that a temperature cycling process is required to separate the double strands and allow further cycles of replication. However, on cooling, reannealing of existing strands is rapid compared to synthesis of new complementary strands, therefore replication driven by temperature cycling is blocked by reannealing. In a very simple model dealing with replication of a single type of plus and minus strand, reannealing inhibits replication at a very low strand concentration [75]. However, the possibility remains that in a diverse mixture of random sequences, reannealing of partially matching strands can lead to configurations in which productive primer extension and ligation can continue to occur [76]. It is known that functional ribozymes can be assembled by ligation

in sequence mixtures that contain all the required fragments of the ribozymes [77,78] and it was suggested [76] that sequence information could be encoded in a mixture if the fragments formed a virtual circular genome. However, recent computer simulations of non-enzymatic synthesis in mixtures of RNA fragments [78] show that functional sequences become scrambled, even in the limit of zero mutational error, and that sequence information cannot be passed on.

Some experimentalists have proposed the use of different nucleic acid polymers generically referred to as XNAs to aid in template copying. There are several of these polymers, each differing from RNA in their base or sugar configuration. Some XNAs have been observed to increase the replication of RNA templates and may also interact with DNA and proteins [74]. Though this makes XNAs useful to an RNA world, it is unlikely that an XNA could replace RNA as the more likely first molecule of life as some have suggested. To date no XNA has been observed to perform catalytic functions in the absence of RNA, and no examples of XNAs in life exist. Therefore it seems unlikely life would create an XNA world, pass all functions to DNA,RNA and proteins, then leave no traces of XNAs in modern life.

1.4 Rolling circles and viroids

Considering the issues facing RNA replication, work done by Andrew Tupper within our group has suggested that the rolling circle mechanism is a likely way to get replication started in the RNA World [75]. In rolling circle replication, multiple copies of a complementary strand are synthesized by repeatedly going around the same template strand. The growing strand contains a self-cleaving hammerhead ribozyme (HHRz) which cleaves the tail off the growing strand at a set position

within the template. The linear strands produced by the cleavage have the ability to re-circularlize and reinitiate rolling circle replication. Thus rolling circles are not inhibited be annealing such as with linear strands. In chapter 2 we present some original work, which expands on the model presented [75]. The primary difference is the addition of annealing between linear segments and an explicit differentiation between linear and circular strands. Ordinary differential equations are used to show that even with the added mechanism of annealing between linear strands, exponential growth of strands occurs in rolling circle replication within realistic concentrations.

The smallest replicating RNAs in modern biology are viroids that use the rolling circle mechanism [79]. Though viroid replication depends on protein polymerases, ribozyme-driven replication on circular templates has also been shown in the laboratory [58,80]. Non-enzymatic strand displacement has also been shown to some extent in the laboratory, but is very slow [81]. Rolling circle replication does not require temperature-cycling to melt double strands, because the old complementary strand is gradually displaced from the template at the same time the new strand is synthesized. If non-enzymatic replication via the rolling circle mechanism can be achieved, as was proposed [75], this would be a likely point of origin of RNA replication, because there would be a direct pathway of evolution from non-enzymatic mechanisms to ribozymes and eventually to protein polymerases. This scenario is supported by recent experiments [82] showing that circular strands can arise during random polymerization of nucleotides.

Many types of self-cleaving ribozymes are known [42,48,83–86] but we are most interested in the hammerhead ribozyme because of its use in naturally-occurring

viroids, and because it is short and simple. The hammerhead is thought to have arisen multiple times in evolution and arises relatively easily from random RNA pools in in vitro selection experiments [87]. If circular strands could be replicated non-enzymatically, then we would argue the first kind of ribozyme that would be required in the history of life would be a hammerhead, which would be much easier to evolve than the more complicated polymerase ribozyme. Because of the limitations of experimental RNA replication, a theoretical model is required to provide support for this hypothesis.

1.5 Theoretical models for gene replication in an RNA world

In chapter 3 we discuss a computational model which assume that non-enzymatic replication of small circles containing a hammerhead ribozyme is possible. We ask whether ribozymes of some beneficial function can be encoded on such circles. The beneficial function could be a polymerase ribozyme which increases the rate of replication above the non-enzymatic rate, or a nucleotide synthetase ribozyme that increases the availability of monomers, or a ribozyme involved in lipid synthesis that increases the availability of membrane lipids. We will call the complimentary sequence of such a ribozyme a gene because it holds the information to make catalytic molecules. We understand that since these coding RNA sequence do not need to be translated like DNA genes they are different, but will refer to them as genes for simplicity. Several computational models for the cooperation of genes of these types have already been studied [88–91]. These models are the latest in a line of theoretical models for replication that started with Manfred Eigen in 1971

[92].

Eigen postulated that low single unit replication fidelity in non-enzymatic replication would limit the information that could be passed down before the origin of life. The lower the fidelity, the smaller the sequence which could be maintained without disappearing, and the simpler the enzymatic molecule that could be made from it. Though this applies to any method of replication which relies on separate information transfer and enzymatic activity molecules, in the context of the RNA world this "error threshold" limits the maximum RNA sequence size which can be maintained by a population. Two issues can arise from this depending on the severity of replication infidelity. The first and most problematic, occurs when fidelity is so low that only very short sequences can be replicated. In this limit single ribozymes cannot be maintained in the population because most exceed dozens of base pairs in length. Unless several polymerases appear from random polymerization at the same time and can act on each other to achieve replication (which is very unlikely), there can be no sustained replication of ribozymes and the RNA world is near impossible.

It has been argued this first scenario would not occur in the RNA world. Ribozyme function is determined by secondary and tertiary structures which can often be fulfilled by a variety of similar sequences [93]. Base-pair mispairing caused by low fidelity can therefore lead to neutral or benign mutations which do not affect or simply lower the activity of a ribozyme. Studies of these "phenotypic" error thresholds depend on the structure of specific ribozymes and have been found to be encouragingly high in some cases [94]. However, knowledge of the structure and stability of a ribozyme is needed to accurately predict individual thresholds

and only a few ribozymes are well studied enough to judge.

However replication fidelity could still be an issue if it is high enough to produce individual ribozymes, but too low to reproduce a long genome containing several [92]. This scenario may lead to fatal competition between ribozymes, in which only one ribozyme or non-coding sequences may survive. Eigen believed that this issues could be avoided by the implementation of hypercycles: a network of molecules which could catalyse each other's formation. In the context of RNA world, replicators can be sets of ribozymes which depended on each other for replication however, replicators are mostly treated as mathematical objects within a complex set of differential equations which compose the hypercycle, and rarely hold any characteristics unique to specific molecules. Despite some interesting qualities, hypercycles have been seen to only succeed if networks of replicators are small. Issues also arise through mutations, since hypercycles are only stable if replicators only catalysed the creation of one other replicator in the network [95]. This means that mutations which could create sequences not included in the hypercycle could disrupt the cycle. Furthermore this limits the use of a polymerase ribozyme in hypercycle models, since polymerases would need to be able to catalyse the replication almost any sequence, not just a select number of sequences in a cycle, to be useful in the RNA world.

The promiscuity of RNA polymerase also highlights the need for stochastic, instead of deterministic, simulation of replicative processes. In stochastic models, reactions occur randomly from a distribution which is determined from some appropriate algorithm (which we will discuss later). Though stochastic models can give similar results to differential equations, there is no master equation guiding

the concentrations of reactants. Instead, each reaction has a chance of occurring at any given point proportional to its rate and the sum of other rates in the near vicinity. Reactions are carried out between ribozymes and other non enzymatic molecules available in a vicinity. This is useful because at a molecule-to-molecule level polymerases act on strands which are closeby, meaning there is a spatial limit to the effect they can have in a population. This is important because we expect that new ribozymes are rare and only appear a few at a time. Therefore with something like a polymerase only some places within a solution would benefit from increased replication at the beginning. This introduces a level of chance regarding the success of the polymerase. For example if polymerases find themselves far from their complimentary sequences, they will act on other strands around them and forgo replicating themselves. This has been shown to kill off replication polymerases in computer models, where short non-coding sequences act as parasites [97,98]. These models have shown the importance of spatial clustering or compartmentalization to the survival of ribozymes like the polymerase [99].

In our own model shown in chapter 3 we use a simple stochastic model to show how errors during rolling circle replication can provide an environment for new ribozymes to spread. We use a Gillespie algorithm to determine the distribution which guides our stochastic model. In the Gillespie algorithm every reaction in a vicinity has a probability of occurring which is given by:

$$\frac{a_j}{\sum_j a_j} \tag{1.1}$$

Where a_j is the rate of the j^{th} reaction and $\sum_j a_j$ is the sum of all reactions. A

random number x on the interval [0,1] is chosen and the smallest j' reaction which satisfies:

$$x \sum_{j} a_j < \sum_{1}^{j'} a_j \tag{1.2}$$

Is chosen to occur. Given that reactions are chosen randomly we can apply the Monte-Carlo probability p for a reaction j over a time interval τ :

$$p(j, \tau) = a_j e^{-\tau \sum_j a_j} \tag{1.3}$$

We can re-arrange this to get the time to a next reaction:

$$\tau = \frac{-1}{\sum_j a_j} ln \frac{a_j}{p} \tag{1.4}$$

where $\frac{a_j}{p}$ can be taken to be a random number on the interval [0,1]. This algorithm is particularly suited for modeling a set of interconnected reactions where fluctuations can have an impact [100].

In our simulation, we suppose that circles are contained in lipid vesicles (protocells) that can divide when the RNA strands are multiplying within them. Beneficial genes that increase the rate of RNA replication by some means will also increase the rate of cell growth and division and will hence be selected. This is because of evolutionary protocell models which show that genes with beneficial functions can be selected in protocells, whereas they would not be selected in a

well-mixed system without compartments [99]. In chapter 3 we use this model to argue that rolling circle replication, through use of the HHRz, can aquire and maintain beneficial mutations.

Experiments exploring the stochastic replication of ribozymes in space, are closely related to a class of models called "stochastic correctors" which were proposed in the late 80's [96]. These models (in the context of the RNA world) simulate prebiotic cells filled with a variety of RNA sequences which divided as the number of sequences inside them grow. Enzymatic and non-enzymatic reactions occur stochastically and as cells divide, their contents are partitioned randomly between two daughter cells. These models are useful in exploring how gene co-operation can overcome issues with replication fidelity and may lead to the creation of genenome. Even though ribozymes can compete within each cell in stochastic corrector models, cells containing a variety of sequences can sometimes divide faster and become dominant in a population. This does not occur through a cycle where ribozymes act on each other, but from the added benefits of different ribozymes contributing to different functions [94]. Overall, these models are better at capturing the random nature of replication at small concentrations than the differential equations used in hypercycles. They can properly explore the evolutionary path of protocells by simulating new ribozymes and may also model spatial effects which can impact replication as previously mentioned [97,98]. In particular recent stochastic corrector models have focused on the relationship between ribozymes with unrelated functions. One paper focusing on lipid synthesis and RNA polymerization [88] showed that compartmentalization in protocells was important for the survival and co-operation between different ribozymes. This

paper showed that diffusion and concentration can disrupt the ability of different ribozymes to benefit from each other, but the effects can be mitigated by encapsulation in protocells. Stochastic models have also been used to simulate the use of rolling circles as replicators of a diverse set of ribozymes [91] and as a means of encoding several ribozymes on a singular genome [89]. In chapter 4, we explore how the model proposed in chapter 3 can be used to show wether rolling circles can aid in genome creation. Before this in chapter 2, we will talk about a basic deterministic model which expands on the work done in [75]. This work shows that annealing does not impede exponential growth of rolling circles in a bulk solution.

Chapter 2

Deterministic Model for Growth Rate of Rolling Circle Replication

Non-enzymatic replication of RNA is a requirement for the creation and propagation of ribozymes in the RNA world. Current methods of non-enzymatic RNA replication use linear templates and have experienced some success. However temperature cycling is required to separate linear templates from products, and long strands made in this way will anneal quickly to templates once separation temperatures are relaxed. This imposes a limit on the length of strands created by linear templation. Tupper and Higgs [75] show with a mathematical model that this inhibition problem causes strand concentration to grow non-exponentially above a certain concentration of strands. This is important because failure to grow exponentially condemns strands to become diluted out or degrade to extinction. Tupper and Higgs [75] propose that this can be avoided through the use of rolling circle replication, in which a circular template is replicated through strand displacement.

This forgoes the need for temperature cycling and avoids product inhibition. Circular templates may also begin replication anywhere and avoid end degradation. We present a slightly more complex version of the work done in [75] which solidifies the claim rolling circles can avoid being slowed critically be strand inhibition.

2.1 Simple equations for rolling circle replication

In [75], Tupper and Higgs propose a simple ordinary differential model for the concentration of strands undergoing replication through strand displacement. The system of equations they propose is:

$$\frac{dP}{dt} = -R_{syn}P + \frac{1}{2}R_{dis}D - R_{ann}PM \tag{2.1}$$

$$\frac{dM}{dt} = -R_{syn}M + \frac{1}{2}R_{dis}D - R_{ann}PM \tag{2.2}$$

$$\frac{dD}{dt} = R_{syn}(P+M) + R_{ann}PM \tag{2.3}$$

where P is the concentration of positive strands, M is the concentration of minus strands and D is the concentration of double strands. The rates are:

1. R_{syn}: rate at which RNA monomer addition and polymerization occurs on circular templates

2. R_{ann}: rate at which strands (measuring 100bp) anneal to each other.

3. R_{dis}: rate of linear strand displacement from a rolling circle by toe-hold interactions

Higgs and Tupper propose that when annealing is negligible the concentration of $D = D_o e^{\gamma t}$ increases proportional to:

$$\gamma = \frac{R_{syn}}{2}\left(-1 + \sqrt{1 + \frac{4R_{dis}}{R_{syn}}}\right) t \tag{2.4}$$

If we take a first order approximation:

$$\sqrt{a^2 + b} \approx \; a + \frac{b}{2a} \tag{2.5}$$

Where $a = 1$ and $b = \frac{4R_{dis}}{R_{syn}}$ then the above equation can be simplified to:

$$\gamma = R_{dis} \tag{2.6}$$

On the other hand when annealing is not negligible Tupper and Higgs propose the concentration grows proportional to:

$$\gamma = \frac{1}{2} R_{dis} \tag{2.7}$$

Which is smaller than before but unlike models of linear templated growth without displacement, is still exponential. Linear templates undergoing strand displacement have a high likelihood of not completely copying a template. Tupper and Higgs argue that on a linear template, new strands will constantly be displaced

by the tail end of the old strand as replication occurs. This is caused by re-annealing of the old strand, which is more favourable than the annealing of the displacing strand due to possessing more base pairs. Tupper and Higgs suggest that this does not occur for displacement on a circular strand, because the growing end is part of the old strand. On a circular template displacement does not push the old strand out, just exposes it into a hanging tail. In rolling circle replication this tail is cleaved off by the hammerhead to create a new strand.

