Template Directed Ligation of RNA Oligomers

MODELING THE EFFECTS OF HYDROLYSIS, POLYMERIZATION AND TEMPLATE-DIRECTED LIGATION ON THE FORMATION OF PREBIOTIC RNA OLIGOMERS

By ERIC TURNER, B.Sc.

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

McMaster University © Copyright by Eric Turner, September 2018

McMaster University MASTER OF SCIENCE (2018) Hamilton, Ontario (Physics)

TITLE: Modeling the Effects of Hydrolysis, Polymerization and Template-Directed Ligation on the Formation of Prebiotic RNA Oligomers

AUTHOR: Eric Turner, B.Sc. (McMaster University)

SUPERVISOR: Professor P. Higgs

NUMBER OF PAGES: ix, 79

LAY ABSTRACT:

The origin of life on Earth is a long-debated question that has been asked by nearly every civilization to have existed. This research addresses the origin of life in the context of the RNA World theory, which proposes that the first kind of replicating molecules were RNA strands, specifically, catalytic RNA sequences, called ribozymes. We carry out computer simulations of the formation and break-up of short RNA strands. Strands can grow by joining together randomly, or due to the action of template strands. We find that, if this process occurs repeatedly, the RNA strands in the mixture move towards states in which groups of sequences that are good templates for one another occur together at high concentrations. By studying the possible states that arise in this reaction mixture, we hope to learn about the first replicating RNA strands that lead to the origin of life.

ABSTRACT:

The key to the RNA world hypothesis is the ribozyme, an information and catalytic agent that preceded proteins and DNA. Prior to ribozymes the sequences of RNA needed to build up to a length that could potentially be a ribozyme. This research focuses on computational modelling of hydrolysis, polymerization, and template-directed ligation to determine sequence patterns and characteristics that may have emerged due to these simple processes. A model containing L- and D-chirality monomers is used that incorporates the advantage of being a uniform chirality to achieve chiral symmetry breaking. Another chirality model is used where being uniform provides no advantage and a symmetry breaking still occurs. Beyond chirality we look at nucleobase models where we use a two letter alphabet containing adenine and uracil to determine symmetry breaking in sequence space. This results in self-complementary sequences dominating this model at all ligation rates but under certain initial conditions including high concentration, other types of sequences can be dominant. If a third base, guanine is added to this model a wobble base is created. In these models the self-complementary sequences containing uracil are the most prevalent due to uracil's ability to pair with both adenine and guanine. Finally, upon adding a fourth base to the model guanine also becomes a wobble pair and the sequences containing uracil and guanine dominate the system for low ligation rates but at higher rates the uniform uracil and guanine sequences dominate. For each model a version is run with the templating reaction scaling linearly with the number of binding sites and without, where all templates are equally good. Generally, the scaling causes symmetry breakings at lower ligation values for each model.

ACKNOWLEDGEMENTS:

Thank you to my Supervisor Dr. Paul Higgs for his enlightening discussions and support throughout my research.

A special thanks to my partner Jamie for being such a delightfully supportive person over the course of this degree.

Finally, a special thanks to the students of the Astrobiology program with whom I had many thought provoking discussions.

Table of Contents

Chapter 1: Introduction	1
1.1 Motivation	1
1.2 A Plausible Prebiotic Pathway to Nucleotides	3
1.3 Beyond Nucleotides	6
1.4 Chirality	9
1.5 Previous Models of Template-Directed Replication	11
1.6 Goals	
Chanter 2: Modeling Template-Directed Replication	15
2 1 Simulating the Formation and Breakun of RNA Strands	15
2.1 5 Hudrolysis	15
2.1.2 Polymerization	16
2.1.2 Torymerization 2.1.3 Template-Directed Ligation	17
2.1.5 Template Directed ingation	
2.2. Deterministic Model	18
2.3 Methods for Simulating Systems of Chemical Reactions	19
2.4 Definition of Templates and Motifs	
Chapter 3: LD Models	
3.1 Chirality Models	25
3.2 Chiral Symmetry Breaking	25
3.3 Templating and Sequences	30
3.4 Alternate Equilibria	40
3.5 Chirality Conclusions	46
Chanter 4: AU Model	48
4 1 All Nucleobase Model	48
4.2 Unique Situations	56
4 3 All Conclusions	59
Chapter 5: Wobble Base Models	
5.1 AUG Model	60
5.2 ATGC/AUGC Model	65
5.3 Wobble Base Conclusions	68
Chapter 6: Conclusions and Future Work	
6.1 Summary	
6.2 Limitations and Future Work	73
References:	76

LIST OF FIGURES AND TABLES

Figures

Figure 2.1 The 3 basic reactions present in each model	15
Figure 2.2 Template-directed ligation for the uniform LD model	22
Figure 2.3 The change in the ligation process from LD-UT to LD-AT	23
Figure 2.4 The change in the ligation process from LD-UT to AU-T	24
Figure 3.1 Bifurcation of L- and D-monomers in LD-UT	26
Figure 3.2 Bifurcation of L- and D-monomers in LD-UM	
Figure 3.3 Bifurcation of L- and D-monomers in LD-AT	29
Figure 3.4 Bifurcation of L- and D-monomers in LD-AM	30
Figure 3.5 Sequences of tetramers at bifurcation in LD-UT	33
Figure 3.6 Sequences of tetramers at bifurcation in LD-AT	37
Figure 3.7 Sequences of tetramers at bifurcation in LD-AM	38
Figure 3.8 The effects of racemization on system dynamics	39
Figure 3.9 Reaction families for the LD models	41
Figure 3.10 LDDL dominant phase for the LD-AM model	43
Figure 3.11 Uniform sequence dominant phase for the LD-AM model	44
Figure 4.1 Reaction families for the AU model	50
Figure 4.2 Sequences of tetramers in the AU-T model	52
Figure 4.3 AAUU dominant phase in the AU-M model	54
Figure 4.4 Codominant uniform sequences in the AU-M model	55
Figure 4.5 Dominant pentamers in the AU-T model with max length 5	56
Figure 4.6 Dominant pentamers in the AU-M model with max length 5	57
Figure 5.1 The AUG model with motifs highlighting domiannt tetramers	64
Figure 5.2 The AUGC model with motifs highlighting domiannt tetramers	67

<u>Tables</u>

Table 2.1 Model abbreviations and their differences	
Table 3.1 Number of unique templates for the LD-UT model	
Table 3.2 Number of unique templates for the LD-UM model	
Table 3.3 Number of unique tempaltes for sequences in the LD-AT model	

LIST OF ABBREVIATIONS AND SYMBOLS

RNA	Ribonucleic Acid
DNA	Deoxyribonucleic Acid
IDP	Interplanetary Dust Particle
HCN	Hydrogen Cyanide
А	Adenine
U	Uracil
G	Guanine
С	Cytosine
Т	Thymine
ppb	parts per billion
NASA	National Aeronautics and Space Administration
L	L-form chirality
D	D-form chirality
LD-UT	LD model with Uniform Templates
LD-UM	LD model with Uniform Motifs
LD-AT	LD model with all Templates
LD-AM	LD model with all Motifs
AU-T	AU model with Templates
AU-M	AU model with Motifs

DECLARATION OF ACADEMIC ACHIEVEMENT

I, Eric Turner, with the guidance of my supervisor, Dr. Paul Higgs conducted this research between September 2016 and September 2018 at McMaster University. It has not been submitted previously for a degree of any kind. I am the author of this work and have been careful to attribute the work of others.

Chapter 1: Introduction

1.1 Motivation

This thesis will illuminate how simple chemical mechanisms influenced the structure of the first RNA oligomers prebiotically. The beginnings of life on Earth is a complex problem with many contributing aspects, RNA oligomer structure being only a small component. Prior to their structure it is important to discuss how a prebiotic Earth could have possibly reached a point where oligomers can form.

The RNA world concept came about in 1962 and provided a possible explanation for the origin of life on Earth (Rich, 1962). Alexander Rich proposed that the RNA world preceded the current DNA, RNA, and protein world due to its ability to carry information like DNA can, and also can act as a catalyst like current proteins (Rich, 1962). Ribozymes are the key to the theory of the RNA world because they provide both the information carrying function as well as the catalyst function; specifically, they would catalyze key steps in the process of replication. Ribozyme are RNA molecules that can act as an enzyme, the first one discovered was a self-splicing intron by Cech in 1982, which could cleave sequences of RNA. It was Gilbert in 1986 that proposed ribozymes as the key to the RNA world and made the connection between their catalytic properties and information storing capabilities (Gilbert, 1986). Further, he is largely credited as the originator of the term "RNA world" (Gilbert, 1986). Moving to present day the shorter of the experimentally found ribozymes to perform the catalytic function are around 200 bases long. Johnston et al., 2001 found a ribozyme 189 bases long that was capable of polymerizing short oligomers around 14 bases long via a template while Wochner et al., 2011 found a ribozyme of similar length that could extend primers by 26 nucleotides. Paul and Joyce had a self-replicating ligase ribozyme 61 bases in length but it lacks the general replication of oligomers possessed by the polymerases of Wochner et al. and Johnston et al.

In a world of DNA, RNA, and proteins the synthesis of long RNA strands occurs via transcription from a DNA template and is catalyzed by RNA polymerase proteins. However, on a prebiotic Earth none of this machinery existed and the creation of the first ribozyme would have to come from many energetically favorable reactions, including the ones discussed in this thesis. Current mechanisms for overcoming thermodynamically unfavorable gradients require complex metabolisms to drive things inside of biological organisms that did not exist on prebiotic Earth. Life today is highly ordered and feeds off of the order of the surrounding environment, a simple example is how we sustain ourselves as humans. We consume food and use its chemical energy to run our metabolism and dispose of disordered waste products. Therefore, it is important to address what kinds of thermodynamic conditions are required to have a reaction move forward at a rate fast enough that the nucleotides and chemical components cannot degrade. Most importantly the reactions leading to the creation of a ribozyme will have to address and satisfy the Gibbs free energy equation which relates the enthalpy of reaction to the entropy of reaction. It is unlikely to find a 61 base long ribozyme, like the one of Paul and Joyce, by randomly connecting nucleotides because the entropy term is too large due to the fact that there are 4⁶¹ possible sequences of that length, so there must be some

conditions or steps that lead up to that point.

1.2 A Plausible Prebiotic Pathway to Nucleotides

The Miller-Urey experiment was the original and most influential experiment to the study of the origins of life (Miller, 1953). The experiment was built on the work of Oparin which centered around the ability of early Earth to create organic molecules out of inorganic reactants (Oparin, 1938). The experiment combined water, hydrogen, methane. and ammonia in an apparatus where it was heated then that evaporate was continuously sparked and condensed and collected. It was an attempt to simulate early conditions on Earth to determine if these inorganic compounds could give rise to organic compounds used in life. The Miller-Urev experiment was only capable of creating some amino acids under these conditions that were thought to exist on early Earth (Miller, 1953). However, their research sparked an interest and others followed up to find they could create the nucleobase adenine from hydrogen cyanide (HCN) and Ammonia in a water solution. As more experiments were done it was shown to be possible to get all of the nucleobases (uracil, guanine, cytosine, and thymine) under early Earth's reducing atmosphere (Ferus et al., 2017). However, there is now some debate about the composition of the Earth's early atmosphere and whether it was reducing or oxidizing (Zahnle et al., 2010). If it is not possible to create the nucleobases on Earth then there must be an extraterrestrial source, as we know from Earth's current state that these nucleobases in fact exist on Earth. It is possible for the nucleobases to have formed on meteorites and interplanetary dust particles (IDPs) (Callahan et al., 2011; Nuevo et al., 2014). Meteorites and IDPs were impacting and settling on the planet frequently during the early history of Earth and depositing the nucleobases in the environment. The number of nucleobases on IDPs is inconsequential but the meteorites are found to have concentrations of 0.25 to 515 parts per billion (ppb) which could potentially last long enough to form the first nucleotides and oligomers (Pearce and Pudritz, 2015; Pearce et al., 2017). An interesting note about meteorites supplying the nucleobases is that they only bring reasonable abundances of adenine, guanine and uracil due to the cytosine being easily broken into uracil via UV radiation (Pearce and Pudritz, 2016). The implications of only having three of the four canonical bases used in RNA will be addressed in chapter 5.

If there is a reasonable concentration of nucleobases, whether or not they are produced on Earth, we still need the other pieces of the nucleotide molecule: the ribose sugar and the phosphate backbone. Then the nucleobase, ribose, and phosphate need to come together to form a nucleotide.

