SIMULATIONS OF NON-ENZYMATIC TEMPLATE DIRECTED RNA REPLICATION

SIMULATIONS OF NON-ENZYMATIC TEMPLATE DIRECTED RNA REPLICATION

By POUYAN CHAMANIAN, B.Sc.

A Thesis Submitted to the School of Graduate Studies in Partial Fulfilment of the

Requirements for the Degree Master of Science

McMaster University | Master of Science (2021)

Hamilton, Ontario | Biology and Astrobiology

TITLE: Simulations of Non-Enzymatic Template Directed RNA Replication

AUTHOR: Pouyan Chamanian, B.Sc. (McMaster University)

SUPERVISOR: Dr. Paul G. Higgs

NUMBER OF PAGES: xi, 84

Lay Abstract

The origin of biological life can be traced back by looking at the common themes between modern cellular processes. The role of RNA polymers seems to be of great importance, making us believe that an RNA world existed leading up to life's origin. During this time, RNA would act as both a genetic material and a catalyst. To examine this theory in more detail, we use computational modeling to recreate and explore the various potential chemistries and conditions on the early Earth. Specifically, we explore the problems that exist for the replication and production of RNA polymers. Our results can be used to guide future theoretical and experimental research of the RNA world.

Abstract

The universal traits of cellular expression and replication in modern life point to the existence of an ancient RNA world. Leading up to the origin of life, this stage of evolution utilized RNA as the genetic material, and as a catalyst in the form of ribozymes. Although it is expected that a polymerase ribozyme was required for the efficient replication of RNA, it is also likely that the earliest form of replication took place under non-enzymatic conditions. There are several problems with the current scenarios depicting non-enzymatic RNA replication, thus we aim to examine them in more detail using computational models. We first consider the relationship between the thermodynamics of RNA base pairing and non-enzymatic nucleotide addition in an attempt to model the rate of primer extension. Our predicted rates reveal the model parameters to be too simple to produce reliably accurate results. For now, we should simply use available experimental rate data, until we have access to more data and less unknown parameters. Nevertheless, the model indicates that the primer extension rate does depend on thermodynamics of base pairing, and a more accurate model can be of great use when creating realistic complex models of RNA world scenarios. In chapter 3, we investigate non-enzymatic RNA replication under temperature cycling using computer simulations. When starting with a diverse mixture of sequences, partially matching sequences can reanneal in configurations that allow continued strand growth. This is in contrast to the case of having multiple copies of matching sequences, where reannealing occurs quickly upon cooling. We find that, starting with short oligomers, strands can grow over multiple cycles to produce long sequences over 100 nucleotides in length. The small strand extension per cycle does not produce replicates of any one specific sequence. This relates to the work done in chapter 4, where we look for the presence of a virtual circular

genome within our simulations. In a virtual circle, short overlapping RNA sequences will make up a mutually catalytic set. Within the diversity of our simulation, virtual circles are rare, and require a specific level of starting mixture diversity along with no input of new sequences. Continued replication of the diverse sequence mixture and emergence of long strands may eventually lead to the creation of rolling circles and ribozymes.

Acknowledgements

I would like to express my utmost thanks to my supervisor, Dr. Paul Higgs, for his guidance and invaluable teachings throughout the years. For me, his academic dedication and enthusiastic attitude created a productive and pleasant research experience.

I would like to thank Andrew, Armin, and Felipe for sharing in my journey and providing many insightful ideas and solutions.

I would also like to give a special thank you to my parents and my dear friends who have constantly supported me through both good and bad times. Without them, this work would not be possible.

Table of Contents

List of Figures and Tables

Figures

Tables

List of Abbreviations and Symbols

LCA	Last common ancestor
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
NADH	Nicotinamide adenine dinucleotide
$FADH_2$	Flavin adenine dinucleotide
CoA	Coenzyme A
ATP	Adenosine triphosphate
2-MeImp	5'-phosphor-2-methyl-imidazolide
A	Adenosine
U	Uracil
C	Cytosine
G	Guanine
NN	Nearest neighbor

Declaration of Academic Achievement

The research presented in this thesis involves two main projects. I contributed to the construction of the theoretical design, model, code, data analysis, and writing of all sections. I presented the work in chapter 3 and 4 at the International Society for the Study of the Origin of Life conference and was awarded distinctions. The work described in this thesis is a result of the combined effort of myself and my supervisor, Dr. Paul Higgs, with theoretical design contributions from Dr. Andrew Tupper.

Chapter 1: Introduction

When discussing the origin and development of primitive life on Earth, it is important to consider the emergence of the first metabolically active cells. In the modern biological world, the cells of all living organisms share a similar biochemistry, suggesting the presence of a last common ancestor (LCA)[1]. The LCA can be comparable to early unicellular microbial life. Thus, it was already quite complicated and was likely preceded by a much simpler kind of protocell. To fully understand how life originated, we must obtain a clear picture of how abiotic processes on the primitive earth could have led to the emergence of such a cell. A defining component of living organisms is their ability to propagate genetic information through an interconnected network of macromolecules. Specifically, DNA stores the genetic information to be passed on through the help of functional proteins, which are synthesized using the genetic code and an RNA intermediate. The complex interconnected DNA-RNA-protein network we study today was likely not present on the early Earth. Alternatively, there is reason to believe that a much simpler RNA world preceded the DNA-RNA-protein world. The transition between some primitive biochemical world and contemporary organisms constitutes the evolution of life. Thus, the development of life is dependent on some biochemical self-contained system capable of evolution. The RNA world theory proposes that the early origins of life revolved around RNA's potential for both storing information, as well as replicating this information while allowing for mutations and evolution through natural selection[2].

1.1 Basis for the RNA World Theory for the Origin of Life

The phylogenetic tree of life connects three domains, these being Archaea, Bacteria, and Eukaryote. The LCA is positioned as the ancestor to the three domains between the Bacteria and Archaea-Eukaryote branches, making Eukaryotes more closely related to Archaea than to Bacteria[3]. This idea allows us to make connections between traits seen within modern biology and be poised in assuming their presence within the LCA. With this in mind, we should start to think about the fundamental characteristics that all modern life relies on. The central dogma of molecular biology explains the flow of information from DNA to functional proteins in modern life[4]. It is a universal trait, thus some form of this should have been present in the LCA. The precise transfer begins at the DNA and is then sent to an RNA intermediate through transcription. RNA is translated to functional and structural proteins which are essential cellular components. From this, we see RNA only being useful as a temporary carrier of information, and rather expendable when compared to DNA. A deeper look, however, reveals several other important roles, especially regarding the translation of proteins.

Indeed, RNA shows a somewhat close connection to protein in modern life. Similar to how proteins hold information which can interact in functional and useful ways, RNA can fold and interact in a variety of useful mechanisms. Its broad range of functions are important for regulating gene expression, and assisting protein translation, and it can enact some of its functions through catalytic properties. These catalytic RNA are referred to as ribozymes. A well-known and universally conserved example of a modern ribozyme is positioned within the ribosome as the active site for catalyzing the synthesis of proteins[5]. Another universally conserved function shown in ribozymes is that of cleavage. Specifically, RNase P is shown as

being vital in translation, catalyzing the cleavage of the pre-tRNA backbone[6]. During DNA replication, it is observed that initiation is done through the synthesis of an RNA primer[7]. This counterproductive step is interesting as the switch to DNA primers would save the many resources required in the removal of primers and additional extension of DNA at the ends of genomes[8]. With more discoveries regarding the various roles of RNA, we can postulate its importance within the cellular ancestry as an important macromolecular precursor.

More recent discoveries show the synthesis of DNA nucleotides depending on the initial synthesis of RNA nucleotides. Enzymatic processes subsequent to ribonucleotide synthesis via specialized proteins must occur to allow for the conversion of the 2' hydroxyl group and uridine to thymine[9]. In addition to DNA nucleotides, many important coenzymes are also derived from RNA, namely NADH, $FADH_2$, CoA, and ATP. All these various traits mentioned point to an important place for RNA as an early and important macromolecule during the origin and development of early life. If we assume that a much simpler molecular interplay existed before the evolution of complex macromolecular cooperation, it is reasonable to assume RNA emerged as the first important macromolecule. Its many connections to both proteins and DNA enable RNA to take on a variety of roles. Also, its role as derivatives of important molecules may signify their absence at an early evolutionary point, in turn making the presence of RNA primers, coenzymes, and even ribozymes in modern biology a sort of relic of an ancient RNA world. It can be postulated that previous forms of ribosomes were made up exclusively from RNA, and later evolved to have surrounding proteins[10]. The hammerhead ribozyme, a modern ribozyme which catalyzes self-cleavage through a transesterification reaction, is another example. This ribozyme has been found to be ubiquitous among genomes of several different kingdoms, and said to be a relic

from the RNA world[11]. So then, the time of the RNA world is made in reference to when RNA acted as both the molecule responsible for storing genes, and catalyzing reactions[10,12].

1.2 The Emergence and Phases of an RNA World

The different stages of the RNA world, as well as the transitions between these stages, each contain their own set of challenges. To start, we must consider the synthesis of nucleosides and their components, which then need to polymerize through *de novo,* and later templated mechanisms. Where we start seeing signs of "life" is when RNA polymers begin self-replication and increase this rate exponentially due to evolutionary selection mechanisms. This can be considered a "chemical evolution" event since it does not yet meet the qualifications for biological life[13].

1.2.1 Synthesis of RNA Nucleotides

Nucleotide synthesis requires an abundance of starting material on the prebiotic Earth capable of forming the key components of RNA, namely ribose and the four nitrogenous bases. There are studies discussing plausible routes for the prebiotic synthesis of these components, but more challenging is the synthesis of nucleotides from these molecules. Although synthesis mechanisms remain unknown, there are two main domains of thought regarding this topic. The top-down perspective argues for the creation of RNA nucleotides through modern mechanisms, except without the initial help from enzymes[14]. The bottom-up perspective argues for a different set of reactions that the ones in modern life, which would have later came about through evolution[15]. The universal method for RNA nucleotide synthesis has been shown to possibly exist under non-enzymatic conditions, given the presence of iron and other metals[16–18]. These reactions can be enhanced later through the help

of ribozymes. In contrast, RNA may have instead used simpler precursors than the ones used universally in modern life. Many studies have shown the synthesis of nucleobases and ribose sugars from simple reagents, as well as their presence in meteorites[19–21]. Some problems remaining with regards to RNA nucleotide synthesis is its impure chemical production, resulting in many undesired side products[22]. Along with this, the problem of chirality must be solved, requiring us to justify the transition to a homochiral right-handed ribose mixture from a heterochiral prebiotic mixture of nucleotides. Although it is generally accepted that RNA nucleotides were scarce on the primitive Earth, there is evidence showing the synthesis of nucleotides through the help of ribozyme activity[23]. However, a major challenge still lies within the context of prebiotic chemistry. Before we can consider the emergence of ribozymes, we must think about how RNA could have been synthesized prebiotically.

1.2.2 Synthesis of RNA Polymers

There are a variety of factors and challenges to be considered for each prebiotic synthesis reaction step of RNA. To start, we need to consider whether RNA was synthesized in a one-pot reaction or a sequence of reactions taking place in different environments[24]. Sequential synthesis of RNA can be considered advantageous since it reduces the likeliness of producing an intractable mixture compared to the one-pot alternative[24]. However, some promising studies have shown success with one-pot reactions, alongside other interesting observations such as multiple beneficial functions emerging from a single molecule, e.g. phosphate[25]. The importance of phosphate plays into the requirement of the RNA world following the synthesis of nucleotides, namely RNA polymerization. We can generally classify RNA polymerization into three classes. First is the synthesis of RNA through spontaneous ligation of nucleotides or oligomers, resulting in random sequences[26]. The

process of forming strands of RNA is generally thought to require activation of monomers. Along with this, there are many other proposed methods for decreasing the activation energy of nucleotide ligation non-enzymatically[25]. For example, the spatial position and orientation of monomers can be a contributing factor. Studies have shown the use of templating to be especially effective at producing polymers. The use of montmorillonite clay as a catalyst has shown success in binding RNA to produce long polymers[27]. Guiding monomers into the optimal position for monomer backbone interactions seems to be one of the most effective ways of lowering the activation energy. The use of a complementary RNA template to polymerize RNA is thought to be the second class of polymerization. The third class is differentiated from the second through the addition of a polymerase ribozyme.

During random spontaneous polymerization, there exists an equilibrium distribution of lengths. Since bond formation is reversible, we can expect a distribution where the concentrations of polymers decrease by a constant ratio for every nucleotide increase in the length[28]. This constant relies on the rate of bond formation and generally predicts a low concentration of long polymers in prebiotic aqueous environments. In fact, without the use of activated monomers, even short polymers are rare[29]. For this reason, alternatives environmental conditions have been proposed. Examples include the previously mentioned use of clay, as well as lipid and salt environments. Lipids have been shown to limit the available space of nucleotides in a dry phase, driving polymerization due to ordering the nucleotides in a desirable orientation. Salt crystals formed in dry phase also have a similar effect. From this, wet-dry phases can be shown to generate polymers up to 300 nucleotides in length without requiring activation of nucleotides[30,31]. Overall, the wet-dry conditions

promote a higher rate of forming long polymers, but at the cost of requiring a less common type of environment.

RNA synthesis using a complementary RNA template sequence classifies an important stage in the emergence of life. For transfer of information to occur, RNA must replicate whether it be through non-enzymatic or ribozyme catalyzed mechanisms. Both have shown to hold their own challenges. One possibility is that random spontaneous RNA synthesis eventually leads to the emergence of a specific ribozyme, at which point replication would begin. This scenario skips past replication in non-enzymatic conditions. Experimental studies have shown functional self-replicating RNA ribozymes. RNA ligase ribozymes have shown replication by interacting with two specific sequence substrates and catalyzing their ligation[32,33]. Similarly, other RNA ribozymes have been discovered which can self-replicate through the assembly of specific oligomers[34,35]. The issue with this type of replication is with their reliance on inputs of specific sequence substrates, which would be rarely formed spontaneously. Polymerase ribozymes have also been worked on for quite some time, aiming to catalyze primer extension reactions similar to modern protein polymerases. Up to now, there have been many advancements in creating a minimized polymerase ribozyme capable of binding generic RNA sequences and achieving high rates of synthesis[36–38]. However, these ribozymes lack the ability to self-replicate, limiting their usefulness. It seems that being able to replicate under non-enzymatic conditions may be necessary to allow for the emergence of different ribozymes and a variety of available sequence substrates. RNA synthesis through non-enzymatic addition of monomers or oligomers has been shown experimentally[39,40], providing a way of quickly synthesizing long strands along with the transfer of information. One main problem is the difficulty in separating double stranded RNA products without a

catalyst[41]. Multiple rounds of replication seem impossible with longer templates but can still be done under conditions that allow for temperature cycling to drive strand separation.

1.3 Mechanisms of Non-Enzymatic Template-Directed RNA Synthesis

The commonly cited reactions involved in the *de novo* synthesis of a single RNA strand are the activation of nucleotides, followed by the interaction between the 3'-hydroxyl and the 5'-phosphate of two RNA molecules[10]. Some of the plausible forms of activated nucleotides include nucleoside 5'-polyphosphates, and nucleoside 5'-phosphoramidates. A major problem to consider with regards to non-enzymatic polymerization is the rate of the ligation reaction when compared to the hydrolysis rate, or the rate at which polynucleotides break down. For example, a polyphosphate activation group would not be able to compete with hydrolysis in a non-enzymatic setting, and thus would not be an ideal activation mechanism in this context. In contrast, the phosphoramidate activated nucleotides, usually in the form of nucleoside 5'-phosphorimidazolides, are more plausible and used frequently in experimental models. As mentioned previously, the use of a templating substrate is likely to have been involved in non-enzymatic polymerization. As mentioned in a review by Szostack, there are eight major challenges standing in the way of an experimentally reproducible prebiotic RNA replication cycle[41]. One of these challenges is that oligomerization of RNA tends to produce more 2'-5' linkages compared to the 3'-5' linkages that we observe in modern RNA. On the other hand, using montmorillonite clay as a template seems to produce mainly 3'-5' linked oligomers, likely because of the way it orients the RNA prior to ligation[27]. Other minerals should also be tested to examine their effectiveness as templates compared to montmorillonite. Additionally, we can consider the use of RNA templates in this regard. The use of a specific activation method using 5'-phosphor-2-methyl-imidazolide (2-

MeImp) has been shown to maintain strict regiospecificity of 3'-5' linkages, although only for a polymerization of oligo-G's on a poly-C template[42]. Metal ions such as Zn^{2+} can also be used as catalysts to drive increased formation of 3'-5' linkages[43]. Nevertheless, obtaining complete regiospecificity is difficult and rare, even under such conditions. Experimental exploration of regiospecificity still remains limited due to the difficulty of producing mixed 2'-5' and 3'-5' linkages. However, experiments on the effect of ribo/deoxyribo backbone heterogeneity on ribozyme structure and function have so far shown minimal deviations[41]. The implication that backbone heterogeneity may not limit prebiotic RNA replication is promising, but the evidence is limited.

A more severe challenge is the high melting temperature of long RNA duplexes, making strand-separation an unlikely event in most primordial Earth conditions. This is one challenge which we will focus on, along with the problems of fidelity and strand reannealing. If we assume that an RNA complementary strand can synthesize on a template with little error, then the resulting duplex product would be at a dead-end unless there is some separation mechanism. Increasing the error rate could act as such a mechanism for strand separation, since more mismatches would mean greater destabilization[41]. On the other hand, propagation of advantageous information relies on replication with good fidelity. The error rate required for this is approximated to be less than the reciprocal of the number of functionally important bases[41]. This would mean that as the genome size increases, the minimum fidelity must also increase. A study by Leu et al. proposes that the transition from RNA to DNA is important since DNA would decrease the error rate by about half[44]. However, with RNA they report an average error rate of about 17%. The topic of fidelity needs to be explored further to understand more robust mechanisms for decreasing RNA

replication error rate. Using well established free energy parameters for duplex stability based on nearest-neighbor interactions, estimates of error rates have been shown to correlate with experimental observations[44]. Additionally, the parameters account for the GU wobble pairs, a mismatch having similarities to Watson-Crick pairing, resulting in a greater base-pair stability compared to other mismatches. An effect of mutations during polymerization is the decrease in RNA addition rate following a mismatch base pair. Studies have shown that further extension is stalled in both enzymatic and non-enzymatic systems[45]. However, this is seen to have a positive impact on overall sequence fidelity. The stalling effect results in faster complete polymerization of strands with fewer errors. Therefore, high fidelity strands are more readily available for future replication, and the average fidelity is increased[41].

With regards to strand separation, it might be possible that temperature cycling in certain primordial Earth environments played an important role. For example, high temperatures could emerge temporarily in a body of water from hydrothermal vents. As the hot water dissociated into neighboring lakes or ponds, a subsequent cooling event would allow the separated strands to bind and synthesize new complementary products[46]. Generally, it is thought that low product strand concentrations must be maintained to allow formation of new strands. Once a high enough concentration of strands is reached, complementary strand reannealing would occur faster than primer binding and extension[47]. A few alternative mechanisms of non-enzymatic RNA replication have been proposed to solve the problem of strand separation and reannealing. These include strand displacement and rolling circle replication. In strand displacement, short oligonucleotides are said to bind and displace a part of the existing RNA duplex, creating a branch point where further extension of the oligomer can create a replicate while displacing the old complement[48]. However, this process can also

be quite slow as the invade oligonucleotide can also be displaced and separated as the existing complement reanneals again. The rolling circle replication mechanism, common to viroids, overcomes this by assuming a circular strand that does not require a new oligomer to invade. The complementary sequence continues to grow and displace itself, eventually being cleaved to produce a new replicate[47].

Regardless of the mechanism, non-enzymatic extension of RNA on a template should be reasonably efficient and processive to compete against hydrolysis and other potential mechanisms of strand loss. Different activation methods have shown improved primer extension kinetics and product yield, like the use of 2-aminoimidazole activated nucleotides[49]. Replication has also been shown using ligation of tetramers instead of monomer addition, allowing the formation of long RNAs quickly[50]. Nucleotides have been used in more recent studies as activating groups. The mechanism involves the formation of an imidazolium-bridged dinucleotide intermediate which binds next to the primer[51]. The primer extends by one nucleotide and displaces the activated nucleotide. This can improve the rate of monomer addition, and potentially the fidelity of replication due to the increased base pairing stability of dinucleotide binding compared to a single nucleotide. Additionally, these mechanisms have been shown to form replicate strands within model prebiotic vesicles[50,52]. For replication within vesicles, the selective membrane barrier would likely only allow oligonucleotides shorter than tetramers to pass through[53]. This means longer strands would have to be synthesized within the vesicle or encapsulated from the outside environment during vesicle formation. The benefits of compartmentalizing sequences within vesicles partly comes from its function in concentrating the components within. This can assist in driving catalytic reactions[54]. The increase in concentration of other substances, such as Mg^{2+} ions have also

been shown to drive both non-enzymatic and ribozyme based replication mechanisms[36,52]. Lastly, the division of a genotype between several vesicles can be beneficial in the repression of short parasitic RNA which can inhibit the replication of genomic sequences, ultimately destroying the genome[55]. Once a genome or a set of useful genetic sequences have emerged, their maintenance would be essential for prolonged replication and the potential for selection and evolution.