2.2 Expanded equations

We expanded on the equations above by more closely applying characteristics of rolling circle replication. To do this we differentiate between linear and circular strands and allow annealing between all kinds of strands. Annealing between a circle and a line creates a circle which can undergo rolling circle replication, annealing between two lines creates a strand which can no longer replicate. We want to see if annealing between the lines causes replication to stall below an exponential level. The system of ordinary differential equations describing rolling circle replication is therefore as follows:

$$\frac{dL_p}{dt} = Q_m R_{dis} + C_p R_{cle+} - L_p R_{circ+} - C_m L_p R_{ann} - L_m L_p R_{ann} \tag{2.8}$$

$$\frac{dL_m}{dt} = Q_p R_{dis} + C_p R_{cle-} - L_m R_{circ-} - C_p L_m R_{ann} - L_m L_p R_{ann} \tag{2.9}$$

$$\frac{dC_p}{dt} = L_p R_{circ+} - C_p(R_{cle+} + R_{syn}) - C_p L_m R_{ann} \tag{2.10}$$

$$\frac{dC_m}{dt} = L_m R_{circ-} - C_m(R_{cle-} + R_{syn}) - C_m L_p R_{ann} \tag{2.11}$$

$$\frac{dQ_p}{dt} = C_p R_{syn} + C_p L_m R_{ann} \tag{2.12}$$

$$\frac{dQ_m}{dt} = C_n R_{syn} + C_m L_p R_{ann} \tag{2.13}$$

Here, L refers to single linear strands, C to single circular strands and Q to double strands undergoing rolling circle replication. This model makes the distinction between plus and minus strands labeled with a subscript p for plus and m for minus. Furthermore several rates are defined below. Strands are expected to measure around 100bp which is taken into consideration for the annealing rate used.

1. R_{circ+}/R_{circ-}: rate at which linear plus/minus strands fold into their circular form through end ribozymes

2. R_{cle+}/R_{cle-}: rate at which the ribozyme in a circular plus/minus strand cleaves into their linear form

The specific values of these constants becomes relevant because certain numbers allow us to simplify the solutions to these equations. The values used are provided

Rsyn [106]	Rdis [81]	Rann [107]	Rcirc(+/-)	Rcle(+/-)
$10^{-2}hr^{-1}$	$10^{-4}hr^{-1}$	$10^{10}hr^{-1}$	$10^{-2}hr^{-1}$ *to* $10hr^{-1}$	$10^{-2}hr^{-1}$ *to* $100hr^{-1}$

TABLE 2.1: Values for constant rates used. R_{ann} is given for a RNA strand on the order of 100bp which is a suitable length containing two complimentary hammerhead rybozymes needed for rolling circle replication. Also notice that the R_{circ} and R_{cle} values are the same for the plus and minus strands, there are no known cleaving and cricularising rates for the rybozyme we have proposed, as such a range of possible values from literature is used

.

in table 2.1, R_{circ} and R_{cle} are given as possible ranges. R_{dis} is taken from [81] and R_{syn} from [106]. R_{ann} is taken from [107]. From our chosen values we can see that $R_{dis} << R_{cle}$ & R_{syn} which means the displacement of the linear component of the rolling circle is the limiting step in the process of creating linear strands from rolling circles. With this assumption the creation of linear strands is driven by the term $Q_m R_{dis}$. We also assume that annealing can only happen between a pair of linear strands or a pair of linear and circular strands. We do not include a term for the annealing of two fully circular strands.

2.3 Simplified equations

In order to analyse the growth that this model might suggest we simplify equations (2.8-2.13) by combining terms for plus and minus strands. Specifically:

1. $R_{circ+} = R_{circ-} = R_{circ}$

2. $R_{cle+} = R_{cle-} = R_{cle}$

3. $L_p = L_m$

4. $\mathrm{L} = L_p + L_m$ (or $\mathrm{L} = 2L_m$)

5. $C_p = C_m$

6. $\mathrm{C} = C_p + C_m$ (or $\mathrm{C} = 2C_m$)

7. $Q_p = Q_m$

8. $\mathrm{Q} = Q_p + Q_m$ (or $\mathrm{Q} = 2Q_m$)

which allows us to ruduce our equations to:

$$\frac{dL}{dt} = QR_{dis} + CR_{cle} - L(R_{circ} + R_{syn}) - \frac{1}{2}CLR_{ann} - \frac{1}{2}L^2R_{ann} - LR_{hyd} \quad (2.14)$$

$$\frac{dC}{dt} = LR_{circ} - C(R_{cle} + R_{syn}) - \frac{1}{2}CLR_{ann} - CR_{hyd} \quad (2.15)$$

$$\frac{dQ}{dt} = CR_{syn} + \frac{1}{2}CLR_{ann} - R_{loss}Q \quad (2.16)$$

notice that for the addition of the $C_mL_pR_{ann}$ and $C_pL_mR_{ann}$ terms we have

$$\begin{aligned} C_m L_p R_{ann} + C_p L_m R_{ann} &= R_{ann}(C_m L_p + C_p L_m) \\ &= R_{ann}(L_m(C_m + C_p)) \\ &= R_{ann}(L_m C) \\ &= \frac{1}{2} L C R_{ann} \end{aligned} \tag{2.17}$$

similarly for $L_m L_p R_{ann}$ and $L_p L_m R_{ann}$

$$\begin{aligned} L_m L_p R_{ann} + L_p L_m R_{ann} &= R_{ann}(L_m L_p + L_p L_m) \\ &= R_{ann}(2 L_m L_p) \\ &= R_{ann}(2 L_m^2) \\ &= \frac{2}{4} L^2 R_{ann} \\ &= \frac{1}{2} L^2 R_{ann} \end{aligned} \tag{2.18}$$

We can simplify the equations further by combining the C and L states into a new state S comprising all single strands, within this state there is some probability an S strand is linear and a probability an S strand is circular. Specifically $L = f_{lin}(S)$ and $C = f_{circ}(S)$ while $f_{lin} = 1 - f_{circ}$.

using these parameters we can write eq (2.14) + (2.15) and (2.16) as:

$$\frac{dS}{dt} = R_{dis} Q - R_{syn} S - f_{circ} f_{lin} R_{ann} S^2 - \frac{1}{2} f_{lin}^2 R_{ann} S^2 \tag{2.19}$$

$$\frac{dQ}{dt} = f_{circ} R_{syn} S + \frac{1}{2} f_{circ} f_{lin} R_{ann} S^2 \tag{2.20}$$

From these equations we can approximate the rate of growth of the strands in three separate phases, and the product concentrations at which there is transition between them. Generally speaking, the characteristics of these phases are controlled by the three negative terms in eq (2.19). In the first phase occurs when the concentration of strands is low. In this case the S^2 annealing terms are too small to contribute to growth, and can be ignored. Here R_{syn} dominates and we can simplify equations (2.19, 2.20) to:

$$\frac{dS}{dt} = R_{dis} Q - R_{syn} S \tag{2.21}$$

$$\frac{dQ}{dt} = S R_{syn} f circ \tag{2.22}$$

We can solve this system of differential equations to find the approximate growth rate in this non-annealing phase. In matrix form this system looks like:

$$\begin{bmatrix} -S R_{syn} & R_{dis} \\ S R_{syn} f circ & 0 \end{bmatrix}$$

the determinant is:

$$\begin{bmatrix} -S R_{syn} - \lambda & R_{dis} \\ S R_{syn} f circ & -\lambda \end{bmatrix}$$

which gives the solution:

$$\begin{aligned} 0 &= (-R_{syn} - \lambda)(-\lambda) - R_{dis}R_{syn}f_{circ} \\ &= \lambda^2 + R_{syn}\lambda - R_{dis}R_{syn}f_{circ} \\ \lambda &= \frac{1}{2}(-R_{syn} \pm \sqrt{(R_{syn}f_{circ})^2 + 4R_{dis}R_{syn}f_{circ}}) \\ \lambda &= \frac{R_{syn}f_{circ}}{2}\left(\frac{-1}{f_{circ}} \pm \sqrt{1 + \frac{4R_{dis}}{R_{syn}f_{circ}}}\right) \end{aligned} \tag{2.23}$$

We take the positive value of λ to be the growth rate of linear and rolling circle strands in the solution of the form:

$$Q = Q_0 e^{\lambda t} \tag{2.24}$$

$$S = S_0 e^{\lambda t} \tag{2.25}$$

When the concentration of strands is below some critical concentration where annealing is negligible.

In the case that $R_{syn}f_{circ} >> 4R_{dis}$, which is likely when $f_{circ} \sim 1$, the equation for λ can be simplified using the Taylor series approximation:

$$-1 + \sqrt{1+x} \approx \frac{x}{2} - \frac{x^2}{8} + \frac{x^3}{16} \tag{2.26}$$

In which case:

$$\lambda \approx \frac{R_{syn}}{2}(\frac{1}{2}\frac{4R_{dis}}{R_{syn}}) = R_{dis} \tag{2.27}$$

This is similar to the approximation for the growth proposed in [75] which we showed in equation (2.6) for a double strand.

2.4 Critical concentration

Transition to the second phase of growth occurs when the annealing terms are no longer negligible. From equation (2.20) we can see the annealing term grows quadraticaly while the synthesis term grows linearly with S. We can define a point of equilibrium for the annealing and synthesis term:

$$\begin{aligned} \frac{1}{2}S^2 R_{ann} f_{lin} f_{circ} &= S R_{syn} \\ S^* &= 2\frac{R_{syn}}{R_{ann} f_{lin} f_{circ}} \end{aligned} \tag{2.28}$$

Where S^* is the critical concentration of single strands where the growth of rolling circles and linear strands goes from being synthesis dominated to annealing dominated. In order to approximate this concentration in terms of Q we can write:

$$\begin{aligned} \frac{S}{Q} &= \frac{R_{dis}}{\lambda + R_{syn}} \\ &\approx \frac{R_{dis}}{R_{syn}} \end{aligned} \tag{2.29}$$

Which lets us modify (2.21) into:

$$Q^* = \frac{2R_{syn}^2}{R_{dis} R_{ann} f_{lin} f_{circ}} \tag{2.30}$$

Because our equations assume that the rolling circle replication process is limited by the displacement rate and not the cleaving of the hanging strand from the rolling circle, $f_{lin} = 0$ is not a value that is within our model since we will always have $0 < f_{lin}$.

2.5 Growth during the annealing dominated phase

At the critical concentration defined in equation (2.30) the annealing between single strands is no longer negligible, and we can drop the synthesis terms and add a constant term for simplicity to get:

$$\frac{dS}{dt} = R_{dis} Q - \alpha S^2 \tag{2.31}$$

$$\frac{dQ}{dt} = \beta S^2 \tag{2.32}$$

where

$$\alpha = (f_{circ} + \frac{1}{2} f_{lin}) f_{lin} R_{ann} \tag{2.33}$$

$$\beta = \frac{1}{2} f_{lin} f_{circ} R_{ann} \tag{2.34}$$

We can approximate the equations for Q, S and S^2 as a series of exponentials:

$$Q = Ae^{\gamma t}(1 + q_1 e^{-\frac{\gamma t}{2}} + q_2 e^{-\gamma t} + ...) \tag{2.35}$$

$$S = Be^{\frac{\gamma t}{2}}(1 + s_1 e^{-\frac{\gamma t}{2}} + s_2 e^{-\gamma t} + ...) \tag{2.36}$$

$$S^2 = B^2 e^{\gamma t}(1 + 2s_1 e^{-\frac{\gamma t}{2}} + ...) \tag{2.37}$$

For this approximation to work we only need the first two terms and drop the rest. The derivative for Q and S in this exponential form are therefore:

$$\frac{dQ}{dt} = A(\gamma e^{\gamma t} + q_1 \frac{\gamma}{2} e^{\frac{\gamma t}{2}}) \tag{2.38}$$

$$\frac{dS}{dt} = B(\frac{\gamma}{2}e^{\frac{\gamma t}{2}} + 0) \tag{2.39}$$

The first two terms of (2.35) and (2.4) can be substituted into (2.31) and (2.32) to get expressions for $\frac{dQ}{dt}$ and $\frac{dS}{dt}$ in terms of γ, t, A and B. We can then equate these expressions to (2.38) and (2.39) to get the equations:

$$A(\gamma e^{\gamma t} + q_1\frac{\gamma}{2}e^{\frac{\gamma t}{2}}) = \beta B^2 e^{\gamma t}(1 + 2s_1 e^{-\frac{\gamma t}{2}}) \tag{2.40}$$

$$B\frac{\gamma}{2}e^{\frac{\gamma t}{2}} = -\alpha B^2 e^{\gamma t}(1 + 2s_1 e^{-\frac{\gamma t}{2}}) + R_{dis}A(e^{\gamma t} + q_1 e^{\frac{\gamma t}{2}}) \tag{2.41}$$

From here we can rearrange to solve for the coefficients that precede the $e^{\gamma t}$ and $e^{\frac{\gamma t}{2}}$ terms in both equations and solve the resulting system of equations for γ:

$$\begin{aligned}\gamma &= \frac{\beta R_{dis}}{\alpha} \\ &= \frac{f_{circ}R_{dis}}{2f_{circ} + f_{lin}}\end{aligned} \tag{2.42}$$

as well as writting out a relationship between B and A:

$$B = \sqrt{\frac{AR_{dis}}{f_{lin}(f_{circ} + \frac{1}{2}f_{lin})R_{ann}}} \tag{2.43}$$

Now we can write a leading term approximation for the concentration of rolling circles and single strands in the annealing dominated phase:

$$Q = Ae^{\gamma t} \tag{2.44}$$

$$S = Be^{\frac{\gamma t}{2}} \tag{2.45}$$

Notice that if we make A the critical concentration equation (2.30) then we can get a full model approximation of the strands up to the point where annealing dominated growth transitions into the next phase of annealing impeded growth. Furthermore if we take the limit where all strands go to circles $f_{lin} = 0 \quad f_{circ} = 1$, then $\gamma = \frac{1}{2}$ we get the results from equation (2.7) for Q. This makes sense because the model proposed in [75] does not take into account linear fragments and has the limit described above.

To determine the concentration at which the annealing rate of linear strands becomes dominant and impedes exponential growth, we look at when the LR_{circ} term becomes smaller than the $\frac{1}{2}L^2R_{ann}$ term in equation (2.14). By equating and rearranging the terms we can see this concentration is:

$$L^* = \frac{2R_{circ}}{R_{ann}} \tag{2.46}$$

Since we are still near a point of constant change, the number of linear strands is proportional to the number of rolling circles meaning:

$$\begin{aligned} \frac{L^2}{Q} &\approx \frac{R_{dis}}{R_{ann}} \\ Q^{**} &= \frac{4R_{circ}^2}{R_{ann}R_{dis}}, \end{aligned} \tag{2.47}$$

Above this concentration the annealing rate of linear strands to each other is greater than the rate of circularisation. This effectively removes the ability of our linear strands to transition between lines and circles resulting in the disequilibrium of these two states. The effect of this is that the growth of the strands is no longer exponential meaning degrading forces could make the circles disappear.

we do not explore the phase where linear strand annealing become relevant, because it requires very high concentrations which exceed the concentrations of modern cells.