Ribose is made abiologically through the formose reaction where formaldehyde and glycoaldehyde, both naturally present on early Earth, are converted into pentoses (Butlerow, 1861; Breslow, 1959). In early experiments there was some question as to whether the ribose is created in great enough concentration due to its rapid degradation (Shapiro, 1988; Benner et al., 2004). However, more recently there has been headway made in stabilizing the ribose through borates so that the ribose can last long enough to form a more stable structure like a nucleotide (Ricardo et al., 2004; Benner et al., 2012).

Phosphate is present in many different organophosphate compounds and can act as donors to produce nucleotide monophosphates from nucleosides (Reimann and Zubay, 1999). Inorganic phosphates can also be used to phosphorylate nucleosides in the presence of urea and ammonium chloride (Lohrmann amd Orgel, 1971). The joining of the ribose and nucleobase to create the nucleoside that is phosphorylated is difficult. There have been small successes but only for the creation of some nucleosides and in small yields, more research is needed to bridge the gap from ribose and nucleobase to nucleoside (Fuller et al., 1972).

To create these necessary components each reaction requires specific environmental conditions that often do not coincide. For example, the molecules HCN, used in the formation of the nucleobases, and formaldehyde, used in the formation of ribose, readily react (Zubay and Mui, 2001) meaning they likely did not form in the same place which could point towards nucleobases having an extraterrestrial source. However, Benner's work attempts to show a plausible chain of reactions and environmental conditions that allow for the formation of nucleotides (Benner et al., 2012). The work of Powner, Gerland and Sutherland, 2009 also shows a chain of reactions that culminates in nucleotides.

Now that there is a hypothetical pathway to nucleotides on the prebiotic Earth it is time to think about the key to the RNA world: ribozymes. The first step to ribozymes is to create long chains of nucleotides, referred to henceforth as oligomers, which eventually build up to a ribozyme or ribozymes capable of catalyzing replication. The catalyzed reaction could be the reaction required for the formation of itself, termed autocatalysis. Alternatively, it could catalyze the reaction to catalyze the formation of other ribozymes that in turn catalyze its formation termed autocatalytic sets. It is important we do not mistake chemical processes like autocatalytic sets for life. The autocatalytic sets of ribozymes can self-replicate and RNA strands with higher reproducibility will outcompete others, which fits the current NASA definition of life, "A self-sustaining chemical system capable of Darwinian evolution". An important difference here, addressed in Higgs 2017, is that this chemical evolution is reproducible and not openended since the chemical system will eventually select the best strands based on their chemical traits whereas life will keep improving indefinitely, creating new and better fitness traits. This research does not claim to have a solution up to the point in the RNA world where nucleotides are abundant. Instead this research focuses on the next step in the process to try and fill in the gap between the introduction of nucleotides on Earth and the construction of ribozymes and to explore some of the possibilities or features that may be present (in the environment) during that step.

1.3 Beyond Nucleotides

Once nucleotides have formed the next step in the creation of a ribozyme is polymerization. Polymerization is the process by which monomers and oligomers join together to create chains of nucleotides. A phosphodiester bond is formed between two nucleotides to join their strands via a dehydration reaction. A dehydration reaction requires a drier environment than that of early Earth since the water molecules naturally want to hydrolyze this bond. This means that there must have been some conditions that allow this reaction to move forwards and create oligomers in the first place. Two of the leading schools of thought on how this may have occurred are through wet-dry cycling and the presence of clays. Wet-dry cycling is a concept where we imagine a warm little pond and on the edge of this pond are nucleotides and oligomers (Deamer et al., 2006; Mamajanov et al., 2014; Damer and Deamer, 2015; Higgs, 2016). As the pond begins to dry, the oligomers and nucleotides on the edge are no longer surrounded by hydrolyzing water molecules, and so the equilibrium shifts in the other direction and water is taken out of the nucleotides and oligomers through a polymerization (dehydration) reaction. Now that a longer oligomer has formed and once the pond level rises again it can move closer to other oligomers and possibly survive hydrolysis until the next time the water level drops and form an even longer oligomer (Ross and Deamer, 2016).

The second possibility is the presence of clays. The most prospective clay is Montmorillonite, a substance thought possible to have been present during abiogenesis due to its formation by the weathering of volcanic ash and the high level of volcanic activity at that time (Ferris, 2005). Montmorillonite, like other clays, promotes polymerization of oligomers due to its chemical structure allowing the nucleobases to bind via van der Waals forces and hold the nucleotides in place (Ding et al., 1996; Himbert et al., 2016). As the clay binds more nucleotides it aligns the backbone so that the phosphodiester bond is more likely to form. Thus, this hypothesis aids in the creation of longer oligomers by aligning nucleotides in an orientation that promotes bond formation.

Both of these mechanisms have one thing in common, the use of a flat surface to promote polymerization. By changing the polymerization environment to a two dimensional surface it takes away the many degrees of freedom available in a three

7

dimensional environment like water. It does this by holding the nucleotides by their nucleobases, due to the surface charge of Montmorillonite, so there are more oligomers in a smaller area with a greater chance of interacting in the proper orientation.

Template-directed ligation is a similar process to polymerization in that it is the joining of two oligomers or monomers except there is a third oligomer that acts to catalyze the reaction. In this reaction there is a longer oligomer that binds to two smaller oligomers via the hydrogen bonding between the complementary monomer units. The two smaller oligomers are then held in place by the template as a phosphodiester bond forms between them to create a larger oligomer. The new oligomer then detaches from the template oligomer and can go on to template for more oligomers. On a prebiotic Earth this process occurs non-enzymatically and potential pathways for this are outlined in Szostak, 2012. Chen and Nowak, 2012 modeled this process and noted that sequences of higher fitness, where the fitness values were randomly assigned, were selected and would dominate the system establishing that evolutionary dynamics emerge once replication occurs. Furthermore, the process of template-directed ligation has been shown to select certain properties of nucleotides like a uniform 3'-5' backbone in RNA sequences (Rohatgi et al., 1996). The reaction could be a step between polymerization and ribozyme catalyzed replication as it has the advantage of not needing a specific ribozyme to perform replication. It also has the advantage of being able to ligate oligomers instead of adding one monomer at a time. The experiments in template-directed ligation are reviewed in Orgel, 2012 but none of the experiments are of a completely replicating system so far.

Polymerization and template-directed ligation often require activated nucleotides. Activated nucleotides being those that are primed by being triphosphorylated or otherwise and thus make polymerization energetically favorable. Activated pyrimidine nucleotides can occur prebiotically under certain conditions according to a set of reactions outlined in Powner, Gerland and Sutherland, 2009. A plausible route to prebiotically activated purines comes from Lohrmann 1975 which reveals a chemical pathway to 5'tetraphosphate nucleosides. In some cases it is possible to undergo polymerization without activation like in the presence of the previously mentioned clays (Rajamani et al., 2008; Himbert et al., 2016). In the models in this thesis all nucleotides undergoing reactions are assumed to be activated including those that have formed into oligomers and then been broken due to hydrolysis or we have assumed that activation is unnecessary.

1.4 Chirality

Furthermore, the process of template-directed ligation has proposed a solution to life's uniform chirality. Uniform chirality is important because the most important macromolecules to life including amino acids, RNA, and DNA are homochiral and require homochirality to function while the chemical reactions that produce these molecules produce both chiralities. Chirality is a property of a molecule that requires us to define its handedness. Chiral molecules have two forms, dextrorotatory and levorotatory (referred to henceforth as D- and L- for labeling molecules' chirality), and each form is identical other than that they are mirror images of each other and do not match when superimposed on each other no matter their orientation. One thing we hope

to show in this thesis is a plausible pathway to uniform chirality. From species to species the chiral molecules are consistent which could suggest that the selection occurred prebiotically and thus the mechanism is of great interest. Template-directed ligation leads to autocatalysis of oligomers of the same chirality. This is called asymmetric autocatalysis, since the catalytic property only applies to a subset and breaks the symmetry. In this scenario, the catalytic property applies to its own matching enantiomer and thus fluctuations can break the symmetry by choosing one enantiomer over the other. The introduction of autocatalysis leads to growth of instabilities in a system of oligomers that leads to one chirality dominating over the other and a symmetry breaking occurring between the chiralities (Frank, 1953). If we look at Plasson et al., 2007 or Frank, 1953 they use the fact that like-handed monomers act as catalysts for the nucleotide formation of the same handedness to find chiral symmetry breaking at the monomer synthesis level. In our model we are interested in chiral symmetry breaking at the oligomer level. If we look at synthesis, template-directed ligation and enzyme-directed formation of oligomers the synthesis (polymerization) is non-selective and will create any mixed chirality oligomer and the enzyme-directed formation replicates any strand without bias. In the template-directed case there is selection occurring by the templates for certain complements since they are more likely to bind to molecules of the same chirality. This was shown in Bolli et al., 1997 where the tetramers of uniform chirality were ligated much more quickly than those of mixed-chirality. Furthermore, it has also been shown recently in our group that template-directed replication also leads to chiral symmetry breaking (Tupper et al., 2017). In their model beyond a certain ligation rate slight fluctuations in the enantiomeric excess are amplified by the asymmetric autocatalysis provided by the template-directed ligation reaction and a chiral symmetry breaking occurs. The template-directed ligation only occurs for uniformly chiral molecules.

Chirality can also emerge in other ways. One such example is that the origins of biomolecules originate on comets and meteorites, as previously discussed, and during this time circularly polarized radiation may have destroyed one chirality over the other (Bonner and Rubenstein, 1987). In the case of autocatalysis, the symmetry breaking only occurs if there is racemization of the monomers. Racemization is the process by which one chirality is converted to the other in an attempt to make a racemic mixture.

In this thesis we wish to simulate reactions of nucleotide solutions in which template directed synthesis can occur and investigate properties and different types of symmetry breaking in models with different chirality and in the case of RNA sequences containing the canonical nucleobases.

1.5 Previous Models of Template-Directed Replication

Previous models have been used to determine how characteristics of the first polymers emerged, including a binary model produced by Tupper et al., 2017, a binary model made by Fellermann, Tanaka, and Rasmussen, 2017, and a dimer model done by Tkachenko and Maslov, 2017. These provided groundwork and motivation for the research within this thesis.

In the binary model by Tupper et al. it is shown that uniform chirality (D) can emerge due to prebiotic selection under the influence of three processes: hydrolysis,

11

polymerization, and template-directed ligation. In their model the only sequences that can be templated for and thus catalyzed are the uniform sequences. Uniform sequences meaning a sequence composed of monomers purely of one type of chirality, backbone, or sugar. Past certain ligation values a bifurcation occurs where one of the chiralities or bonds or sugars begins to dominate over the other type resulting in a symmetry breaking phase transition.

The binary model by Fellermann, Tanaka, and Rasmussen contains the same basic reactions as the binary model but instead of restricting catalyzation to uniform sequences they allow each sequence to act as a catalyst to its own formation. The second way the model differs from the Tupper et al. model is that the template must be the same size as the resultant polymer whereas Tupper et al., 2017 allow larger polymers to template for smaller ones as well, provided there is still a binding site. The final difference between these two models is that directionality of the strand does not matter in the model by Fellermann, Tanaka, and Rasmussen. These are subtle but important differences in the models as will be evident throughout this thesis. The binary model found interesting selections of sequences which favor alternating sequences above all else and certain families of sequences over others.

Finally, the dimer model is similar to the polymer model, complementarity and directionality are taken into account when undergoing the template-directed ligation. The difference between this model and the polymer model comes from the use of dimers as the base unit while Tupper et al., 2017 use monomer units. The model finds that most of the dimers become extinct due to being outcompeted by other dimers, reducing

12

information entropy for the sequences. Essentially there is a kind of symmetry breaking in sequence space for the dimer model.

Each of the models use the same processes that will be used in this work: hydrolysis, polymerization, and template-directed ligation. Furthermore, we will be taking the best parts of each of these models as the base model and expanding on them. We will incorporate templates for all sequences as in the Fellermann, Tanaka, and Rasmussen model but keep in mind the direction of the strands as in the Tkachenko and Maslov model while also allowing strands of a longer length to act as templates for shorter ones as in the Tupper et al. model. The model will be explained in detail in the following chapter.