1.4 Virtual Circular Genomes

The idea of a virtual circular genome was proposed by Zhou et al. with the aim of solving some of the persisting difficulties facing long genomic sequences in the RNA world[56]. Namely, we have already mentioned uncertainties regarding strand elongation over long template sequences, strand separation, and strand reannealing. Furthermore, a constant input of defined RNA primers is required for continuous replication, even for short strands. In a circular genome, this is not an issue since there is no specific replication starting point. Though, circular strands require mechanisms for cleavage and re-circularization, usually requiring the help of a catalyst or ribozyme[57]. In response, Zhou et al. propose a virtual circular genome, where the circular sequence on both complementary strands is split into a set of overlapping oligonucleotide sequences comprising all or most possible start and stop sites[56]. Genome replication could occur through template-directed primer elongation mechanisms. We can likely assume a decrease in the concentration of strands as RNA length increases, but this interestingly results in an advantageous property for the virtual genome hypothesis. It can be shown that elongation of all sequences in the mixture using genomic templates by as little as one nucleotide may result in replication of the entire set of genomic strands.

Replication is still limited to environments that can allow for strand separation, namely through temperature cycling. Over many cycles, strands can bind in various conformations eventually leading to at least a small increase in length for all genomic RNA sequences. The presence of a specific set of complementary sequences means every sequence will always have a template for continued growth, which also drives the potential for full replication of the genomic set. Although this process is intriguing, it is important to contemplate how such a mixture could emerge or be maintained. The authors mention two potential ideas with regards to the emergence of a virtual circular genome. First, encapsulation of a high concentration of RNA polymers within vesicles could eventually lead to spontaneous emergence of a genomic set. Second, the set of oligonucleotides could emerge over time using a long circularized physical sequence as a template[56]. With regards to maintenance of the genome, it was proposed that this would be dependent mainly on the copy number of the different sequences. A high copy number would make it quite unlikely for parts of the genome to be lost in the case of vesicle division.

Within this scheme, RNA sequences encoding ribozymes would have to be assembled from the set of shorter strands in order to function. The main mechanism mentioned by Zhou et al. involves fast ligation of oligonucleotides assisted by complementary splints[56]. Evolution of ribozymes is mentioned to be dependent on copy number of genomic strands. A high copy number would make it difficult for random mutations to manifest their effects. However, it is stated that over several random sequence segregation events during vesicle division, mutant sequences could become fixed and demonstrate their impact[56]. Lastly, the emergence of an efficient polymerase ribozyme over the evolutionary period could result in the transition away from a virtual circular genome, possibly to a more sophisticated physical genome structure.

The complexity of this hypothesis poses a challenge for experimental investigation. Synthesis of virtual genome sets, and construction of the appropriate vesicles and environmental conditions can be quite tedious and expensive. Using a theoretical modelling approach can guide future experiments by providing insight into potential outcomes.

1.5 Realistic Computational Modelling of RNA Replication

Computational modelling can be of great help in uncovering uncertainties regarding non-enzymatic RNA replication which would be difficult to do through experiments. This is mainly resulting from the lower requirement for time and cost, producing predictive insights which can lead to future productive experiments. We will be focusing on constructing realistic models of RNA replication. These models are generally built upon parameters provided through experimental data. They can provide understandable results which can guide short-term research, overall becoming a powerful form of utility for RNA world research as the amount of experimental data increases. Simulations are constructed mainly following the rules of thermodynamics and chemical kinetics. Randomized algorithms such as the Monte Carlo or Gillespie algorithms can be used to generate stochastic reaction events involved in RNA replication, such as base pairing, ligation, polymerization, strand separation, and hydrolysis. The probability of reaction events depends on their rates which can be directly obtained or predicted from experimental research.

Thermodynamic parameters for RNA folding stabilities are available from studies that use nearest neighbor prediction methods[58]. These predictions are mostly derived from optical melting data from experiments, and account for the stabilizing effect of stacking interactions between adjacent base pairs[59]. They have been determined for free energy and enthalpy

changes of Watson-Crick helices and GU pairs[60], which can be used for predicting the probability of nucleotide base pairing during non-enzymatic template-directed RNA polymerization. These thermodynamic parameters can be useful in modelling primer extension rates during non-enzymatic RNA replication. Alternatively, there are available data for experimentally measured primer extension rates by single nucleotides, which can be directly implemented[44,61]. Though, this could limit the outcome of simulations based on the experimental error or constrained amount of data.

1.6 Thesis Aims

Within this thesis, we aim to construct computational models to investigate non-enzymatic RNA replication and provide insights regarding the emergence of long RNA polymers. Generally, we are focusing on the transition from the non-enzymatic to the enzymatic RNA world, so we assume the availability of nucleotides containing the four nitrogenous bases present in modern RNA: adenosine (A), cytosine (C), guanosine (G), and uracil (U). Specifically, we look at models of non-enzymatic template-directed RNA synthesis, as a potential source of creating long RNA sequences and propagating information through replication.

In chapter 2, we consider the relationship between RNA primer extension rate on a template and thermodynamic parameters of base pairing. We utilize a database containing nearest neighbor free energy parameters for Watson-Crick and GU base pairs[60]. These thermodynamic parameters are used in our model to predict non-enzymatic primer polymerization rates, depending on the base being added. We assume that the stability of base pairing of the incoming nucleotide is the limiting factor during polymerization. Using

experimentally measured nucleotide addition rates, we tried to fit our unknown parameters. We find that extension rates do depend on thermodynamics, but we cannot be certain about the accuracy of our model to predict RNA primer extension rates. Simply, there are too many unknown parameters and not enough experimental rate data to construct a reliable predictive model.

In chapter 3, we construct computer simulations of non-enzymatic RNA replication starting with a diverse mixture of short oligomers. Our goal involves revisiting the strand reannealing problem during RNA replication when using temperature cycling. We find that in our more complex model, reannealing does not occur due to the high sequence diversity and numerous potential helix configurations. Small extension of strands over many cycles produces long RNA sequences, even at low polymerization and high mutation rates. This illustrates a possible scenario where sequences long enough to act as ribozymes can emerge.

In chapter 4, we discuss the hypothesis put forth by Zhou et al. regarding virtual circular genomes and contemplate their plausibility within the context of our simulations. We find that our simulations are unlikely to converge onto a specific set of sequences capable of encompassing a virtual genome. If we start with a random mixture of sequences, the diversity needs to be just right in order to create a specific path of overlapping sequences. Too much diversity creates many branching paths unindicative of a specific virtual sequence, and too little diversity cannot create long enough genomes capable of carrying useful information. The conditions for achieving and maintaining a virtual genome seem to be too idealistic, leading us to favor a physical genome structure, such as rolling circles.

Chapter 2: Predictive Model of Non-Enzymatic Primer Extension using Thermodynamics of Base Pairing

2.1 Methods

Considering an RNA template-primer duplex, the model assumes a primer extension rate governed by hydrogen bond formation of incoming nucleotides to the complementary template. The regions of importance included in the model are the last base pair of the helix and the potential base pair formed through addition of a nucleotide to the first open template base. We consider two main parameters, the fraction of time the incoming nucleotide is base-paired, and the ligation reaction rate. Let the net rate of extension, $R_{ext}(X|ZW, Y)$, be equal to the product of the kinetic and thermodynamic terms:

$$R_{ext}(X|ZW\ ,\ Y) = R_{lig}(ZW\ ,\ XY) * f(ZW\ ,\ XY),$$

where ZW is pair 1, XY is pair 2, $R_{lig}(ZW, XY)$is the ligation rate given that X is annealed to the template, and $f(ZW\ ,\ XY)$ is the fraction of time X is annealed to the template. Assuming a template strand with a primer growing in the 5' to 3' direction, ZW is the final bound nucleotide pair and XY is the incoming nucleotide (X) onto the open template base (Y), adjacent to W. This results in a (ZW , XY) base stacking interaction which contributes to the thermodynamic stability of the base pair. $f(ZW,XY)$ is determined using an updated nearest-neighbour (NN) stacking energy dataset[60]. $R_{lig}(ZW\ ,\ XY)$ is assumed to be some constant which we will approximate using experimental data.

We assume that monomer binding and unbinding is much faster than ligation, creating an equilibrium between the off and on rates of monomers binding next to the end of the primer. Therefore, the fraction of time monomer X is bound to the template is approximated as:

$$f(X|ZW\ ,\ XY) \approx \frac{[X]e^{\left(-\frac{\Delta G^0_{ZW,XY}}{RT}\right)}}{1+[X]e^{\left(-\frac{\Delta G^0_{ZW,XY}}{RT}\right)}},$$

where [X] is the concentration of nucleotide X in solution, R is the universal gas constant, T is the temperature in Kelvin, and $\Delta G^0_{ZW,XY}$ is the standard free energy of monomer binding. The NN stacking energies are known for every correct Watson-Crick set of pairs ZW and XY, as well as for GU pairs. However, the observed free binding energy is roughly one kcal less stable, meaning less negative, than the NN estimate[62], and so we will refer to this difference as $\Delta G^0_{Partial(ZW,XY)}$. This gives us:

$$\Delta G^0_{ZW,XY} = NN(ZW,XY) + \Delta G^0_{Partial(ZW,XY)}.$$

We consider the current model to have one free parameter for fitting the $R_{lig}(ZW\ ,\ XY)$ constant. This means that R_{lig} is a single constant k_{lig} which is independent of ZW and XY. In other words, the incoming nucleotide base will not determine the rate of the phosphodiester bond formation reaction. Here, we assume that the net extension rate is being dominated by thermodynamics, and that variations between the Watson-Crick and mismatched base pairs are accounted for in the free energies of stacking. Also, increasing the number of parameters runs the risk of overfitting and minimizing the real-world application of the model. We would not want to fit the noise and experimental error of the observed rates, but rather a universal

constant. Since the NN stacking free energies are unavailable for non-Watson-Crick base pairs, we also left this as a variable parameter to be fit using experimental data. We used data from the results of two studies that measured the extension rate for each possible XY pair using similar experimental procedures[44,61]. These experimental rates are shown in Table 2.1. We assigned values for the variable parameters based on the minimum residual sum of squares between the experimental and theoretical extension rates. To compare our predicted theoretical rates, we plotted them on a log scale against the experimental rates from both studies. Our theoretical outputs took into consideration the experimental conditions[44]. Free energy calculations were converted from the original temperature of 37 to 22 degrees Celsius. We also used the same primer-template complex base pairs and included four times the U base nucleotide concentration. For the calculation of change in free energy, the neighbouring base pair ZW was set to GC.

2.2 Results

Using available data of RNA base pair stacking energies and experimentally obtained primer extension rates, we created a model to predict primer extension rates based on thermodynamic parameters. The rates were calculated for every possible base pair. Our theoretical extension rates are compared to the experimental rates in figure 2.1 for each base pair between the incoming nucleotide and the template. A perfect prediction of the experimental rates is shown by a line where each measured rate equals the theoretical rate. Many of the measured experimental rates fall below the line of reference. The unlabelled group of points to the left are the mismatched base pairs for which thermodynamic free energies were unknown. A single parameter value was set for these mismatched pairs. The four unlabelled mismatches with higher predicted rate values are the base pairs where U was

the incoming base and was set to four times the concentration of other nucleotides. Values for extension rates reported in the two studies[44,61] and our estimated rates are listed in table 2.1 corresponding to each unique base pair interaction.

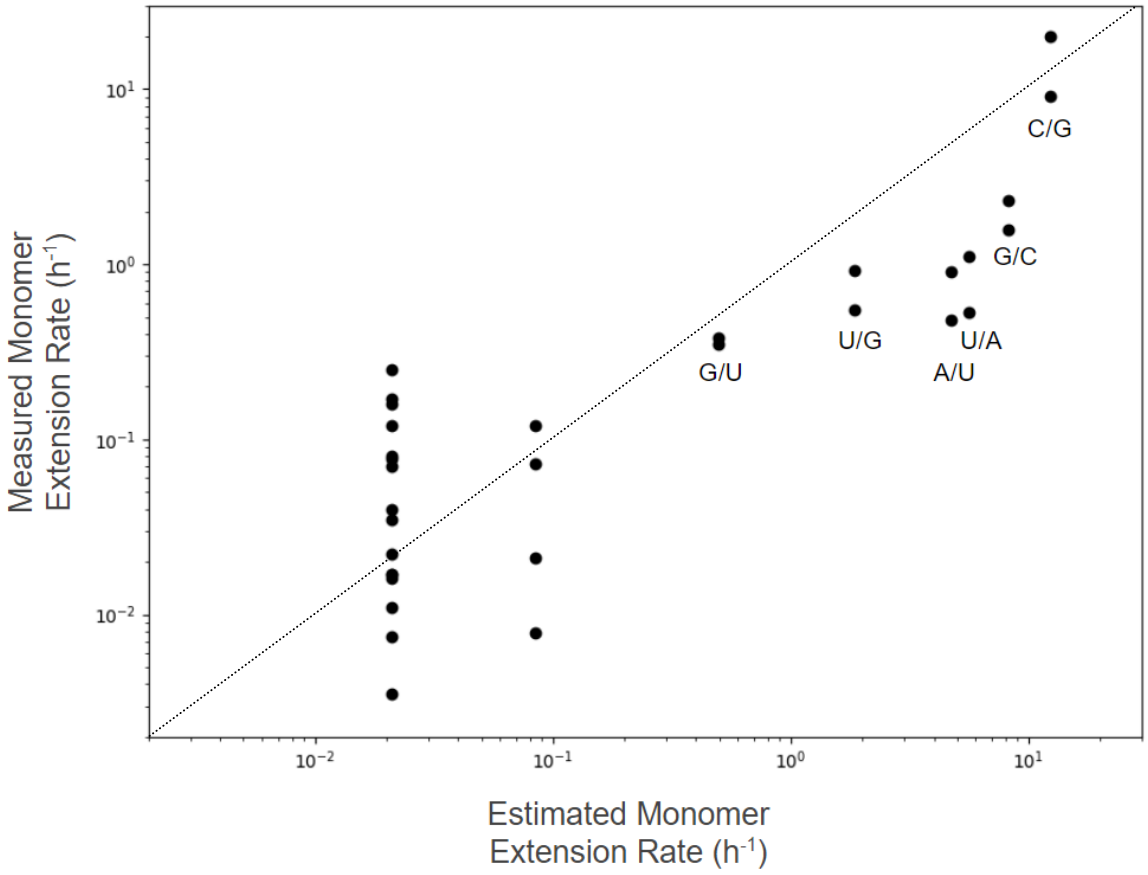

Figure 2.1. Comparison of experimentally measured and theoretical monomer extension rates (h^{-1}) obtained from our model centered on thermodynamic base pairing stability. Data points are labelled with incoming nucleotide and template nucleotide bases, respectively. Unlabelled points include all non-Watson-Crick mismatched base pairs except for GU pairs. The dotted line indicates a perfect fit and acts as a point of reference, where estimated and measured rates are identical.

For the Watson-Crick and GU base pairs, the difference in rate values do not follow a similar trend between the estimated rates compared to the measured rates. GU and AU base

pairs have comparable rates in the measured case but are quite different in the estimated rates specifically with G binding to U being much slower. C binding to G is also at a relatively high rate in the measured case, whereas its only slightly higher than the G to C binding case in the estimated rates. In both cases, the increase in extension rate follows the same trend, except for the switch between the U/G and A/U base pairing. The model predicted a higher extension rate when A binds to U compared to U binding G, whereas the opposite is true for the measured rates. Overall, the extension rates predicted by thermodynamics are generally higher for the Watson-Crick and GU base pairs compared to the experimental rates. Lastly, in the case of our fitted thermodynamic parameter used for mismatched base pairs, we observe a higher extension rate when U is the incoming nucleotide base. The measured rates do not show such a drastic increase in the rate for U binding.

Table 2.1. Comparing predicted primer extension rates after fitting of free parameters with experimental data of extension rates from studies by Leu et al.[44] *and Bapat et al.*[61].

Template	Incoming Nucleotide	Theoretical Rate (h^{-1})	Experimental Rate (h^{-1}) from Leu et al.	Experimental Rate (h^{-1}) from Bapat et al.
C	G	8.2258	2.3	1.58
C	U	0.0843	0.0078	0.021
C	C	0.0211	0.0035	0.022
C	A	0.0211	0.0075	0.017
G	C	12.324	20	9.03
G	U	4.7219	0.91	0.48
G	G	0.0211	0.25	0.16
G	A	0.0211	0.035	0.04
A	U	5.606	1.1	0.53
A	C	0.0211	0.078	0.08
A	G	0.0211	0.12	0.17
A	A	0.0211	0.016	0.07
U	A	1.8618	0.55	0.92
U	C	0.0211	0.011	0.38
U	G	0.4987	0.35	0.12
U	U	0.0843	0.073	0.017

2.3 Discussion

In general, our results show that based on the thermodynamic parameters used, estimated extension rates tend to be higher than the experimental rates for Watson-Crick and GU base pairs. These base pairs are the only ones for which we were able to obtain the NN thermodynamic parameters. Our fitted free energy stability parameter for the non-Watson-

Crick pairs produced an extension rate which is somewhat close to the average measured rate for mismatches, which is about $0.065h^{-1}$. However, the mismatched measured rates fall within quite a large range, with the lowest being about $0.0035h^{-1}$ for C to C binding and the highest at $0.55h^{-1}$ for A to U binding. In a realistic model, this variation should be accounted for. Thus, we would require the thermodynamic parameters to be provided as they have now for GU pairs, or we would need to include extra parameters to account for the rate variation. Note that for this model, we only allowed one parameter k_{lig} to account for the rate of ligation between bound nucleotides. One way to obtain an overall closer fit to the experimental data would be to increase the number of parameters that define the ligation rate. If we increase the number of free parameters to two, we could account for a different rate of ligation depending on whether the bound base pair is a Watson-Crick base pair or not. This assumes that Watson-Crick base pairs sit differently compared to mismatches and would be in a more optimal position to drive the ligation reaction. Following this idea, the most complex case would include a different parameter for each base pair and each neighboring base pair, resulting in 256 free parameters. This case would not be possible since we would require 256 observed rates within the experiment. Additionally, when we increase the number of parameters, we end up creating a model which would only be good at predicting the rates within the experimental condition. Rather, we would like to generalize the model be used in simulations which contain a variety of sequences and conditions not used in the experiment. Also, there are quite a few uncertainties regarding the thermodynamic parameters, which could likely account for the differences in the measured rates.

The individual NN parameter predictions for Watson-Crick and GU base pairs are reported for dimer base pairs. Since we consider monomer addition in our model, it is

probable that the NN stacking energies would not directly translate to this case. There are other NN parameters which are reported, like for the initiation of a duplex, or when dealing with helices ending in an AU base pair[59]. These parameters were not considered since it was uncertain if they would be relevant in the case of monomer addition. We mainly wanted to test a simple model to understand the dependence of extension rate on the thermodynamics of base pairing. It seems that this relationship does exist, but without the inclusion of more parameters from a greater amount of experimental data, the applicability of this model remains inconclusive. It may be true that the predictions can get close to the experimental data if we were to fit a greater number of parameters. However, we should consider that implementing the experimental rate data directly within a realistic computational model would be simpler and more accurate.

It is also clear that the rates obtained from the experimental studies do not encompass a fully reliable sample. Apart from the low number of data points, there seems to be large differences in the rates obtained between the studies. Both studies follow the same experimental protocol and conditions, which means there may have been some considerable room left for experimental error due to the protocol design. They used ImpN activation of their RNA nucleotides, and specific concentrations to achieve detectable primer extension. Fitting extra parameters to only this sample set would not allow the model to achieve a universal applicability. Overall, the experiments have their limitations, and more data is necessary to make a reasonable model of this nature.