2.6 Results

We can use these equations to show and approximate the solutions to a system of rolling circles. Equations (2.19) and (2.20) where approximated using numerical model to find the concetrations of Q and S. A Radau implicit Runga-Kutta method was used for this numerical solution because it dealt with the large difference in constants we used. A scenario was modeled using parameters as described in Table 1 with a logarithmic y-axis. The simulation also used $R_{circ} = 1$ $R_{cle} = 1$. The concentration given by the numerical solution are shown in black for Q and red for S. For each graph eight other lines are included. The black boxes labeled "non-annealing", "annealing" and "sub exponential growth" show phases described

by equations (2.30) and (2.47) these points show were the phase changes discussed above happen. Each graph also has four diagonal dotted lines, these show the expected exponential rate of growth that we approximated in equations (2.27) and (2.42) for Q and S.

This graph shows a similar conclusion to the work done in [75], which postulated that exponential growth occurred in rolling circle replication even at high concentrations. Here we have shown that even with a slightly more complex model which takes into account linear and circular strands as well as annealing between linear strands, exponential growth still occurs below expected prebiotic concentrations. Specifically we can see from the three graphs that in all combinations of circularisation and linearisation, the transition of Qs into the third non-exponential phase occurs above a concentration of 10^{-6} which is a concentration that is only expected in modern cells. Because of this it seems likely that rolling circles could have experienced exponential growth at prebiotc concentrations.

In this chapter we have proposed a deterministic model for rolling circle replication without mutations. Our goal was to determine if annealing would impede the replication of strands, our findings suggest that this only becomes an issue at concentrations too high to expect at the origin of life. In the next chapter we present a stochastic model for rolling circle replication, we use it to explore how a gene can become fixed in a population cells containing rolling circles which are allowed to mutate. Since the conclusions of this chapter show annealing is not an issue, we exclude it from the model in the next chapter.

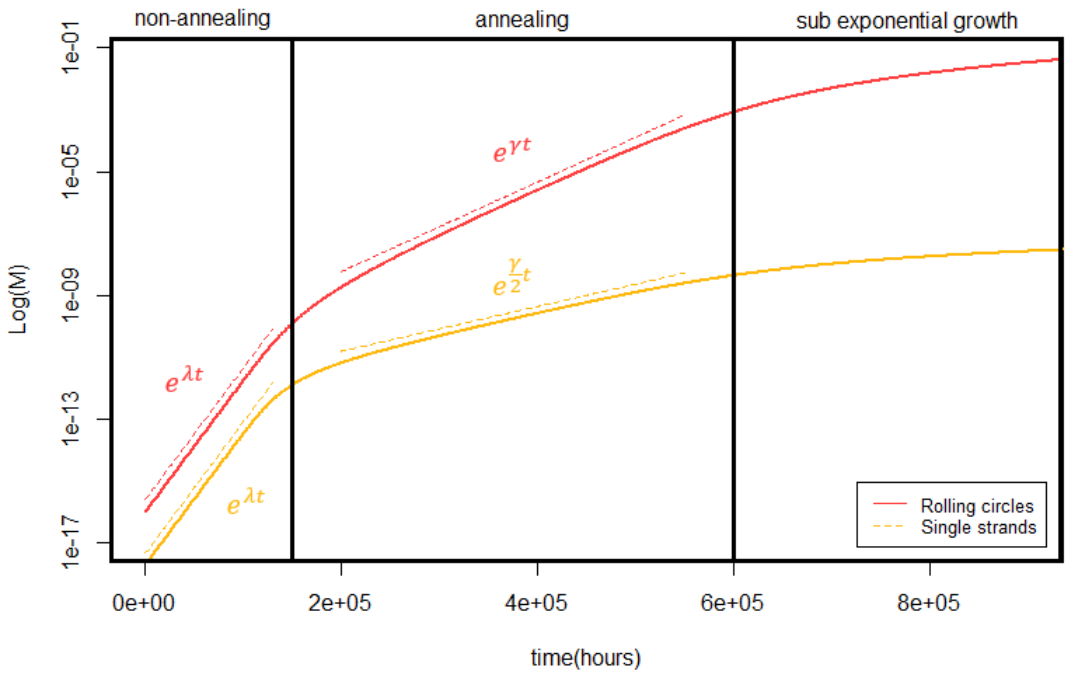

FIGURE 2.1: Graphs showing the growth of rolling circles. Growth is always exponential for circles Q unles concentration rises past 10^{-4} which is higher than what we think is possible in an RNA world

Chapter 3

Rolling Circles as a Means of Encoding Genes in the RNA World

3.1 Summary

The case of ribozymes on circular chromosomes was studied in [89], where it was argued that circular chromosomes have advantages over linear chromosomes because replication can start at any point in the sequence and because the ends of linear chromosomes tend to degrade. Our own reasons for favoring circular chromosomes are somewhat different. Others in our group have argued in the past that temperature cycling mechanisms to separate double strands are not efficient, either because they are inhibited by reannealing [75], or they cannot replicate sequence information [78], and therefore, that a strand displacement mechanism is required that does not require temperature cycling. The work done by Tupper

and Higgs [75], showed that strand displacement from a linear template is very difficult because the short growing complementary strand is repeatedly displaced by the pre-existing complete complementary strand before it can reach completion. Replication on a circular template avoids this problem. Therefore, in the model we will show, we assume that only circular strands can replicate. Linear strands which are cleaved from the rolling circle have some probability of re-circularizing and becoming templates and some probability of forming folded structures that cannot replicate, but which can be functional ribozymes.

Selection for rapid replication tends to reduce template length to the minimum, and therefore acts against encoding beneficial genes, as in previous models [88,89]. However, we show that this is countered by an inherent mechanism in rolling circles that leads to length doubling. Cleavage of the tail from a rolling circle can only occur if the sequence of the hammerhead ribozyme is correctly copied from the template. If a sequence error occurs during copying of this ribozyme, the tail grows longer until the same point is reached on the subsequent cycle. This time, if the sequence is correctly copied, a double-length strand is released, which can form a double-length circle. We firstly show results from a paper under review which investigates the distribution of sequence lengths expected under the action of this doubling mechanism. The paper then discuss the requirements for a longer circle containing a beneficial gene to become fixed within a population.

3.2 Methods

3.2.1 Basic mechanism of rolling circle replication

We simulate a population of RNAs with differing lengths and sequences. Rather than store full RNA sequences at the single base level, we represent sequences as strings of characters where each character represents an RNA section of length roughly 25 bases. The hammerhead ribozyme (HHRz) is represented as two characters AZ, where A and Z are the 5' and 3' parts of the ribozyme that are formed when it cleaves. A HHRz is typically 50 nucleotides long; hence our choice of 25 for the length represented by one character. We use the * character to indicate an RNA section with no particular sequence or function. Any sequence, such as ****AZ**, which contains the AZ motif can cleave between the A and the Z. The HHRz is able to re-ligate, leading to the formation of a new circular strand [48]. In our model, any sequence, such as Z*******A, with Z at its 5' end and A at its 3' end can circularize. Copying A and Z motifs creates their complementary sequences, denoted by a and z. Copying a and z recreates A and Z. Replication proceeds in the reverse direction (3' to 5' on the template), therefore the complement of AZ is za. We assume that za is not itself a HHRz, but it encodes a sequence that is complementary to the ribozyme. For repeated replication to occur, the sequence must contain both AZ and za; therefore, the minimal replicating circle has sequence ZzaA.

If we begin copying the template at z and proceed in the reverse direction around the circular template, then we generate a complementary sequence ZzaAZzaAZzaA.... which can cleave at each occurrence of the AZ motif to give multiple copies

of ZzaA. This can re-circularize and continue replication. This is illustrated in Figure 3.1. The ZzaA sequence is self-complementary at the level of the character representation, but it need not be at the single nucleotide level. For example, there may be two HHRz sequences, A1Z1 and A2Z2 with equivalent function but different base sequences. In this case, there would be plus and minus strands Z1z2a2A1 and Z2z1a1A2, illustrated by blue and green strands in Figure 3.1. Copying each of these gives rise to the other.

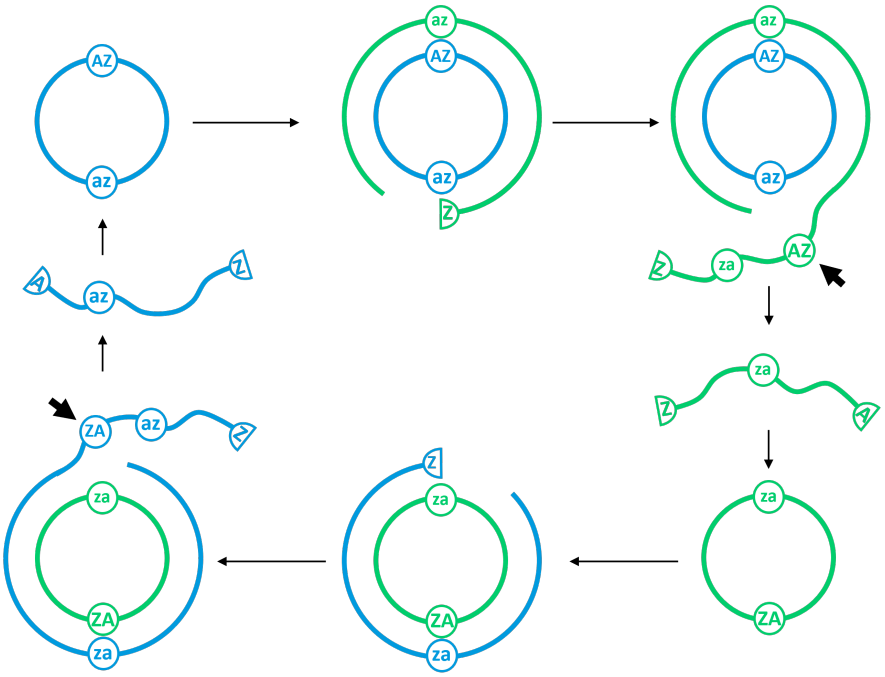

FIGURE 3.1: Mechanism of non-enzymatic rolling circle replication. Blue and green strands are complimentary plus and minus strands, each of which contains a ribozyme unit AZ and its complement za. After one circuit around the template, a double strand is created. After another circuit, a tail is produced. When the AZ motif is exposed in the tail, cleavage occurs, creating a linear strand that can circularize and begin the cycle anew.

3.2.2 Model details

Each RNA strand in the model is stored as a string of characters, as described above. A rate r_i is assigned to each sequence i in the model, which is the rate of the next event that can happen to that sequence. Events are of different types but there is always only one possible next event for each sequence. The method of Gillespie [99] is used to simulate the stochastic occurrence of events as described in chapter 1.

A strand may either be linear or circular. A linear strand is able to fold to a structure at rate R_{fold}; therefore we set the rate r_i to be R_{fold} for linear strands. For a linear strand that begins with Z and ends with A, the folding event brings the two ends together in the appropriate arrangement for ligation. When this event occurs, the linear strand becomes a single-stranded circle. A linear strand that does not have both Z and A at its ends cannot form a circle. When folding of such a strand occurs, it forms a folded linear strand that does not undergo any further events. In this case, we set $r_i = 0$, so it is never selected for further events.

The rate r_i for a single stranded circle is set to R_{syn} , the rate of synthesis of a single character in the complementary strand. When this event occurs, a random position is chosen on the template circle and a complementary sequence is initiated with the character that is complementary to the template character. An example is shown in Table 3.1, Event 1. The sequence before the colon denotes the template and the sequence after the colon denotes the complementary strand. The variable last is used to store the position on the template of the last character added. In this case the A that has been synthesized is complementary to the a in the template, which is at position 3 in the template; therefore last = 3.

TABLE 3.1: Example of a series of events occurring on a circular template. Sequences before the colon are template circles. Sequences after the colon are complementary strands. Underlined sequences in the complement are single-stranded tails.

Event	From state	Last	To state	Last	Rate
1	ZzaA	-	ZzaA:A	3	R_{syn}
2	ZzaA:A	3	ZzaA:AZ	2	R_{syn}
3	ZzaA:AZ	2	ZzaA:AZz	1	R_{syn}
4	ZzaA:AZz	1	ZzaA:AZza	4	R_{syn}
5	ZzaA:AZza	4	ZzaA:AZzaA	3	R_{dis}
6 cleavage	ZzaA:AZzaA	3	ZzaA:AZzaAZ A + ZzaA:ZzaAZ	2 2	R_{dis}
7 cleavage	ZzaA:ZzaAZzaA	3	ZzaA : ZzaAZzaAZ ZzaA + ZzaA:ZzaAZ	2 2	R_{dis}

When the complementary strand is shorter than the length of the template circle, we suppose that synthesis of the next complementary character occurs at the same rate $r_i = R_{syn}$. When this event occurs, an extra character is added to the complementary sequence, as in Event 2 in Table 3.1. As synthesis goes in the reverse direction on the template, last is decreased by 1. If last was 1 prior to the event, then it is set to the length of the template (because the template is circular). Synthesis can proceed by several events until the length of the complementary strand reaches the length of the template circle (Events 3 and 4).

At this point, further addition to the 3' end of the growing strand requires displacement of the 5' end. We set the rate to $r_i = R_dis$ in this case. We assume that the rate of synthesis involving strand-displacement, R_{dis}, is slower than the rate of synthesis on an unobstructed single strand, R_{syn}, as discussed by Tupper and Higgs [75]. The complementary strand produced after Event 5 has a single-stranded tail emerging from the circle. This is the initial A at the 5' end of the complementary sequence (underlined in Table 3.1). After one further event,

the tail has grown to AZ. We assume that the ribozyme cleaves instantaneously whenever the AZ motif arises in the free tail. In this case a single A is cleaved off. This A is a linear strand which cannot circularize. It will fold to form a folded linear strand that plays no active role. The complementary strand on the circle can continue to grow by strand displacement. After several more steps, Event 7 will occur, which leads to cleavage of a complete ZzaA, which will circularize and initiate replication. Thus the cycle is complete.

The first sequence to be cleaved off is usually an incomplete fragment that cannot circularize, as with the single A in this case. The length of this fragment will depend on the position at which the first synthesis event occurs on a new circle. Only if the new strand is initiated with the Z will the first sequence formed be complete. However, once one fragment is released, subsequent sequences cleaved will always be complete (i.e. equal length to the template).

3.2.3 Mutations

In the above examples, we assumed perfectly accurate sequence replication, so that each character in the template always created its complementary character in the complementary strand. However, in the simulations below, we allow deleterious mutations to occur with probability u per character. Characters A, a, Z, and z represent specific sequences with structure and function. Each of these is accurately replicated to its complementary character with probability $1-u$. With probability u, a deleterious mutation occurs. The new character is *, which represents a sequence with no function. If the template character is *, this is always copied to another *. We ignore mutations that create functional characters from

* because they are expected to be much rarer than deleterious mutations.

Mutations in the hammerhead motifs can create strands that have interesting properties. Suppose we are copying the template ZzaA, and a mutation occurs when either the z or the a is being synthesized. The resulting sequence is a 4-mer Z*aA or Zz*A. Both of these sequences can circularize because they have intact Z and A. However, if replication begins on these templates, the complement will not contain the AZ motif, therefore the tail will never cleave. Thus if a mutation occurs in the za motif, this leads to circles whose tails can never cleave. We call these non-cleaving circles (as shown in Figure 3.2).