1.6 Goals

It is difficult to experimentally determine if it is possible to build a ribozyme from nucleotides since the mechanisms and chemical reactions that resulted in oligomers long enough to act as ribozymes most likely took place over millions of years. This research does not aim to find the first ribozyme but instead looks at the beginning of building small oligomers. This research turns to simulations which allow for the varying of initial conditions and the evolution of the system over long periods of time to determine the outcome in a fraction of the time. The simulations focus on a few non-biotic processes, hydrolysis, polymerization, and template-directed ligation, and the effects they have on nucleotide sequences and lengths. These processes are chosen because they are mechanistically simple and are likely to have taken place in one form or another provided there is water present and require no specific circumstances. Using these simple mechanisms this thesis explores the following:

- The effects incurred by these processes on a system that contains only two types of monomer units, an L-chirality and a D-chirality monomer. In this model there occur self-complementary uniform sequences LLLL and DDDD, since L and D pair with themselves.
- Understand symmetry breaking transitions in templating reactions including chiral symmetry breaking but also symmetry breaking in sequence space. Not all sequences are equal due to the template-directed ligation increasing fitness of certain sequences and decreasing others.
- 3. The effects of varying chemical reaction rates and initial conditions on two of the canonical bases of RNA, (A)denine and (U)racil. In this model A pairs with U so AAAA would complement for UUUU while AUAU and UAUA would be self-complementary. Limiting the model to two bases makes for a relatively simple model without adding the degrees of freedom that come with having all four bases and allows the exploration of the properties/effects of complementarity, length, and sequence frequency.
- 4. Introducing (G)uanine and (C)ytosine to the adenine and uracil model to understand how the wobble base, uracil, influences the dynamics of the oligomers. In this model U can pair with A or G, and G can pair with U or C.

Chapter 2: Modeling Template-Directed Replication

2.1 Simulating the Formation and Breakup of RNA Strands

The models in this thesis are considering simulations of nucleotide solutions of several different types. There will be a chirality model where each monomer will be either an L or D monomer to represent their chirality and which will complement themselves. There will be a model that uses two nucleobases as the monomers, A and U to represent adenine and uracil, that complement for each other. Finally, there will be models that incorporate 3 or 4 nucleobases that include a wobble pair that can complement for two other monomers. Each of these models will be subject to the processes of hydrolysis, polymerization, and template-directed ligation as seen in figure 2.1. A final process called interchange will also be used which varies slightly between models. Each of these processes is outlined below.



Figure 2.1 The 3 basic reactions present in each model: Hydrolysis, Polymerization and Template-directed Ligation. The chirality model is represented here with each box representing an L- or D-monomer.

2.1.1 Hydrolysis

Hydrolysis occurs when a water molecule cleaves the phosphodiester bond between two nucleotides via a hydration reaction, resulting in one oligomer being split into two. Hydrolysis generally occurs when RNA is in a single-stranded and unfolded state. This model does not address folding since it will only be dealing with relatively short oligomers and the RNA-duplex formation and breaking is handled in a single step so only one rate of hydrolysis will be used. The rate of hydrolysis used in these models scales linearly with the length of the oligomer (i.e. an oligomer of length six will be twice as likely to undergo hydrolysis as an oligomer of length three). The types of monomers on either side of the bond have no impact on its chance of breaking. Water was likely necessary for the transportation of the chemicals needed to create life and required in many of the chemical reactions. Due to this and the fact that Earth was almost entirely covered in water 3.9-4.4 billion years ago save for sparse islands life undoubtedly developed in a wet environment. This justifies the hydrolysis reaction used in the models in this thesis.

2.1.2 Polymerization

Polymerization is the mechanism responsible for joining monomers and oligomers together. This mechanism acts in the exact opposite way of hydrolysis as can be seen in figure 2.1. Unlike the hydrolysis reaction, in this model, the polymerization rate does not scale with any characteristics like length because a monomer unit cannot be attached to multiple chains of oligomers. There may be some scaling with length due to diffusion scaling with length, but that scaling makes no appearance in this model. There is no folding in the model nor is it taken into account that it might be more difficult to make a bond between different types of monomers. Thus the rate is constant for any strands being polymerized with the exception of template-directed ligation as explained below.

2.1.3 Template-Directed Ligation

Template-directed ligation is a similar process to polymerization with the exception that there is a third oligomer involved to act as the template. The binding of each oligomer to the template and unbinding all happen instantaneously in the models, if the strands met the criteria, that they were complementary, then the result is 2 strands or it fails and the 3 strands are unchanged. Since it is modeled as an instantaneous reaction it overlooks binding and stacking energies associated with forming the hydrogen bonds and does not consider the on/off rate of the oligomers in forming the helix. These factors are absorbed into one rate, which is the ligation rate. It is assumed that all of the oligomers have and will form 3' and 5' phosphodiester bonds and thus the strands will consider directionality and run 3'-5' with complements going 5'-3' or vice versa.

2.1.4 Interchange

The final parameter will be a process called interchange which is a generalization of racemization for the L and D chiral models that we can apply to the canonical bases models. The process converts one monomer to another in an effort to create a racemic mixture, in the chiral case this occurs between L and D. When moving into the models with the nucleobases the interchange mechanism isn't the same, but we could imagine some process that would cause a nucleobase to be replaced by another. The interchange allows for competition over the resources (monomer units in these models) since on a prebiotic Earth one might imagine there wouldn't be an endless supply of nucleotides. There is no interchange occurring in the oligomers, only in the monomer units.

2.2 Deterministic Model

The primary model in this research is the deterministic model, which incorporates each of the parameters discussed above to create a system of equations that govern the simulation. The equations incorporate the hydrolysis rate (k_H) , polymerization rate (k_P) , ligation rate (k_L) , and interchange rate (k_{int}) , as well as the concentration of sequences (C_i) , and the time step (dt). The time step, whose criteria will be discussed later on, is chosen to be very small to simulate a continuous reaction in this well-mixed case. Each of the reactions is handled as outlined below.

For the hydrolysis reaction consider a sequence k, which has length L_k nucleotides. The net rate of change in concentration of sequence k by Hydrolysis is:

$$R_{k}^{hyd} = k_{H}\left(\sum_{j} (C_{j,k} + C_{k,j}) - (L_{k} - 1)C_{k}\right)$$
(1)

Here *j.k* and *k.j* denote sequences that are concatenations of sequences *j* and *k* with *k* at either the 5' or 3' end. The sum is over all possible subsequences *j*. Equation 1 says that *k* is produced by hydrolysis of any of the concatenated sequences, and that there are (L_k-1) bonds in sequence *k* that can be hydrolyzed.

The net rate of change concentration of sequence k by Polymerization is

$$R_k^{pol} = k_P \left(\sum_{i:j=k} C_i C_j - \sum_j C_j C_k - \sum_j C_k C_j \right)$$
(2)

Here the first sum is over all possible pairs of sequences *i* and *j* whose concatenation is *i*.*j* = k.

The net rate of change concentration of sequence *k* by Ligation is

$$R_{k}^{lig} = k_{L} \left(\sum_{i,j=k} C_{i}C_{j}C_{i,j}^{temp} - \sum_{j} C_{j}C_{k}C_{j,k}^{temp} - \sum_{j} C_{k}C_{j}C_{k,j}^{temp} \right)$$
(3)

where $C_{i,j}^{temp}$ is the concentration of strands that are templates for *i.j.* Templates are explained in detail in section 2.3.

The overall rate equation is

$$\frac{dC_k}{dt} = R_k^{hyd} + R_k^{pol} + R_k^{lig} \tag{4}$$

If k is a monomer, then we also need to account for interchange between the types of monomers. We assume a rate of interchange k_{int} from each monomer to each other type. If there are n kinds of monomer, we have

$$R_k^{int} = k_{int}(C_{mon} - nC_k) \tag{5}$$

This extra term needs to be added to $\frac{dC_k}{dt}$ when k is a monomer.

2.3 Methods for Simulating Systems of Chemical Reactions

The simulation was done by solving the rate equations for every sequence up to the max length at each time step and updating the concentrations of the sequences based on the solution to equation (4). The total concentrations, which vary from 10 to 80 in our models, are usually split between the two monomers to initialize the simulation but we explore other initializations as well. The time steps are a microsecond (10⁻⁶) but this is equivalent to the hydrolysis rate (k_H) which we set to unity, therefore we can think of all other rates as being scaled to the hydrolysis rate and a single time step is on the scale of a hydrolysis reaction. The simulation is complete when the stopping criteria is met. The stopping criteria is checked after each time step and met when each of the concentrations have not changed by more than 10^{-10} over the past 100 000 time steps, at which point the system is deemed to be at equilibrium.

A length limit is introduced in an effort to reduce computation time since the number of sequences at length n is 2^n and understand the reactions prior to sequences folding. Once the sequences start folding we lose the template reaction and hydrolysis is less consistent along the entire sequence. In all cases this limit is set at 4 or 6, meaning any sequences longer than that do not occur. The two reactants drop in concentration as the product increase in concentration, each by the same amount.

The values of the rates: hydrolysis (k_H) , polymerization (k_P) , ligation (k_L) , and interchange (k_{int}) are not well known and the rates incorporate many steps (like binding and unbinding for ligation) so instead this model uses scaled rates to understand what happens in cases where the rates are faster or slower than others. In this model the interchange rate (k_{int}) , hydrolysis rate (k_H) , and polymerization rate (k_P) are set to unity and the ligation rate (k_L) is varied. The natural speeds of the reactions are also dependent on the total concentration since the polymerization scales as $[c]^2$ while the ligation scales as $[c]^3$ thus if the total concentration is low the ligation will be slower and if they are high then the ligation will be much faster. Increasing the concentration will have similar effects as increasing the ligation rate but they are different in that the concentration will affect the other terms as well just to a lesser degree since hydrolysis~[c], polymerization~[c]² and ligation~[c]³ whereas increasing the ligation rate only effects the ligation.

Model Name	Which motifs can be	Catalytic rate depends on
	catalysts?	the concentration of
LD-UT	Only uniform L or uniform	Template strands
	D motifs	
LD-UM	Only uniform L or uniform	Motifs in the templates
	D motifs	
LD-AT	All LD motifs	Template strands
LD-AM	All LD motifs	Motifs in the templates
AU-T	All AU motifs	Template strands
AU-M	All AU motifs	Motifs in the templates

Table 2.1 Model abbreviations and their differences to simplify referencing which model we are discussing in later chapters.

2.4 Definition of Templates and Motifs

In this thesis we study several models that differ in the way that templates are defined. These models are listed in Table 2.1 The abbreviation LD denotes models for chirality, with two kinds of monomers L and D that pair with themselves. The abbreviation AU denotes models for complementary sequences, with two nucleotides A and U that pair with each other.

The model studied by Tupper et al., 2017 is denoted LD-UT in Table 2.1. In this model the 'U' denotes that both *i* and *j* must be uniform oligomers (all L or all D) of the same kind, and hence t(i,j) is also uniform, see figure 2.2. For example, if i = LL and j = L, then i,j = LLL, and t(i,j) is also LLL. Example sequences that are templates for LLL include LLL, LLLLL, DLLLDD etc. In Model LD-UT all the templates are weighted equally.



Figure 2.2 Template-directed ligation for the uniform LD model.

For two oligomers *i* and *j* that are ligated to form *i.j* we can define a template motif t(i,j) that forms a double strand with *i.j* and acts as a template. We begin by defining model LD-AT. The 'A' denotes that all sequences have a template motif, and the 'T' denotes that all template strands contribute equally to catalysis. For example, if *i* = LL and *j* = D, then *i.j* = LLD. As L and D pair with themselves, t(i,j) is DLL. We assume that the two strands pair in opposite directions (as in Fig 2.3), hence in this model, the template t(i,j) is always the reverse of *i.j*. Any sequence *k* is a template for formation of *i.j* if it contains t(i,j) at least once. For example, DLL, DDLLL, DLLDLL, *etc.* all contain DLL. In this model, all template strands are assumed to be equally good catalysts. Hence, in equation (3), $C_{i,j}^{temp} = \sum_k C_k$, where the sum is over all sequences *k* that contain the motif t(i,j).



Figure 2.3 The change in the ligation process from last model to this model, it is important to realize that the direction of the strand did not matter in the previous model but in this model it does due to being able to template for any strand.