Being able to predict primer extension rates with reasonable accuracy can allow for complex modelling of non-enzymatic RNA replication. Realistic models could be constructed involving base pairing configurations different than those measured in the experiments. This

could let us make predictive observations of RNA replication outcomes in a variety of theoretical conditions and on a larger time scale compared to experiments. For example, free energy of base pairing can be extended to dimer or oligomer binding, allowing the prediction of extension rates by these longer strands. The prediction of monomer addition can be used in complex models involving polymerization via monomer addition. These scenarios would include mixtures of various random sequences giving rise to different base pairing configurations and would be difficult to produce experimentally. A simulation of RNA template-directed synthesis using such mechanisms is shown in chapter 3, but experimental rates are implemented directly. An accurate predictive model could allow for quantitative and realistic results from the simulation, given that the other parameters are reasonably set according to experimental or early prebiotic Earth conditions. Overall, this study has accentuated the necessary interplay between theoretical and experimental frameworks required for the future of RNA world research.

Chapter 3: Computer Simulations of Non-Enzymatic Template-Directed RNA Synthesis Driven by Temperature Cycling

3.1 Methods

We carry out stochastic simulations which follow the steps of RNA synthesis in a finite volume of solution containing monomers and RNA strands. The simulation incorporates inflow of monomers and short oligomers created by random polymerization, pairing of complementary strands, strand growth via primer extension, nucleation of new primers from monomers on a template, ligation of neighbouring strands bound to the same complementary strand, melting of paired strands driven by temperature cycling, and loss of strands via outflow. At any given point, the program stores the sequences of all strands present, the positions of all helices connecting strands, and the number of available single monomers of each type of nucleotide. Rates of all possible reaction steps are calculated, and the Gillespie algorithm is used to select one possible reaction step with a probability proportional to its rate. A summary of the simulation parameters, along with their standard value used is shown in Table 3.1.

The system begins with an initial number of monomers, N_0 = 5000, of each nucleotide A, C, G and U, and an initial number, N_{init} = 500, of oligomers of lengths in the range 4-10. These oligomers are assumed to be generated by random polymerization without a template and consist of random sequences of the four nucleotides. We suppose the oligomer length *n* has an exponential distribution, with a probability distribution $P(n) = A\lambda^{n-4}$, where λ = 1/2 and *A* is the constant necessary to normalize the distribution. This is the equilibrium length distribution likely to be achieved under spontaneous random RNA polymerization[28]. We cut

off this distribution at a maximum of 10, because longer sequences are very rare, and to make clear that sequences longer than 10 originate by templating reactions, not random polymerization. We assume that length l_0 = 4 is the shortest length of primer that is stable long enough to initiate new strand growth. Dimers and trimers will also be present in the mixture, but we ignore them in this simulation because we assume that their rate of detaching from a template will be high, and that detaching is likely to occur before addition of new monomers to the primer. If dimers and trimers were included, there would be many on and off reaction steps that would change nothing but would slow down the program considerably. The mean length of the initial oligomers is

$$n_{init} = \frac{\sum_{n=4}^{n=10} n\lambda^{n-4}}{\sum_{n=4}^{n=10} \lambda^{n-4}} = 4.94. \quad (1)$$

Table 3.1. Standard values of the parameters used in the simulations.

Parameter	Standard Value	Meaning
N_0	5000	Initial number of monomers of each type (A, C, G and U)
N_{init}	500	Initial number of random oligomers
ϕ	0.05	Fractional flow rate = fraction of strands lost at the end of each cycle
N_{inflow}	25	Number of random oligomers flowing in per cycle
l_0	4	Minimum possible length of primers and helices
λ	0.5	Parameter controlling exponential distribution of primer lengths
k_{nuc}	0.1h^{-1}	Nucleation rate constant per hour at the initial nucleotide concentration.
k_{ann}	10 h^{-1}	Annealing rate constant per hour per window pair
k_{add}	10 h^{-1}	Rate constant for monomer addition at the initial nucleotide concentration.
k_{lig}	1 h^{-1}	Ligation rate constant
k_{melt}	1 h^{-1}	Melting rate of a helix of length l_0
$\Delta G/kT$	2	Stacking free energy per base pair, relative to kT.
e	0.01	Rate constant for each incorrect base addition as a fraction of k_{add}
T_{grow}	6 h	Length of growth phase

Strands in the mixture may be connected by double stranded regions (helices). New helices may be formed by nucleation of monomers on an existing strand, or by annealing two existing strands with complementary sequences. The minimum allowed helix length is $l_0 = 4$. The probability of a new helix nucleating on a given strand is proportional to the number of unpaired windows of length l_0 on this strand. For a sequence i of length n_i the number of windows is $w_i = n_i - l_0 + 1$ at the beginning of each cycle when there are no other helices already present. This number is reduced when helices form because each helix blocks the formation of further overlapping helices. The program keeps track of the number of available windows on each strand. For each available window, the nucleation rate is

$$r_{nuc} = k_{nuc} \frac{N_1 N_2 N_3 N_4}{N_0^4}, \quad (2)$$

where k_{nuc} is the nucleation rate constant, and N_1, N_2, N_3, and N_4 are the number of available monomers of the required type to form the new tetramer. The tetramer is assumed to be exactly complementary to the template strand. At the beginning of the simulation, the number of each monomer in the system is N_0. Thus, according to equation 1, the nucleation rate is equal to k_{nuc} per window when the monomers are at their initial concentration, and it decreases in proportion to monomer concentration to the power 4 when the monomers are used up by the polymerization process. When a new tetramer is formed, the corresponding monomers are removed from the count of free monomers.

To form a new helix by annealing existing strands, there must be an available window of length 4 on two different strands. We define the attempted rate of annealing per pair of windows as $r_{ann} = k_{ann}/N_0$. For any given available window, let the total number of other

windows in the system that are potential partners for pairing be w_{tot}. The total attempted rate of annealing to the first window is

$$r_{ann}^{tot} = k_{ann} w_{tot} / N_0. \quad (3)$$

There are initially N_0 monomers of each type of nucleotide. If all these monomers were turned into tetramers, then w_{tot} would be N_0. Thus, by scaling the annealing rate by $1/N_0$ we keep the total rate of annealing to one window proportional to the total concentration of sequence windows in the system. Note that the above rate is an *attempted* rate of annealing. We choose random pairs of available windows at this attempt rate, but the annealing only occurs if the sequences are complementary.

If a helix forms in the middle of two sequences, the ends of the sequences are single-stranded tails which cannot grow (as in Figure 3.1A). When the 3' end of a strand is at the end of a helix (as in Figure 3.1B, orange sites), monomer addition can occur (*i.e.*, primer extension occurs). Monomer addition is directional and only occurs at 3' ends. When the 5' end of a strand is at the end of a helix (as in Figure 3.1C, green sites), monomer addition cannot occur, but this 5' end can be ligated to the 3' end of another strand if the end of the other strand grows to be adjacent to this point.

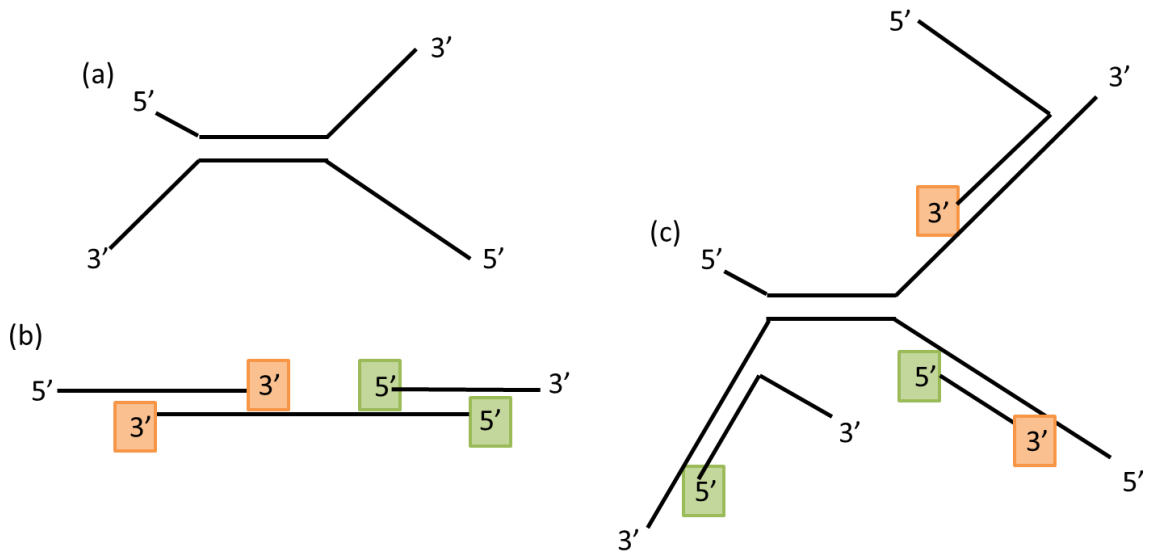

Figure 3.1. Examples of structures that form via annealing of strands. (A) If a helix forms in the middle of the strands, so that the ends of the strands are in single-stranded tails, then no growth is possible at these ends. (B) If the 3' end of a strand is the last base in a helix (orange squares), this is a site for monomer addition. If the 5' end of a strand is the last base in a helix (green squares), we assume that monomer addition cannot occur, but this is a potential site for ligation, if the 3' end of another strand grows to be adjacent to this site. (C) Connection of multiple strands forms branching clusters with many tails and many potential points of sequence growth. We assume that formation of an additional helix is not possible between strands that are already in the same cluster. This prevents unrealistic loops and knots forming within a cluster.

Whenever there is a 3' end that is paired and the next site on the template is unpaired, monomer addition can occur at a rate

$$r_{add} = k_{add} N_1 / N_0, \tag{4}$$

where N_1 is the number of bases in solution of the type that are complementary to the template. We scale the rate by the initial number of monomers, N_0. Thus, the addition rate is k_{add} when the monomer concentration is equal to the initial monomer concentration and decreases in proportion to the monomer concentration that remains available when the

monomers get used up. Initially, we consider the limiting case where there is perfect pairing between template and the growing complementary strand, *i.e.*, there is zero error rate. In this case, the only type of nucleotide that can be added is the one complementary to the template. We later allow for non-zero error rates of monomer addition where a templated strand can grow by a non-complimentary base at a lower rate.

Whenever there is a 3' end that is paired and the next site on the template is paired to the 5' end of another strand, then ligation can occur at rate k_{lig}. This rate is constant, independent of the concentration of monomers and strands.

We keep track of clusters of sequences that are connected by helices. At the beginning of each growth phase, each strand is defined to be in its own cluster. When two sequences are connected, they are placed in the same cluster. We impose the restriction that a new helix can be formed between two strands only if they are not already in the same cluster. This means that all the clusters that form have a branching tree structure (as in Figure 3.1C). In an early version of this program in which this restriction was not imposed, we found that multiple connections formed between sets of sequences that were already in the same cluster. This resulted in a dense structure of entangled knots and loops. Such a structure would be impossible to achieve in three-dimensional space with real molecules because of excluded volume restrictions and the finite length and flexibility of strands. Our simulation does not account for excluded volume and three-dimensional coordinates of the strands. We impose the branching cluster rule as a simple way of preventing the formation of unrealistic loops and knotted structures.

Melting of helices can occur during the growth phase. When this occurs, the two strands forming a helix separate, and the cluster containing the helix is divided into two

separate clusters. The melting rate for a minimum-length helix of length $l_0 = 4$ is k_{melt}. The melting rate for a longer helix of length l is

$$r_{melt}(l) = k_{melt} \exp\left(-\frac{(l-l_0)\Delta G}{kT}\right), \quad (5)$$

where ΔG is the average stacking free energy per additional base pair in the helix. For simplicity, we treat stacking free energy using a single average value, and we do not consider sequence-specific stacking parameters.

Each growth phase lasts for a time of $T_{grow} = 6$ hours. At the end of this time, the high temperature phase occurs. All helices are melted immediately. Flow occurs into and out of the system with a fractional flow rate ϕ. Each strand is lost from the system with a probability ϕ. A fraction ϕ of monomers of each type are lost from the system. New monomers and primer strands are then added at the initial concentration. Thus, ϕN_0 monomers of each type are added, and $N_{inflow} = \phi N_{init}$ new oligomers are added with the same exponential length distribution as the initial sequences.

3.2 Results

We wanted to understand whether template-directed RNA synthesis mechanisms could lead to the creation of longer sequences, rather than just being limited to the replication of the longest template strand. Figures 3.2 and 3.3 show the results of four separate simulation with four different values of the monomer addition rate k_{add}. All other parameters take the standard values in Table 3.1. The data is plotted at the end of each growth phase before melting of helices and loss of strands via outflow. We observe an increase in mean and

maximum sequence length over time due to primer extension and ligation reactions. The mean length begins at n_{init} = 4.94, which is about the mean length of inflowing oligomers. There is a constant increase in mean length until a stationary value where the loss of old (long) strands is balanced by the inflow of new (short) oligomers. For the highest k_{add}, the mean length of sequences reaches up to 50 nucleotides in length, which is substantially higher than n_{init}. For the lowest value of k_{add}, the mean length still increases significantly to about 9 (Figure 3.2A). Increase in the monomer addition rate clearly increases the sequence length via a higher rate of RNA extension, but with greatly diminishing returns. We also show the maximum sequence length in the population at each cycle (Figure 3.2B). The maximum lengths reach a substantially higher scale compared to the mean lengths. Limited by memory storage in the program, we store the sequence length in an array with a maximum length of 300, so sequences longer than this are not permitted. This limit is briefly reached in the highest k_{add} case. The lowest k_{add} value shows a peak maximum sequence length of about 90, and as k_{add} increases, we can see this peak reach a length of over over 200 nucleotides. The increase in the peak maximum length also seems to correlate with an increase in the range of maximum lengths over the different cycles. We also measured the number of sequences in the mixture, N_{seq} (Figure 3.2C). In each run, N_{seq} begins at the initial value N_{init} = 500. It rises rapidly for the first few cycles and then falls to a steady state value where the increase in strands due to inflow and nucleation is balanced by the outflow. The lower k_{add} value correlates with a higher number of sequences within the system. Note that new sequences can only be added through nucleation on an open template or inflow of random oligomers, which is at a constant rate $N_{inflow} = \phi N_{init}$.

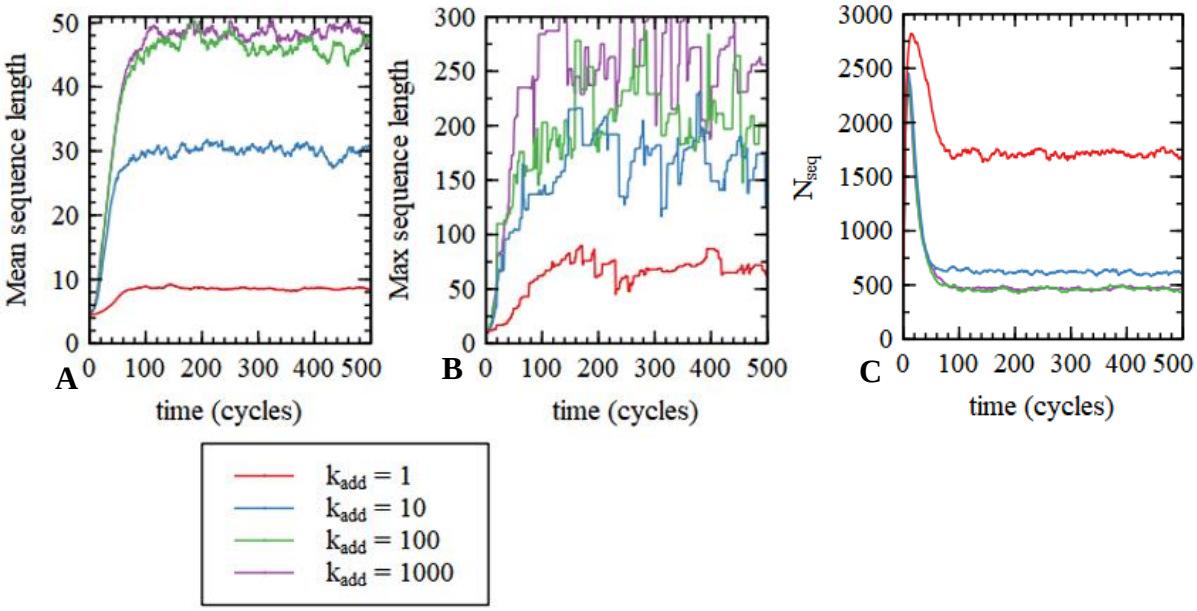

Figure 3.2. Comparing the change in (A) mean and (B) maximum sequence lengths, and (C) total number of sequences over 500 cycles of four separate simulations varying the monomer addition rate value k_{add}.

To better understand the distribution of the RNA sequence lengths, we measured the lengths of each strand at the end of the growth period of each cycle. In Figure 3.3A, we averaged this data over the last 250 cycles of the simulation runs, which lasted 500 cycles. Note that all the data points for the rest of Figure 3.3 were also averaged using the last 250 cycles. We show the steady state length distribution *N(n)* over the lengths up to 100 (Figure 3.3A). Past the 100-length mark, the sequences had a low sample size which resulted in large fluctuations of their distributions. The measured distribution of lengths is decreasing with *n* somewhat at a negative exponential rate. Yet, this rate is much slower than our set length distribution of starting sequences arising from random polymerization. Also, it does not seem to correspond to a single exponential. In the three higher monomer addition rate values, we see a considerable number of long strands in the 50-100 nucleotide range. The smallest k_{add}

value still produced some long sequences as well, but mainly resulted in a much larger population of short to medium length strands.

We were interested in knowing what the growth rate per cycle looked like in terms of the specific change in sequence length. We calculated the mean increase in length of a strand in one cycle as Δn, defined as the length at the end of the cycle minus the length at the start of the cycle. We show these data as a function of the length at the start of the cycle (Figure 3.3B). Expectedly, Δn is larger overall for higher values of k_{add}. Regardless, Δn is quite small even for the largest addition rates. For k_{add} = 1000 h^{-1}, Δn is close to 5 for short sequences (tetramers and pentamers) and decreases to around 1 for longer sequences of length 100. For k_{add} = 1 h^{-1}, Δn is around 0.3, and is almost independent of n. We were also interested to see the number of helices being formed on templates, since we were allowing multiple strands to bind and form clusters. We measured the mean number of separate helices $h(n)$ in which a sequence is bound and show it as a function of the length of the sequences at the end of the growth period. We observe an $h(n)$ close to 1 for short sequences, followed by a linear increase with n. For the two higher k_{add} values of 100 h^{-1} and 1000 h^{-1}, $h(n)$ falls close to 3 at n = 100. Lower k_{add} values show higher number of helices up to about 5-7 for medium to long strands. Overall, the number of helices increases proportional to the strand length, meaning that longer strands bind several helices and the mean length of each helix stays constant.

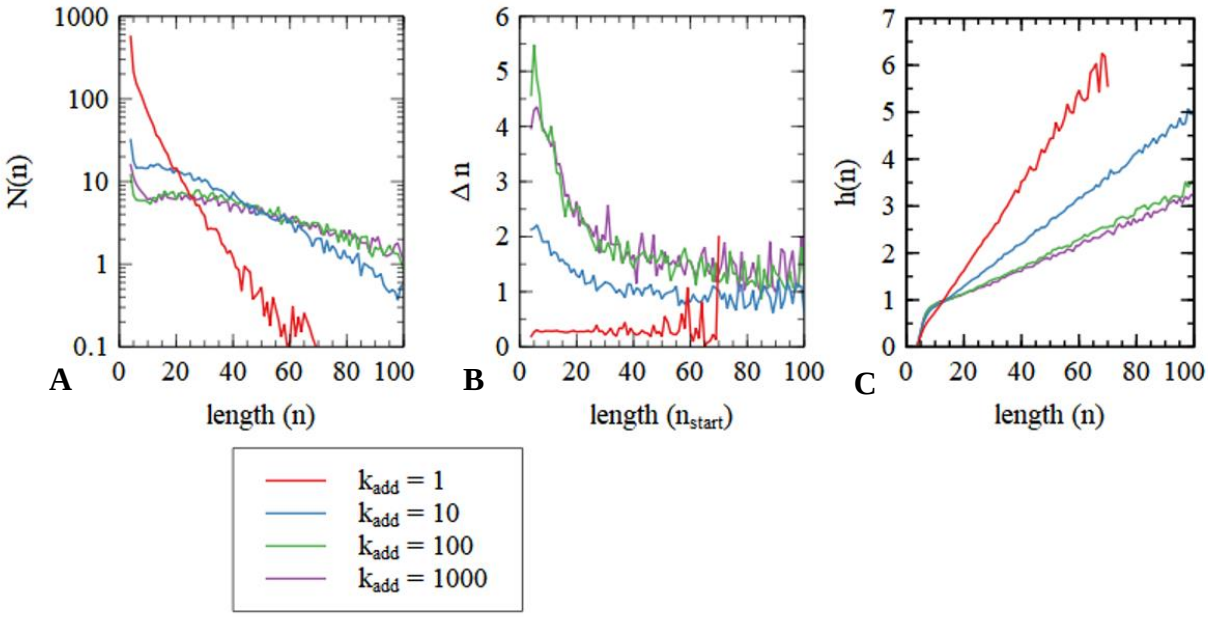

Figure 3.3. Distribution of (A) sequence lengths, (B) mean increase in length per cycle and (C) mean number of helices per cycle as a function of RNA strand length for four simulations varying by the monomer addition rate value k_{add}.