On the other hand, if a mutation occurs when the AZ motif is being synthesized, this produces ZzaA* or Zza*Z in the growing tail. These tails do not cleave at the appropriate point because they do not contain AZ, however if there is no mutation on the next time around the circle, then we arrive at a tail such as ZzaA*zaAZ, which cleaves to give ZzaA*zaA. This is a double-length strand (an 8-mer) that can circularize. It contains only one functional AZ motif but two za motifs. We call this a halving circle (see Figure 3.2), because if replication begins on this 8-mer, it will cleave in two places, producing two 4-mers. If no further mutations occur, replication of ZzaA*zaA will produce the intact 4-mer ZzaA, and the non-cleaving sequence Z*aA.

If a second mutation occurs during replication of the halving sequence, then one of the two cleaving motifs can disappear. This produces an 8-mer such as Zza*Z*aA, that can accurately replicate, because it contains one AZ and one za. This sequence could also have been produced if two mutations occurred at once

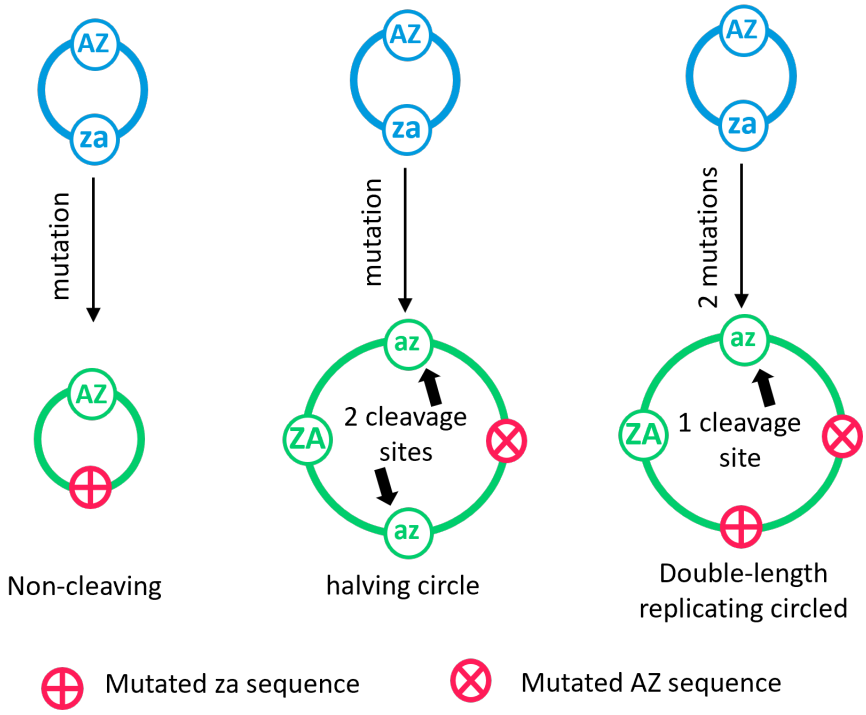

FIGURE 3.2: Mutations when copying minimal length circles give three kinds of strands with different behavior. A mutation in the za motif (either z*, *a, or **)gives a non-cleaving circle which produces a complementary strand that can never cleave. A mutation in the AZ motif creates a halving circle which goes on to produce circles of half its own length. A mutation in both motifs produces a circle of double the original length that can stably replicate.

during replication of the initial 4-mer. The remaining half motifs in the 8-mer will soon disappear by further mutation because there is no longer any selection acting on them. This would yield Zza****A. We see that mutations in the hammerhead motifs give a built-in mechanism by which longer-length rolling circles can be created. All possible multiples of the original 4-mer can be made. For example, 12-mers can be made directly from a 4-mer if there are mutation two passages around the circle but not on the third. 16-mers can be made by doubling an 8-mer, and so on. A minimal length 4-mer encodes only the ribozyme motifs necessary for

its own replication. However, if longer sequences arise containing non-functional * regions, it is possible for beneficial functional sequences to eventually arise in these regions by de novo mutation. This means that the initial 4-mers could evolve from simple minimal replicators to chromosomes that encode genes that are of useful function to the cell that contains them. We investigate the possibility of encoding beneficial genes further below, after first considering replication of circles without any other function.

3.2.4 Protocell compartments

We consider a system in which circles are replicating inside protocell compartments. Monomers are supplied from outside. The protocell membranes are assumed to be permeable to monomers but not RNA strands. Cell division is coupled to replication of the rolling circles. When the number of strands in any one cell reaches a maximum S_0, the cell divides in two, with strands distributed randomly between the two daughter cells. There are N memory slots defined for cells, which allows up to N cells to be present in the population. When a cell division occurs, one of the daughter cells remains in the memory slot of the parent, and the other one is placed in a random memory slot, overwriting whatever was in that slot. This insures that cells compete for limited resources. There is also a small loss rate R_{loss} at which cells become empty at random. We include this loss rate in addition to maintaining the limit of N cells in order to ensure that cells that cannot replicate are destroyed eventually. This guarantees the existence of an error threshold, i.e. a maximum mutation rate above which the whole population of cells dies out because the average cell division rate becomes slower than R_{loss}.

Compartments are often introduced into models of RNA replicators as a means of preventing the invasion of parasitic sequences. In models for polymerase ribozymes, parasites will destroy the polymerases in a well-mixed system, but the polymerases survive when replication occurs in compartments [31,98,100]. Selection at the level of compartments selects for cooperatively replicating sequences which enable rapid cell growth and division. This can overcome selection at the level of individual sequences, which favors rapidly replicating parasites. In the current model, however, we are assuming that replication is non-enzymatic, and it does not depend on cooperative polymerase ribozymes. A population of rolling circles replicating non-enzymatically would survive even in a well-mixed system, if the mutation rate were not too high to prevent replication of the hammerhead ribozyme. However, we are also interested here in the possibility of selection for beneficial genes encoded on the rolling circles. These genes cannot survive in a well-mixed system because any benefit they create would be diluted across the whole system and would give a very small benefit to all sequences equally, whether or not they encoded the beneficial gene. If protocell compartments are present, then the effects of a beneficial gene apply only to sequences in the same cell. The presence of a beneficial gene therefore benefits the sequence that encodes the gene relative to sequences in other cells that do not possess the gene. As the existence of protocells is essential for the evolution of beneficial genes, we will introduce protocells into this model right from the beginning.

The model was run with parameters shown in Table 3.2. We used values reported in the literature to approximate the orders of magnitude between R_{dis} and

TABLE 3.2: Standard values for model parameters

Reaction	Rate	Description
R_{dis}	$1hr^{-1}$	Rate of copying a single character when strand displacement is required
R_{syn}	$100hr^{-1}$	Rate of copying a single character when synthesis is possible without strand displacement
R_{fold}	$100hr^{-1}$	Rate at which a linear strand folds into its secondary structure
R_{loss}	$10^{-3}hr^{-1}$	Rate at which a cell dies from lack of materials
N	200	Number of protocells in population
S_o	20	Maximum number of strands per protocell

R_{syn}. The rate of non-enzymatic addition of a 25-base character with strand displacement, R_{dis}, was defined as our point of reference with a value of $1hr^{-1}$, and we assumed that the R_{syn} rate (which does not require strand displacement) is 100 times faster. When probing for an error threshold in Figure 3.4, the rate of cell death R_{loss} is defined as $10^{-3}hr^{-1}$, about 100 times less than the rate of cell reproduction observed later in Figure 3.6. This corresponds to one cell death for every 100 new cells created, which is reasonable for a healthy population. R_{loss} is otherwise set to 0 when u is small and we are not close to an error threshold because replication is much faster than R_{loss}. Similarly, we expect folding of a strand into its secondary structure to be a relatively fast event, so we also set R_{fold} to $100hr^{-1}$. We do not aim to create a time-accurate model with our parameters but aim to correctly identify fast and slow steps. Though we use hours as our time unit for simplicity, it is only the relative size of the rates that effects the outcome of our model.

We begin with a single circular ZzaA sequence per cell. After allowing the system to reach a steady state, the mean number of strands per cell of each length

and type was counted and averaged over time. Strands are classified as one of four types, indicated by colours in Figure 3.3. Reproducing strands are those which have one complete AZ motif and one complete za. Non-Cleaving strands have one complete AZ motif and no za. Halving strands are those which have one complete AZ and more than one complete za. Fragments are those which do not have Z and A at the ends, and which therefore cannot circularize.

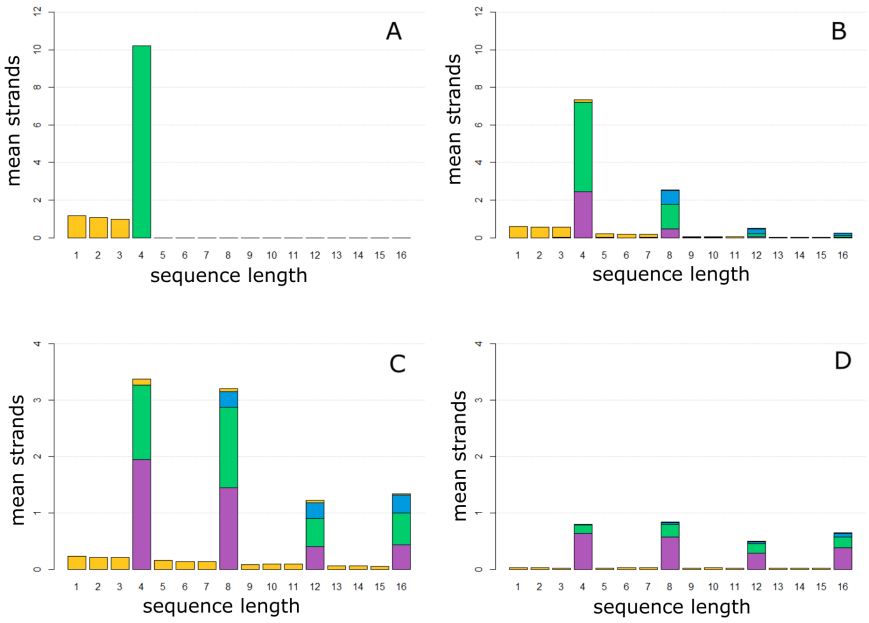

FIGURE 3.3: Mean number of strands of each length per cell. Colours indicate reproductive strands (green), non-cleaving strands (magenta), halving strands (blue), and non-circular fragments (yellow). (A) mutation probability $u = 0$. (B), mutation probability $u = 0.15$. (C) mutation probability $u = 0.3$. (D(mutation probability $u = 0.45$. Distribution above length 20 too small to show. $u_{indel} = 0$ for all graphs.

As the mutation probability increases as shown in Fig 3.3C ($u = 0.3$) and Fig 3.3D ($u = 0.5$), the number of reproductive circles decreases and the relative proportion halving and non-cleaving strands increases until the system meets an

error threshold where replication can no longer keep up with the loss rate, as shown in Figure 3.4. At this point, all types of strands disappear, because they cannot be maintained without continued replication of the reproductive circles. The observed error threshold of $u = 0.6$ is the probability of a deleterious mutation in a single character (A, Z etc). As each character represents a length roughly 25 bases,$u = 25u_{per-base}f_{del}$, where $u_{per-base}$ is the point mutation rate per base, and f_{del} is the fraction of mutations that are deleterious, i.e. the fraction of mutations within the structural region of the hammerhead that destroy the function of the hammerhead. This region contains loops whose sequence may not be important and paired regions in which compensatory changes may occur. Thus, f_{del} may be considerably less than 1. If we assume $f_{del} = 0.5$, then an error threshold of 0.6 in u corresponds to a threshold of 0.048 in $u_{per-base}$, which is not unreasonably small. The threshold value of 0.6 is dependent on the loss rate of cells, and would be higher if the loss rate were lower.

3.2.5 Insertions and deletions

The 4-mer ZzaA is the shortest sequence having both the hammerhead and its complement. It is therefore the most rapidly replicating type of circle in this model. There is no reason why the self-cleaving ribozymes should initially appear on a minimal length sequence. However, we expect there to be strong selection for speed of replication; therefore if the hammerhead motifs initially appear on a longer circle, we expect relatively quick evolution toward the minimal-length circle due to deletions of the non-functional parts of the circle. Insertions and deletions (indels) can occur via the slipped-strand mispairing mechanism [101]. This tends to produce short tandem repeats which are often seen in viral genes [101,102].

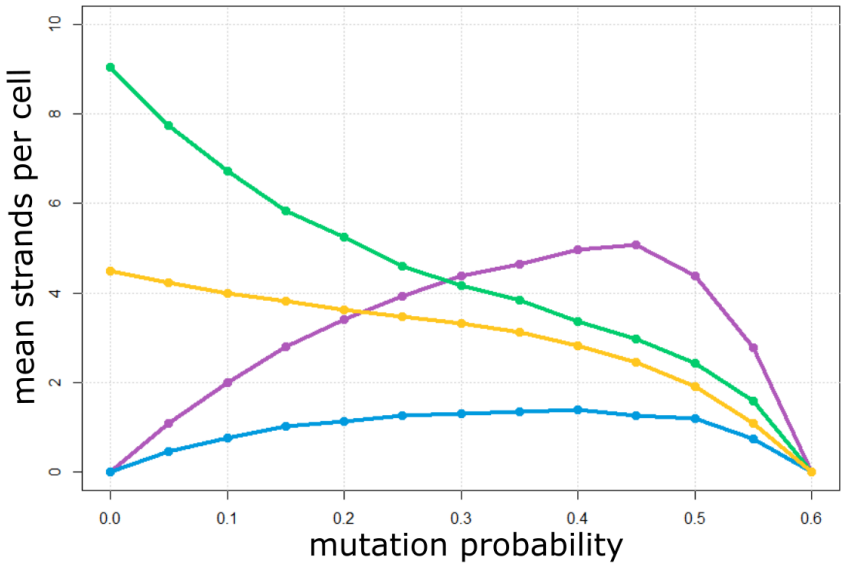

FIGURE 3.4: Mean number of strands per cell of the four types as a function of mutation probability u. Colours indicate reproductive strands (green), non-cleaving strands (magenta), halving strands (blue), and non-circular fragments (yellow). Strands of different lengths of each type are combined.

Indels are introduced into the model in the following way. Each time a character is copied, an insertion occurs with probability u_{indel}, a deletion occurs with probability u_{indel}, or neither with probability $1-2u_{indel}$. In the case of an insertion, a * character is inserted in the growing sequence after the character that was just copied. In the case of a deletion, the character that was just copied is deleted, but the position of growing end on the template (indicated by last) still advances by one.