Model LD-UM and LD-AM differs from LD-UT and LD-AT respectively in that each motif in the template strand is weighted equally, rather than each template strand being weighted equally. For example, in the LD-AM case DLLDLL contains the motif DLL twice and in the LD-UM case LLL appears three times in LLLLL, but only once in LLL and DLLLDD. The total motif concentration is

$$C_{i,j}^{mot} = \sum_{k} N_{t(i,j)}^{k} C_{k}$$

where $N_{t(i,j)}^{k}$ is the number of times that the template motif appears in sequence k. For model LD-AM, the following equation replaces equation 3

$$R_{k}^{lig} = k_{L} \left(\sum_{i,j=k} C_{i}C_{j}C_{i,j}^{mot} - \sum_{j} C_{j}C_{k} C_{j,k}^{mot} - \sum_{j} C_{k}C_{j} C_{k,j}^{mot} \right)$$
(6)

For the AU models, the template motif is the reverse complement, not simply the

reverse sequence (see figure 2.4). For example, i = AA and j = U, then i.j = AAU, t(i,j) is AUU. For Model AU-T, all template strands are weighted equally, whereas for more AU-M, strands are weighted in proportion to the number of occurrences of the motif.



Figure 2.4 The change in the ligation process from the LD model to this AU model. Similar to the nonuniform LD model the direction of the strand will play a large role in determining templates and autocatalytic sequences/sets. The AU model introduces opposite complementary pairing, resulting in a largely different model.

Chapter 3: LD Models

3.1 Chirality Models

The LD models will begin by confirming the results of Tupper et al., 2017 in the LD-UT model and move on to the other chirality models: LD-UM where the number of motifs are taken into account, LD-AT where the templating reaction changes so that any sequence can be templated for and act as a template, and LD-AM where the number of motifs are taken into account in the LD-AT model. We will discuss chiral symmetry breaking and then move onto symmetry breaking in sequence space. Each of the models in this chapter and others only change by manipulating how the template-directed ligation works as discussed in section 2.3.

3.2 Chiral Symmetry Breaking

We used the deterministic method of section 2.2 to simulate the LD models starting from concentrations of 5.1 for L and 4.9 for D monomers (total concentration of 10) and with a max length of 6. We begin with the case where only uniform strands can be templates, the LD-UT model. The best way to understand if a single chirality is chosen over the other is through the enantiomeric excess defined as

$$EE_L = \frac{L-D}{L+D} = -EE_D \tag{7}$$

the difference in the amount of monomers, contained within both monomers and polymers, divided by the total number where $EE_{L or D} = 1$ represents completely chiral

sequences. Therefore, a large and consistent value for this signifies a chiral symmetry breaking.



Figure 3.1 Bifurcation of L- and D-monomers in the LD-UT model that initially start at 51:49 split occurring at $k_L = 4.6$ where $k_H = k_P = 1$ and the max length was set to 6. Enantiomeric excess quickly rises representing a phase transition to one chirality.

As can be seen in figure 3.1 a symmetry breaking occurs in line with that of Tupper et al., 2017. At a ligation rate beyond the bifurcation at $k_L = 4.6$, there are two solutions with positive or negative EE, according to whether D or L becomes the dominant chirality. Note that there is nothing in the model that favors one over the other. The prediction of this theory (and of other models that involve asymmetric autocatalysis) is that the 50:50 mixture is unstable and that either one or the other chirality will dominate with equal probability. The initial 51:49 ratio is to counteract the computational precision and introduce a bias that represents a fluctuation.

Next we look at the LD-UM model which simply adds motifs to the LD-UT
model. In the previous model we observed chiral symmetry breaking at a value of $k_L = 4.6$ where $k_H = k_P = 1$. In that model each of the templates were equally good whereas one might expect in reality a fully uniform hexamer would most likely be a better template than a dimer for joining two monomers, there are more places to bind and the molecule itself is larger increasing the chance of collision. It is at this point we introduce motifs, as mentioned in chapter 2. Each templates efficacy on the ligation rate now scales linearly with the unique number of motifs and we use equation 6 with $C_{i,j}^{mot}$ instead of equation 3 with $C_{i,j}^{temp}$. Under these conditions we have now essentially made each of the templates better or at least as good as before, this results in a lower ligation rate at which bifurcation occurs. As can be seen in figure 3.2 the bifurcation is lowered to a ligation value of 2.1 and the bifurcation changes to a smoother curve.



Figure 3.2 Bifurcation of L- and D-monomers that initially start with a 51:49 split occurring at $k_L = 2.1$ where $k_H = k_P = 1$ and the max length was set to 6 in the LD-UM model.

The chiral models previously mentioned are easily the best case scenario for chirality to emerge; specifically, because only uniform strands are acting as templates and inhibition in not being accounted for in the motifs model. However, life is messy, and having some monomers in a sequence of opposing chirality may not entirely discourage templating. In this LD-AT model we explore the worst-case scenario where anything can template provided the other strand is complementary i.e. an L-monomer matches with an L-monomer and a D with a D taking into account direction of each strand. This is the worst-case scenario since the structural differences in the monomers make it more difficult for the templating reaction to occur since it would change how the monomers orient and align with each other for bonding. This is the first major change in the template directed ligation step.

There is still a bifurcation around a ligation value of 40.4 for this model (see figure 3.3), a value nearly 10 times that of the previous model with no motifs. The bifurcation occurred for a starting configuration with a ratio of 90:10 for the monomers and was absent at this value for the normal ratio of 51:49 (after running for 100 times longer than normal) which suggests that for the bifurcation to occur at this value a large perturbation is required in one direction. The bifurcation for the starting configuration with a ratio of 51:49 occurred at a ligation value of 90 which means it is still possible to have this bifurcation but it



Figure 3.3 Bifurcation of L- and D-monomers that initially start with a 90:10 split occurring at $k_L = 40.4$ where $k_H = k_P = 1$ and the max length was set to 6 using the LD-AT model.

may require either a large perturbation or a large ligation value or both. The importance of the fact that the monomers complement themselves cannot be overstated. Without this fact the perturbations would not be able to build and grow to cause the bifurcation as it provides a very powerful feedback loop. The final model we look at is the LD-AM model which is simply the previous model but incorporating motifs. Just as in the case of the uniform models, the bifurcation occurred at a lower ligation value when we applied motifs to this non-uniform model. As can be seen if figure 3.4, a bifurcation occurs at a ligation value around 8.4 for this model, a significantly lower value from the LD-AT model and from a 51:49 staring ratio of the monomers as opposed to the 90:10 required in the LD-AT model.



Figure 3.4 LD-AM model with a 51:49 split in the monomers and a max length of 6. There is a bifurcation occurring at a ligation value of 8.4.

3.3 Templating and Sequences

These bifurcations are the result of feedback loops created by the ligation term. In the case of the LD-UT model only uniform products can be formed. That means under a max length of 6 there are only 5 oligomers that can be templated for per chirality, which is easily enumerated in the table below.

		TEMPLATES							
Р		Dimer	Trimer	Tetramer	Pentamer	Hexamer	Total		
R	Dimer	1	3	8	19	43	74		
O D U C T	Trimer	0	1	3	8	20	32		
	Tetramer	0	0	1	3	8	12		
	Pentamer	0	0	0	1	3	4		
	Hexamer	0	0	0	0	1	1		

Table 3.1 The number of unique sequences that provide templates for the production of uniform oligomers. The total number of templates have been broken down into the oligomers that provide the templates.

A pattern can clearly be seen in the table. For the templating of uniform oligomers

we can determine the number of templates from the following formulas,

$$T_{i} = 1 + \sum_{n=i+1}^{n_{max}} S_{n}$$
 (8)

$$S_n = 2^{n-i-1} - S_{n-i-1} + 2S_{n-1}$$
(9)

where *i* is the product length, *n* is the template length and therefore T_i is the total number of templates for length *i* products and S_n is the number of templates of length *n*. It is fairly intuitive to derive this when we consider the unique oligomer combinations. Moving from one oligomer length to the next, we can account for all of the unique combinations by simply adding an L-monomer to each of the previous oligomers combinations and then doing the same with a D-monomer, resulting in double the unique combinations of the previous length. Therefore, we will have at least twice as many of the templates we had at the previous length which results in the $2S_{n-1}$ term. The other length of sequences we care about is *i* times ago, since the only way to get new sequences that contain uniform sections is by adding that letter to the start *i* times in a row. Obviously, for example in the case of templating a trimer, after two additions of L-monomers all of those will be new templates so the number of trimers we had will be templates by the time we build up to pentamers. Knowing that the number of *n*-mers is simply 2^n there are 2^{n-i} possible new templates for a uniform product of length *i* from oligomers of length *n*. However, we are only interested in the half that start with the opposing monomer since the ones that already start with the correct monomer will be accounted for in the $2S_{n-1}$ term since they should become a template at the previous length hence $\frac{2^{n-i}}{2}$ or 2^{n-i-1} . We also care about the sequences of the length below since any sequence that was a template at that length will be part of the half that we assumed will be new templates due to their starting with the opposing monomer. No new templates can form from the addition of the opposing monomer but old ones still work which results in the subtraction of S_{n-i-1} to prevent double counting.

We see from the tables that there are many templates coming from each length, it is important to not let this mislead you as the concentrations of each of the templates matters as well. As a general result for these models, an increase in the template-directed ligation rate causes a shift in the distribution of lengths towards longer oligomers. At low ligation rates most of the monomers will be in the dimers and trimers while at higher ligation rates there will be more monomers in the pentamers and hexamers. As well as changing length distributions the sequences associated in each length are not all equal in concentration due to the ligation term. If ligation was absent then all sequences would be equal in concentration due to random polymerization. By introducing the templatedirected ligation for uniform strands there is a bias to increase the concentration of strands that contain uniform sections. This is evident in the figure below where we see strands of higher uniformity at higher concentration while alternating strands have the lowest concentration.



Figure 3.5 Sequences of tetramers over a range of ligation rates where $k_H = k_P = 1$ and the max length was set to 6 in the LD-UT model. Multiple splitting of sequence symmetries occurring at the bifurcation point of $k_L = 4.6$, in this case L is the dominant chirality. Legend is in order of decreasing concentration after the bifurcation.

The graph shows that at a ligation of zero all of the sequences are equally abundant and therefore the only factor creating the sequence splitting comes from the ligation. As ligation increases the splitting of sequence space becomes more and more pronounced. Between a ligation value of 0 and 4.6, at which the bifurcation occurs, the more uniform sequences are favored which puts the completely uniform sequences at the highest concentration and the alternating sequences at the lowest concentration. We can see there are 6 unique concentrations for the 16 sequences which is due to symmetries within sequence space. The first obvious symmetry can be seen in the legend of the figure where we simply interchange L and D monomers for each of the sequences bringing it down to 8 unique concentrations, call this chiral sequence symmetry. The other sequence symmetries come from reading the sequences backwards as in the case of LLDL and LDLL and finally backwards and chirally as in the case of LDDD and LLLD, bringing unique concentrations down to 6. Once the bifurcation point is reached the chiral sequence symmetry is broken and the 6 unique concentrations are split into 10. This is best illustrated with an example, take the uniform tetramers, prior to the bifurcation point they are exactly equal in concentration but after one emerges the other concentration rapidly drops. If it is the chiral symmetry being broken one might expect there to be 12 unique concentrations, however, the alternating sequences are equally bad, due to containing no uniform sections and the LLDD-DDLL pair are equally good as they contain the same uniform sections.

In the case of the LD-UM model where we have included motifs we can easily figure out the number of templates in a manner similar to equation 9, except we do not have to worry about double counting and instead are left with the simple form of

$$S_n = (n - i + 1)2^{n - i} \tag{10}$$

The factor in front comes from the fact that you double the amount of templates you had previously since you simply add a unique new motif to each of the sequences that were already templates. The exponential term comes from doubling the templates as well, but this time it accounts for the templates created from adding the letter *i* times since after *i* times everything is guaranteed to be a template and since we do not care about double counting we do not need to subtract a term like previously.

	TEMPLATES								
P R O D U C T		Dimer	Trimer	Tetramer	Pentamer	Hexamer	Total		
	Dimer	1	4	12	32	80	129		
	Trimer	0	1	4	12	32	49		
	Tetramer	0	0	1	4	12	17		
	Pentamer	0	0	0	1	4	5		
	Hexamer	0	0	0	0	1	1		

Table 3.2 The number of motifs that provide templates for the production of uniform oligomers. The total number of motifs have been broken down into the oligomers that provide the motifs.

The sequences follow the same patterns we saw in figure 3.5 for the chirally uniform product model with the exception that the separation is slightly larger for the uniform strands in this case and the splitting occurs at a lower bifurcation value.