To further understand the simulation outcomes with regards to varying parameters, we ran four separate simulations differing only in the nucleation rate k_{nuc} and keeping all other parameters as standard. This is similar to our analysis above, except we wanted to observe how varying the nucleation rate would impact our measured variables. We later repeat these analyses with only varying the strand annealing rate k_{ann}. When looking at changes in mean lengths of sequences over time, we observe that a lower primer nucleation rate results in an increased mean sequence length (Figure 3.4A). At a k_{nuc} value of 0 h^{-1}, the mean sequence length is increased up to nearly a length of 40 nucleotides. This is the most a change in k_{nuc} can contribute to increasing sequence length, but also comes at the cost of never synthesizing more sequences than the starting population number. On the other hand, increasing k_{nuc} steadily lowers the mean length increase over time. With regards to the change in maximum sequence length, there is the same correlation to varying k_{nuc} as seen with the

mean length, but the difference between the results of k_{nuc} values is not as significant (Figure 3.4B). With increasing k_{nuc} values, the number of sequences present over time also increases, both at the initial peak and steady state concentrations (Figure 3.4C). It is important to note that at a k_{nuc} value of 0 h^{-1}, the number of sequences is maintained roughly at the initial strand number N_{init} of 500.

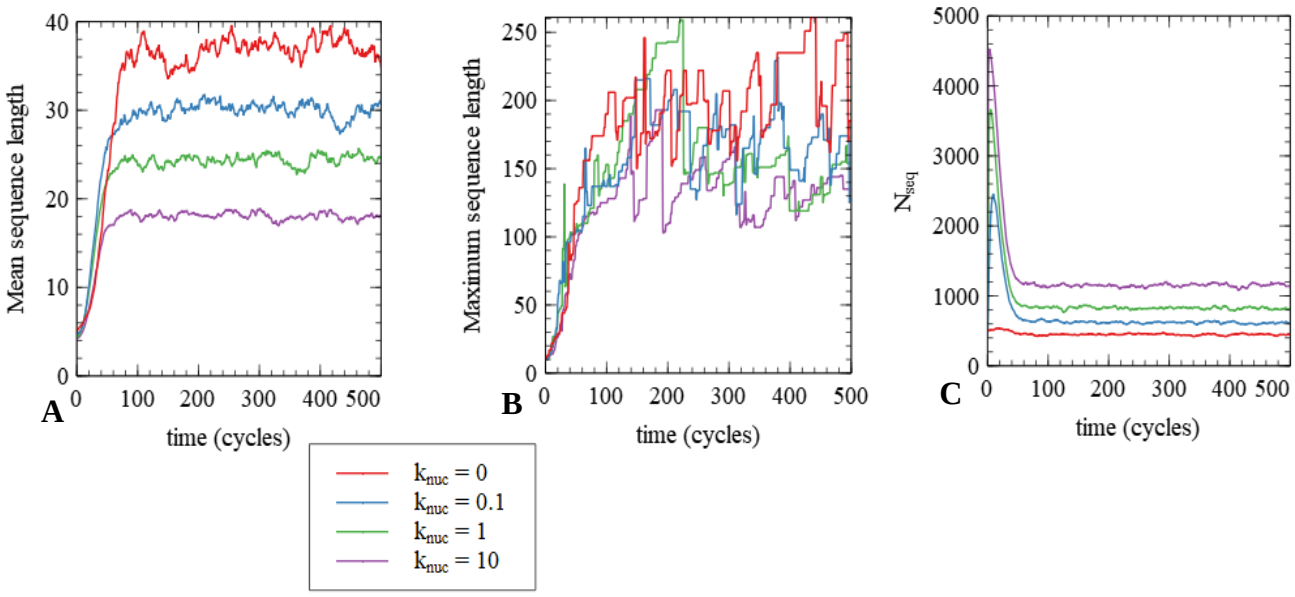

Figure 3.4. Comparing the change in (A) mean and (B) maximum sequence lengths, and (C) total number of sequences over 500 cycles of four separate simulations varying by the nucleation rate value k_{nuc}.

Figure 3.5 shows the results for the variables measured over sequence lengths up to 100, as previously shown with varying k_{add} values. We show the average length distributions, length increases, and helices formed. With regards to the sequence length distributions, the difference in outcome between varying nucleation rates is seen prior to the sequence length of 40 (Figure 3.5A). There is a major increase in the number of short oligomers of length 10 and under as k_{nuc} is increased. Eventually, at longer lengths, the length distributions do not seem to

depend much on the rate of nucleation. This trend is also somewhat apparent when looking at the mean increase in lengths per cycle over varying starting lengths. In figure 3.5B, we see that the average length increase Δn is higher for lower k_{nuc} values, but the four lines begin to converge at longer sequence lengths above 60. At these greater lengths, the Δn values are roughly in the 0.5-1.5 range. At the highest k_{nuc} value of 10 h^{-1}, the mean increase in length per cycle is less than one across almost all lengths. Note that the fluctuations in Δn at higher lengths are due to the smaller sample population of these strands resulting in less reliable averages. As we increase the k_{nuc} value we also observe a steady increase in the average number of helices formed, with an increasing disparity between the results as we move to longer length strands (Figure 3.5C). The impact of nucleation rate is not as significant for short to medium strands, but at sequence lengths of 80 and above, we see an $h(n)$ value of about 3-4 at the lowest k_{nuc} value of 0 h^{-1} and an $h(n)$ of about 6-7 for the highest k_{nuc} value of 10. Overall, from this case of varying primer nucleation rates, and the previous case of varying monomer addition rates, we see a relationship between sequence length, average length increase, and average number of helices formed per cycle. Smaller average increases in length seem to correlate with production of shorter sequence lengths, but an increase in the number of helices produced per cycle. Interestingly, this relationship does not appear as evidently when looking at varying strand annealing rates.

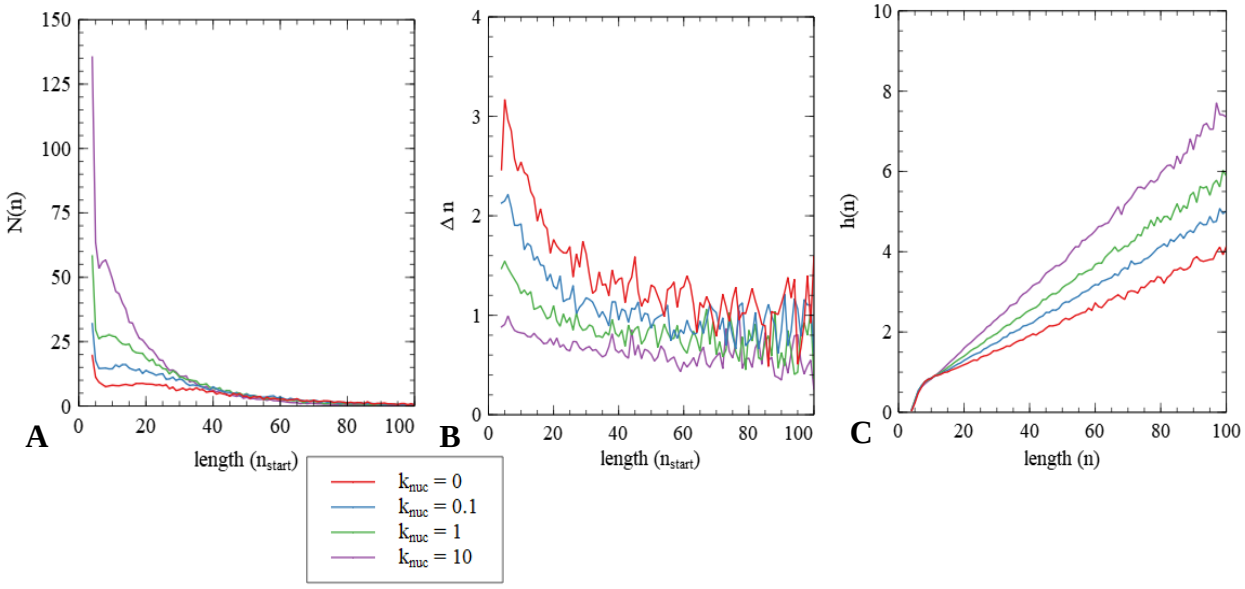

Figure 3.5. Distribution of (A) sequence lengths, (B) mean increase in length per cycle and (C) mean number of helices per cycle as a function of RNA strand length for four simulations varying by the nucleation rate value k_{nuc}.

Initially, we consider the changes in mean and maximum sequence length, and number of sequences present over time while varying the strand annealing rate k_{ann}. Over time, these variables follow the same general trend as seen with previous results when varying the k_{add} and k_{nuc} parameters. A clear observation with the annealing rate is that strands do not grow longer than the initial sequence lengths set at the beginning of the simulation. This is evident from the unchanging mean and maximum sequence lengths for a k_{ann} equal to 0 h^{-1} (Figure 3.6A-B). The mean length of strands stays fairly constant at the n_{init} value of 4.94 and the maximum length is maintained at exactly 10, which was the limit set for the length of starting oligomers produced by random polymerization. At increased k_{ann} values, we see an increase in the mean sequence length, as well as an increase in the rate at which the sequences reach their equilibrium mean length. This is evident by the increase of the slope for lines representing the higher k_{ann} values (Figure 3.6A). In contrast to the steady change in mean

lengths when varying the nucleation rate, the increase in mean lengths seems to diminish as the annealing rate reaches higher values. The highest k_{ann} value of 100 h^{-1} results in a mean sequence length of about 35. Although a k_{ann} of 1 h^{-1} is lower than the value we report in our standard parameters, we still see a significant growth in sequence length with the mean length reaching up to 18 and the maximum length reaching almost 200 nucleotides long at its peak. The maximum sequence lengths for the higher k_{ann} values are mostly similar, except for a large peak for the highest k_{ann} value which temporarily reaches the maximum allowed length of 300 (Figure 3.6B). It can be said that, although the effect of sequence annealing rate on the maximum length of strands is not great, it is still significant. With regards to the number of sequences, N_{seq}, we see an increase in all four simulations, and higher k_{ann} values correlate with a greater increase and steady state number of sequences over time. Note that N_{seq} does not have the dip that occurs in all our previous simulations when k_{ann} is set to 0 h^{-1}. Also, the increase in k_{ann} increases the rate at which the dip in N_{seq} occurs. There seems to be a strong correlation between the changes in mean sequence length and number of sequences over time between the four simulations. This is the case for the simulations varying in k_{add}, k_{nuc}, and k_{ann}.

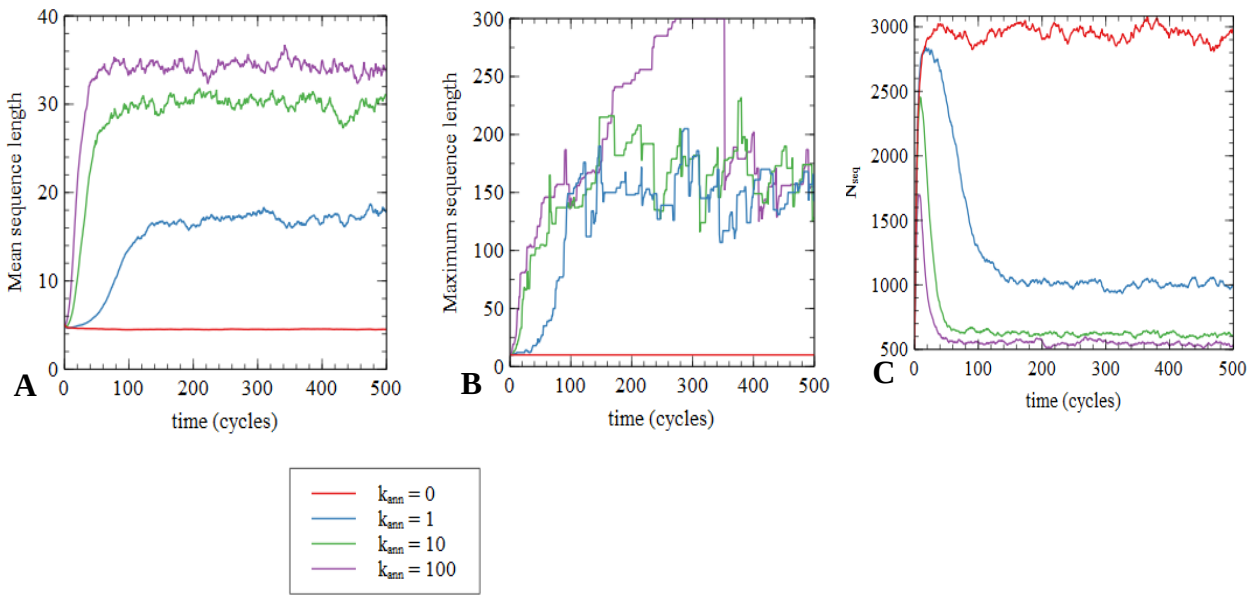

Figure 3.6. Comparing the change in (A) mean and (B) maximum sequence lengths, and (C) total number of sequences over 500 cycles of four separate simulations varying by the annealing rate value k_{ann}.

The sequence length distributions for varying k_{ann} values show a somewhat different outcome compared to previous length distribution results. Firstly, at a k_{ann} of 0 h^{-1}, the length distribution is almost the same as the distribution initialized for the randomly polymerized oligomers at the start of the simulation. In this case, there are no sequences that exceed the starting maximum length of 10 (Figure 3.7A). The number of sequences exceed 1000 for the shortest sequence lengths, but we restrict the scale to better analyze the results for the higher annealing rate simulations. The two middle values of k_{ann} of 1 h^{-1} and 10 h^{-1} show a negative exponential distribution as we have previously seen. However, for a k_{ann} value of 100 h^{-1}, we see an interesting difference for the length distribution of shorter sequences below the length of 15 nucleotides. The peak of the distribution has shifted from the normal starting length to roughly 15 nucleotides in length, and there is a steep negative exponential distribution of lengths as strands get shorter (Figure 3.7A). In this regime, the population of the shortest

strands (length 4-5) are comparable to that of long strands of about length 70. From the length of 20 and onwards, *N(n)* follows the same trend as the other non-zero annealing rate results. When looking at the mean length increase per cycle, there does not seem to be a dependence on the sequence annealing rate for longer sequences. Yet, for sequences of roughly length 10 and smaller, a higher k_{ann} greatly increases the average Δn (Figure 3.7B). The Δn is always 0 when k_{ann} is 0 h^{-1} and sits fairly consistently at around 1 for the other k_{ann} values after length 30. The trend is slightly different for a k_{ann} value of 1 h^{-1}, since the Δn for the smallest sequences under length 10 is lower than for the longer lengths. With a sequence annealing rate of 0 h^{-1}, we observe a mean number of helices formed per cycle of roughly 0.1 for lengths up to 10. There is an almost linear increase of mean helices formed as length increases and this trend is similar for the three non-zero annealing rate simulations (Figure 3.7C). The lines for these three simulations seem to slightly converge, in contrast to the previous simulations varying addition and nucleation rates, where the lines diverged. Overall, a higher k_{ann} value increases the average number of helices made for lengths up to about 70. However, past a k_{ann} of 10 h^{-1}, *h(n)* does not seem to significantly dependent on this parameter.

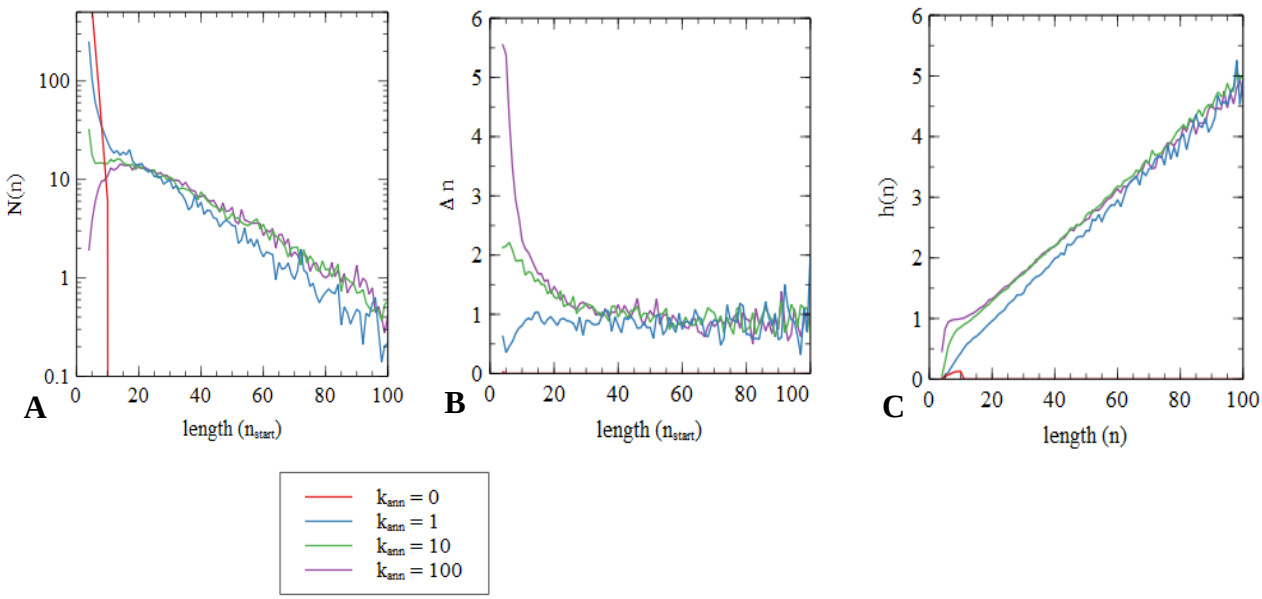

Figure 3.7. (A) Distribution of sequence lengths, (B) mean increase in length per cycle and (C) mean number of helices per cycle as a function of RNA strand length for four simulations varying by the annealing rate value k_{ann}.

We explored the parameters of monomer addition, primer nucleation, and sequence annealing in the simplest version of our simulation to understand whether template-directed RNA synthesis was plausible using our theoretical model. More specifically, we wanted to address a range of chemical reaction rates to account for the ambiguous conditions present on the early Earth environment. Within the mechanism of non-enzymatic RNA replication, we must certainly also account for the occurrence of mutations. We were curious whether the results from our simulations using standard parameters would change significantly with the inclusion of error rates. We included the error rate parameter *e* as the rate of monomer addition for a non-complimentary nucleotide determined as a fraction of the non-error addition rate, k_{add}. We also included a simple condition to account for the stalling of RNA growth as a result of a mutation. In this version, we just prevent any further growth of a strand once it incurs a mutation. In a later set of simulations, we also look at a case where

growth is still allowed after a mutation at a much lower rate. The same six variables as with previous results were measured for four separate simulations only varying in the error rate *e*. All other parameters were set to the standard parameters summarized in table 3.1. Overall, these simulation results followed trends like the case of varying k_{add}.

The mean and maximum sequence lengths over time decreased with increasing error rate, while the number of sequences increased (Figure 3.8). However, this is not the case when *e* is equal to 0.01, which translates to a 1% error rate for each incorrect nucleotide base pair. In terms of growth, the mean length is decreased significantly at *e* values of 0.05 and 0.1, but the changes in maximum length show a lot of overlap, thus weakening their correlation to error rate. The mean length increases at the same rate in all four cases but reaches a steady state length of 20 when there is a 10% error rate, compared to length 30 for the non-error and 1% error rate simulations (Figure 3.8A). Although the maximum length peaks are lower for *e* values of 0.05, and 0.1, an *e* of 0.01 shows a higher peak than the non-error simulation with maximum lengths of about 260 and 240, respectively (Figure 3.8B). The number of sequences in the chamber over time show the same spike up to 2500 for all four simulations. After about 60 cycles, N_{seq} reaches its equilibrium value of roughly 600 for an *e value* of 0 and 0.01, 700 for an *e* of 0.05, and 850 for an *e* of 0.1 (Figure 3.8C).