Figure 3.5 considers a case where the first replicating circle that arises happens to be an 8-mer Zza****A. Each cell in the population begins with one copy of this 8-mer. The mutation rate is $u = 0.15$ and the indel rate is $u_{indel} = 0.015$. After

a relatively large number of time steps in the Gillespie algorithm $T = 2 \times 10^4$, shown in Fig 3.5A, 7-mers have evolved from this by deletion, and double-length circles have arisen from both the 8-mer and the 7-mer by the doubling mechanism described above. After a much longer number of steps $T = 22 \times 10^6$, shown in Figure 3.5B, reproducing circles of lengths 4, 5 and 6 have also evolved by deletion. The length distribution converges to a state in which the 4-mers are most frequent, together with multiples of the 4-mer created by the doubling mechanism, and small numbers of 5, 6, and 7-mers which are replenished by the insertions occurring in the 4-mer. In the case where $u_{indel} << u$, the 5, 6 and 7-mers will be much less frequent than the 4-mers because selection favours rapid replication. In this case, the length distribution will consist of the minimal-length 4-mer and its multiples that arise via the doubling mechanism, as was already shown in Figure 3.3B.

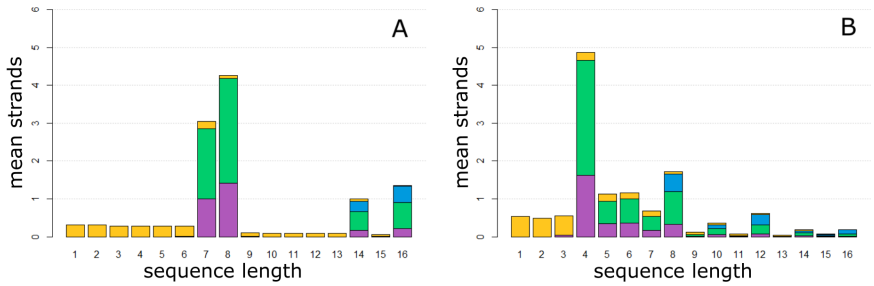

FIGURE 3.5: Mean number of strands of each length per cell, beginning from a single 8-mer per cell, including point mutations with probability $u = 0.15$ and indels with probability $u_{indel} = 0.015$. Colours indicate reproductive strands (green), non-cleaving strands (magenta), halving strands (blue), and non-circular fragments (yellow). (A) Simulation steps elapsed $T = 2 \times 10^4$ (B) Simulation steps elapsed $T = 2 \times 10^6$.

Although we do not know with any certainty the rates of the mutation and selection processes that would have applied in the RNA World, we expect that

point mutation rates associated with non-enzymatic replication would be quite large, as measured experimentally [39,67]. On the other hand, indel rates might be much smaller than this. We expect selection for increased replication rate to be strong, and therefore to operate on a short time scale of a few cell divisions. Under the assumption that insertions and deletions were rare and occurred with roughly equal frequency, then rare deletions would spread rapidly due to selection, whereas rare insertions would be eliminated by selection as fast as they could originate. This would result in a population dominated by minimal-length circles, which would have no additional non-coding regions that might evolve to encode beneficial genes. In order for a substantial numbers of longer sequences to be maintained in the population (with space available for beneficial genes), there must be a mechanism of lengthening circles that operates sufficiently rapidly to counter selection for rapidly-replicating, minimal-length circles. The doubling mechanism that we have seen here, does indeed work rapidly, because it occurs on the time scale of single point mutations that render the HHRz non-functional. Thus, we argue that the doubling mechanism is important for creating a population containing circles substantially longer than the minimum length.

3.2.6 Beneficial genes

Up to this point, we have only considered circles that replicate but encode no additional function. We now introduce a ribozyme with beneficial function for the cell, represented by character B. Copying B gives its complement b, and vice versa. Mutations may occur with probability u to give a non-functional character *. We assume the B ribozyme is functional only when it is in a folded linear strand. It is not functional when it is part of a circular strand being used as a template, or part

of the complementary strand that is being synthesized on a rolling circle. The b character is also assumed to be non-functional. The simulation keeps track of the number of functional ribozymes n_B in each cell (i.e. the number of occurrences of the B character in folded linear sequences). Each functional ribozyme gives an increase in the rates of polymerization. For a cell with n_B ribozymes, the rate of strand-displacement is increased to $R_{dis}(1+\beta n_B)$.We are assuming that beneficial ribozyme speeds up polymerization by contributing in some way to the metabolic reactions in the cell. Replication of all strands in the cell is benefitted equally, not just the strand on which the B gene is present. The B ribozyme does not represent a polymerase that binds to one specific template at a time, and it does not require to bind to a template in order to give the beneficial effect.

With the rules of the model as specified above, when a new linear strand is cleaved from a rolling circle, it folds at a rate R_{fold}. Cleavage can only occur between the A and Z characters. Therefore every sequence ends with A. If the strand begins with a Z character (arising from the previous cleavage or from starting synthesis with this character), then the strand can circularize. We have assumed that folding of the sequence represents formation of the structure that brings to two ends together and allows circularization. A strand beginning with Z always forms a circle. A fragmentary strand not beginning with Z cannot form a circle and always forms a folded linear strand. When B genes are not present in the model, a folded linear strand plays no role, but when B genes are included, a folded linear strand becomes a beneficial ribozyme whenever it contains a B character. Thus with these rules of the model, the B gene can only have a useful function if it arises in a fragmentary sequence produced from the first incomplete cycle of a rolling circle.

B genes in complete copies of the circle, always end up as new circular templates without contributing to the function of the cell.

This suggests that we need a new kind of hammerhead ribozyme whose rate of circularization is tunable. We represent the tunable ribozyme by the characters AY, and their complement ya. Whenever AY appears in the tail of a rolling circle, cleavage of the strand is assumed to be immediate (as with AZ). This produces an unfolded linear strand. Folding of all linear strands occurs at the same rate R_{fold}. If the strand begins with Y, then when folding occurs, a circle is produced with probability f_{circ} and a folded linear strand with probability $1 - f_{circ}$. Strands beginning with Z always form a circle ($f_{circ} = 1$). Fragmentary strands that begin with neither Y nor Z always form a folded linear strand ($f_{circ} = 0$).

Adding the f_{circ} parameter introduces a trade-off between circularization, which produces a new template, and folding to a linear strand, which allows the expression of a beneficial ribozyme if it is present on the sequence. A wide variety of hammerhead ribozymes are known with various distinct configurations, sequences, and rates [103]. Metallic ion concentration has also been suggested to change catalytic activity in the HHRz [84]. Thus it is likely that such tunability is evolutionarily possible. As a minimal-length 4-mer without B genes will always benefit from forming a circle, there should be strong selection for rapidly circularizing ribozymes on minimal circles. Therefore we began with Z-type ribozymes (with $f_{circ} = 1$) on the minimal circles. Y-type ribozymes (with $f_{circ} < 1$) can only be beneficial when B genes also exist.

We now wish to determine the range of parameters for which the presence of

a B gene increases the reproduction rate of the cells which contain it. The most common type of sequence arising, other than the minimal 4-mer is the double-length 8-mer. Therefore we suppose that the B gene arises initially on an 8-mer with sequence YzaB***A, which we call a plus strand because it encodes a functional B ribozyme. The complementary minus strand (accounting for circularity) is Zya***bA. The minus strand always benefits from circularization; therefore, we assume a Z-type ribozyme on the minus strand. The plus strand has a trade-off between circularization and expression of the B genes; therefore, we assume a Y-type ribozyme on the plus strand.

We ran simulations in which each cell began with one copy of the 4-mer plus strand. After reaching a steady state, the simulation was run for a period of 10^6 simulation steps and the average number of divisions per cell per unit time was measured. This is shown in Figure 3.6. For comparison, we also measure the division rate of cells that begin with the minimal 4-mer ZzzA (shown as a dashed line in Fig 3.6). The mutation and indel probabilities u and u_{indel} were set to zero in these simulations, so the only replicating sequences that arise are the 8-mers or 4-mers that we begin with, and there is only one type of replicating sequence in each cell.

When there is no benefit of the B gene ($\beta = 0$ in Fig 3.6A), cells containing the 8-mers always reproduce more slowly than cells containing the 4-mers. In this case, the division rate of the 8-mer cells is maximum when f_{circ} is 1. There is no advantage to not forming a circle if the B gene does not produce a benefit to the cell. When $\beta > 0$, there is an optimal value of f_{circ} in the range 0.3 to 0.5 for the parameters shown. The trade-off favors increasing the expression of the B gene at

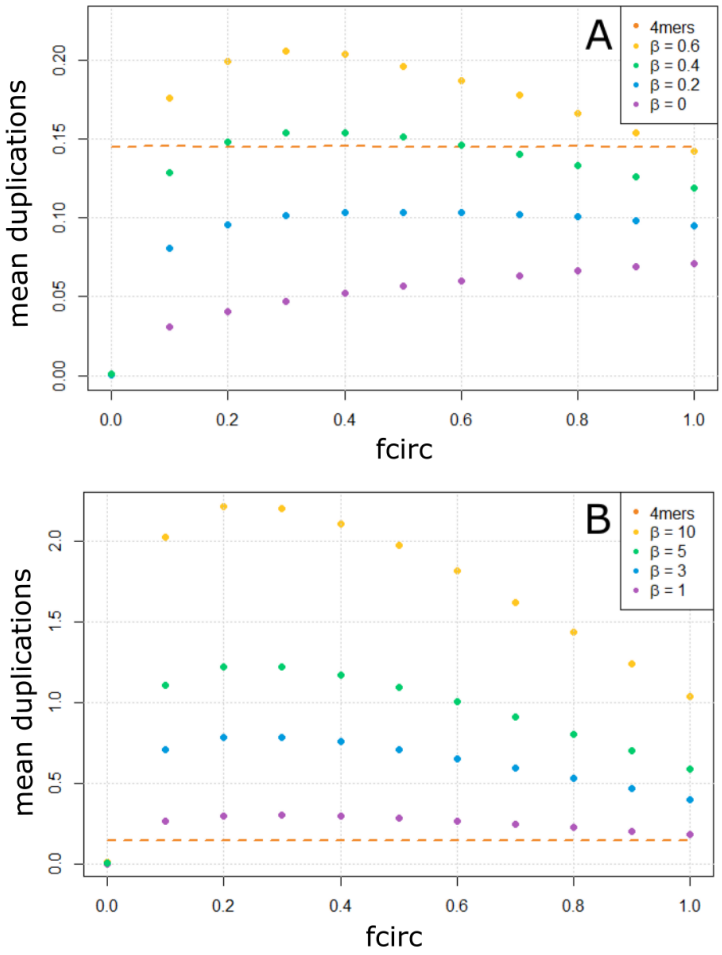

FIGURE 3.6: Comparison of reproductive rates of cells containing 8-mers with a beneficial gene and cells containing minimal-length 4-mers. The reproduction rate of the 8-mer cells depends on the size of the beneficial effect, β, and the circularization probability of the hammerhead, f_{circ}. There is no mutation in this Figure: $u = 0$ and $u_{indel} = 0$. (A) shows lower values of β, where the reproduction rate of the 8-mer cells is comparable to that of the 4-mer cells, or less. (B) shows higher values of β, where the reproduction rate of the 8-mer cells is much higher than the 4-mer cells.

the expense of reducing the number of templates. For $\beta = 0.2$, the division rate of the 8-mer cells remains lower than the 4-mer cells across the whole range of f_{circ}.

For $\beta = 0.4$, the division rate of the 8-cells just exceeds the 4-mer cells when f_{circ} is close to its optimal value. For 0.6, the division rate of the 8-cells exceeds the 4-mer cells across most of the range of f_{circ}. Higher values of β are shown in Fig 3.6B. For $\beta > 1$, the 8-mer cells reproduce much faster than the 4-mer cells. It can also be seen that even when $f_{circ} = 1$ (when a Y gene always circularizes), the division rate of the 8-mer cells still increases with β. So there is some benefit given by the B genes even when the only functional B ribozymes are on the incomplete fragments.

3.2.7 Spread of a beneficial gene

So far we have shown that cells containing double-length circles with a beneficial genes can sometimes out-compete cells containing minimal-length circles. However, we assumed above that the longer circle with the beneficial gene was already established in a separate cell from the cells containing minimal 4-mers. More realistically, the beneficial gene is likely to first arise as a single copy inside a cell that also contains minimal 4-mers. The longer circle is at a disadvantage relative to minimal 4-mers in the same cell because it replicates more slowly. Spread of the beneficial 8-mer requires the selection at the cell level to exceed the disadvantage at the molecular level.

After introduction of the single B gene, we track how many cells contain circles with the B gene or its complement b. Cells can be divided into four types: those with the beneficial genes, B or b, and no 4-mers, those with 4-mers and no beneficial gene, those with both, and those with neither. 8-mer circles without B or b will arise from duplication of the 4-mers or mutation of B and b characters to *. These

circles are counted as having neither a 4-mer or a beneficial 8-mer. Multiples of the 8-mer which have a B gene will arise by duplication of the 8-mers, and strands with the benefit smaller than 8 can arise from u_{indel}, these are included as having a benefit but no 4-mers.

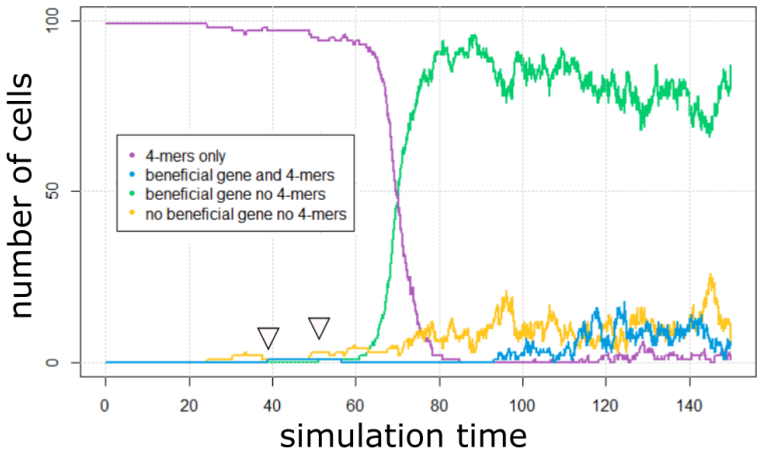

FIGURE 3.7: Graph showing number of cells in a 100 cell population divided into those with beneficial genes and no 4-mers, those with 4-mers and no beneficial gene, those with both, and those with neither. Time is in simulation hours. The first arrow shows the point at which the beneficial gene was added with $\beta = 5$. The second arrow shows the point at which the first cell appears containing beneficial 8-mers but no 4-mers. $u = 0.15$ and $u_{indel} = 0.015$ in this graph,

The results of such a simulation are shown in Figure 3.7. The first arrow (at time close to 40h) shows the point at which the single B gene was added with $\beta = 5$. Cells which contain both 4-mers and beneficial 8-mers remain rare for a long time after this. The second arrow (at time close to 50h) shows the point at which the first cell containing beneficial 8-mers and no 4-mers appears. The appearance of the first cell of this type requires a cell-division event in which all beneficial 8-mers from a mixed cell segregate to one daughter cell while all 4-mers segregate to the

other daughter. This is relatively rare, but once cells of this type are created, they rapidly multiply. In Figure 3.7, cells containing only beneficial 8-mers become the dominant type by about time 90h. Mixed cells disappear at around time 90h because they are out-competed by the cells with only beneficial 8-mers. However, this simulation also includes indels occurring with rate $u_{indel} = 0.05$; therefore 4-mers can also be created by deletions occurring in 8-mers. This recreates mixed cells later in the simulation (around time 100h), however these mixed cells do not take over the population, because selection against them is quite strong. The scenario seen in Fig 3.7 shows how a beneficial gene arising on a longer sequence can eventually spread to a high frequency in the population. This requires the creation of a cell that contains only beneficial 8-mer circles without any minimal circles, which is relatively rare. It is more likely that the beneficial 8-mers will disappear whilst they are still rare in the population, either due to deleterious mutations in the B gene or due to death of the mixed cells containing the B genes. If the beneficial 8-mers disappear before the creation of a cell containing only beneficial 8-mers, then they will not spread through the population. More examples of this scenario are given in Figure 3.8.