In the LD-AT model there is no general formula for calculating the templates for any given sequence due to the multitude of non-uniform patterns, the reason it was possible before was due to the uniformity of the products resulting in it only being necessary to track how many monomer units had been added to a sequence and not the sequence itself.

Table 3.3 The number of templates for specific sequences of oligomers. The total number of oligomers have been broken down into the oligomers that provide the templates.

	TEMPLATES									
Р		Dimer	Trimer	Tetramer	Pentamer	Hexamer	Total			
R	LL	1	3	8	19	43	74			
0	LD	1	4	11	26	57	99			
D	DL	1	4	11	26	57	99			
U	DD	1	3	8	19	43	74			
C	LLL	0	1	3	8	20	32			
I	LLD	0	1	4	12	31	48			
	LDL	0	1	4	11	27	43			
	LDD	0	1	4	12	31	48			
	DLL	0	1	4	12	31	48			
	DLD	0	1	4	11	27	43			

M.Sc. Thesis – E. Turner; McMaster University – Physics and Astronomy

DDL	0	1	4	12	31	48
DDD	0	1	3	8	20	32

As can be seen in the table above the uniform strands no longer have the most templates when compared to the others. In fact, the uniform strands boast the lowest number of templates, even lower than the other self-complementary alternating strands. The introduction of the ability for mixed sequences to be templated resulted in some nonuniform sequences being self-complementary, for example LDL. However, the uniform strands have the advantage of being self-complimentary at every length while the other self-complimentary strands like LDL are only self-complimentary at their own length. This is why past a ligation value of 40.4 the uniform strands dominate, as can be seen in figure 3.6. Interestingly, at ligation values lower than the bifurcation value the higher concentration sequences are those of the alternating sequences and the lowest are those of the uniform sequences. This is not intuitive since as we have seen the alternating sequences have neither the most templates nor are they the most autocatalytic yet they hold the highest concentration prior to bifurcation. The reason for this is most likely due to the higher concentration of dimers and the large number of templates for the mixed dimers which results in polymerization of more alternating strands. This shows that the system is complex and doesn't simply rely on the templates to determine the dominant sequences, however they are a good indicator.



Figure 3.6 Sequences of tetramers over a range of ligation rates where $k_H = k_P = 1$ and the max length was set to 6 in the LD-AT model. Multiple splitting of sequence symmetries occurring at the bifurcation point of $k_L = 40.4$, in this case L is the dominant chirality. Legend is in order of decreasing concentration after the bifurcation.

Finally, for each of the sequences of a given length in the LD-AM there are the exact same number of templates available. The number of templates can be determined by equation 10 just as in the case with uniform model with motifs (clearly, since the uniform strands have the same templates as all the others at that length). This produces an interesting result where the sequences have the exact same concentration prior to the bifurcation point seen in figure 3.7 but then split after that point to reflect the chiral symmetry breaking.



Figure 3.7 Sequences of tetramers over a range of ligation rates where $k_H = k_P = 1$ and the max length was set to 6 in the LD-AM model. Multiple splitting of sequence symmetries occurring at the bifurcation point of $k_L = 8.4$, in this case L is the dominant chirality. Legend is in order of decreasing concentration after the bifurcation.

While there has been a lot of focus on ligation values and its ratio to the hydration and polymerization rate we have neglected the interchange rate which also has some effect on the dynamics. Increasing the interchange rate causes a more rapid change from one monomer to the other, what this means is that when one of the chiralities gains an advantage over the other by having more templates it will use up the shared resources of the monomers but only of its own type and that resource will be replenished much more quickly with higher interchange rates. The overall equilibrium concentrations do not change however, as can be seen in figure 3.8 below, the speed with which they reach equilibrium is faster for the larger interchange. For the remainder of the models a k_R of 10 is used to keep consistent while speeding up the models that compete over resources. It is important to note here that the concentration of free monomers of L and D is always equal at equilibrium while the concentration of L and D monomers in oligomers is not after bifurcation occurs. Therefore, there can be an enantiomeric excess in the oligomers but not in the monomers. The importance of the interchange process is highlighted here because without it and starting with 50:50 there will never be an enantiomeric excess. The presence of the interchange reaction allows the instability to occur and the enantiomeric excess to arise in the oligomers but is also responsible for the equal number of monomers since they directly influence each other and the process acts to balance the two.



Figure 3.8 The fraction of oligomers that are mixed or uniform in either chirality over time for two different values of interchange with $k_L = 4.6$, and $k_H = k_P = 1.0$.

3.4 Alternate Equilibria

In the LD-AT and LD-AM models an interesting difference comes about in the feedback due to the change in ligation. Previously each of the templates was selfcomplementary since reading 3' to 5' or 5' to 3' doesn't matter for a uniform strand as they will always be matching but by introducing templating that is no longer uniform this property is broken for many of the other sequences. Furthermore, there are some sequences that, similar to the uniform sequences, are self-complementary; for example, the trimer 3'-LDL-5' or 3'-DLD-5'. In this model the self-complementary sequences are those that are symmetric about their middle, which includes the uniform strands. This does not end up being the case for models outside of this chapter since in this case monomers of the same type complement each other whereas each of the canonical bases, which are used as the monomer units in Chapters 4 and 5, do not complement themselves. The differences between self-complimentary feedback and non-self-complimentary feedback is most easily shown through a diagram linking the templates. In the diagram, we can easily see uniform sequences template themselves and their "food" while the nonuniform strands cannot catalyze their own food and instead form an autocatalytic set with their compliments. More interesting is the non-uniform autocatalytic sequences, in the diagram we take the alternating sequences LDL and DLD as our example but there are many others such as LDDL, DLLD, LLDLL, LDLDL, etc. but each of these templating diagrams look fairly similar, characterized by a mix of self-complementary and non-selfcomplementary sequences in the reaction family. Extending what we see in figure 3.7(c)to the pentamer LDLDL that becomes self-complementary again but at the next length we

M.Sc. Thesis – E. Turner; McMaster University – Physics and Astronomy

find LDLDLD which again is no longer self-complementary. Each of the families of selfcomplementary sequences do not follow the same pattern of complementary followed by non-complementary as the length increases like we see in the alternating family. This is easily shown by considering LDDDDL.



Figure 3.9 Three different types of reaction families: (a) Fully autocatalytic uniform sequences, (b) and (c) Autocatalytic set containing some autocatalytic sequences.

We have established that given an unremarkable starting configuration, starting with nearly even numbers of monomers, the final results for each of the models tend to reach the same conclusion, one chirality is chosen over the other. However, if we look at more exotic starting configurations to the models there are a variety of different equilibrium values occurring. These alternative equilibria are the results of the selfcomplementary sequences discussed earlier that contain sequences in their family that aren't all self-complementary as is the case with the uniform sequences. If there is a high enough fluctuation into one of these sequences it could be enough to trap the system in that state. The trapping is due to the hydrolysis being unable to keep up with the rate of ligation and polymerization and therefore unable to free the monomers to be used in the usual dominant templates. i.e. uniform templates.

We use the LD-AM model (making all sequences have equal templates) which will encourage trapping of secondary states. There will also be a cut off to a max length of 4 in hopes of simplifying the reaction families. The monomer concentration was also increased to 80, not only to increase the chance of finding these other states but because the situations in which these states might arise would likely be in pockets of high concentration, whether it be due to spatial effects, vesicles or other effects. To begin, the most likely sequences to be trapped as alternate equilibria state would be the selfcomplementary sequences. In these scenarios our starting configuration has all of the oligomers in whichever sequence we are trying to trap. As can be seen in figures 3.10 and 3.11 below, each of the self-complementary tetramers (LDDL, DLLD, LLLL and DDDD) reach an equilibrium where they are the dominant sequence so long as they are beyond a certain ligation value. In figure 3.10 the starting configuration is $C_{LDDL} = 10.1$, $C_{DLLD} = 9.9$ and there are a couple transitions based on the ligation value. First at low ligation values there is the completely symmetric phase where all sequences are at the same concentration because of the motifs. Following that at a ligation value of 2-4 the uniform sequence dominates, in this case DDDD since LDDL is slightly higher than



Figure 3.10 LD-AM model starting with $C_{LDDL} = 10.1$, $C_{DLLD} = 9.9$. Up to a ligation value of 4 the uniform sequence DDDD dominates. For a ligation value of 5 and 6 the alternating sequence dominates and beyond a ligation value of 6 LDDL dominates.

DLLD thus more DD are made than LL and sways it in that direction for the symmetry breaking. Next, we see a range of ligation values from 5-6 where the alternating sequences are the dominant sequences just as they were in the LD-AT model (figure 3.6) prior to bifurcation. Finally, the LDDL or DLLD (depending on which has the higher starting concentration) dominate at higher ligation rates. If we only begin with one of LDDL or DLLD there is no intermediate state where the alternating sequences dominate and instead it transitions sooner. The other chirally symmetric pair, LLLL and DDDD, which we saw dominate in the bifurcation graphs are still able to dominate under these new conditions (see figure 3.11) showing it is not simply the shift in environmental conditions that has caused the LDDL or DLLD sequence to dominate. Given a large

enough fluctuation one will emerge over the other for a range of ligation values. We can see there are other states present depending on the ligation value in the case of figure 3.10 which implies that the rate of ligation will matter when determining what can dominate.



Figure 3.11 LD-AM model starting with $C_{LLLL} = 12$, $C_{DDDD} = 8$. Beyond a ligation value of 2 a single uniform sequence as we found to be the case when we were looking at bifurcation earlier.

Furthermore, it is possible to trap certain autocatalytic sets provided that ligation is high enough. In figure 3.12 the starting concentration is entirely of LLDL and by a ligation value of 11 it and its complement LDLL are the dominant sequences. These sets require a higher ligation rate than those that contain self-complementary sequences. In this scenario, prior to ligation rates high enough to trap the system, the dominant state was the uniform strands which is in contrast to what was found for the LDDL+DLLD starting configuration in which there was a state where alternating strands dominated. Obviously, just as every other model in this section, it has to do with the starting configurations of each and what sequences they can template. In the LDDL+DLLD configuration the tetramers both act as a template to the formation of LD, DL while only 1 of them acts as a template for DD or LL at the dimer level. LLDL acts as a template to LD, DL, and LL at the dimer level. In the LDDL+DLLD scenario there are more templates for the alternating dimers than uniform ones so the alternating sequences build up whereas in the LLDL scenario it equally templates for the uniform dimer and each of the alternating ones, and given that the uniform sequences are better autocatalytically they dominate. These properties extend to larger max lengths as well, as we increase the max length to 6 we can then predict the states that can form by determining the self-



Figure 3.12 LD-AM model starting with $C_{LLDL} = 20$. At a ligation value below 10 the uniform strand dominates, specifically LLLL due to the larger number of L monomers in LLDL. Above a ligation value of 10, LLDL and LDLL co-dominate the system.

complementary sequences. While there are many of these states they are not expected to emerge in a more realistic scenario since there would be selection pressure that isn't accounted for in this model. For instance, the key to the templating is to hold the reactants in place and promote formation of its own sequence but it is difficult if molecules are skewed in different directions because of their chirality and aren't in the correct position to bond.

3.5 Chirality Conclusions

The chirality models above showed clearly that templating could have a vital role in the emergence of a single chirality used in all of life. Beyond confirming the results found in Tupper et al., 2017 we found that the bifurcation occurred in all 4 models tested, even in the worst case scenario, the LD-AT model, which do not give uniform sequences the benefits that being uniformly chiral should like in the LD-U(T/M) models. There is undoubtedly a large selection pressure on the chirality provided by the templating reaction. The templating reaction can also result in the emergence of uniform bases and backbones as is shown in the Tupper et al., 2017 paper where they use a model that operates in the exact same way as this LD model except instead we switch the L and D for an X and R in the case of a base or differentiate by bonds in the case of a backbone. All of the results above can therefore be applied to those models as well. We found that there is no sequence splitting at low ligation rates for the LD-AM model while there is splitting in the LD-AT model. Furthermore, it is possible to reach states other than those

M.Sc. Thesis – E. Turner; McMaster University – Physics and Astronomy

dominated by purely uniform sequences if we start with a sufficient concentration of certain sequences.

Chapter 4: AU Model

4.1 AU Nucleobase Model

Last chapter we established that the templating reaction can account for the emergence of uniform chirality. In this model we push further to see if templating causes any symmetry breaking to occur in sequence space once a uniform chirality, base and backbone have been chosen. To begin there will only be a pair of complementary bases, A(denine) and U(racil), and in the next chapter will move to including 3 and then 4 canonical bases. Using only 2 bases first allows for a simpler model to analyze and provides a base level of understanding for when we move to the 3 and 4 letter alphabets, it also saves the trouble of dealing with the wobble of uracil which can bond to adenine or guanine, not following Watson-Crick base pair rules.