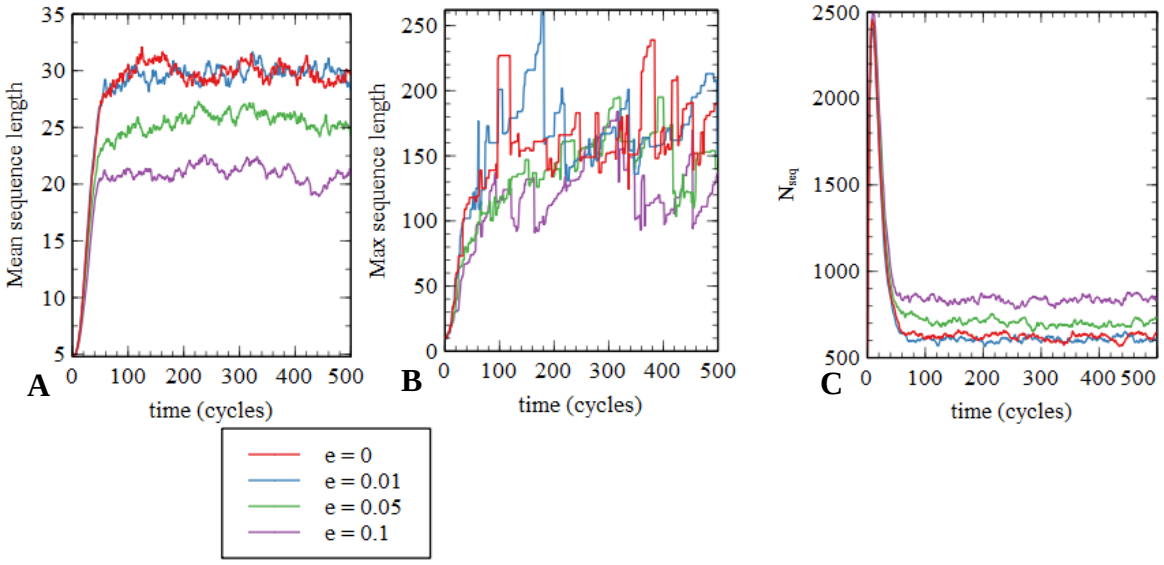

Figure 3.8. Comparing the change in (A) mean and (B) maximum sequence lengths, and (C) total number of sequences over 500 cycles of four separate simulations varying by the error rate value e.

With regards to distribution of lengths, we see comparable results from all four simulations past a sequence length of about 30. For shorter sequences below length 30, there are significantly more sequences as *e* increases to 0.05 and 0.1 (Figure 3.9A). Generally, the results for the non-error and 1% error simulations are almost non-distinguishable for the variables measured in figure 3.9. Up to a sequence length of about 70, the simulation results all show the same downward trend for mean increase in sequence length per cycle. For an *e* value greater than 0.01, the mean length increase Δn is lowered as *e* is increased. This Δn is quite small regardless, ranging from 0.5 to 2 in all cases combined (Figure 3.9B). After a length of 70, there are large fluctuations in Δn due to lower sample size of sequences, making it difficult to judge its correlation with regards to error rate. A higher error rate showed an increase in the mean number of helices formed per cycle, mainly because of a greater rate of increase of *h(n)* over increasing lengths (Figure 3.9C). For the highest *e* value of 0.1, *h(n)* was

about 6 for long sequences over length 80, compared to about length 4 in the non-error and 1% error cases. All simulations followed the same linear relationship of increasing average helix number for longer length sequences. Overall, it seems that more helix formation correlates with a lower average length increase per cycle, which was also observed in the previous simulation results.

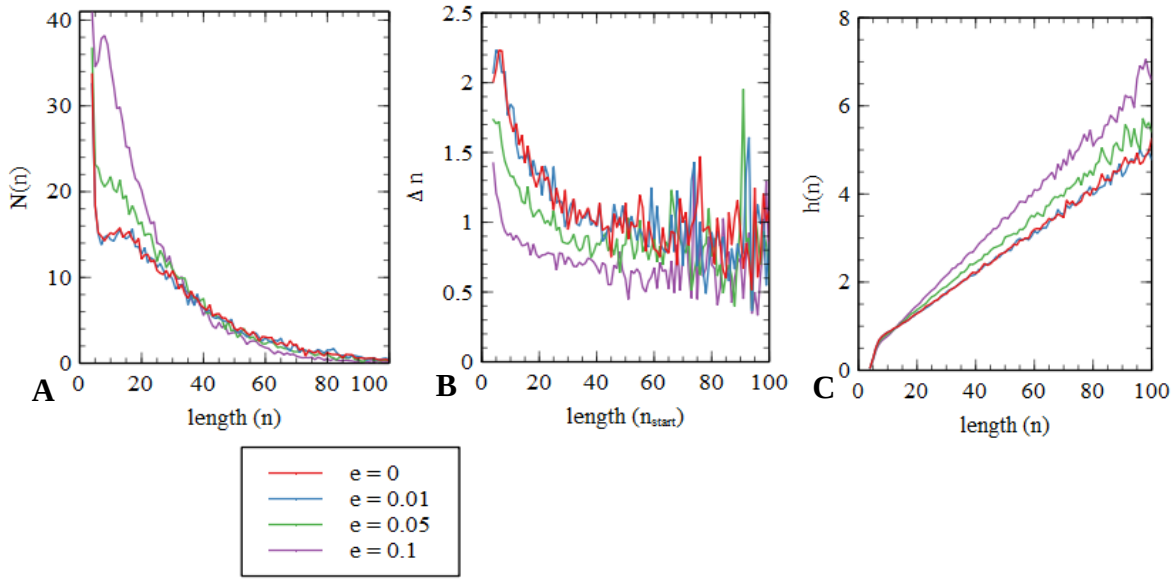

Figure 3.9. (A) Distribution of sequence lengths, (B) mean increase in length per cycle and (C) mean number of helices formed per cycle as a function of RNA strand length for four simulations varying by the error rate value e.

The inclusion of mutation rates allowed us to observe the outcome of our RNA synthesis model with greater complexity and accuracy. We wanted to continue implementing components which we thought could be important factors for non-enzymatic template-directed RNA synthesis. So far, we included theoretical monomer addition rates which were only dependent on whether the incoming nucleotide was a correct match or not. We wanted to understand the impact of having varied rates for each nucleotide, rather than the same rates like in our simpler models. In the next set of simulations, we replace the theoretical addition

rates with experimentally obtained rates for each possible nucleotide base addition. The rates are averaged from two separate studies[44,61], previously referred to in chapter 2, and the averages used are listed in Table 3.2. Due to the four times increase in concentration of U nucleotides in the experiments, the rates used in our simulation are divided by four when U is being added. We initially consider a case where the occurrence of a mismatch stops any further growth of the respective strand. We then consider the case where growth can continue after an error at a constant combined addition rate, $k_{mismatch}$ = 1 h^{-1}, for all four possible nucleotides, meaning an effective addition rate of each nucleotide base of 0.25 h^{-1}. However, two errors in a row will still terminate growth of the respective strand. In this regime, we also allow mismatches to occur during annealing of two strands at a maximum of one consecutive error. The results of this post-error growth implementation are only shown for the simulations that use experimental rates.

Table 3.2. Average experimental primer extension rates from two separate studies varying upon nucleotide base pairing.

		Base added			
		A	C	G	U
Template base	A	0.043 h^{-1}	0.079 h^{-1}	0.145 h^{-1}	0.20375 h^{-1}
	C	0.01225 h^{-1}	0.01275 h^{-1}	1.94 h^{-1}	0.0036 h^{-1}
	G	0.0375 h^{-1}	14.515 h^{-1}	0.205 h^{-1}	0.17375 h^{-1}
	U	0.735 h^{-1}	0.014 h^{-1}	0.365 h^{-1}	0.024125 h^{-1}

It is important to note that these rates are much higher for CG base pairs compared to AU base pairs. To compare our results from these experimental rates to those using the standard theoretical rates, we scaled up the experimental rates to match the average correct standard nucleotide addition rate of 10 h^{-1}. The average rate of Watson-Crick base additions

in the experiments is 4.35 h^{-1}, therefore we multiply all the rates by a factor of 10 / 4.35 = 2.3. This way, we can isolate and examine the effects of non-uniform addition rates. We present simulation results which compare scaled experimental growth rates with and without growth after error, and standard uniform theoretical growth rates (k_{add} = 10 h^{-1}) at 1% and 10% error rates.

The experimental rates have a substantial overall impact on the RNA growth results. The sequences in the experimental rates case still managed to increase from a starting point of around 5 nucleotides to about 8 nucleotides in mean length. However, this is a lot less than what we saw in previous simulations where the mean length reaches about 30 for standard growth rates and an error rate *e* of 0.01. When growth is permitted after a mismatch in the experimental growth rates case, the mean length is increased up to about 13 nucleotides in length which is a significant improvement (Figure 3.10A). The maximum sequence lengths are also decreased greatly in the scaled experimental rates simulations compared to the standard rates. Using the experimental growth rates, the maximum length peaks just below 50 nucleotides, whereas the simulations using the standard theoretical rates with an *e* of 0.01 show strands with lengths of over 200 nucleotides (Figure 3.10B). In the simulation with growth permitted after error, the maximum sequence lengths overlap quite a lot with the results of the simulation with standard rates and 10% error. Growth after error allows for sequence lengths to increase up to 150 nucleotides. It even overlaps somewhat with the standard rates case with 1% error. Compared to the mean length variable, the inclusion of growth after error is much better for saving the maximum length growth after switching to experimental monomer addition rates. The number of sequences N_{seq} follows the opposite trend of mean lengths for each simulation. After the initial peak in N_{seq}, the number falls to

just under 2000 for the case with experimental rates, about 1200 with growth after error, and under 1000 for the simulations using standard growth rates (Figure 3.10C).

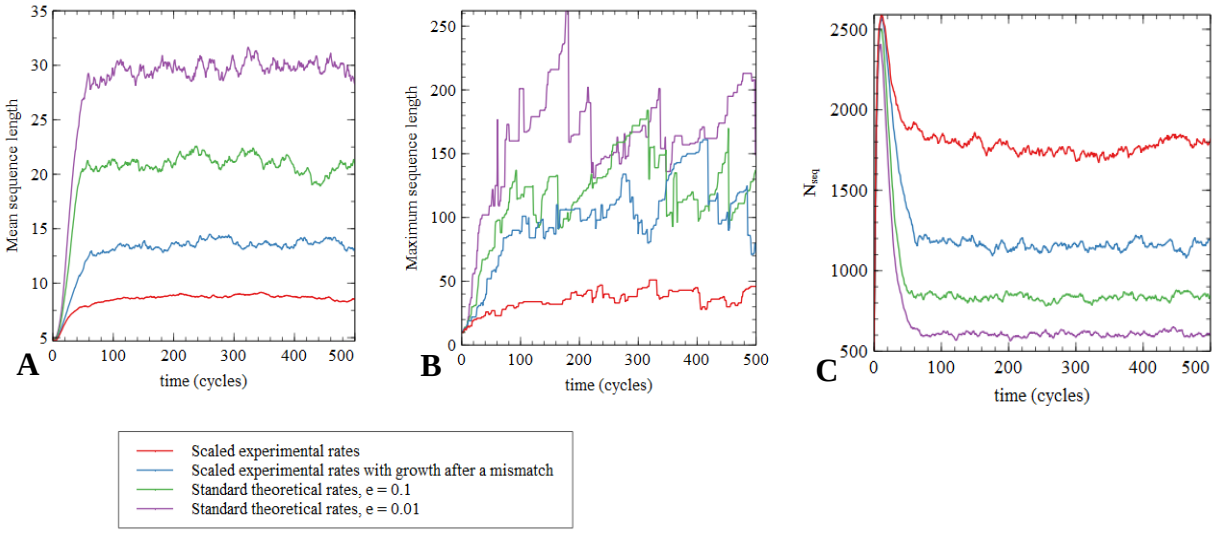

Figure 3.10. Change in the (A) mean and (B) maximum sequence lengths, and (C) total number of sequences over 500 cycles of four separate simulations. Simulations vary by monomer addition rates, error rates, and whether growth after an error was permitted.

The length distribution measured from the experimental rates simulation shows most of the sequences being shorter than 20 nucleotides, with longer length sequences having a significantly decreased *N(n)* value compared to the other simulations. The growth after error case shows a distribution close to the simulations using standard rates for lengths 20 and higher, but a higher *N(n)* for sequences shorter than 20 (Figure 3.11A). The mean increase in length per cycle becomes considerably low when using the experimental rates at around 0.1-0.2 for medium length strands. This is increased to about 0.5 in the growth after error case but is still significantly lower than the Δn values for the simulations using the standard theoretical rates (Figure 3.11B). The mean number of helices formed per cycle does not seem to differ between the experimental rates when allowing or disallowing growth after error. This can

only be said for sequences of about length 40 and lower, since there are little to no sequences of greater length made in the case without growth after error. With standard theoretical rates, we see a decrease in the number of helices formed per cycle (Figure 3.11C).

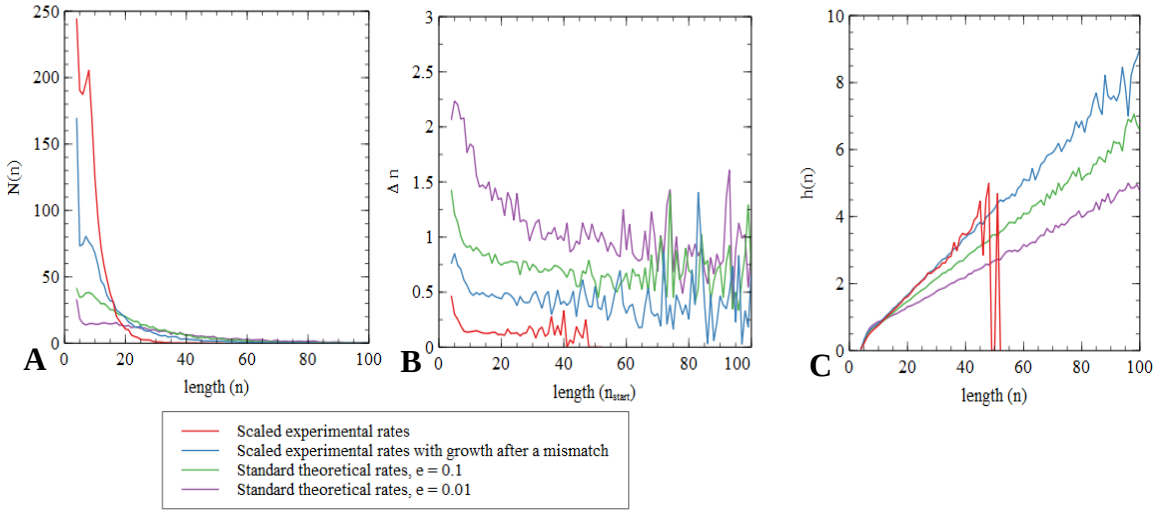

Figure 3.11. (A) Distribution of sequence lengths, (B) mean increase in length per cycle and (C) mean number of helices per cycle as a function of RNA strand length for four simulations. Simulations vary by monomer addition rates, error rates, and whether growth after an error was permitted.

One main difference between our standard rates and the experimental rates is that the experimental rates are quite skewed compared to the uniform theoretical rates. We were curious to see how this impacted the presence of different bases in the sequences being made. We decided to measure the number of nucleotides present in the system at the end of each cycle. It would follow that, the more nucleotides that remained in the system, the less that nucleotide base was being incorporated into the growing sequences. We ran three simulations: one using the standard theoretical monomer addition rates, one with scaled experimental addition rates without growth after error, and one with scaled experimental

addition rates including growth after error. We expected to see a significant difference between these simulation with regards to the number of different nucleotides.

All three simulations follow a similar trend with a sharp decrease of the nucleotide number within the first 20 cycles. In the simulation using standard addition rates and a 1% error rate, all four nucleotide bases have near identical numbers per cycle (Figure 3.12A). After the initial dip, they rise back slightly to their steady state populations of about 1200, which are balanced by their use in growing strands and the constant inflow of new monomers. When experimental addition rates are used, the steady state nucleotide numbers are all different except for G and A. U base monomers are much higher at about 2600, while C base monomers are the lowest at about 1400 (Figure 3.12B). G and A base monomers are slightly higher at about 1600. The inclusion of growth after error only decreases the steady state number of U base monomers to about 2400. The C base monomer numbers do not change, but A base monomers are decreased in number to match the C base monomers at about 1300. The G base monomers are not affected (Figure 3.12C). It is important to note that the number of monomers in the system has a negative correlation with their use in growing RNA sequences.

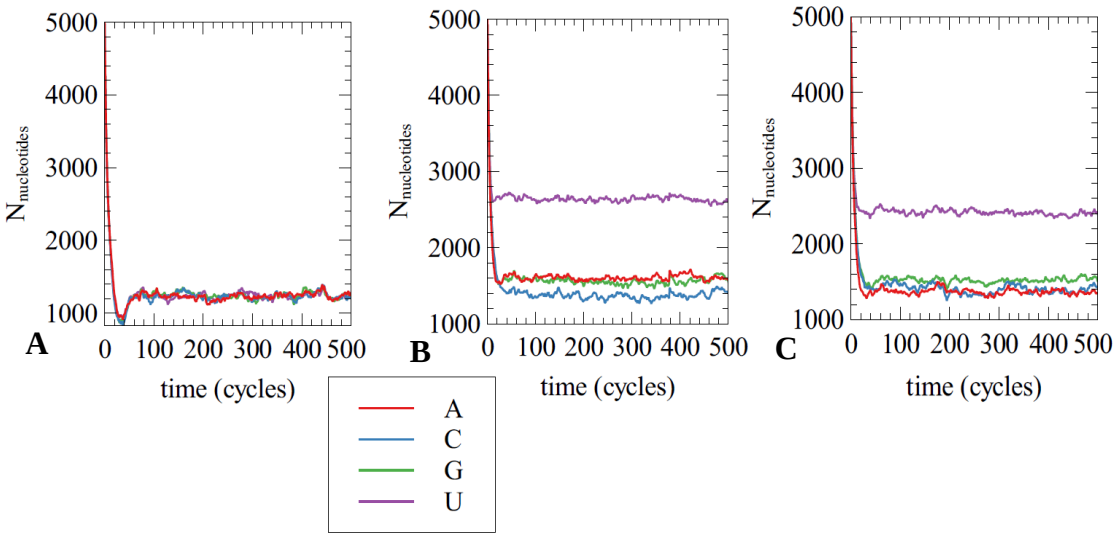

Figure 3.12. Comparison of the total number of nucleotides of each of the four Watson-Crick bases over 500 cycles in three separate simulations. Simulations either used (A) the standard theoretical rates and an e value of 0.1, or (B) used the scaled experimental rates without growth after error and (C) with growth after error.

3.3 Discussion

Our RNA synthesis model assumes simple non-enzymatic reactions between pre-existing RNA nucleotides and oligomers, and environmental temperature cycling. We created and executed simulations that followed these reaction events. One of our goals was to understand whether templated RNA replication could ever lead to sequences which exceeded the length of the longest template. Previously, it was revealed that reannealing of complementary sequences would not allow for this outcome[47]. The fast rate of reannealing during the cool phase of temperature cycling creates dead-end products and limits the number of long available templates due to the low rate of separation. We hypothesized that a diverse sequence mixture would allow a greater chance of producing productive binding conformations. These conformations would lead to the lengthening of existing strands or new nucleated primers. In addition, we wanted to test the idea of increasing the maximum length

of the sequences in the starting pool through rare, staggered binding conformations. We initially produced the simplest version of our simulation that allowed us to explore these ideas. This version included all the reactions that could allow RNA growth, but assumed no chance of mutations during sequence annealing, nucleation, and monomer addition. With this simple version, our goal was to observe an outcome under idealistic conditions and see whether varying reaction rates would greatly impact the results.

Since we do not have precise measurements of all the rates for the simulation reactions, the results we obtained using the standard set of parameters should be interpreted through a qualitative lens. We varied three parameters which would likely have the biggest impact on the measured variables. The three parameters were rates for monomer addition, primer nucleation, and strand annealing. Changing any of these rates impacted the growth of RNA strands. From our results, we show that RNA sequences grow under all the varying rate conditions, except when the annealing rate is zero. A low monomer addition rate also had a major impact on decreasing the amount of RNA growth. The growth is limited due to the finite time that each sequence remains in the reaction chamber, which is typically $1/\phi$, meaning 20 cycles in our case. On the other hand, a high monomer addition rate is also going to be limited by the finite number of monomers available. Therefore, having a low monomer addition rate is still going to produce long sequences if there is limited loss of existing strands and a reliable supply of nucleotides. The effect of changing nucleation on RNA strand growth is not as significant. Mainly, we show that a higher nucleation rate will limit RNA growth since it will lower the overall available template space for binding of existing strands. Successful nucleation of a short primer is generally thought to occur at a low rate compared to annealing of two existing strands. This is due to the low stability of monomers, dimers, and

trimers when bound to a template. However, primer nucleation is important for RNA replication for the maintenance and evolution of useful sequences. Also, a nucleation rate of zero does not show enough of a benefit to RNA growth for it to act as an advantage in an early non-enzymatic phase of the RNA. Even in the case of high nucleation rate, it seems that long sequence length is achieved, but significantly more short sequences are created. This is shown by the results measuring maximum length and distribution of lengths.