We measured how often this occurs by running our simulation multiple times. In each run, a single B gene was introduced on an 8-mer. The simulation was continued until one of two stop criteria was reached: either (i) the beneficial gene disappeared completely; or (ii) the number cells containing beneficial 8-mers and no 4-mers reached at least 90% of the population. The percentage of runs in which the beneficial 8-mer cells spread to high frequency is plotted against mutation probability u in Figure 3.9 for different values of the benefit parameter β.

Only a small percentage of the runs lead to spread of the beneficial gene, and high values of β are required in order to get appreciable probabilities of spread. For $\beta = 5$, the maximum probability of spread is only about 2.3%. In comparison with Figure 3.6B, we see that the reproduction rate of 8-mer cells with $\beta = 5$ and $f_{circ} = 0.3$ is approximately 8 times higher than for a 4-mer cell. A beneficial gene that gave an immediate 8-fold increase in fitness would have a very high probability of fixation in the usual approximation for the fixation rate used in population genetics [104]. However, the usual theory for the fixation probability does not apply in our case, because there are multiple strands in each cell. The reproduction rate of a cell depends on the mixture of genetic strands that it contains and also on the number of copies of folded ribozymes, which is variable. Mixed cells containing both 4-mers and 8-mers with the B gene do not spread to high frequencies. The spread of the B gene only occurs if a cell is established that contains the beneficial gene but no 4-mers (as shown in Figure 3.7). For this reason, the probability of spread is quite small, even when the benefit is large. To check that the B gene cannot spread due to spread of mixed cells, we ran additional simulations in which the run was stopped on three separate conditions: either (i) the beneficial gene dies out; (ii) cells containing the B gene and no 4-mers reach 90% of total; or (iii) mixed cells with both the B gene and 4-mers reach 90%. It was found that the runs never stopped due to criterion (iii). Thus, it was never observed that mixed cells reach high proportion.

For each value of β in Figure 3.9, a peak occurs in the probability of spread at around $u = 0.15$. When u is very low, the 4-mers replicate accurately. There are few 8-mer (or longer) circles created. If a B gene arises on an 8-mer, it is in a cell

that contains almost entirely 4-mers. It therefore has a low chance of spread. As u increases, replication of the 4-mers is less accurate, and most of the cells also contain significant numbers of 8-mers and longer circles. If a B gene arises on an 8-mer in this case, it has fewer 4-mers to compete with, and it is less likely to die out before the creation of a cell that contains only beneficial 8-mers. Thus the spread probability of the B gene is larger. If u is too high, however, B genes disappear due to deleterious mutations. We assumed that the mutation probability u of B to * is the same as that of A and Z.

The results in Figure 3.7, 3.8 and 3.9 have zero indel rate, $u_{indel} = 0$. We supposed that the B gene arises on an 8-mer because 8-mers are the most common type of longer sequence. However, if u_{indel} is not zero, then shorter circles containing the beneficial gene can also arise by deletion. If the initial beneficial 8-mer is YzaB***A, as before, then deletions of the * characters can occur, giving 7-mers and 6-mers and, eventually, the 5-mer YzaBA. The B gene can also be deleted, giving the original minimal 4-mer YzaA. Thus, the long-term survival of the B gene depends on competition of the 5-mer and the 4-mer. We investigated this case by beginning with a population of cells each containing the 8-mer YzaB***A, and allowing the simulation to reach a steady state. When the indel rate is small ($u_{indel} = 0.015$) and the benefit is fairly large ($\beta = 5$), 5-mers YzaBA become dominant, alongside multiples of the 5-mer which also contain B genes (shown in Fig 3.10A). If the indel rate is too large, however, or if the benefit is too small, the minimal 4-mers arise. Figure 3.10B shows the final steady-state distribution when $u_{indel} = 0.015$ and $\beta = 1$. The B gene has been lost, and we have a distribution of the 4-mer and its multiples.

In summary - in order for the B gene to survive, it has to maintain itself against deleterious mutations, deletions and selection favoring shorter sequences. These results show that this is fairly difficult, and it by no means occurs every time. However, at least sometimes, a beneficial gene becomes established. Thus, there is a route that leads from minimal, non-functional replicators towards replicating strands that encode beneficial functions.

3.3 Discussion

We have shown that the rolling circle mechanism is a feasible way of maintaining replication of RNA strands in a population of protocells. Rolling circles have the unusual property that point mutations that prevent the operation of the HHRzs give rise to doublings of lengths of the circles, and yet the ribozyme is not eliminated by the mutation because there is a second chance of correctly copying the same template on the next passage of the circle. Under the assumption that these point mutations are more frequent than short deletions that eliminate non-essential parts of the circles, then we expect a broad distribution of sequence lengths to arise, whereas in absence of this doubling mechanism, we would expect small deletions plus selection for rapid replication to lead to almost entirely minimal length circles. We have assumed that beneficial genes can occasionally arise in non-coding regions of circles. But the de novo appearance of beneficial genes is presumably very rare, so a mechanism is required that maintains appreciable frequencies of circles that are significantly longer than the minimal length.

As argued in the introduction, our principal reason for considering circular templates is that we require a strand displacement mechanism to avoid product

inhibition and this is only likely to work on a circular template. But this leads to several other points that are relevant for the evolution of chromosomes carrying useful genes. We assumed that the first circuit of the rolling circle occurs at a more rapid rate than subsequent circuits because the first one does not require strand displacement, whereas subsequent circuits do. For this reason, a circular template is almost always in a double-stranded state. This has important consequences for the stability of the genetic molecule, because it is known that double stranded RNAs are much less prone to degradation than single strands. Another advantage of circular strands as templates might be increased processivity of polymerase ribozymes. Recently developed polymerases have a clamp domain that wraps around the template [58]; hence if the template is circular, the polymerase can proceed multiple times around the same template.

In RNA World models there is always an apparent conflict between the need of a sequence to act both as a gene and a ribozyme. Presumably the folding of a strand to a functional ribozyme structure prevents its operation as a template. The rolling circle mechanism leads to an immediate distinction between double-stranded circles that are used as templates and folded linear single strands that function as ribozymes. We have pointed out that the relative rate of formation of folded strands to new circular templates is a tunable, evolvable property of the HHRzs (modelled by the f_{circ} parameter). If the circle does not encode beneficial genes, then it should always re-circularize as fast as possible ($f_{circ} = 1$). The same is true for the negative strand of a circle encoding a beneficial ribozyme. But the positive strand requires a balance between folding and re-circularizing, meaning that there is an optimal value of f_{circ}, as shown in Fig. 3.6. The independent

evolution of f_{circ} on plus and minus strands allows the relative numbers of plus and minus circles and folded ribozymes to be optimized to increase overall replication rate. The division of labour between template and catalyst has been discussed previously in the context of the origin of DNA [105], but in the current model this arises naturally in the RNA World without the need for a second kind of genetic polymer.

In the current paper, we have only considered circles with a single beneficial gene, but rolling circles could in principle encode multiple types of beneficial genes, as proposed in [89]. Placing multiple genes on a chromosome strand is beneficial from the point of view of coordinating gene replication, but it introduces the need for a mechanism to allow folding of functional ribozymes on separate linear strands – i.e. a need to distinguish transcription of a single ribozyme from replication of a chromosome. In the rolling circle mechanism, different beneficial genes on the same strand could be separated by copies of HHRzs, which would allow some strands to be cleaved into separate single folded ribozymes while other strands re-circularize and become templates. This might avoid the need to evolve separate transcriptional start and stop signals for each gene. We suspect that this would only work for a relatively small number of genes, however, because it will be necessary at least sometimes to complete replication of the whole circle before cleavage occurs at intermediate positions. This problem would increase the advantage of separate smaller circles encoding a single gene relative to longer circles with multiple genes.

In a single cell, shorter circles always replicate faster than longer ones. When considering the origins of the 8-mer circles containing the beneficial gene in Fig 3.7, we showed that the longer circle only spreads after it eventually manages to

get into a cell that does not contain any 4-mers. A similar issue would arise when considering competition between separate circles containing one beneficial gene each and longer circles with multiple genes.

In summary, the occurrence of circles in abiotic RNA polymerization [82], the ease with which self-cleaving ribozymes arise de novo [87], the ability of polymerase ribozymes to copy circular templates [58,80], and the natural occurrence of circular viroids [79] all point to the importance of circular templates in RNA replication. Developments in experimental methods for non-enzymatic and ribozyme-catalyzed replication may soon make it possible to study the evolution and replication of circular templates in experiments and potentially advance our understanding of the origins of life.

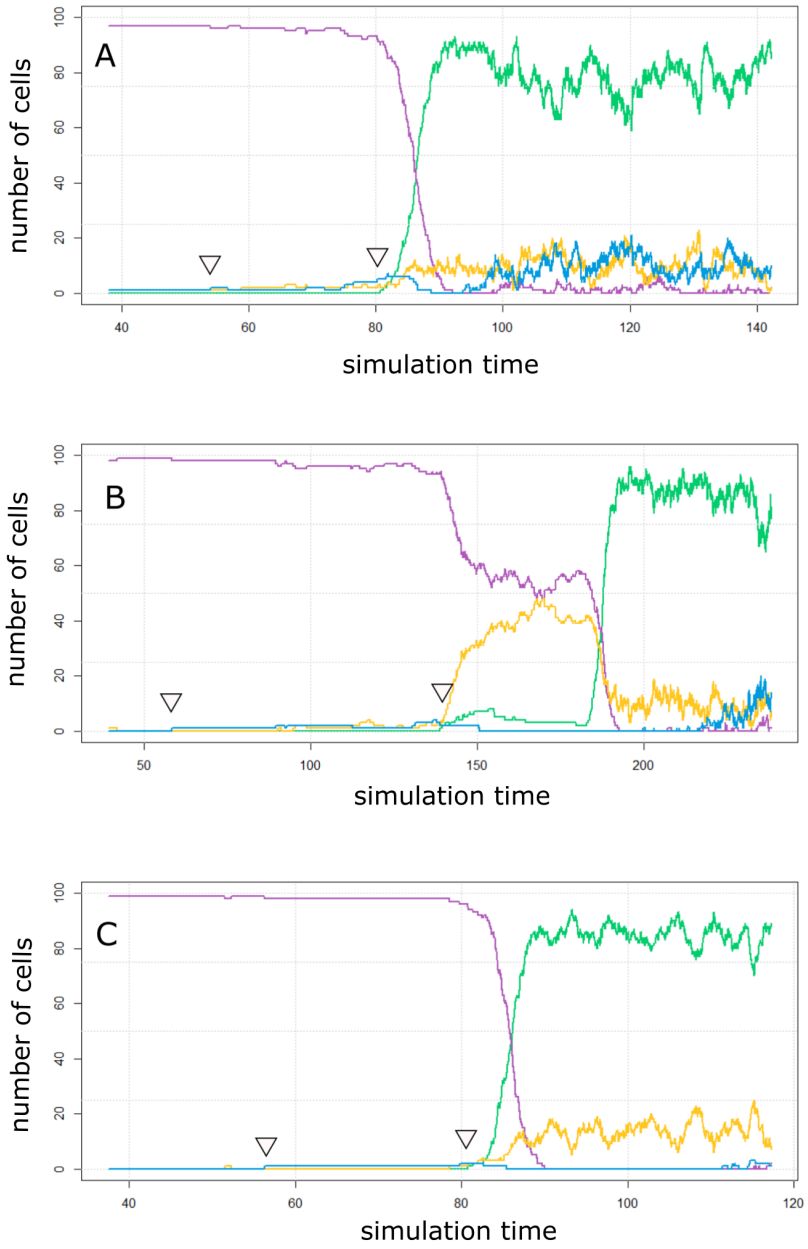

FIGURE 3.8: Graph shows the same conditions as Figure 3.7. (A) looks almost identical to 3.7 (B) shows a situation where the beneficial cells beat a mixed 4-mer/mutant population (C) shows a simulation where only 3 types of cells survive till the very end. These highlight how under different circumstances the benefit still has to end up in a cell with no 4-mers to spread

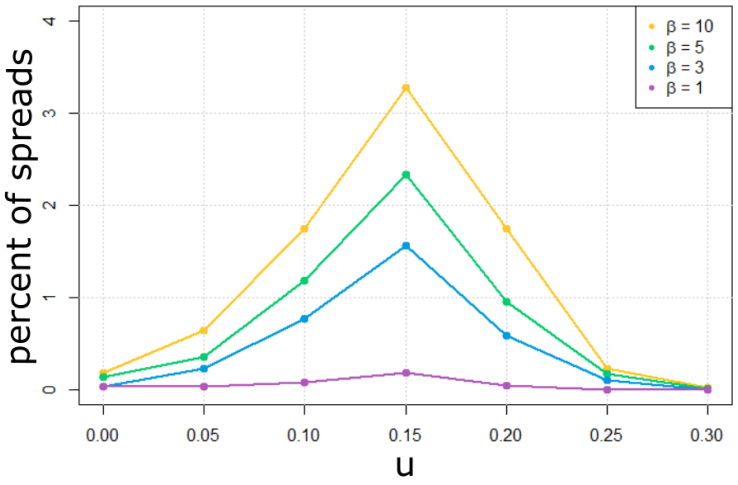

FIGURE 3.9: Percentage of (2×10^4) runs that result in takeover by the beneficial gene in a population of 100 cells. $u = 0.15$ and $u_{indel} = 0$ for all runs.

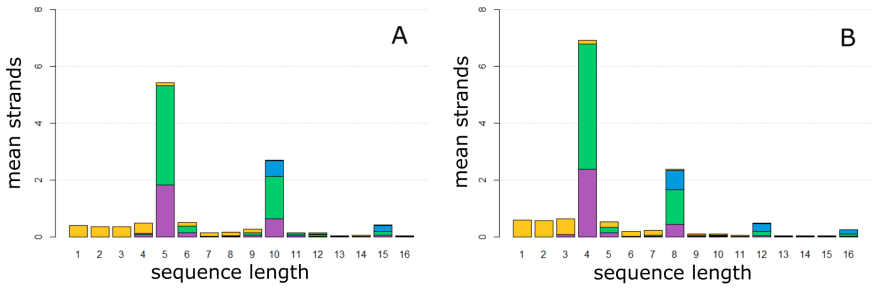

FIGURE 3.10: Steady state distribution of lengths beginning from 8-mer sequences YzaB***A, with $u_{indel} = 0.015$ and $u = 0.15$. (A) $\beta = 5$. (B) $\beta = 1$. Colours indicate reproductive strands (green), non-cleaving strands (magenta), halving strands (blue), and non-circular fragments (yellow).