An immediate and clear result of having the monomers complement their opposites instead of themselves as in the LD model is that the system becomes much more stable to fluctuations in the ratio of monomers. As was mentioned numerous times in the previous models, the key to the symmetry breaking was that monomers were selfcomplementary so that when one chirality gained an advantage it would build on that advantage because it would use the same monomer type to build larger oligomers putting that monomer at a deficit and thus causing the interchange to refill those lost with the other monomer. In this model we allowed interchange between A and U in the same way as we allowed interchange between L and D in the LD model. However, there is no instability between A and U in this case because if a temporary excess of A arises in the oligomers, this will synthesize more oligomers with U. Thus there is a negative feedback that prevents differences in A and U building up, whereas there is a positive feedback in the LD case that causes the chiral instability. Even though there is no symmetry breaking between the total A and U concentration, we show below that there are many other kinds of instability that arise in the AU model. These cause symmetry breaking in the sequence space-i.e. different sequence of the same length have different concentrations.

Another change due to the alteration to what makes a sequence complementary is obviously how we look for self-complementary sequences. In previous models the uniform sequences gained huge advantages because each of the products were selfcomplementary but now a uniform strand would act as a template for the production of the opposite uniform strand simply acting as a catalytic set (see figure 4.1(a)). Our rules for complementary strands in the LD model were that they need to be symmetric across the middle of the sequence but this will not work for the AU model since if it were symmetric the ends of the sequence would be the same and hence the template would have the opposites on both ends. In fact, for this model the first and last monomer in the sequence need to be opposite for them to be a possible template. There is no simple rule like what we found in the LD model where we look for symmetry, or at least not a simple symmetry. In this case the sequences need to be chirally symmetric across the middle of the molecule. For example, if the first half of the molecule is AUU then the second half would need to be AAU to have it be self-complementary. It is not as simple to find these complements as it was in previous models but it is a helpful tool nonetheless. For sequences of *n*-mers there are unique self-complementary sequences equal to the number of unique $\frac{n}{2}$ -mer sequences provided that n is even. This is evident in table 4.1 where

there are 4 dimers and 4 self-complementary tetramers and 8 trimers and 8 selfcomplementary hexamers. This extends to higher length sequences as well due to the sequence being chirally symmetric across the center, therefore there can only be a number of unique halves equal to the number of unique $\frac{n}{2}$ -mer sequences. The same rule applies for the LD model but the symmetry across the center is different.



Figure 4.1 Two different types of reaction families: (a) Autocatalytic set containing no autocatalytic sequence, (b) Autocatalytic set containing some autocatalytic sequences (c) Autocatalytic set requiring only a single tetramer. No completely autocatalytic reaction families exist like those we saw in the LD model with the uniform strands.

Figure 4.1 highlights the reaction families similar to the examples in the LD model except that there are now only two reaction families. 4.1(a) is simply an autocatalytic set with no autocatalytic sequences while 4.1(b) is similar to 3.9(c) which are both autocatalytic sets containing some autocatalytic sequences. You will notice that there is no longer a fully autocatalytic family containing all autocatalytic sequences like 3.9(a) (the uniform sequences in the LD model) and that is due to there being no autocatalytic sequences in odd length oligomers like trimers or pentamers, thus 4.1(b) is as autocatalytic as is possible in this model. 4.1(c) also contains autocatalytic sequences but contains less than that of 4.1(b) at the dimer level and when building up from monomers the autocatalytic dimers have the highest influence on dominant sequences. The reason for this lack of autocatalytic sequences is due to the rule discussed above in which it states that the sequences need to be chirally symmetric. In an odd length oligomer there is a center monomer which will always complement its opposite and cannot be split in two to satisfy the symmetry, thus there are no autocatalytic odd length oligomers.

From what we know of the self-complementary sequences and with some general intuition it is impossible for a symmetry breaking to emerge between the two monomers in this case, for one cannot template without the other. However, there does appear to be evidence that the sequences should split and certain sequences should have a higher fitness than others and thus dominate. In the figure below there are 6 unique concentrations with the dominant sequence being the alternating sequence. Recall that the LD model also had alternating sequences being dominant prior to bifurcation due to

51

having some autocatalytic oligomers in its family while also not having a low number of templates. In this case, while the number of templates has not changed, the alternating sequences are the most autocatalytic, therefore we expect to see them dominate at all ligation values given an unremarkable starting configuration. This is evident in figure 4.2 as the ligation increases to a large value and there is still no change in dominant sequence



Figure 4.2 The concentration of the tetramer sequences in the AU-T model over increasing ligation rates starting with monomers in even concentrations totaling a concentration of 10. The alternating tetramers increase their dominance as ligation increases while the uniform sequences become less abundant.

with a starting configuration of evenly split monomers. Just as before there are symmetries that cause only 6 different concentrations to occur. The symmetries are the same as in the LD model except the sequence symmetries are broken in different ways. In fact, if the LD model had exactly even monomers and did not have interchange it would look identical to this model for the usual case while operating on entirely different principles due to the symmetries. If interchange were set to 0 in this model there wouldn't be too much of a difference since any self-complementary strands have to have an equal amount of each of the monomers and for the strands that are not self-complementary the templates require the opposing monomer.

In the case of the LD-M model, each of the sequences has the exact same amount of templates and since small fluctuations in the AU model are quickly dissipated due to the characteristic of pairing with the opposing monomer the sequences all maintain the same concentration starting from nearly even monomers as we did in figure 4.2. The model is more interesting in the cases where the starting configuration is not evenly distributed in the monomers, which will be discussed next.

Self-									
Complimentary									
Dimers		Trimers		Tetramers		Pentamers		Hexamers	
LD	AU								
Model	Model								
LL	AU	LLL		LLLL	UUAA	LLLLL		LLLLLL	UUUAAA
DD	UA	LDL		LDDL	UAUA	LLDLL		LLDDLL	UUAUAA
		DLD		DLLD	AUAU	LDLDL		LDLLDL	UAUAUA
		DDD		DDDD	AAUU	LDDDL		LDDDDL	UAAUUA
						DLLLD		DLLLLD	AUUAAU
						DLDLD		DLDDLD	AUAUAU
						DDLDD		DDLLDD	AAUAUU
						DDDDD		DDDDDD	AAAUUU

Table 4.1 A list of each of the self-complementary oligomers contrasting the LD and AU model.

The AU and LD model diverge most drastically on their self-complementary sequences and the structures of their autocatalytic families. However, as in the LD model

it is possible to trap the system in different states and predict these states based on the self-complementary sequences like AAUU in figure 4.3. In these cases, we limit the max length to 4 and increase the concentration to 80 as we did in the chirality model. When looking at the LD-AT and LD-AM models there was skepticism that some of the states were not possible due to detrimental structural properties created by the differing chiralities; in this model there is no chirality to detriment structure since these are bases that exist and do not hinder each other. It would be easy to imagine in a spatial model that



Figure 4.3 AU-M model starting with a concentration of 20 in AAUU. AAUU sequences dominate beyond a ligation value of 4 and below a ligation of 4 the concentrations are all equal. The sequences are listed in order of decreasing concentration at a ligation of 40.

self-complementary sequences may be in higher concentration pockets where they are repeatedly templating oligomers around themselves. If there was then an encapsulation in a vesicle it is plausible that these sequences would emerge as the dominant oligomer. Furthermore, ligation is very fast compared to polymerization and hydrolysis under certain conditions and if these oligomers were the first to be made by chance and proliferated quickly enough they could dominate the system. These dominant oligomers are not restricted to the self-complementary sequences and can occur for non-complements as well like UUUU and AAAA in figure 4.4. The incorporation of motifs in these models allows the transition to the abnormal state to occur at a lower ligation rate because it increases the number of times it is templated, essentially inflating the ligation value.



Figure 4.4 AU-M model starting with $C_{AAAA} = 10.1$, $C_{UUUU} = 9.9$. At ligation rates below 5 the system is completely symmetric and all sequences have the same concentration. Beyond a ligation of 5 the system is completely dominated by the autocatalytic set of AAAA and UUUU.

4.2 Unique Situations

At the beginning of this chapter we discussed that the odd-length oligomers in the AU model have no self-complementary sequences. If the max length is set to an odd length like 5 there are no predicted dominant sequences for that length. As can be seen in figure 4.5 the sequences UUAAU and its compliment AUUAA dominate at high ligation rates while AUAUA and UAUAU dominate at lower ligation rates. These are the two self-complementary tetramers with an extra base added to either end. It is unclear why the dominant of these two types of self-complementary sequences changes. It would appear



Figure 4.5 AU-T model with max length 5 over a range of ligation values starting from equal concentrations of monomers with a total monomer concentration of 80. At ligation values between 9 and 39 the alternating strands AUAUA and UAUAU dominate while above those values UUAAU and AUUAA dominate.

that the odd length oligomers merely act as a source of food for the even length oligomers and are entirely dependent on which of those even length sequences is dominating. It should be possible to predict from the results of the 4 and 6 max length limits what will be the dominant odd length oligomers. This theory is supported in figure 4.6 where we begin with all of the concentration in the tetramer AAUU, a solution to the max length 4 system which resulted in UUAAA and UUUAA dominating at the pentamer level.



Figure 4.6 AU-M model with max length 5 over a range of ligation values starting from 20 concentration in the tetramer UUAA. Prior to a ligation value of 10 the sequences have even concentrations. At ligation values of 10-12 there is a state where the uniform pentamers are dominant but at higher ligation values UUAAA and UUUAA dominate.

However, in the max length 4 system starting with the same configuration immediately gave only the AAUU solution for all ligation values and in the figure we find there is a different solution at lower ligation where uniform pentamers dominate. This most likely has to do with concentration effects since there is a total concentration of 80 monomers in both but in the max length model there is a portion of the concentration in the newly existing set of pentamers.

If we expand on the idea that for different ligation rates there will emerge different dominant sequences then at the extreme end with very high ligation rates polymerization should be negligible. If the polymerization rate is set to 0 and we look at a single individual tetramer it will break down into its reaction family and reach an equilibrium with those lower length sequences where they each exist with reasonable concentrations. No other sequences will be made since polymerization introduces the variety. Introducing other tetramers results in competition over the lower length sequences, just as it occurred in the LD models over the monomers. Tetramers with shared reaction families compete over the smaller sequences resulting in the better sequence, as in the case of AUAU over AUAA or the sequence with a higher fluctuation, as in the case between AAUU and UUAA, to siphon off of the other as it breaks down until it is extinct. In the case where we do not consider motifs one of the alternating strands is the only extant tetramer at equilibrium. The fact that no other tetramers exist show how complex and interconnected this model is, where the concentration of one will affect the concentration of the entire rest of the sequences. Upon introducing motifs where everything has an equal number of templates, if kept exactly even the tetramers will all remain, but under a slight perturbation they cascade once again. However, in this case one of AAUU or UUAA selfcomplementary tetramers prevail over the others. This is relevant to the pentamer case discussed above which saw the alternating strands dominate at low values and these

sequences dominate at higher values. There would appear to be a correlation between the rate of ligation and which of these self-complementary sequence dominates.

4.3 AU Conclusions

In these AU models there is no instability between the two bases like what was seen in the chirality models; A and U have equal concentrations in all cases. We have seen that alternating sequences dominate the system in the AU-T model given unremarkable starting configurations but this is not the case for the AU-M model which remains with all concentrations equal under the same conditions. There are several kinds of phases in the AU-M model: (i) where a single self-complementary sequence dominates Ex. UUAA, (ii) where two mutually complementary sequences dominate Ex. AAAA and UUUU, and (iii) where two equivalent self-complementary sequences dominate Ex. AUAU and UAUA.

Chapter 5: Wobble Base Models

5.1 AUG Model

The final variation of this model we will discuss is the introduction of the wobble base. Now that we have determined the most likely sequences to succeed between a pair of nucleobases we introduce the other nucleobases to determine if they will carry over to the canonical bases. Among the canonical bases adenine, uracil, guanine and cytosine there is a wobble base pair that does not follow the Watson-Crick pairing rules between uracil and guanine. This breaks the previous symmetry from the AU model where each of the nucleobases was equal in every way but now uracil can pair with either adenine or guanine. Furthermore, if cytosine is included in the model guanine will also be able to pair with two nucleobases while adenine and cytosine can still only pair with one.