Sequence annealing rate seems to be almost as important as monomer addition rate in producing long sequences. In this model, sequences can only grow via template-directed monomer addition. Without annealing, you would only be able to grow nucleated primers, and sequences would never grow longer than the longest strand in the system. Meanwhile, if the rate of annealing was greater, we can expect more RNA growth because more sequences can be given a chance to bind with the desired conformation. In our simulation, we require perfect complementary binding of at least four nucleotides for annealing. One could envisage a case where annealing occurs at a higher rate depending on the helix stability and tolerance for imperfect binding, in turn allowing a higher rate of RNA growth.

From our results, we should note that a greater rate of RNA growth corresponds to a lower sequence number. Having high sequence number can be beneficial in some cases, specifically early in the RNA world where greater diversification can result in the emergence of functional strands more quickly.

There are several factors which can influence the number of sequences. In our model, new sequences are introduced through nucleation, or inflow of new random strands. They are lost only through outflow. Since inflow of new strands is constant, we can say that the number of sequences is only affected directly by the rate of nucleation. We consistently

observe a lower steady state sequence number for simulations with lower RNA growth. This can be explained by how RNA nucleotides are distributed in different simulations. Generally, as the concentration of free monomers decreases, so does the nucleation rate and in turn, the number of sequences. This explains the initial rapid increase in sequence number at the start of the simulation when the monomer concentration is high. In the case of high monomer addition rate, the steady state nucleotide concentration is going to be lower and result in a lower number of sequences. We can clearly see the affect of nucleation rate on number of sequences when the nucleation rate is set to zero. Here, there is a balanced inflow and outflow of sequences with not other inputs, resulting in a stable number of sequences over time equal to the starting population. We also see the impact of nucleation rate on sequence number through the increased peak number of sequences as nucleation rate is increased. A lack of RNA growth shown in the case where annealing rate is zero shows no decrease in the number of sequences after the initial increase, since nucleotides are not being used for monomer addition. An increase in the monomer supply essentially increases the rates for both monomer addition and primer nucleation. The degree in which it affects these reactions can be further explored through simulations varying in monomer concentration and inflow. A change in binding stability could impact the nucleation rate since it would affect our l_0 parameter for minimum possible primer and helix length. Yet, this could affect annealing rate as well which indirectly affects the monomer concentration. This is a limitation of our current simulations, but some interesting outcomes may come from the implementation of a more complex binding and separation model.

Increasing the number of sequences is not fully in line with sequence diversification. Full replication of an existing sequence following primer nucleation simply results in the

complementary sequence, making it difficult to see how a low diversity sequence population could ever diversify without introducing new random sequences. Indeed, quick diversification seems to depend mainly on the starting population diversity, the rate that new sequences are introduced, and the mutation rate. However, given that the initial sequence space is diverse enough, creating new RNA strands via nucleation can also lead to increased sequence diversity. Nucleated primers should only grow to replicate a part of the template sequence, after which they can separate and bind to a different template in the next cycle. This way, the initial nucleated primer would create a unique sequence from the combination of existing ones. Nucleation can be quite important for diversification, especially in cases where inflow of new sequences is limited. Drawing back to the results of our simulations, we should consider that greater RNA growth should not always be the desired outcome if it greatly restricts nucleation. Interestingly, the growth of long RNA strands must not necessarily depend on high monomer addition or annealing rates, as shown by measuring the average increase in sequence lengths per cycle.

In most of our simulation results, we find that a substantial number of long sequences are formed, even though the increase in length in any once cycle is quite small. This infers the growth of long sequence to come as a result of continued presence in the reaction chamber over many cycles. We can consider a general case for sequence growth assuming an exponential length distribution, $N(n) = A\mu^n$, for some μ. If the increase in length were Δn, independently of n, then the number of sequences of length n at the end of a growth phase would be equal to the number of sequences of length n-Δn at the beginning. As in the case of our simulations, if we assume a fraction ϕ is lost at the end of a cycle, we need $A\mu^{(n-\Delta n)}(1-\phi) = A\mu^n$ to maintain a constant sequence number. Thus, $\Delta n = \frac{\ln(1-\phi)}{\ln \mu}$. If we consider a

vesicle as the reaction chamber, where the contents double before the cell splits, then the dilution factor is $\phi = ½$ and $\mu = 2$, giving us a $\Delta n = 1$. However, the exponential distribution assumed here is not the case for our simulations. Regardless, this shows the possibility for generating long strands through temperature cycling and small increases in length per cycle, given that the sequence remains in the reaction chamber over many cycles. This point clarifies the significant growth of strands in cases of small monomer addition rate and high mutation rate.

In the study by Tupper and Higgs[47] outlining the reannealing problem, similar conditions in our simulations would result in less potential growth of sequences and lower sequence number. However, we can see that in a diverse enough mixture of sequences, there is a high chance for partially matching helices to form. These binding configurations create the potential for long sequences to grow, as well as allowing more replication through nucleation. This idea is supported by our results showing an average helix number per cycle that increases for longer RNA strands. Short sequences do not have the template space to allow for formation of multiple helices. In comparison, long sequences can allow for branching clusters of helices, which is what we see in all our simulation results. This contrasts with forming long duplex structures as seen in the aforementioned study. With regards to the different parameters, we see a major impact on helix formation coming from varying nucleation and monomer addition rates. We can say that helix formation is dependent on the balance between these two rates. A higher nucleation rate will form more helices, and a low polymerization rate will limit the template space taken up by sequences. It is important to note when annealing rate is zero, there is a close to zero value for average helices per cycle. This shows that annealing is important for short strands to form helices, specifically

with longer templates. It also explains why the number of sequences seems to increase with length initially (up to length 20) in the case of high annealing rate. Whereas higher polymerization rate increases the overall growth of strands, higher annealing rate would specifically increase the rate of growth for shorter sequences. In this case, the increased growth rate comes at the cost of nucleating new strands since template space is being taken by existing strands. Nevertheless, long sequences will still grow at a rate independent from the annealing rate. This shows that achieving the binding configuration required for long sequences to grow is not the limiting step, but most likely the polymerization rate is.

In the most ideal and simple version of the simulation, we test a variety of rates to gauge the potential for this RNA synthesis model to produce long strands. There are many factors which can influence the reaction rates, one being the presence of error during replication. Error rate can be quite detrimental for passing on useful information through RNA replication. In our case, we are more concerned with producing a diverse set of long strands from which useful and functional RNA could first emerge from. Still, it has been observed that mismatches result in a stalling effect, decreasing the rate of subsequent monomer addition compared to that following a Watson-Crick base pair[45]. Although this has been said to help improve overall replication fidelity, the post mismatch stalling naturally limits the growth of sequences in the present scenario. Here we assume a mismatch will completely stop any further monomer addition. Even at quite high error rates, the decrease in growth is not as severe as having a low monomer addition rate. When there is a 1% error rate for each mismatch, meaning an overall error rate of 3%, we see no change from the case without error. This draws back to the low requirement for Δn per cycle. Even if a sequence grows by one nucleotide and stops due to a mismatch, the sequence can still grow to long

lengths if it remains within the system over many cycles. Indeed, we see long sequences over 100 nucleotides even when the overall error rate is 30%. On average, experiments have shown a replication error rate of about 17%[44], which we show to have only a small impact on our simulation results. This gives us some hope that high error rates would not be detrimental to synthesis of long strands in the early non-enzymatic RNA world. Implementing this simple error case helped in exploring whether the present model would be limited under less ideal conditions, which would more closely match the conditions present on the early Earth or in RNA replication experiments.

In this pursuit, we discovered a major limitation when implementing experimentally obtained rates from two non-enzymatic primer extension studies[61,63]. The rates were scaled by a factor of about 2.3 because they were much lower on average than our standard monomer addition rate. However, we cannot ascertain an accurate set of rates due to the limited amount of data, as shown in our chapter 2 results. The experimental rates are dependent on the experimental conditions. The real rates could certainly vary from these in part due to their dependence on nucleotide concentrations. We have considered various parameter values in the first part of the results to account for a range of outcomes. With the experimental errors, the goal was to compare a set of data with considerable variability between different addition rates. We choose to scale the rates so that the average rate of a correct pair addition is k_{add}, allowing a comparison to be made with our constant rates. Relative to the case with constant rates, a large variance in rate shows a slowing down of RNA growth.

The RNA growth is even lower when using the experimental rates compared to using the lowest monomer addition rates of 1 h^{-1}. Nevertheless, it is likely that the monomer addition rates were also dependent on the base of the incoming nucleotides during templated

non-enzymatic replication before the origin of life. If we assume an even number of each of the four bases present in the sequence space of the starting set of randomly synthesized RNA, then the template-directed strand growth will depend on an even chance between the addition of each nucleotide base. From the experimental rates used in our simulations, we can see that even after scaling, every monomer addition rate for the correct Watson-Crick base pairs is lower than our standard k_{add} of 10 h^{-1} except for the addition of C binding to G. This means that once there are little to no strands which can grow by a C base, the overall monomer addition rate is going to suffer. Additionally, the error rate for AU binding is much higher than CG binding, even though the average factor of error overall is about 0.0241. In the case with no growth after an error, if there are more potential monomer additions for A and U bases, then there is a much slower rate of addition alongside a higher chance of mismatch stalling compared to the case with uniform rates. This detrimental affect on RNA growth is saved considerably by allowing a small rate of growth after an error. This is likely due to decreasing the impact of high experimental error rates on limiting strand growth. The unequal incorporation of different nucleotides, specifically the U nucleotide in our case, shows a preference towards strands that contain less of the slow adding nucleotide base. It may be the case that in a system where there is only loss of strands, but no inflow of random oligomers, the sequence space would shift towards having more of the fast-growing nucleotide bases. As a result, overall growth rate would increase. This condition can be tested in future theoretical studies. Regardless, it is apparent that the many non-idealistic scenarios leading up to the origin of life make it difficult to ascertain the possibility of expanding a non-enzymatic RNA world. New and more accurate quantitative experimental measurements of non-enzymatic RNA synthesis reactions can help in exploring realistic outcomes for theoretical simulations

like these. These results could in turn allow for an understanding of necessary components of the stages leading up to the origin of life, such as environmental cycles, compartmentalization, or the presence of other important molecules.

Chapter 4: Exploring the Possibility of Virtual Circular Genomes

4.1 Methods

As discussed within chapter 1, the presence of virtual circular genomes has been proposed to be of relevance to non-enzymatic RNA replication. Thus, we look at the sequences generated by our simulations in chapter 3 to see if anything resembling a virtual circle arises. One way to look for the presence of virtual circular genomes is to consider sequences of length *k* (*k*-mers) within the RNA formed in the simulation. The number of possible *k*-mers is 4^k, and each *k*-mer can be labelled by an integer *i* from 0 to 4^k - 1. Let m_i be the number of occurrences of *k*-mer *i* in a mixture of sequences, counting all the overlapping *k*-mers within sequences of length *k* or longer. The frequency of *k*-mer *i* in the mixture is $f_i = m_i / \sum_j m_j$. We define the diversity of *k*-mers in a mixture of sequences, D_k, as the number of *k*-mers that are present at least once in the mixture.

We make use of 5-mers specifically. A particular 5-mer *i* can be written $i = n_1 n_2 n_3 n_4 n_5$, where each of the *n*'s is a nucleotide base A, C, G or U. We will say that a word *j* follows from word *i* if the first 4 letters of *j* are the last four letters of *i*. Hence $j = n_2 n_3 n_4 n_5 n_6$. There are four possible following words from any given word, as there are four possibilities for n_6. We define a transition matrix T_{ij} such that $T_{ij} = 1$ if word *j* follows word *i* and both *i* and *j* are present in the mixture, and $T_{ij} = 0$ otherwise. We say there is a path of length *n* steps from *i* to *j* if there is a series of *n* words that follow from each other that are all present in the mixture and gradually transform *i* into *j*. We define a path matrix, such that $P_{ij}^n = 1$ if such a path exists, and $P_{ij}^n = 0$ if no such path exists. Clearly, $P_{ij}^1 = T_{ij}$. The path

matrices for larger numbers of steps can be calculated by straightforward iteration: $P_{ij}^{n+1} = 1$ if $\sum_k P_{ik}^n T_{kj} > 0$, otherwise $P_{ij}^{n+1} = 0$.

For any given *k*-mer word *i*, the fraction of words in the mixture that are accessible by a path of length *n* is $\sum_j f_j P_{ij}^n$. We define the connectivity function $X(n)$ as the probability that two randomly chosen words from the mixture are connected by a path of length *n*:

$$X(n) = \sum_i \sum_j f_i f_j P_{ij}^n.$$

If $P_{ii}^n = 1$, there is a circular path that returns to word *i* after *n* steps. We define the circularity function $C(n)$ as the fraction of words in the mixture which are part of a circular path of *n* steps:

$$C(n) = \sum_i f_i P_{ii}^n.$$

Furthermore, we define $P_{ii}^* = 1$ if word *i* is part of a circular path of any length (*i.e.,* there is at least one *n* for which $P_{ii}^n = 1$). The fraction of words that are part of a circular path of any length is

$$p_{circ} = \sum_i f_i P_{ii}^*.$$

The relative frequency of the most common word that is accessible from word *i* by a path of length *n* in comparison to the total frequency of all accessible words is $\frac{\max_j f_j {P_{ij}}^n}{\sum_j f_j P_{ij}^n}$. If there is only one word that is accessible after *n* steps, then this relative frequency is 1. Thus, we say the initial word is completely specific of the word that follows. If there are many

accessible words, the initial word has low specificity. We define the specificity function of the mixture, $S(n)$, as the mean value of this relative frequency for all words in the mixture:

$$S(n) = \sum_i f_i \frac{\max_j f_j P_{ij}^n}{\sum_j f_j P_{ij}^n}.$$

Suppose there is a virtual circular genome of length 10. There are 10 5-mers that can be formed from each strand of this genome. Thus, the diversity of 5-mers is $D_5 = 20$, assuming that there are no repeated words in the sequence and that there are no words that are present in both strands of the genome. Figure 4.1A shows an example of one strand of this virtual circular genome, and a circular path of 10 words that are taken from this genome. Consider a perfect virtual circle mixture that contains all the words of this genome at equal frequency and no words that are not part of the genome. In this case, all words in the mixture are part of circular paths, so $p_{circ} = 1$, and all the circular paths are length 10, so $C(10) = 1$. There are also circular paths for any *n* which is a multiple of 10, but not for other lengths. Thus $C(n) = 0$ when *n* is not a multiple of 10. For any word in this mixture there is exactly one word that can be reached by a path of length *n* steps. Therefore, the relative frequency of this word is 1. The perfect virtual circle mixture has specificity $S(n) = 1$ for all *n*. Since there is only one word accessible by a path of length *n* from any starting word, the probability that a randomly chosen word is accessible is 1/20. Thus $X(n) = 1/20$, for all *n*. It is important to note, however, that our simulation results in chapter 3 would not result in a virtual genome, but rather a fully connected graph. With a high enough diversity where every possible 5-mer is present, you would get $C(n) = 1$ for every length *n* of 5 and greater. The same trend would

be present for $X(n)$ values. Meanwhile, the specificity of the mixture $S(n)$ would be 0 for all lengths of 5 or greater.

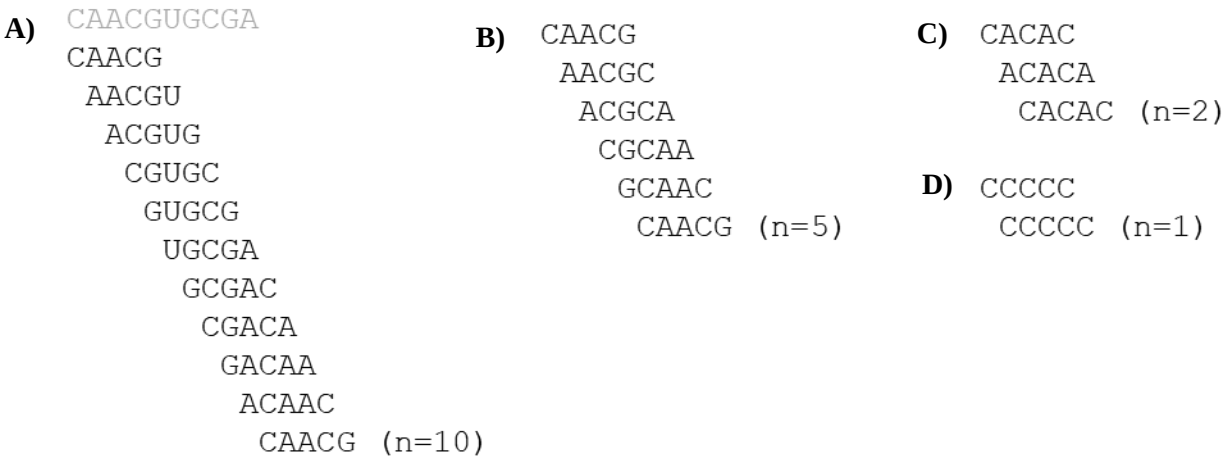

Figure 4.1. (A) A circular path of 10 steps formed from words taken from a virtual circular genome of length 10. (B) A circular path of length 5 existing in a mixture in which all 5-mers are present. (C) and (D) Short circular paths arising when there is a repeating structure in the 5-mer sequence.

We used varying starting conditions and parameters of our simulation and measured the connectivity function $X(n)$, the specificity function $S(n)$, and the circularity function $C(n)$ from the mixture of sequences produced after 500 cycles. We also constructed visual connectivity graphs for all the possible 5-mer sequences in the mixture. Nodes were set as nucleotide bases or sequences and lines were set as path connections with arrows indicating the path direction. Nodes and lines were marked red if the nucleotide was part of a circular path, and grey if it was part of a linear path.

4.2 Results

Using the simulations from chapter 3, we hoped to understand whether virtual circular genomes could emerge through our proposed RNA synthesis model. We run simulations

using the standard set of parameters, but with varying starting conditions. We began with a case which would yield measurements corresponding to the presence of virtual circles. In this case we start with a mixture of 100 sequences in which a virtual circle of length 10 along with its complement are present. The circle also contains no repeated overlapping 5-mer sequences in either of its complementary regions. We compared two cases where the inflow of new random strands was either turned on or off. It seemed likely that introducing random strands would affect the maintenance of the virtual circle by increasing the sequence diversity. Without the inflow of new random sequences, we observed perfect maintenance of the virtual circle, showing a similar outcome as with the example given in the methods. The results show a specificity function (S) value of 1 and a connectivity function (X) value of 0.05 for every path length. The circularity function (C) shows a value of 1 for path lengths which are multiples of 10, and a value of 0 for all other lengths. The connection graphs show that there are only two virtual circles which exist and are complementary (Figure 4.2A). In contrast, the simulation which permitted the inflow of random strands resulted in a connection graph showing a scramble of paths. No clear virtual circle is present, although there are intertwined circular paths. We observe a decreasing S value as lengths increase which levels off close to 0 past sequences of length 10. Meanwhile, the X and C values increase to 1 for all lengths of about 16 and longer (Figure 4.2B).