Chapter 4

Next Steps: A model for a Genomic Rolling Circle

4.1 Models for gene co-operation

RNA cells need to grow in complexity in order to progress towards the organisms we see in modern life. Several ribozymes are required for this however it is unclear whether they could co-exists in the RNA world. This is because if more than one gene was held within the same cell, their different sequences may compete for limited nucleotides during replication or get separated during cell division. Some of this might be mitigated through the co-operation of genes, a process in which selective pressures choose cells which have a variety of genes instead of just a few. This only works if co-operation can overcome the cost of maintaining several genes, such as the increased need for nucleotides or the the tendency of each gene to acquire mutations. Another solution is the linking of genes in a genome, genes held together on one strand are more likely to be replicated together and be inherited

together. However there are some challenges which complicate the existence of a genome in the RNA world. First off, the benefits of holding genes together would need to outweigh the increase in replication time associated with having a longer genomic strand. Furthermore, unlike modern genomes where replication and expression of the genes is separate (i.e. DNA replication is separate from DNA translation into mRNA which initiates expression of the genes) an RNA genome has the same mechanism for expression and replication. In other words, any time the genome is replicated, its contents must either become ribozymes or part of the genome template which will continue to be replicated. Here we discuss how the model proposed in chapter 3 would need to be changed in order to account for gene co-operation, and improve on past models for genomes involving rolling circles.

The model in chapter 3 concerned itself with the replication of a single rare beneficial mutation within a system of rolling circles. This mutation could be considered a gene in the context of the RNA world. We argue minimal sequences cannot co-exist with long strands containing a gene, and separation must occur in cells for the gene to spread. This is ok because the short strands in this model are undesired because they benefit off the gene, but do not contribute to the long-term improvement of the population. However, the issue of strand co-existence becomes important and nuanced when more than one beneficial sequence is present.

Past models have shown that spatial separation, allows individual genes to benefit from each other and spread together [88,90,91]. In these models, ribozymes from separate genes contribute to different functions, like lipid synthesis, RNA polymerisation or nucleotide synthesis. Clusters which contain a host of different

functions replicate faster than those that don't contain any, or only some of the genes. Therefore, spatial clustering provides a method of selection on the population. The stronger the clustering (for example through the use of a protocell) the better the co-operation between genes [88]. This is not the same as our results in chapter 3, in which B genes could not spread without the HHRz and co-operation was therefore required. In that case compartmentalization in cells provided protection from stifling minimal sequence, not an incentive for co-operation.

Two papers have looked at the use of rolling circles in scenarios of gene co-operation and concluded that rolling circles may aid in decreasing degradation of genes and the formation of genomes [89,91]. In these papers, genes are held on the same [89] or separate [91] strands, and code for ribozymes with non-related functions. One of these models [89], proposes the use of rolling circles as a means of linking separate ribozymes in a genome. This is beneficial because genes in a genome get passed on together, increasing the chances for co-operation and allowing for new functions to be picked up. In this model, several ribozyme-coding sequences exist on an antisense strand, which constitutes a genome. Complimentary HHRz sequences separate each ribozyme on this strand. This allows separate ribozymes to be created through cleavage when the sense strand is replicated. Sometimes cleavage does not occur between the ribozymes and the whole sense strand replicates. From here, it can circularise, become a template, and continues the cycle of rolling circle replication. The cleavage rate between ribozymes therefore determines whether a genome is expressed or replicated. If HHRzs cleave too fast, the compliment to the genome will break into ribozymes and not become a template for further replication of the genome. However if HHRzs cleave too

slowly, ribozymes are not produced and the benefits of keeping all genes together does not outweigh the cost of having a long genome. This model is good but fails to consider the dynamic diversity which rolling circles experience. Particularly by only allowing the fully replicated genome to circularise and continue rolling circle replication, the model introduces a bias towards maintaining a genome.

In Figure (3.2) we describe how different mutations can give rise to differently cleaving circles of equal length. Depending on the position of HHRzs and their compliments, different strands will be made in subsequent replications. To illustrate why this is important we will discuss the model proposed in [89] within the language established in chapter 3, and show that this model only explores the easiest way to maintain a genome on a rolling circle.

We introduce a second beneficial mutation C which performs some catalytic function when present in a folded linear strand. The compliment to a C is c. If we restrict the genome proposed in [89] to two genes for simplicity, its sequence could have a variety of configurations. We discuss 3 possibilities, ZzabzacA, ZzabAZzacA and ZzabzaczaA.

4.2 Diversity in genomic sequences leads to different expressed ribozymes

The first strand, ZzabzacA, is the simplest. Its compliment is ABZACazZ. Using the | symbols to denote where cleavage can occur we see that some of the different strands that can occur in ABZ|ACazZ are ABZ, ACazZ and the functional compliment ABZACazZ. Table 4.1 has a summary of this. Of the possible

strands created, ABZ and ACazZ are able to become ribozymes. One concern in this replication might be that the difference in length might cause differences in folding strengths. Though this would also be dependant on the folding affinity of each individual ribozyme it may constrain which kinds of ribozymes can be encoded in C. Furthermore the sequence ACazZ has the capacity to circularize, and can continue to undergo rolling circle replication because of the az motif. The sequence ABZ can also circularise but will not become an independently replicating rolling circle because it lacks an az. This provides the C ribozyme a pathway to become separated from B, effectively breaking up the genome. In the model proposed in [89] only the functional compliment ABZACazZ can circularise and continue rolling circle replication, avoiding the issue altogether.

Other sequences may be proposed to alleviate some of these problems, but themselves run into issues. For example, the sequence ZzabAZzacA adds a AZ motif between both ribozymes. The strands ABazZ and ACazZ can now be created, and have an equal advantage, but both can now replicate separately. This makes it less likely one ribozyme will outcompete another, but further decreases the chances of linking all genes on a genome. As a final example, the sequence ZzabzaczaA can also hold a genome. Here the ribozymes can only be made as ABZ and ACZ strands. This avoids genes becoming separate rolling circles however, the sequence AazZ also appears and introduces the issues we discuss in chapter 3 into the system.

One solution would be to propose that each HHRz can only circularize with the pair it cleaves with. Using a number notation to show pairs, this would look like $Zz_1a_1bz_2a_2cA$. Replication of this strand gives A_1BZ_2 and A_2CazZ_1 which

TABLE 4.1: Example of strands which might make a genomic rolling circle and the strands they produce. The | symbol represents places on the strand where cleavage can occur

Genome Sequence	Exact Compliment	Possible Strands
ZzabzacA	ABZ|ACazZ	ABZ, ACazZ, ABZACazZ,
ZzabAZzacA	ABazZ|ACazZ	ABzaZ, ACazZ, ABazZACazZ
ZzabzaczaA	ABZ|ACZ|AazZ	ABZ, ACZ, AazZ, AazZABZACZ

cannot circularize because the A and Z at each end come from different pairs. This would give us a model equivalent to that shown in [89], because only the functional compliment $A_1BZ_2A_2CazZ_1$ can circularize. Given that the results of [89] show a genome can be maintained on a rolling circle, we hypothesize that on the limit where circularisation between misspaired HHRzs is 0, rolling several genes can be encoded on the same ribozyme.

4.3 Modified stochastic model for gene co-operation in rolling circles

However it is unclear how likely this scenario is and it would be worth exploring if a genome can be maintained in scenarios where circularisation between misspaired HHRzs is reduced but not 0. Furthermore, the problems we have presented above raise a lot of questions about how co-operation between genes might behave in a genomic rolling circle. For example, in the solution we gave using $A_1BZ_2 and A_2CazZ_1$, one ribozyme has an inherent difference in size. One interesting question is how this might impact expression of a ribozyme or the addition of new ribozymes. Genes on a longer strand may be at a disadvantage if it takes longer to replicate, or if the extra unused sequences impact its ability to fold and

function. We have also shown that several different sequences are possible for a genomic rolling circle, but are not sure if these sequences can be achieved through mutations (as shown in Figure 3.2) or mechanisms like deletion and insertion. Furthermore, if two ribozymes exist which can each independently cleave, but cannot re-circularize with each other (i.e. $Z_1A_1 and Z_2A_2$) it might be assumed their rates would differ. It would be interesting to see how this might impact the expression of the genes they help replicate.

These questions could be explored with a modified version of the model proposed in chapter 3. In this new model the basic method of replication would stay the same, however we would incorporate a method of distinguishing between paired HHRz as shown by the subscript system above. In the prior model, cleavage always occurred at a HHRz, however the cleavage of each HHRz would need to become a rate, because sometimes cleavage should not occur between genes to get full replication of the genome. This also allows us to explore different cleavage rates for different HHRzs and compare how this impacts co-operation. The rate of different HHRz pairs could also be modeled, for example having a slow or zero rate between mismatching pairs (like Z_1A_2 and Z_2A_1). Finally, more ribozyme functions would need to be defined in the model, as opposed to a generic benefit to displacement. New ribozymes like C could function on the stability of a proto-membrane, or the abundance of nucleotides in a cell.

These next steps would allow us to better explore how rolling circles could have contributed to the genetic maturity of early life. These findings might be useful in motivating more work in the replicative characteristics of circular RNA which has been gaining attention recently. Though we may find that this model does not

account for many of the properties of rolling circles that such work might discover, it may highlight some key difficulties in creating genomes and encourage others to propose novel ways to achieve gene replication in the RNA world.

Chapter 5

Citations

1. Chyba, C.F.; Hand, K.P. ASTROBIOLOGY: The Study of the Living Universe. Annual Review of Astronomy and Astrophysics 2005, 43, 31–74

2. Higgs, P.G. Chemical Evolution and the Evolutionary Definition of Life. Journal of Molecular Evolution 2017, 84, 225–235

3. Baross, J.A. Planets and Life | Evolution: a defining feature of life. Cambridge University Press, New York 2007

4. Cleaves II, H.J. The Origin of the Biologically Coded Amino Acids. Journal of Theoretical Biology 2010, 263, 490–498

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

6. Chen, Y.; Ma, W. The Origin of Biological Homochirality along with the Origin of Life. PLOS Computational Biology 2020, 16, e1007592

7. Sengupta, S.; Higgs, P.G. Pathways of Genetic Code Evolution in Ancient and Modern Organisms. Journal of Molecular Evolution 2015, 80, 229–243

8. Bernier, C.R.; Petrov, A.S.; Kovacs, N.A.; Penev, P.I.; Williams, L.D. Translation: The Universal Structural Core of Life. Molecular Biology and Evolution 2018, 35, 2065–2076

9. Deamer, D. Membranes and the Origin of Life: A Century of Conjecture. J Mol Evol 2016, 83, 159–168,

10. Orgel, L.E. Prebiotic Chemistry and the Origin of the RNA World. Critical Reviews in Biochemistry and Molecular Biology 2004, 39, 99–123

11. Miller, S.L. A Production of Amino Acids under Possible Primitive Earth Conditions. Science 1953, 117, 528–529.

12. Miller, S.L. The Mechanism of Synthesis of Amino Acids by Electric Discharges. Biochimica et Biophysica Acta 1957, 23, 480–489

13. Rushdi, A.I.; Simoneit, B.R. Lipid Formation by Aqueous Fischer-Tropsch-Type Synthesis over a Temperature Range of 100 to 400 Degrees C. Orig Life Evol Biosph 2001, 31, 103–118

14. Oró, J. Synthesis of Adenine from Ammonium Cyanide. Biochemical and Biophysical Research Communications 1960, 2, 407–412

15. Shanker, U.; Bhushan, B.; Bhattacharjee, G.; Kamaluddin Formation of Nucleobases from Formamide in the Presence of Iron Oxides: Implication in Chemical Evolution and Origin of Life. Astrobiology 2011, 11, 225–233

16. Cleaves, II J.H.; Nelson, K.E.; Miller, S.L. The Prebiotic Synthesis of Pyrimidines in Frozen Solution. The Science of Nature 2006, 93, 228-231

17. Ritson, D.; Sutherland, J.D. Prebiotic Synthesis of Simple Sugars by Photoredox Systems Chemistry. Nat Chem 2012, 4, 895–899

18. Parker, E.T.; Cleaves, H.J.; Dworkin, J.P.; Glavin, D.P.; Callahan, M.; Aubrey, A.; Lazcano, A.; Bada, J.L. Primordial Synthesis of Amines and Amino Acids in

a 1958 Miller H2S-Rich Spark Discharge Experiment. Proceedings of the National Academy of Sciences 2011, 108, 5526–5531

19. Sagan, C.; Khare, B.N. Long-Wavelength Ultraviolet Photoproduction of Amino Acids on the Primitive Earth. Science 1971, 173, 417–420

20. Takahashi, J.; Hosokawa, T.; Masuda, H.; Kaneko, T.; Kobayashi, K.; Saito, T.; Utsumi, Y. Abiotic Synthesis of Amino Acids by X-Ray Irradiation of Simple Inorganic Gases. Appl. Phys. Lett. 1999, 74, 877–879

21. Kobayashi, K.; Kaneko, T.; Saito, T. Characterization of Complex Organic Compounds Formed in Simulated Planetary Atmospheres by the Action of High Energy Particles. Adv Space Res 1999, 24, 461–464

22. Powner, M.W.; Gerland, B.; Sutherland, J.D. Synthesis of Activated Pyrimidine Ribonucleotides in Prebiotically Plausible Conditions. Nature 2009, 459, 239–242

23. Kasting, J.F.; Brown, L.L. The Early Atmosphere as a Source of Biogenic Compounds. In The Molecular Origins of Life: Assembling Pieces of the Puzzle; Brack, A., Ed.; Cambridge University Press: Cambridge, 1998; 35–56

24. Walker, J.C.G.; Klein, C.; Schidlowski, M.; Schopf, J.W.; Stevenson, D.J.; Walter, M.R. Environmental Evolution of the Archean-Early Proterozoic Earth; 1983; pp. 260–290

25. Holland, H.D. The Chemical Evolution of the Atmosphere and Oceans; Princeton University Press, 2021;

26. Trendall, A.P. Carbon Dioxide in the Precambrian Atmosphere. Geochimica et Cosmochimica Acta 1966, 30, 435–437

27. Ehrenfreund, P.; Charnley, S.B. Organic Molecules in the Interstellar Medium, Comets, and Meteorites: A Voyage from Dark Clouds to the Early Earth. Annual

Review of Astronomy and Astrophysics 2000, 38, 427–483

28. Pearce, B.K.D.; Pudritz, R.E.; Semenov, D.A.; Henning, T.K. Origin of the RNA World: The Fate of Nucleobases in Warm Little Ponds. PNAS 2017, 114, 11327–11332

29. Russell, M.J.; Hall, A.J.; Martin, W. Serpentinization as a Source of Energy at the Origin of Life. Geobiology 2010, 8, 355–371

30. Robertson, M.P.; Joyce, G.F. The Origins of the RNA World. Cold Spring Harb Perspect Biol 2012, 4, a003608

31. Higgs, P.G.; Lehman, N. The RNA World: Molecular Cooperation at the Origins of Life. Nat Rev Genet 2015, 16, 7–17

32. Copley, S.D.; Smith, E.; Morowitz, H.J. The Origin of the RNA World: Co-Evolution of Genes and Metabolism. Bioorganic Chemistry 2007, 35, 430–443

33. Segré, D.; Ben-Eli, D.; Deamer, D.W.; Lancet, D. The Lipid World. Orig Life Evol Biosph 2001, 31, 119–145

34. Kempes, C.P.; Krakauer, D.C. The Multiple Paths to Multiple Life. J Mol Evol 2021, 89, 415–426

35. Ban, N.; Nissen, P.; Hansen, J.; Moore, P.B.; Steitz, T.A. The Complete Atomic Structure of the Large Ribosomal Subunit at 2.4 A Resolution. Science 2000, 289, 905–920

36. Maizels, N.; Weiner, A.M. The Genomic Tag Hypothesis What Molecular Fossils Tell Us about the Evolution of TRNA. Gesteland, R F , Cech, T R , Atkins, J F Cold Spring Harbor Monograph Series; The RNA world, Second edition 1999.