As mentioned in the introduction there is some evidence that the bases could have come from meteorites which did not contain the nucleobase cytosine. Under these circumstances it is prudent to explore the implications a 3 letter alphabet may have in sequence space, prior to the introduction of cytosine.

In these models interchange is treated similar to how it is treated in the LD and AU model except there is now at least a third and possibly a fourth monomer. The interchange's goal is still to equalize the monomer concentrations so the monomer's concentration is subtracted by the average concentration of the others to determine if it will increase or decrease and by how much. From this we can see there is no bias in the

creation of the monomers which translates to there being no bias in the amount of each sequence being made.

With the introduction of a third monomer the number of sequences and consequently self-complementary sequences is expanded. Now the unique number of nmers is 3^n , and instead of the usual 16 tetramers that have been used as our usual subjects of analysis, there are now 81 unique tetramers. The rule for finding self-complementary *n*-mers discussed in the AU model still applies to these but no longer encompasses all cases, in this model there are more than the number of unique dimers due to the wobble. While there are still only 9 unique first halves of the tetramers in this 3 letter alphabet the ability for uracil to pair with either of the other monomers expands the constrictions on the second half of the sequence while keeping chirally symmetric across the middle. If the first half contains a uracil monomer it will contribute $2^{N_{Uracil}}$ unique self-complementary sequences, where N_{Uracil} is the number of uracil monomers in the first half of the sequence, instead of one due to its ability to pair with either A or G. That means to determine how many self-complementary tetramer sequences are present in this AUG model we look at the 9 dimers of which 4 do not contain uracil, 4 contain 1 uracil and 1 contains 2 uracil, resulting in 16 unique self-complementary tetramers as follows:



-AUAU -AUGU -GUGU -GUAU -UAUA -UAUG -UGUG -UGUA



The wobble pair further changes things by introducing more than a single template for oligomers at their own length, take for example the tetramer UAUA, in the AU and LD model there was only one complement for it at the tetramer length but in the AUG model it can be templated by UAUA, UAUG, UGUA, and UGUG. In this way there are 4 families of tetramers that can self-complement and complement each other highlighted in the above lists by colour. Furthermore, the sequences containing uracil will have more templates than those without which extends to the model when incorporating motifs due to the symmetry breaking caused by uracil's pairings. There is still however a general rule for calculating the number of templates when incorporating motifs. Recall equation 10 $(S_n = (n - i + 1)2^{n-i})$ from section 3.2 for calculating motifs from oligomers of length *n* in the LD model, in this model the motifs can be calculated using

$$S_n = 2^U (n - i + 1)3^{n - i} + S_{n - 1}$$
(11)

where U is the number of uracil monomers in the product, i is the length of the product and n is the length of the template. The solution for this case is not as useful as it is not close formed. It is useful in highlighting the fact that having more uracil in your sequence results in many more templates. From this solution we can see it also breaks the symmetry we saw in the AU model where all of the sequences were at equal concentration prior to any bifurcation when incorporating motifs. Now the concentrations are only equal between each of the sequences containing the greater number of uracil monomers.

Given an initialization where each of the monomers is equal the alternating tetramers containing uracil (the middle column in the list above) share dominance with
equal concentrations while neglecting motifs. In half of these sequences the uracil is in the 2 and 4 position while in the other half they are in 1 and 3, all 8 sequences are not complementary to each other and it is probable that at a very high ligation one of these groups will beat the other. The reason it doesn't happen at low ligation is because both of these families use the same trimers as food. If we were to consider the case where ligation is much faster than polymerization like the case in the AU model where we turned polymerization to 0, given a fluctuation one of these would win over the other. If we apply motifs and run the same simulation, then at low ligation values the uniform uracil tetramer is dominant and at higher ligation rates the left and right columns of the above list are dominant.

To determine what would happen in a scenario where only two monomers were present and then a third slowly introduced we started with an even concentration of A and U monomers and no G. In figure 5.1 the uniform wobble base dominates the system with its complements (AAAA, AAGA, GAGA, etc.) maintaining a secondary concentration. However, at a ligation value around 15 the system transitions to alternating U tetramers (AUAU, UGUA, etc.). This is nearly the exact opposite behavior to what we saw in the AU model, where alternating strands dominated at low ligation and only at high ligation could the autocatalytic set AAAA and UUUU coexist. However, when we consider the templates that the wobble has introduced it is clear that this is the reasonable progression of the system as the tetramer UUUU has 16 templates at its own length, with the next highest being the self-complementary alternating strands containing U with only 4 templates at the tetramer level. As the ligation increases there is a point where the rate of ligation allows the self-complementary strands to win over the more templated UUUU.



Figure 5.1 The AUG model starting with an adenine concentration and uracil concentration of 40 while incorporating motifs. Complements to UUUU include all tetramer combinations containing exclusively A and G monomers.

These are not the only possible outcomes from this model, just as in the AU model if we begin with more abnormal starting concentrations different sequences can dominate beyond these 2 states. If the system begins with all of the concentration within the tetramer AAAA we inevitably get the first state we see in the figure where UUUU dominates and each of its templates including AAAA are at a secondary concentration. There may be more self-complementary tetramers in this model but the patterns remain the same as the AU model. Starting with entirely AAUU strands results in AAUU, AGUU, GAUU and GGUU being dominant which means it is possible to at least trap the system in each of the self-complementary states and the highest template state. Beyond

these states it is possible to trap other autocatalytic sets but it requires larger ligation values. All of the equilibriums in this model take much longer to reach than those in the AU model due to the greater number of interacting sequences. In scenarios where the system begins with all of its concentration in one tetramer like AAUU, while at equilibrium each of AAUU, AGUU, GAUU and GGUU are at equal concentrations for a majority of the time AAUU is dominating. It is here we must be mindful of what we should consider a reasonable system state because there is obviously constraints on this model like a max length and staying in such a stable system to reach equilibrium. In reality these systems are likely never going to reach equilibrium as there will be constant change in the oligomers of longer lengths influencing them as they grow.

While there have been many changes to the number of self-complementary sequences and the dynamics, even with a wobble in both this model and the next it is impossible to get self-complimentary odd length oligomers. The fact that none of the bases complement themselves remains and thus the middle monomer cannot be the same for the template and product. The only scenario where odd-length oligomers are self-complementary are when a monomer is able to pair with itself.

5.2 ATGC/AUGC Model

Finally, we introduce the final monomer, cytosine, to our model. Prior to looking at the effects that cytosine has on the AUG model we look at the ATGC model to possibly provide insight into the question of why does RNA swap out uracil for thymine in DNA. In this ATGC model there are no wobble base pairings. By eliminating the wobble pair the model takes on characteristics of a combination of the LD model and the AU model. The same states and sequence patterns from the AU model can occur for each the AT and GC pairs but if we allow interchange between each of the bases as we have been doing there can occur a fluctuation in one of the pairs that ends up eliminating the other just as in the LD model, thus resulting in a two monomer system. If we simply do not allow interchange, then we basically have two AU models with different names for the bases running at the same time since they do not interact without interchange due to the lack of wobble base pairs. One new thing that the ATGC model adds is a new self-complementary base that contains all of the bases. The sequences GATC and AGCT are self-complementary and we would likely find there to be a state that exists where each of these dominates.

The introduction of the fourth canonical RNA base further differentiates from each of the previous models containing A, U and G because it allows not only uracil to pair with 2 bases but guanine as well. In a similar fashion to the AUG model just discussed this upsets the symmetry and equality normally provided by adding motifs for the number of templates for each sequence. Instead of only considering how many uracil monomers are in the sequence we must count the number of guanine monomers as well to get an accurate number from equation 11. Overall, the effect is that sequences with more U and G monomers will have more templates and be more likely to dominate just as the uniform uracil tetramer dominated in the previous model.

In the case of no motifs the usual suspects are dominating the system, just as in the AUG model, the alternating strands containing U are dominant. One difference is that

now that there are two wobble bases the dominant alternating strands are limited to UGUG and GUGU. The other alternating sequences form tiers at lower concentrations that depend on the amount of G and U. For example, UGUG is the top concentration followed by CGUG at a lower concentration and CGCG at an even lower one.



Figure 5.2 AUGC model starting from even concentrations with motifs and a total concentration of 80 with max length 4. The uniform wobble bases dominate at all ligation values but at values above 10 there become tiers of sequences based on how many of the wobble monomers they contain. The sequences cannot contain the other wobble pair else they fall into the "other tetramers" category.

In figure 5.2 we have initialized nearly equal concentrations of each monomer and incorporated motifs to find an interesting result. At low ligation values we see what we expect from the increased amount of templates, the tetramers with only U and G monomers dominating all at equal concentration. At high ligation values uniform U and G tetramers dominate but the next highest concentrations are sequences containing UUU and GGG where the fourth monomer is not a U or a G. The reason for this is because of

the uniform tetramers being most dominant. It appears that the secondary, tertiary and so on are the templates for the uniform tetramers but split into tiers based on how many U or G monomers it has because they can match with more templates, and based on how uniform it is. An example for the latter criteria being the sequence UUUA is more uniform than UUAU. This comes from the fact that there will be many uniform trimers being templated for and acting as food for the uniform tetramers, so their concentration is likewise increased.

5.3 Wobble Base Conclusions

The introduction of the wobble base pair causes similar dominant sequences as in the AU model but only with the U sequences since there are more templates for U containing sequences. Adding the 4th base into an ATGC model causes symmetry breaking between pairs AT and GC and introduces new self-complementary tetramers containing all 4 bases. In the case of adding the 4th base to the wobble base pair model to create an AUGC model tetramers with only U and G dominate at low ligation values and at higher ligation values uniform U and G take over. This is due to the addition of C making G a wobble pair as well as U. There are many cases we did not have time to test in these wobble models as computation time increases exponentially with the number of bases in the model.

Chapter 6: Conclusions and Future Work

6.1 Summary

Introducing the ability for non-uniform templating in the LD model represented a worst-case scenario for the templating reaction to achieve chiral symmetry breaking. We say the worst case scenario due to the structural problems sequences of mixed chirality would have when aligning to be templated. The incorporation of motifs allowed for templates with more binding sites to be better templates as they should be. Introducing motifs resulted in lowering the ligation value required to achieve symmetry breaking. While, as expected, introducing the ability for any sequence to be templated increased the ligation value required for a symmetry breaking to occur. Nonetheless, even in the worst case scenario with no motifs and templating for every sequence a chiral symmetry breaking occurred and in the best case of uniform templating only and motifs it occurred at the low value of 2.1. The worst case scenario would not be probable since the chirality of each of the monomers would not allow them to line up and bond especially in the case of more alternating sequences, but we would expect to see some of the sequences that are close to being uniform be templated. Similarly, there would not be a linear scaling between template sites and reaction rates due at least to inhibition.

The 'M' models, where motifs are weighted equally, assume that the catalytic ability of a template strand is proportional to the number of places in which the two oligomers can bind to the template. This seems more realistic than the 'T' models, in which all strands are weighted equally. However, if the motifs on a template are overlapping, they cannot operate at the same time, so the M models are also an

approximation. There should be some intermediate instance, which allows for inhibition and motifs to make a template more effective but the problem lies in the toy model itself. The ligation reaction takes place in a single step and to incorporate both inhibition and motifs would require adding a binding and unbinding step, which is discussed in section 6.2. This model can still use the 2 instances to understand generally what will happen if some templates are better than others without being exact and we instead expect the real ligation value to be somewhere in-between these extremes but it is at least possible at both ends. It is left to future research to add the extra steps and inhibition.

After confirming the symmetry breaking results presented by Tupper et al., 2017 and expanding on them to ensure that given any scenario it can occur, we looked at numerous models to try and determine if there was similar symmetry breaking in sequence space. We see the emergence of a single chirality in life today so determining a plausible pathway to that point is useful but in the case of sequences there is no direct result to predict except the sequence states. However, if the results produce some sequences in greater quantities than others it may lead to structure prediction in ribozymes, where we see a common repeating pattern that creates a certain structure when folding.