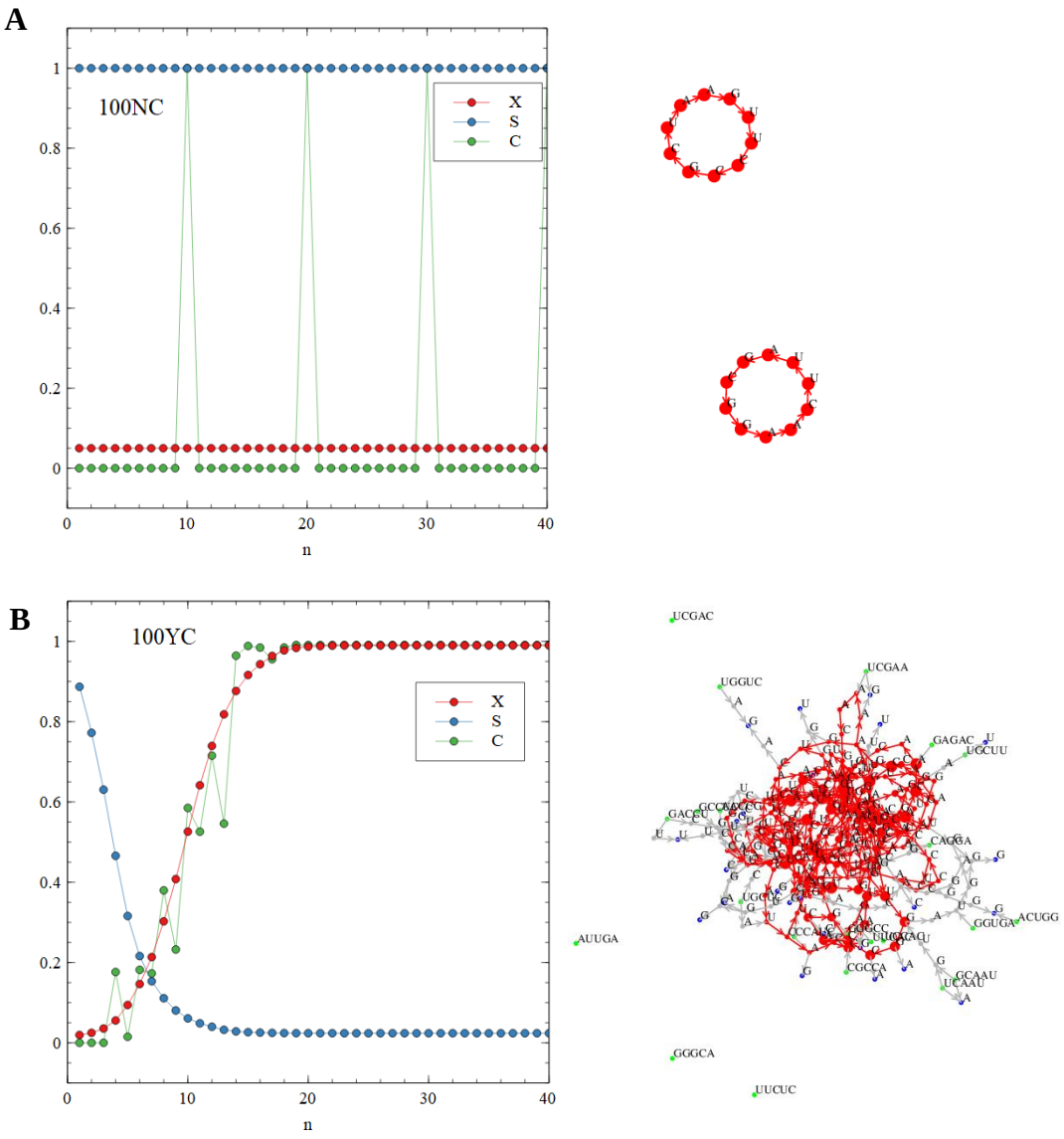

Figure 4.2. XSC functions and connection graphs of two simulations starting from a perfect virtual circle mixture of 100 strands. The simulations differ by either (A) disallowing or (B) allowing the inflow of new random sequences.

In the next set of simulations, we set the starting mixture of 100 sequences to be generated randomly. We wanted to observe whether virtual circles could emerge after 500 cycles of RNA replication and growth. In this case, we disallowed the inflow of random sequences. Many of the simulations resulted in connection graphs like the one in Figure 4.3A.

There are several linear paths that connect 5-mers, but a lack of a diverse or long virtual circular genome. We see circular paths consisting of only a G or a C and circles of alternating bases. The longest unique sequence that is part of a circular path is of length four. We see these trends reflected in the C values fluctuating at lengths which are multiples of four. These peaks also include the circles at lengths of one and two. Interestingly, the X value here is constant and low, like the case in Figure 4.2A. The S value dips down to meet the peaks of the C values, which is again a trend we observed in the case starting with perfect virtual circles. We also show an example of a simulation which produced a connection graph showing several longer virtual circles of length 14, 18, 22, 26, and then for every multiple of 2 afterwards, as shown by the C values (Figure 4.3B). The circle is shown as having a base 14 nucleotide long sequence with added repeats of sequences in between, shown as the outwards rectangular paths. The S value decreases, and the X value increases with path length, starting to plateau once the lengths reach that of the virtual circle.

Lastly, we looked for virtual circles starting from a randomly produced mixture of 500 sequences. Starting with this more diverse mixture, connection graphs showed many intertwined circular paths not dependent on the inflow of new strands. In the case without inflow of strands, there is visually a lower diversity of sequences compared to the case with inflow. Both cases show similar trends with regards to the X, S, and C values. There is a decrease of the S value, which initially begins at a higher point and decreases slower in the case without inflow of new sequences. Both cases result in an S value leveling off at 0. The X and C values increase together from 0 until 1 in both cases, with the rate of increase being slower in the case without sequence inflow (Figure 4.4). Overall, there are no specific virtual circles produced in these cases starting with 500 sequences, or in any cases which inflow of

random sequences was permitted. Starting with a lower mixture population and not introducing any new strands seemed to sometimes lead to the emergence of unique virtual circles.

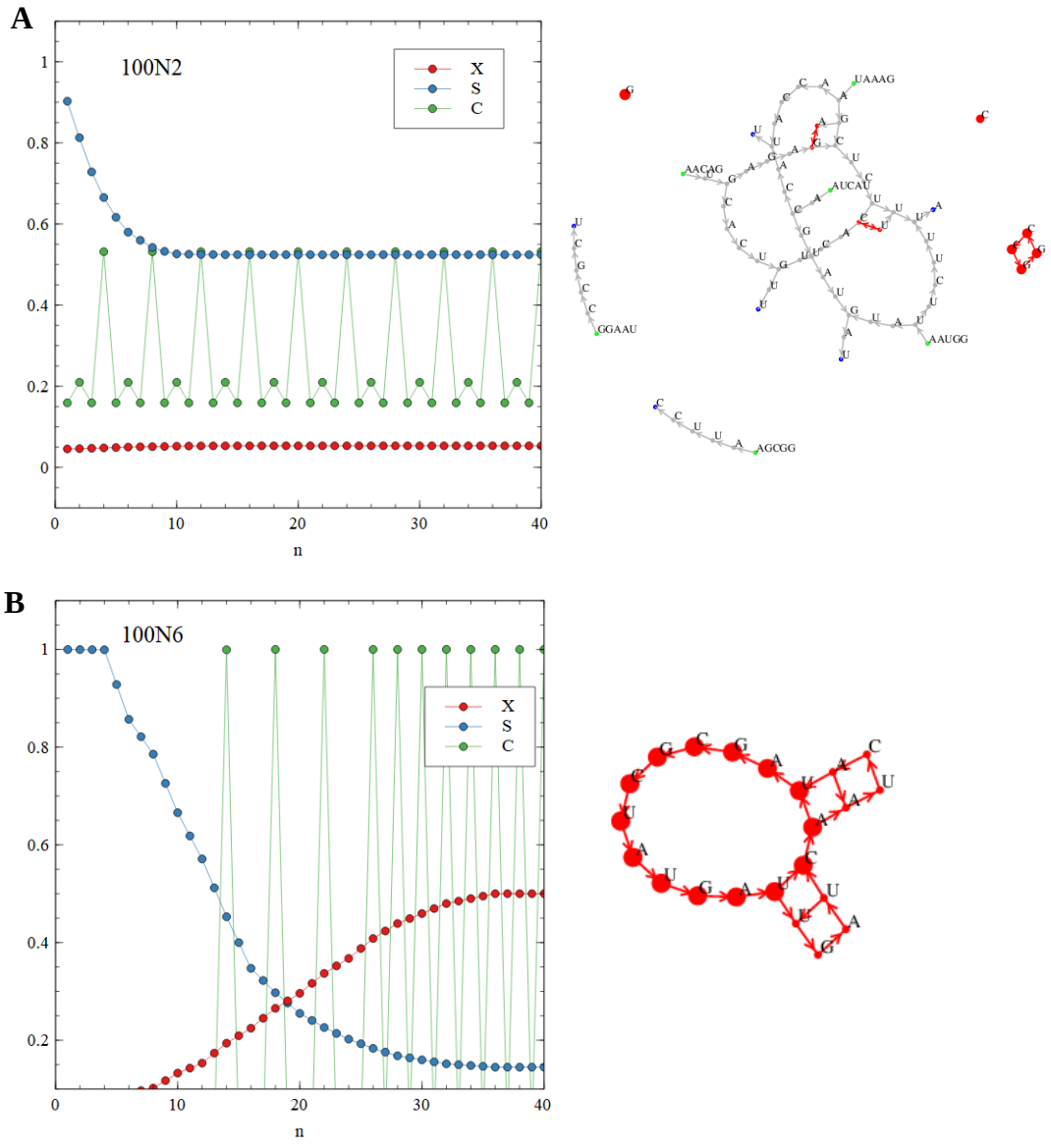

Figure 4.3. XSC functions and connection graphs of two simulations starting from a random mixture of 100 strands. Inflow of new strands was not allowed in either case.

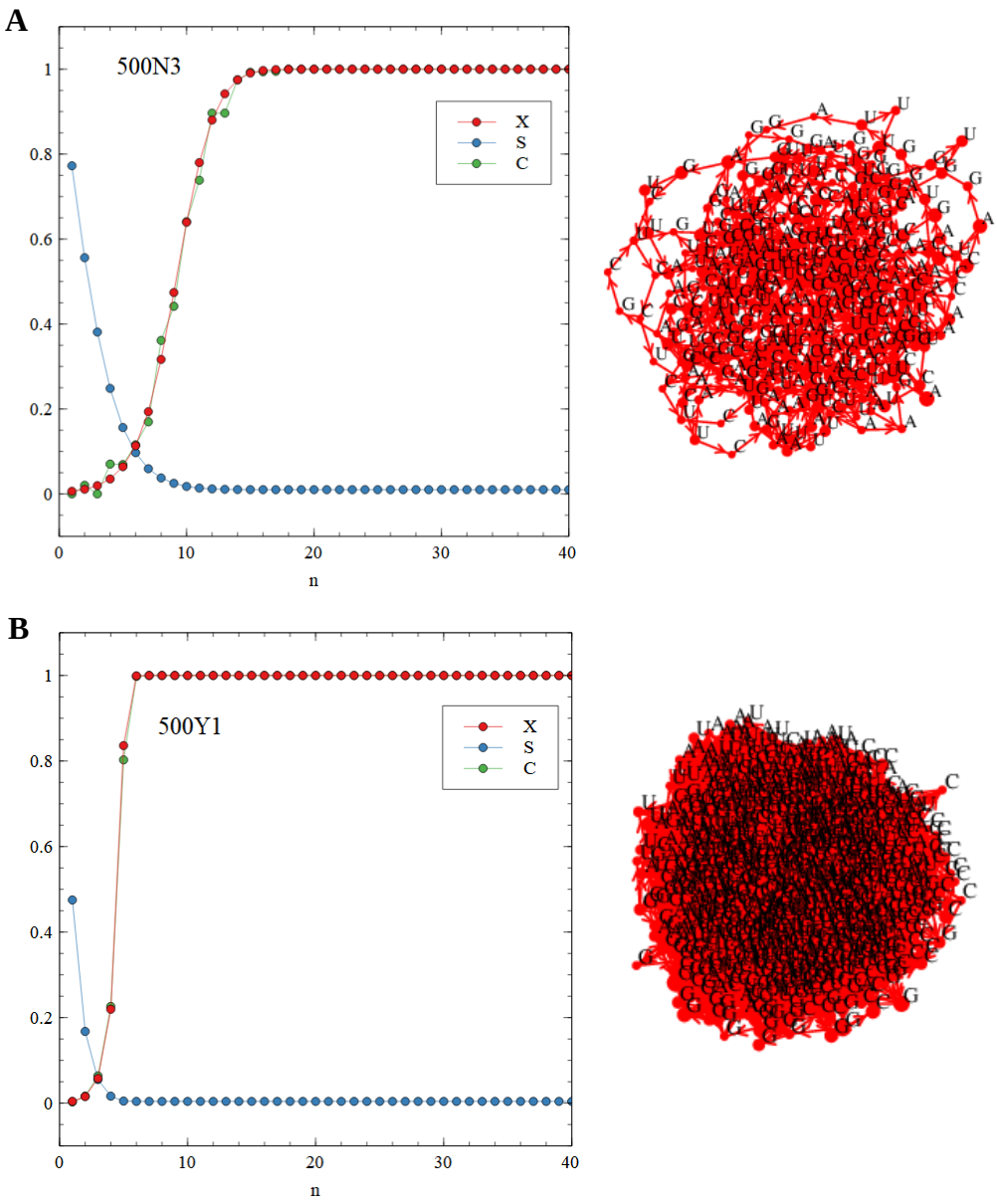

Figure 4.4. XSC functions and connection graphs of two simulations starting from a random mixture of 500 strands. The simulations differ by either (A) disallowing or (B) allowing the inflow of new random sequences.

4.3 Discussion

The template-directed synthesis of RNA under the assumed mechanisms presented in the previous chapter shows potential for emerging long sequences against a strand loss rate. Achieving this increases the possibility that an RNA sequence will develop a useful function. We can think of an RNA genome to make up several of these useful genes, but there are some differing theories about how this genome is stored. Specifically, Zhou et al. has proposed the idea that the genome might not necessarily be stored in one long strand, but rather a collection of shorter strands[56]. These shorter strands would collectively contain all the sequences present in the genome. In addition, the idea of a circular genome has gained popularity mainly due to its advantages in replication[64]. Thus, the collection of overlapping sequences making up the genome are referred to as the virtual circular genome. The advantage of this type of genome is said to come from its ease of replication. Within a situation similar to that considered in our simulations, Zhou et al. states: "Assuming that shorter oligonucleotides were more abundant than longer ones, replication of the entire genome could occur by the growth of all oligonucleotides by as little as one nucleotide on average[56]." Although we can agree that a doubling of sequences can occur through small increases in length per cycle, genome replication implies the presence of a specific set of sequences. In our case, it was likely that the sequence mixture contained a diverse mixture of random strands without any specific virtual circle.

Indeed, it seems that sequence diversity is an important factor influencing the presence of virtual circles. There are several ways in which sequence diversity could vary. One point that seems clear is that the initial sequence diversity directly determines the possibility of a virtual circle. A large starting population of random sequences generally leads

to overlaps between many sequences, meaning there are no specific paths. If we consider a fully diverse mixture where every 5-mer is present with equal frequency, a circular path is possible with 5 steps from any initial word. This outcome is seen to some degree in Figure 4.4. It seems that virtual circles can sometimes occur if the starting population has a low enough diversity. However, if it were possible for new strands to enter the mixture, this would again increase the diversity and eliminate the specificity of circular paths. Nevertheless, one could envisage a scenario in the early RNA world where a low diversity sequence population is encapsulated within a lipid vesicle impermeable to oligomers. We have shown that under our ideal simulation conditions, a pre-existing virtual circle mixture could be maintained, although its emergence may be a rare occurrence.

Under the condition where the initial diversity is low enough and is maintained, it seems that the emergence of specific linear paths is quite likely, as well as simple short circular paths. This explains the difficulty of emerging a virtual circular genome. Too diverse of a mixture creates many non-specific and overlapping circular paths, while a non-diverse mixture makes it difficult for circular paths to exist. Nevertheless, over numerous emergent mixtures and after a long enough time, we cannot rule out the potential emergence of a specific virtual circular genome. Also, we have seen the presence of specific linear paths, suggesting that a virtual genome may not have to be circular. A single circular genome strand benefits greatly in replication compared to a linear strand because it is not limited to a replication starting point. This improves the rate of replication greatly in the case of a circular strand since all the replicates will be complete. However, this is not an important factor in a virtual genome that is broken up into overlapping linear strands. In either case, you replicate a genome by increasing the lengths of all existing strands using genomic templates. The rate of

replication is not as dependent on whether the genome is circular or linear. Due to its more common occurrences in low diversity mixtures, we should certainly consider the possibility of emerging a virtual linear genome. Multiple virtual linear sequences and their complements are produced as seen in Figure 4.3, showing the possibility for random emergence of long functional RNA in a mixture of short oligomers that can replicate reliably. Later, it would be interesting to see connection graphs done at different strand lengths, different time points, and to measure the sequence copy numbers. This would allow a better understanding of how well the sequence is maintained and amplified under ideal conditions.

With regards to circular paths, it can be observed that the rarity may come from the difficulty in creating overlapping paths both to and from a starting sequence. This explains the great level of circular paths created from high diversity mixtures. Indeed, there is a minimum level of diversity required to even potentiate the emergence of a long and specific circular sequence. It should also be noted that longer sequence will have a greater chance of having repeated sequences at separate regions, ultimately allowing for shortcuts, and decreasing specificity. This outcome is seen in Figure 4.3B. The longest unique circular sequence length in this example is 22, but shortcuts are made between repeating sequences leading to a shorter 14 length sequence. Also, small circles of length 4 are made from a portion of the 5-mer sequences. These branching paths are also seen within linear virtual sequences. Ultimately, the branching paths introduce alternative sequences which could also be useful and worth maintaining. Though it seems unlikely that short or repeating sequences will be of much use. Nevertheless, a degree of branching paths can mean the emergence of alternative virtual sequences, whether circular or linear, and is the much more probable case compared to getting long non-branching virtual circles from random mixtures.

So far, we mentioned that virtual sequences could appear when the sequence diversity does not start high or increase over time. Although it may be plausible that mixture diversity is maintained within a semi-permeable lipid vesicle where new oligomer sequences cannot be introduced, it is not as certain whether diversity can be maintained following RNA mutations. Random mutation will create new sequences, and high enough mutation rates can cause new overlapping paths. This would decrease the specificity of the virtual paths, leading to similar results as seen when allowing inflow of new random strands. In general, if the mutation rate, or even inflow of new sequences is low compared to the copy number of genomic strands, then diversity could be maintained through some mechanism of strand loss. For example, vesicle splitting could remove new sequences from the mixture before they continue to mutate and amplify. A small mutation rate would be required for evolution of the genome, allowing a potential beneficial change in the genome sequence. In the future, it would be interesting to see the results of our virtual circle measurements in simulations including error rates during monomer addition. Our simulations naturally assumed the occurrence of a diverse sequence mixture in which strands can grow and replicate consistently but result in random strands with no specific virtual sequence. Limiting the diversity of the mixture as mentioned seems to sometimes allow for the emergence of virtual sequences, but it would be worthwhile to see how the restrictions impact the replication and growth of strands. Finally, even with ideal conditions in place, we should not expect the occurrence of specific virtual circular sequences, but rather virtual sequences that may be linear, and have branching paths creating a set of virtual sequences.

Chapter 5: Conclusions

In this thesis, we concerned ourselves with non-enzymatic replication in the RNA world, mainly investigating the possibility of RNA growth through template-directed polymerization. We specifically addressed the reannealing problem, which was thought to prevent the continuous replication of longer strands even with the presence of environmental cycling[47]. We found that the reannealing case was dependent on the type of RNA mixture, and that without sequence specificity, reannealing can still produce configurations where primer extension occurs meaning growth is not blocked. These diverse sequence environments were also able to produce long RNA sequences over time. The non-enzymatic replication mechanisms involved in this process have been discussed before in the context of a virtual circular genome[56]. Although, from our simulations it seemed that the diversity required for continuous replication and growth would not allow for such virtual genomes. In general, using computational modeling, we managed to expand our understanding of the outcomes of non-enzymatic replication in the RNA world.

Our aim in chapter 2 was built on the understanding that the addition of nucleotides next to a primer where likely based heavily on thermodynamic properties. Specifically, we hypothesized that different bases would extend at rates depending on their stability when paired on the template. We explored this relationship using thermodynamic parameters for base pairing predicted through experiments and observed our hypothesized relationship by comparing predicted rates to experimentally measured primer extension rates. The results were limited because of too many unknown parameters. We predicted rates that were not comparable enough to real rates in order for the model to be applicable to realistic

simulations. In the future, the model can be improved with the discovery of updated thermodynamic parameters, including accurate free energy of monomer binding and mismatch base pairs. This type of model can become quite useful for upcoming quantitative theoretical RNA world research.

Moving forward, we decided to approach the non-enzymatic RNA world with a more qualitative model, mainly to address the case where continuous rounds of RNA replication are driven by temperature cycling. Although temperature cycling is a plausible mechanism for strand separation on the early Earth, rapidly reannealing of complementary sequences is said to pose a great challenge. In chapter 3, we look at a detailed case using a complex computer simulation of non-enzymatic RNA replication. Considering a large mixture of random sequences, the possibility for nucleation of strands from templates, and the inflow of new random sequences, the numerous configurations of helices continuously allow for the extension of polymers. This eventually leads to the emergence of long RNA polymers through template-directed synthesis. It was mainly thought that long strands would emerge only through random spontaneous polymerization, specifically using wet-dry cycling mechanisms[65], and that only amplification of long strands would be done through templated replication. Our findings show that templated replication can create long strands instead, but this is likely at the expense of maintaining a specific template sequence. Since sequences grow by only a few nucleotides each cycle, they usually will grow on multiple templates over time. This prevents the passing on of sequence information.