37. Reichard, P. The Evolution of Ribonucleotide Reduction. Trends in Biochemical Sciences 1997, 22, 81–85

38. Domingo, E.; Holland, J.J. RNA Virus Mutations and Fitness for Survival.

Annu Rev Microbiol, 1997, 51, 151-78

39. Leu, K.; Obermayer, B.; Rajamani, S.; Gerland, U.; Chen, I.A. The Prebiotic Evolutionary Advantage of Transferring Genetic Information from RNA to DNA. Nucleic Acids Res 2011, 39, 8135–8147

40. Micura, R.; Höbartner, C. Fundamental Studies of Functional Nucleic Acids: Aptamers, Riboswitches, Ribozymes and DNAzymes. Chemical Society Reviews 2020, 49, 7331–7353

41. Morrison, D.; Rothenbroker, M.; Li, Y. DNAzymes: Selected for Applications. Small Methods 2018, 2, 1700319

42. Cech, T. The Chemistry of Self-Splicing RNA and RNA Enzymes. Science 1987, 236, 1532–1540.

43. Chandra, M.; Silverman, S.K. DNA and RNA Can Be Equally Efficient Catalysts for CarbonCarbon Bond Formation. J. Am. Chem. Soc. 2008, 130, 2936–2937

44. Kruger, K.; Grabowski, P.J.; Zaug, A.J.; Sands, J.; Gottschling, D.E.; Cech, T.R. Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena. Cell 1982, 31, 147–157

45. Guerrier-Takada, C.; Gardiner, K.; Marsh, T.; Pace, N.; Altman, S. The RNA Moiety of Ribonuclease P Is the Catalytic Subunit of the Enzyme. Cell 1983, 35, 849–857

46. Doudna, J.A.; Cech, T.R. The Chemical Repertoire of Natural Ribozymes. Nature 2002, 418, 222–229

47. Prody, G.A.; Bakos, J.T.; Buzayan, J.M.; Schneider, I.R.; Bruening, G. Autolytic Processing of Dimeric Plant Virus Satellite RNA. Science 1986, 231, 1577–1581.

48. Ferré-D'Amaré, A.R.; Scott, W.G. Small Self-Cleaving Ribozymes. Cold Spring Harb Perspect Biol 2010, 2, a003574

49. Lau, M.W.L.; Cadieux, K.E.C.; Unrau, P.J. Isolation of Fast Purine Nucleotide Synthase Ribozymes. J. Am. Chem. Soc. 2004, 126, 15686–15693

50. Cech, T.R.; Zhang, B. Peptide Bond Formation by in Vitro Selected Ribozymes. Nature (London) 1997, 390, 96–100

51. Li, N.; Huang, F. Ribozyme-Catalyzed Aminoacylation from CoA Thioesters. Biochemistry 2005, 44, 4582–4590

52. Coleman, T.M.; Huang, F. RNA-Catalyzed Thioester Synthesis. Chemistry Biology 2002, 9, 1227–1236

53. Zaher, H.S.; Watkins, R.A.; Unrau, P.J. Two Independently Selected Capping Ribozymes Share Similar Substrate Requirements. RNA 2006, 12, 1949–1958

54. Lau, M.W.L.; Unrau, P.J. A Promiscuous Ribozyme Promotes Nucleotide Synthesis in Addition to Ribose Chemistry. Chemistry Biology 2009, 16, 815–825

55. Schultes, E.A.; Bartel, D.P. One Sequence, Two Ribozymes: Implications for the Emergence of New Ribozyme Folds. Science 2000, 289, 448–448.

56. Horning, P.D.; Joyce, F.G. Amplification of RNA by an RNA Polymerase Ribozyme | PNAS, 2016, 113(35)

57. Attwater, J.; Raguram, A.; Morgunov, A.S.; Gianni, E.; Holliger, P. Ribozyme-Catalysed RNA Synthesis Using Triplet Building Blocks. eLife 2018, 7, e35255

58. Cojocaru, R.; Unrau, P.J. Processive RNA Polymerization and Promoter Recognition in an RNA World. Science 2021, 371, 1225–1232

59. Attwater, J.; Wochner, A.; Holliger, P. In-Ice Evolution of RNA Polymerase Ribozyme Activity. Nat Chem 2013, 5, 1011–1018

60. 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

61. 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

62. Himbert, S.; Chapman, M.; Deamer, D.W.; Rheinstädter, M.C. Organization of Nucleotides in Different Environments and the Formation of Pre-Polymers. Sci Rep 2016, 6, 31285

63. Dass, A.V.; Wunnava, S.; Langlais, J.; Esch, B. von der; Krusche, M.; Ufer, L.; Chrisam, N.; Dubini, R.C.A.; Gartner, F.; Angerpointner, S.; et al. RNA Polymerisation without Catalyst from 2',3'-Cyclic Nucleotides by Drying at Air-Water Interfaces. ChemRxiv 2022

64. Walton, T.; Zhang, W.; Li, L.; Tam, C.P.; Szostak, J.W. The Mechanism of Nonenzymatic Template Copying with Imidazole-Activated Nucleotides. Angewandte Chemie International Edition 2019, 58, 10812–10819

65. Deamer, D. Liquid Crystalline Nanostructures: Organizing Matrices for Non-Enzymatic Nucleic Acid Polymerization. Chem. Soc. Rev. 2012, 41, 5375–5379

66. O'Flaherty, D.K.; Zhou, L.; Szostak, J.W. Nonenzymatic RNA-Templated Synthesis of N3'→P5' Phosphoramidate DNA. Bio Protoc 2020, 10, e3734

67. Bapat, N.V.; Rajamani, S. Effect of Co-Solutes on Template-Directed Nonenzymatic Replication of Nucleic Acids. J Mol Evol 2015, 81, 72–80

68. Prywes, N.; Blain, J.C.; Del Frate, F.; Szostak, J.W. Nonenzymatic Copying of RNA Templates Containing All Four Letters Is Catalyzed by Activated Oligonucleotides. eLife 2016, 5, e17756

69. O'Flaherty, D.K.; Kamat, N.P.; Mirza, F.N.; Li, L.; Prywes, N.; Szostak, J.W. Copying of Mixed-Sequence RNA Templates inside Model Protocells. J. Am. Chem. Soc. 2018, 140, 5171–5178

70. Sosson, M.; Pfeffer, D.; Richert, C. Enzyme-Free Ligation of Dimers and Trimers to RNA Primers. Nucleic Acids Research 2019, 47, 3836–3845

71. Sosson, M.; Richert, C. Enzyme-Free Genetic Copying of DNA and RNA Sequences. Beilstein J Org Chem 2018, 14, 603–617

72. Li, L.; Prywes, N.; Tam, C.P.; O'Flaherty, D.K.; Lelyveld, V.S.; Izgu, E.C.; Pal, A.; Szostak, J.W. Enhanced Nonenzymatic RNA Copying with 2-Aminoimidazole Activated Nucleotides. J. Am. Chem. Soc. 2017, 139, 1810–1813.

73. Szostak, J.W. The Eightfold Path to Non-Enzymatic RNA Replication. Journal of Systems Chemistry 2012, 3, 2

74. Kim, S.C.; Zhou, L.; Zhang, W.; O'Flaherty, D.K.; Rondo-Brovetto, V.; Szostak, J.W. A Model for the Emergence of RNA from a Prebiotically Plausible Mixture of Ribonucleotides, Arabinonucleotides, and 2-Deoxynucleotides. J. Am. Chem. Soc. 2020, 142, 2317–2326

75. Tupper, A.S.; Higgs, P.G. Rolling-Circle and Strand-Displacement Mechanisms for Non-Enzymatic RNA Replication at the Time of the Origin of Life. Journal of Theoretical Biology 2021, 527, 110822

76. Zhou, L.; Ding, D.; Szostak, J.W. The Virtual Circular Genome Model for Primordial RNA Replication. RNA 2020, rna.077693.120

77. Wachowius, F.; Holliger, P. Non-Enzymatic Assembly of a Minimized RNA Polymerase Ribozyme. ChemSystemsChem 2019, 1, 1–4

78. Chamanian, P.; Higgs, P.G. Computer simulations of template-directed RNA synthesis driven by temperature cycling in diverse sequence mixtures. PLOS Comp

Biol 2022, under review.

79. Flores, R.; Gago-Zachert, S.; Serra, P.; Sanjuán, R.; Elena, S.F. Viroids: Survivors from the RNA World? Annu Rev Microbiol 2014, 68, 395–414

80. Kristoffersen, E.L.; Burman, M.; Noy, A.; Holliger, P. Rolling Circle RNA Synthesis Catalyzed by RNA. eLife 2022, 11, e75186

81. Zhou, L.; Kim, S.C.; Ho, K.H.; O'Flaherty, D.K.; Giurgiu, C.; Wright, T.H.; Szostak, J.W. Non-Enzymatic Primer Extension with Strand Displacement. eLife 2019, 8

82. Hassenkam, T.; Deamer, D. Visualizing RNA Polymers Produced by Hot Wet-Dry Cycling. Sci Rep 2022, 12, 10098

83. Altman, S.; Baer, M.; Guerrier-Takada, C.; Vioque, A. Enzymatic Cleavage of RNA by RNA. Trends in Biochemical Sciences 1986, 11, 515–518

84. Conaty, J.; Hendry, P.; Lockett, T. Selected Classes of Minimised Hammerhead Ribozyme Have Very High Cleavage Rates at Low Mg2+ Concentration. Nucleic Acids Research 1999, 27, 2400–2407

85. Hammann, C.; Luptak, A.; Perreault, J.; de la Peña, M. The Ubiquitous Hammerhead Ribozyme. RNA 2012, 18, 871–885

86. O'Rourke, S.M.; Scott, W.G. Chapter Seven - Structural Simplicity and Mechanistic Complexity in the Hammerhead Ribozyme. In Progress in Molecular Biology and Translational Science; Teplow, D.B., Ed.; Academic Press, 2018; Vol. 159, pp. 177–202.

87. Salehi-Ashtiani, K.; Szostak, J.W. In Vitro Evolution Suggests Multiple Origins for the Hammerhead Ribozyme. Nature 2001, 414, 82–84.

88. Ma, W.; Hu, J. Computer Simulation on the Cooperation of Functional Molecules during the Early Stages of Evolution. PLoS ONE 2012, 7(4): e35454.

89. Ma, W.; Yu, C.; Zhang, W. Circularity and Self-Cleavage as a Strategy for the Emergence of a Chromosome in the RNA-Based Protocell. Biol Direct 2013, 8, 21.

90. Kim, Y.E.; Higgs, P.G. Co-Operation between Polymerases and Nucleotide Synthetases in the RNA World. PLOS Computational Biology 2016, 12, e1005161

91. Roy, S.; Sengupta, S. Evolution towards Increasing Complexity through Functional Diversification in a Protocell Model of the RNA World. Proceedings of the Royal Society B: Biological Sciences 2021, 288, 20212098

92. Eigen, M. Selforganization of Matter and the Evolution of Biological Macromolecules. Naturwissenschaften 1971, 58, 465–523

93. Schuster, P.; Fontana, W.; Stadler, P.F.; Hofacker, I.L. From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures. Proceedings of the Royal Society of London. Series B: Biological Sciences 1994, 255, 279–284

94. Szathmáry, E. The Origin of Replicators and Reproducers. Philos Trans R Soc Lond B Biol Sci 2006, 361, 1761–1776

95. Hofbauer, J. A Difference Equation Model for the Hypercycle. SIAM J. Appl. Math. 1984, 44, 762–772

96. Szathmáry, E.; Demeter, L. Group Selection of Early Replicators and the Origin of Life. Journal of Theoretical Biology 1987, 128, 463–486

97. Matsumura, S.; Kun, Á.; Ryckelynck, M.; Coldren, F.; Szilágyi, A.; Jossinet, F.; Rick, C.; Nghe, P.; Szathmáry, E.; Griffiths, A.D. Transient Compartmentalization of RNA Replicators Prevents Extinction Due to Parasites. Science 2016, 354, 1293–1296

98. Shah, V.; de Bouter, J.; Pauli, Q.; Tupper, A.S.; Higgs, P.G. Survival of RNA Replicators Is Much Easier in Protocells Than in Surface-Based, Spatial Systems.

Life 2019, 9, 65

99. Bianconi, G.; Zhao, K.; Chen, I.A.; Nowak, M.A. Selection for Replicases in Protocells. PLOS Computational Biology 2013, 9, e1003051

100. Gillespie, D.T. A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions. Journal of Computational Physics 1976, 22, 403–434

101. Macey, J.R.; Larson, A.; Ananjeva, N.B.; Papenfuss, T.J. Replication Slippage May Cause Parallel Evolution in the Secondary Structures of Mitochondrial Transfer RNAs. Molecular Biology and Evolution 1997, 14, 30–39

102. Hancock, J.M.; Chaleeprom, W.; Chaleeprom, W.; Dale, J.; Gibbs, A. Replication Slippage in the Evolution of Potyviruses. J Gen Virol 1995, 76 (Pt 12), 3229–3232

103. Boots, J.L.; Canny, M.D.; Azimi, E.; Pardi, A. Metal Ion Specificities for Folding and Cleavage Activity in the Schistosoma Hammerhead Ribozyme. RNA 2008, 14, 2212–2222

104. Kimura, M. On the Probability of Fixation of Mutant Genes in a Population. Genetics 1962, 47, 713–719.

105. Takeuchi, N.; Hogeweg, P.; Koonin, E.V. On the Origin of DNA Genomes: Evolution of the Division of Labor between Template and Catalyst in Model Replicator Systems. PLOS Computational Biology 2011, 7, e1002024.

106. Leu, K.; Obermayer, B.; Rajamani, S.; Gerland, U.; Chen, I.A. The Prebiotic Evolutionary advantage of transferring genetic information from RNA to DNA. Nucleic Acids Res 2011 (18);8135-47

107. Rouzan, B.; McMichael, E.; Cave, R.; Sevcik, L.R.; Ostrosky, K.; Whitman,

E.; Stegemann, R.; Sinclair, A.L.; Serra, M.J.; Deckert, A.A. Kinetcis and Thermodynamics of DNA, RNA, and Hybrid Duplex Formation. Biochemistry 2013 (52) 765-772