With the AU model we immediately discovered that there is some favoritism towards alternating strands and with higher ligation the more exaggerated that favoritism becomes. With the incorporation of motifs each of the sequences were equally templated for creating a symmetry at low ligation values where everything had equal concentrations. The symmetry was broken by perturbations and higher ligation rates to sway the system

into certain states determined by the degree of autocatalysis. The easiest of the states to reach were those where the self-complementary sequences dominated like the alternating strands and the AAUU and UUAA strands. However, it was also possible at large enough ligation values and with enough perturbation to reach states that were simply autocatalytic sets where multiple sequences could co-dominate like the UUUU and AAAA state. These states of mixed sequences like the alternating strands are plausible, in contrast to the LD model, because the monomers do not have a chirality difference and instead are just representative of the bases which do not cause detrimental structures for the templating reaction.

Upon adding a third base as might have been the case in a scenario where meteorites were the suppliers of the nucleobases, the motif symmetry was broken. The ability for the uracil monomer to wobble and pair with both adenine and guanine gave it the advantage of having more templates and the previous template equality provided by motifs was no longer present. Because of this there was a state where uniform uracil oligomers dominate the system due to their superior number of templates. However, there were still other states including the self-complementary ones found in the AU model that could overcome the uniform state at higher ligation values. In the case of the four letter alphabet both uracil and guanine had two pairing options which caused oligomers containing a combination of them to be favored at low ligation values in equal amount before transitioning to the uniform sequences as we saw in the AUG model. The final variation where we swap uracil for thymine caused there to be a symmetry breaking between the 2 pairs where one could win over the other provided there was a large enough fluctuation.

The discovery of the dominant species in each of these cases could prove important when looking for the first ribozymes. If ligation is much faster than polymerization, particularly from a bioinformatics point of view, we may expect to see these sequence patterns frequently occurring in ribozyme structures. As one of the purposes of this research was to provide a possible pathway or mixture for the formation of the first ribozyme, thus no ribozyme is present for the templating reaction. Or understanding what kinds of structures these patterns fold to when repeated could give hints to the overall structure of the first ribozyme.

Many of the states from the models required high concentration, high fluctuation and high ligation rates. To determine if these states can form in a real world scenario requires experimental rates for the ligation value (and hydrolysis and polymerization). The models represent a predictive tool for when the ligation rate is experimentally discovered to predict what sequences may emerge out of a set of monomers or whatever sequences the experiment is initialized with. As for the high concentration and fluctuations, there is no way of knowing exactly how concentrated the nucleotides could have been in a warm little pond or hydrothermal vent but we can imagine them aggregating in specific areas to achieve this, just as salts and minerals build up at the water line of geothermal springs. High fluctuations could occur due to similar effects where a single template creates a population of itself and its complement around it or prebiotic vesicles may encapsulate a long template and replicate to high concentration

before bursting. It is in these scenarios that the more exotic of the states, that required high initial concentrations in one sequence, could form.

6.2 Limitations and Future Work

In this work we have split the chirality and the nucleobase models. On early Earth there were most likely sequences of nucleobases of different chiralities polymerizing and interacting and competing. In fact, the monomers would not be limited to variation in only these properties as there could be XNAs, 2'-5' bonding, etc. that RNA is chosen from. It would be interesting to assign all of these properties to monomers and see what emerges from the prebiotic clutter. We must also consider the implications of the sequences growing while all of this is happening because perhaps the first uniform chirality strand was one of the sequences we saw in the nucleobase models that required a more exotic starting perturbation.

Something these models did not consider is the error rate or if the entire sequence is required to match for a templating reaction to occur. The basis of the templating reaction is to pair up with its complement and hold it in place until the other part of its complement pairs up and then the two complements ligate together and detach under the proper conditions resulting in the complete complement to the template. In the case of longer strands there may be sequences that are almost an exact complement except for a single pairing but there may be enough energy from the other pairs to be able to overcome it and stay bound regardless. Alternatively, there is likely a threshold where given the length of an oligomer there are a certain number of pairs required to stay bound to a template but there could be overhang like a hexamer bonding to another hexamer but only three of the bases are bound and the rest hang off leaving room for another oligomer to bond to the second half of the template. This way the template could even act as a catalyst for the formation of larger oligomers than itself. For the most part it will depend on binding energy and the energy due to structures like bulge loops. These kinds of models would require stochastic methods because there would be a very large number of possible reactions at higher lengths. To build a model that incorporates these scenarios requires breaking the ligation into steps. Each of the models discussed lumped the complex steps of the binding and unbinding and ligation into a single rate. By breaking it down into the binding of each of the food oligomers to the template and then ligating and then unbinding, there become many more specific rates from the one we used here. By incorporating those rates, it would also be possible to address the issue with the linearly scaling motifs alluded to earlier. Having the ligation rate scale linearly with number of motifs only makes sense if the reaction is instantaneous and therefore there is no inhibition, these steps and other factors are important to the rate and we have lumped all of that into one number. The steps would allow for inhibition to take place and make some templates better than others without moving back to the symmetry we found in the AU model where everything had the same number of motifs for templating. The rates can be determined through the Gibbs free energy of each of the structures at each step in the process of ligation. The model quickly becomes messy with broken ladder binding if incorporating overhang and determining what can hydrolyze if an oligomer is currently attached to a template, but the broken ladder could be the answer to creating long enough

M.Sc. Thesis – E. Turner; McMaster University – Physics and Astronomy

chains to create the first ribozyme and is worth the attempt. The work here provides the foundations for that research as it has shown that there is bias within the templating reaction.

References:

Benner, S., Kim, H.-J. and Carrigan, M., 2012. Asphalt, water, and the prebiotic synthesis of ribose, ribonucleosides, and RNA. *Accounts of chemical research*, 45(12), pp.2025–2034.

Benner, S., Ricardo, A., and Carrigan, M., 2004. Is there a common chemical model for life in the universe? *Current Opinion in Chemical Biology*, 8(6), pp.672-689.

Bolli, M., Micura, R. and Eschenmoser, A., 1997. Pyranosyl-RNA: chiroselective selfassembly of base sequences by ligative oligomerization of tetra nucleotide-2',3'cyclophosphates. *Chemistry & Biology*, 4(4), pp.309–320.

Bonner, W., and Rubenstein, E., 1987. Supernovae, Neutron Stars and Biomolecular Chirality. *Biosystems*, 20, pp.99-11.

Breslow, R., 1959. On the mechanism of the formose reaction. *Tetrahedron Letters*, 1(21), pp.22–26.

Butlerow, A., 1861. Formation of Monosaccharides from Formaldehyde. *Comptes Rendus*, 53, pp.145–167.

Callahan, M. et al., 2011. Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proceedings of the National Academy of Sciences*, 108(34), pp.13995–13998.

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

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

Deamer, D., Singaram, S., Rajamani, S., Kompanichenko, V., and Guggenheim, S., 2006. Self-assembly processes in the prebiotic environment. *Philosophical Transactions of the Royal Society of London series B Biological Science*, 361, pp.1809–1818.

Ding, P., Kawamura, K., and Ferris, J., 1996. Oligomerization of uridine phosphorimidazolides on montmorillonite: a model for the prebiotic synthesis of RNA on minerals. *Origins of Life and Evolution of the Biosphere*, 26, pp.151-171.

Ferus, M. et al., 2017. Formation of nucleobases in a Miller–Urey reducing atmosphere. *Proceedings of the National Academy of Sciences*, 114(17), pp.4306–4311.

Fellermann, H., Tanaka, S. and Rasmussen, S., 2017. Sequence selection by dynamical symmetry breaking in an autocatalytic binary polymer model. *Physical Review E*, 96(6), p.062407.

Fuller, W. and Sanchez, R., 1972. Studies in prebiotic synthesis: VI. Synthesis of purine nucleosides. *Journal of molecular biology*, 67, pp.25-33.

Gilbert, W., 1986. Origin of life: The RNA world. Nature, 319(6055), p.319618a0.

Higgs, P., 2016. The Effect of Limited Diffusion and Wet–Dry Cycling on Reversible Polymerization Reactions: Implications for Prebiotic Synthesis of Nucleic Acids. *Life*, 6(2), 24.

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

Himbert, S., Chapman, M., Deamer, D. and Rheinstädter, M., 2016. Organization of Nucleotides in Different Environments and the Formation of Pre-Polymers. *Scientific reports*, p.31285.

Johnston, W. et al., 2001. RNA-Catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension. *Science*, 292(5520), pp.1319–1325.

Lohrmann, 1975. Formation of nucleoside 5'-polyphosphates from nucleotides and trimetaphosphate. *Journal of Molecular Evolution*, 6(4), pp.237–252.

Lohrmann, R. and Orgel, L., 1971. Urea-inorganic phosphate mixtures as prebiotic phosphorylating agents. *Science*, 171(3970), pp.490–494.

Mamajanov, I., et al., 2014.Ester Formation and Hydrolysis during Wet-Dry Cycles: Generation of Far-from-Equilibrium Polymers in a Model Prebiotic Reaction. *Journal of the American Chemical Society*, 47, pp1334-1343.

Miller, S., 1953. Production of Amino Acids Under Possible Primitive Earth Conditions. *Science*, 117(3046), pp.528-529.

Nuevo, M., Materese, C. and Sandford, S., 2014. The Photochemistry of Pyrimidine in Realistic Astrophysical Ices and the Production of Nucleobases. *The Astrophysical Journal*, 793(2), p.125.

Oparin, A., 1938. The Origin of Life.

Orgel, L., 2004. Prebiotic chemistry and the origin of the RNA world. *Critical reviews in biochemistry and molecular biology*, 39(2), pp.99-123.

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

Pearce, B. and Pudritz, R., 2015. Seeding the Pregenetic Earth: MeteoriticAbundances of Nucleobases and Potential Reaction Pathways. *The Astrophysical Journal*, 807(1), p.85.

Pearce, B. and Pudritz, R., 2016. Meteorites and the RNA World: A Thermodynamic Model of Nucleobase Synthesis within Planetesimals. *Astrobiology*.

Pearce, B., Pudritz, R., Semenov, D. and Henning, T., 2017. Origin of the RNA world: The fate of nucleobases in warm little ponds. *Proceedings of the National Academy of Sciences*, 114(43), pp.11327–11332.

Plasson, R. et al., 2007. Emergence of homochirality in far-from-equilibrium systems: Mechanisms and role in prebiotic chemistry. *Chirality*, 19(8), pp.589–600.

Powner, M., Gerland, B. and Sutherland, J., 2009. Synthesis of activated pyrimidine ribonucleotides in prebiotically plausible conditions. *Nature*, 459(7244), p.239.

Rajamani, S. et al., 2008. Lipid-assisted synthesis of RNA-like polymers from mononucleotides. *Origins of Life and Evolution of the Biosphere*, 38(1), pp.57-74.

Reimann, R. and Zubay, G., 1999. Nucleoside Phosphorylation: A feasible step in the prebiotic pathway to RNA. *Origins of Life and Evolution of the Biosphere*, 29, pp.229-247.

Ricardo, A., Carrigan, M., Olcott, A., and Benner, S., 2004. Borate minerals stabilize ribose. *Science*, 303(5655), pp.196–196.

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

Rohatgi, R., Bartel, D.P. and Szostak, J.W., 1996. Nonenzymatic, template-directed ligation of oligoribonucleotides is highly regioselective for the formation of 3'- 5' phosphodiester bonds. *Journal of the American Chemical Society*, 118(14), pp.3340–3344.

Ross, D. and Deamer, D., 2016. Dry/Wet Cycling and the Thermodynamics and Kinetics of Prebiotic Polymer Synthesis. *Life*, p.28.

Shapiro, R., 1988. Prebiotic ribose synthesis: A critical analysis. *Origins of Life and Evolution of the Biosphere*, 18(1-2), pp.71–85.

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

Tkachenko, A. and Maslov,S 2017. Onset of natural selection in auto-catalytic heteropolymers. *arXiv preprint arXiv:1710.06385*. [online] Available at: .

Tupper, A., Shi, K. and Higgs, P., 2017. The Role of Templating in the Emergence of RNA from the Prebiotic Chemical Mixture. *Life*, 7(4), p.41.

Wochner, A., Attwater, J., Coulson, A. and Holliger, P., 2011. Ribozyme-Catalyzed Transcription of an Active Ribozyme. *Science*, 332(6026), pp.209–212.

Zahnle, K., Schaefer, L. and Fegley, B., 2010. Earth's Earliest Atmospheres. *Cold Spring Harbor Perspectives in Biology*, 2(10), p.a004895.

Zubay, G. and Mui, T., 2001. Prebiotic Synthesis of Nucleotides. *Origins of Life and Evolution of the Biosphere*, 31(1/2), pp.87–102.