One way that sequence information can be maintained within this regime is through virtual circular genomes. However, our results from chapter 4 indicate that it would be unlikely for a random mixture to converge to a mixture containing a specific virtual genome

path. Even if we start with a virtual circle mixture, random mutations and inflow of new sequences would eventually increase the diversity and lead to a mixture of random sequences. One could make the argument that semi-permeable vesicles would not allow for inflow of random oligomers, but only monomers. Regardless, proposing the spontaneous emergence of a virtual circle mixture from random sequences seems extremely rare. It may be possible that a virtual circular genome emerged from a physical one, after replication of many short segments. This would be a case where continuous information transfer occurs with temperature cycling. This may still not be the most efficient route. Instead, we propose that useful genomic strands emerged as long physical sequences and were replicated through rolling circle or strand displacement mechanisms which do not require temperature cycling. This way, genome replication is not limited to only the environments where temperature or other environmental fluctuations take place.

In a way, our findings point towards the importance of lipid vesicles since it would be beneficial for long sequences to eventually move to a lower diversity environment. This way, sequence information can be maintained within a confined space for a longer period, allowing for possible selection and evolution of beneficial strands. Although the naturally complex origin of life scenario provides many challenges, we should continue to expand our understanding piece by piece. Non-enzymatic replication was likely an important stage within the RNA world preceding the emergence of ribozymes. Through combined discoveries, we are forming a more complete picture of the ways in which this stage could have progressed and transformed leading up to life's origins.

References

1. Orgel LE. The origin of life-a review of facts and speculations. Trends in Biochemical Sciences. 1998;23(12):491–495. doi:10.1016/S0968-0004(98)01300-0

2. Gilbert W. Origin of life: The RNA world. 1986. https://elearning.uniroma1.it/pluginfile.php/867900/mod_resource/content/1/Origin-of-life-The-RNA-world1986Nature.pdf

3. Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences. 1990 [accessed 2021 Dec 19];87(12):4576–4579. https://www.pnas.org/content/87/12/4576. doi:10.1073/PNAS.87.12.4576

4. Crick F. Central Dogma of Molecular Biology. Nature 1970 227:5258. 1970 [accessed 2021 Dec 19];227(5258):561–563. https://www.nature.com/articles/227561a0. doi:10.1038/227561a0

5. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science (New York, N.Y.). 2000 [accessed 2021 Dec 19];289(5481):920–930. https://pubmed.ncbi.nlm.nih.gov/10937990/. doi:10.1126/SCIENCE.289.5481.920

6. Lai LB, Vioque A, Kirsebom LA, Gopalan V. Unexpected diversity of RNase P, an ancient tRNA processing enzyme: challenges and prospects. FEBS letters. 2010 [accessed 2021 Dec 19];584(2):287–296. https://pubmed.ncbi.nlm.nih.gov/19931535/. doi:10.1016/J.FEBSLET.2009.11.048

7. O'Donnell M, Langston L, Stillman B. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. Cold Spring Harbor perspectives in biology. 2013 [accessed 2021 Dec 19];5(7). https://pubmed.ncbi.nlm.nih.gov/23818497/. doi:10.1101/CSHPERSPECT.A010108

8. Blackburn EH, Greider CW, Szostak JW. Telomeres and telomerase: the path from maize, Tetrahymena and yeast to human cancer and aging. Nature medicine. 2006 [accessed 2021 Dec 19];12(10):1133–1138. https://pubmed.ncbi.nlm.nih.gov/17024208/. doi:10.1038/NM1006-1133

9. Poole AM, Logan DT, Sjöberg BM. The evolution of the ribonucleotide reductases: much ado about oxygen. Journal of molecular evolution. 2002 [accessed 2021 Dec 19];55(2):180–196. https://pubmed.ncbi.nlm.nih.gov/12107594/. doi:10.1007/S00239-002-2315-3

10. Robertson MP, Joyce GF. The Origins of the RNA World. [accessed 2020 Nov 30]. http://cshperspectives.cshlp.org/. doi:10.1101/cshperspect.a003608

11. De La Peña M, García-Robles I. Ubiquitous presence of the hammerhead ribozyme motif along the tree of life. RNA. 2010 [accessed 2020 Nov 30];16(10):1943–1950. /pmc/articles/PMC2941103/?report=abstract. doi:10.1261/rna.2130310

12. Bernhardt HS. The RNA world hypothesis: the worst theory of the early evolution of life (except for all the others)a. Biology Direct. 2012 [accessed 2021 Dec 19];7:23. /pmc/articles/PMC3495036/. doi:10.1186/1745-6150-7-23

13. Higgs PG. Chemical Evolution and the Evolutionary Definition of Life. Journal of Molecular Evolution. 2017;84(5–6):225–235. doi:10.1007/s00239-017-9799-3
14. Ralser M. An appeal to magic? The discovery of a non-enzymatic metabolism and its role in the origins of life. Biochemical Journal. 2018 [accessed 2021 Dec 19];475(16):2577. /pmc/articles/PMC6117946/. doi:10.1042/BCJ20160866

15. Orgel LE. Prebiotic chemistry and the origin of the RNA world. Critical reviews in biochemistry and molecular biology. 2004 [accessed 2021 Dec 19];39(2):99–123. https://pubmed.ncbi.nlm.nih.gov/15217990/. doi:10.1080/10409230490460765

16. Keller MA, Kampjut D, Harrison SA, Ralser M. Sulfate radicals enable a non-enzymatic Krebs cycle precursor. Nature ecology & evolution. 2017 [accessed 2021 Dec 19];1(4). https://pubmed.ncbi.nlm.nih.gov/28584880/. doi:10.1038/S41559-017-0083

17. Keller MA, Turchyn A V., Ralser M. Non-enzymatic glycolysis and pentose phosphate pathway-like reactions in a plausible Archean ocean. Molecular Systems Biology. 2014 [accessed 2021 Dec 19];10(4):725. https://onlinelibrary.wiley.com/doi/full/10.1002/msb.20145228. doi:10.1002/MSB.20145228

18. Messner CB, Driscoll PC, Piedrafita G, De Volder MFL, Ralser M. Nonenzymatic gluconeogenesis-like formation of fructose 1,6-bisphosphate in ice. Proceedings of the National Academy of Sciences of the United States of America. 2017 [accessed 2021 Dec 19];114(28):7403–7407. https://pubmed.ncbi.nlm.nih.gov/28652321/. doi:10.1073/PNAS.1702274114

19. Ferus M, Pietrucci F, Saitta AM, Knížek A, Kubelík P, Ivanek O, Shestivska V, Civiš S. Formation of nucleobases in a Miller-Urey reducing atmosphere. Proceedings of the National Academy of Sciences of the United States of America. 2017 [accessed 2021 Dec 19];114(17):4306–4311. https://www.pnas.org/content/114/17/4306. doi:10.1073/PNAS.1700010114/-/DCSUPPLEMENTAL

20. Ricardo A, Carrigan MA, Olcott AN, Benner SA. Borate minerals stabilize ribose. Science (New York, N.Y.). 2004 [accessed 2021 Dec 19];303(5655):196. https://pubmed.ncbi.nlm.nih.gov/14716004/. doi:10.1126/SCIENCE.1092464

21. Pearce BKD, Pudritz RE. Seeding the Pregenetic Earth: Meteoritic Abundances of Nucleobases and Potential Reaction Pathways. Astrophysical Journal. 2015 [accessed 2021 Dec 19];807(1). http://arxiv.org/abs/1505.01465. doi:10.1088/0004-637X/807/1/85

22. Joyce GF. The antiquity of RNA-based evolution. Nature. 2002 [accessed 2021 Apr 24];418(6894):214–221. www.nature.com/nature. doi:10.1038/418214a

23. Unrau PJ, Bartel DP. RNA-catalysed nucleotide synthesis. Nature. 1998 [accessed 2020 Nov 30];395(6699):260–263. https://pubmed.ncbi.nlm.nih.gov/9751052/. doi:10.1038/26193

24. Benner SA, Kim HJ, Carrigan MA. Asphalt, Water, and the Prebiotic Synthesis of Ribose, Ribonucleosides, and RNA. Accounts of Chemical Research. 2012 [accessed 2020 Nov 30];45(12):2025–2034. https://pubmed.ncbi.nlm.nih.gov/22455515/. doi:10.1021/ar200332w

25. Higgs PG, Lehman N. The RNA World: Molecular cooperation at the origins of life. Nature Reviews Genetics. 2015 [accessed 2020 Nov 30];16(1):7–17. www.nature.com/reviews/genetics. doi:10.1038/nrg3841

26. Higgs PG. Three Ways to Make an RNA Sequence : Steps from Chemistry to the RNA World. Handbook of Astrobiology. 2018 Jun 3 [accessed 2021 Dec 19]:395–407. https://www.taylorfrancis.com/chapters/edit/10.1201/b22230-28/three-ways-make-rna-sequence-paul-higgs. doi:10.1201/B22230-28

27. Ferris JP. Mineral Catalysis and Prebiotic Synthesis: Montmorillonite-Catalyzed Formation of RNA. Elements. 2005;1(3):145–149. doi:10.2113/gselements.1.3.145

28. Higgs PG. The Effect of Limited Diffusion and Wet–Dry Cycling on Reversible Polymerization Reactions: Implications for Prebiotic Synthesis of Nucleic Acids. Life. 2016 [accessed 2021 Dec 19];6(2). /pmc/articles/PMC4931461/. doi:10.3390/LIFE6020024

29. Giovanna C, Pino S, Ciciriello F, Di Mauro E. Generation of Long RNA Chains in Water. The Journal of Biological Chemistry. 2009 [accessed 2021 Dec 19];284(48):33206. /pmc/articles/PMC2785163/. doi:10.1074/JBC.M109.041905

30. Rajamani S, Vlassov A, Benner S, Coombs A, Olasagasti F, Deamer D. Lipid-assisted synthesis of RNA-like polymers from mononucleotides. Origins of Life and Evolution of Biospheres. 2008 [accessed 2021 Apr 24];38(1):57–74. https://link.springer.com/article/10.1007/s11084-007-9113-2. doi:10.1007/s11084-007-9113-2

31. Da Silva L, Maurel MC, Deamer D. Salt-Promoted Synthesis of RNA-like Molecules in Simulated Hydrothermal Conditions. Journal of Molecular Evolution. 2015 [accessed 2021 Apr 24];80(2):86–97. https://link.springer.com/article/10.1007/s00239-014-9661-9. doi:10.1007/s00239-014-9661-9

32. Kim DE, Joyce GF. Cross-Catalytic Replication of an RNA Ligase Ribozyme. Chemistry & Biology. 2004 [accessed 2021 Dec 20];11(11):1505–1512. http://www.cell.com/article/S1074552104002777/fulltext. doi:10.1016/J.CHEMBIOL.2004.08.021

33. Paul N, Joyce GF. A self-replicating ligase ribozyme. Proceedings of the National Academy of Sciences of the United States of America. 2002 [accessed 2021 Dec 20];99(20):12733–12740. https://pubmed.ncbi.nlm.nih.gov/12239349/. doi:10.1073/PNAS.202471099

34. Draper WE, Hayden EJ, Lehman N. Mechanisms of covalent self-assembly of the Azoarcus ribozyme from four fragment oligonucleotides. Nucleic Acids Research. 2008 [accessed 2021 Dec 20];36(2):520. /pmc/articles/PMC2241849/. doi:10.1093/NAR/GKM1055

35. Hayden EJ, Von Kiedrowski G, Lehman N. Systems Chemistry on Ribozyme Self-Construction: Evidence for Anabolic Autocatalysis in a Recombination Network. Angewandte Chemie International Edition. 2008 [accessed 2021 Dec 20];47(44):8424–8428. https://onlinelibrary.wiley.com/doi/full/10.1002/anie.200802177. doi:10.1002/ANIE.200802177

36. Attwater J, Wochner A, Holliger P. In-ice evolution of RNA polymerase ribozyme activity. Nature Chemistry. 2013;5(12):1011–1018. doi:10.1038/nchem.1781

37. Attwater J, Raguram A, Morgunov AS, Gianni E, Holliger P. Ribozyme-catalysed RNA synthesis using triplet building blocks. eLife. 2018;7. doi:10.7554/ELIFE.35255

38. Horning DP, Joyce GF. Amplification of RNA by an RNA polymerase ribozyme. Proceedings of the National Academy of Sciences of the United States of America. 2016;113(35):9786–9791. doi:10.1073/pnas.1610103113

39. Orgel LE. Unnatural selection in chemical systems. Accounts of chemical research. 1995 [accessed 2021 Dec 20];28(3):109–118. https://pubmed.ncbi.nlm.nih.gov/11542502/. doi:10.1021/AR00051A004

40. Prywes N, Blain JC, Del Frate F, Szostak JW. Nonenzymatic copying of RNA templates containing all four letters is catalyzed by activated oligonucleotides. eLife. 2016;5(JUN2016). doi:10.7554/eLife.17756

41. Szostak JW. The eightfold path to non-enzymatic RNA replication. Journal of Systems Chemistry. 2012 [accessed 2020 Nov 30];3(1):1–14. https://link.springer.com/articles/10.1186/1759-2208-3-2. doi:10.1186/1759-2208-3-2

42. Inoue T, Orgel LE. A nonenzymatic RNA polymerase model. Science. 1983 [accessed 2020 Nov 30];219(4586):859–862. https://science.sciencemag.org/content/219/4586/859. doi:10.1126/science.6186026

43. Bridson PK, Orgel LE. Catalysis of accurate poly(C)-directed synthesis of 3′-5′-linked oligoguanylates by Zn2+. Journal of Molecular Biology. 1980;144(4):567–577. doi:10.1016/0022-2836(80)90337-X

44. Leu K, Obermayer B, Rajamani S, Gerland U, Chen IA. The prebiotic evolutionary advantage of transferring genetic information from RNA to DNA. Nucleic Acids Research. 2011 [accessed 2020 Nov 30];39(18):8135–8147. https://academic.oup.com/nar/article/39/18/8135/1090041. doi:10.1093/nar/gkr525

45. Rajamani S, Ichida JK, Antal T, Treco DA, Leu K, Nowak MA, Szostak JW, Chen IA. Effect of stalling after mismatches on the error catastrophe in nonenzymatic nucleic acid replication. Journal of the American Chemical Society. 2010 [accessed 2020 Nov 30];132(16):5880–5885. /pmc/articles/PMC2857888/?report=abstract. doi:10.1021/ja100780p

46. Ricardo A, Szostak JW. Origin of life on earth. Scientific American. 2009 [accessed 2020 Nov 30];301(3):54–61. https://pubmed.ncbi.nlm.nih.gov/19708528/. doi:10.1038/scientificamerican0909-54

47. Tupper AS, Higgs PG. Rolling-Circle and Strand-displacement Mechanisms for Non-enzymatic Replication in the RNA World. 2020. doi:10.1017/CBO9781107415324.004

48. Zhou L, Kim SC, Ho KH, O'Flaherty DK, Giurgiu C, Wright TH, Szostak JW. Non-enzymatic primer extension with strand displacement. eLife. 2019 [accessed 2020 Jan 15];8. https://elifesciences.org/articles/51888. doi:10.7554/eLife.51888

49. Li L, Prywes N, Tam CP, Oflaherty DK, Lelyveld VS, Izgu EC, Pal A, Szostak JW. Enhanced nonenzymatic RNA copying with 2-aminoimidazole activated nucleotides. Journal of the American Chemical Society. 2017 [accessed 2021 Dec 21];139(5):1810–1813. https://pubs.acs.org/doi/full/10.1021/jacs.6b13148. doi:10.1021/JACS.6B13148/SUPPL_FILE/JA6B13148_SI_001.PDF

50. Zhou L, O'Flaherty DK, Szostak JW. Template-Directed Copying of RNA by Non-enzymatic Ligation. Angewandte Chemie. 2020 [accessed 2021 Dec 21];132(36):15812–15817. https://onlinelibrary.wiley.com/doi/full/10.1002/ange.202004934. doi:10.1002/ANGE.202004934

51. Walton T, Szostak JW. A Highly Reactive Imidazolium-Bridged Dinucleotide Intermediate in Nonenzymatic RNA Primer Extension. Journal of the American Chemical Society. 2016 [accessed 2021 Dec 21];138(36):11996–12002. https://pubs.acs.org/doi/full/10.1021/jacs.6b07977. doi:10.1021/JACS.6B07977/SUPPL_FILE/JA6B07977_SI_001.PDF

52. O'Flaherty DK, Kamat NP, Mirza FN, Li L, Prywes N, Szostak JW. Copying of Mixed-Sequence RNA Templates inside Model Protocells. Journal of the American Chemical Society. 2018 [accessed 2020 Dec 1];140(15):5171–5178. https://pubs.acs.org/sharingguidelines. doi:10.1021/jacs.8b00639

53. Mansy SS, Szostak JW. Thermostability of model protocell membranes. Proceedings of the National Academy of Sciences of the United States of America. 2008 [accessed 2020 Nov 30];105(36):13351–13355. www.pnas.orgcgidoi10.1073pnas.0805086105. doi:10.1073/pnas.0805086105

54. Strulson CA, Molden RC, Keating CD, Bevilacqua PC. RNA catalysis through compartmentalization. Nature Chemistry 2012 4:11. 2012 [accessed 2021 Dec 21];4(11):941–946. https://www.nature.com/articles/nchem.1466. doi:10.1038/nchem.1466

55. Bansho Y, Ichihashi N, Kazuta Y, Matsuura T, Suzuki H, Yomo T. Importance of Parasite RNA Species Repression for Prolonged Translation-Coupled RNA Self-Replication. Chemistry & Biology. 2012;19(4):478–487. doi:10.1016/J.CHEMBIOL.2012.01.019

56. ZHOU L, DING D, SZOSTAK JW. The Virtual Circular Genome Model for Primordial RNA Replication. RNA. 2020 [accessed 2021 Dec 16];27(1):rna.077693.120. http://rnajournal.cshlp.org/content/early/2020/10/07/rna.077693.120. doi:10.1261/RNA.077693.120

57. Flores R, Grubb D, Elleuch A, Nohales MÁ, Delgado S, Gago S. Rolling-circle replication of viroids, viroid-like satellite RNAs and hepatitis delta virus: variations on a theme. RNA biology. 2011 [accessed 2021 Dec 21];8(2). https://pubmed.ncbi.nlm.nih.gov/21358283/. doi:10.4161/RNA.8.2.14238

58. Turner DH, Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Research. 2010 [accessed 2021 Dec 21];38(Database issue):D280. /pmc/articles/PMC2808915/. doi:10.1093/NAR/GKP892

59. Xia T, SantaLucia J, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson - Crick base pairs. Biochemistry. 1998;37(42):14719–14735. doi:10.1021/bi9809425

60. Chen JL, Dishler AL, Kennedy SD, Yildirim I, Liu B, Turner DH, Serra MJ. Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. Biochemistry. 2012 [accessed 2020 Nov 30];51(16):3508–3522. https://pubs.acs.org/sharingguidelines. doi:10.1021/bi3002709

61. Bapat N V., Rajamani S. Effect of Co-solutes on Template-Directed Nonenzymatic Replication of Nucleic Acids. Journal of Molecular Evolution. 2015 [accessed 2020 Dec 1];81(3–4):72–80. https://link.springer.com/article/10.1007/s00239-015-9700-1. doi:10.1007/s00239-015-9700-1

62. Izgu EC, Fahrenbach AC, Zhang N, Li L, Zhang W, Larsen AT, Blain JC, Szostak JW. Uncovering the thermodynamics of monomer binding for RNA replication. Journal of the American Chemical Society. 2015 [accessed 2020 Dec 1];137(19):6373–6382. https://pubs.acs.org/sharingguidelines. doi:10.1021/jacs.5b02707

63. Chen JL, Dishler AL, Kennedy SD, Yildirim I, Liu B, Turner DH, Serra MJ. Testing the nearest neighbor model for canonical RNA base pairs: Revision of GU parameters. Biochemistry. 2012;51(16):3508–3522. doi:10.1021/bi3002709

64. Meng X, Li X, Zhang P, Wang J, Zhou Y, Chen M. Circular RNA: an emerging key player in RNA world. Briefings in bioinformatics. 2017 [accessed 2021 Dec 16];18(4):547–557. https://pubmed.ncbi.nlm.nih.gov/27255916/. doi:10.1093/BIB/BBW045

65. Higgs P. The Effect of Limited Diffusion and Wet–Dry Cycling on Reversible Polymerization Reactions: Implications for Prebiotic Synthesis of Nucleic Acids. Life. 2016 [accessed 2021 Apr 24];6(2):24. http://www.mdpi.com/2075-1729/6/2/24. doi:10.3390/life6020024