Winter 1-31-2019

The Effect of Dynamic Kinetic Selection on an Evolving Ribozyme Population

Patrick David Poletti
Portland State University

Let us know how access to this document benefits you.

Follow this and additional works at: https://pdxscholar.library.pdx.edu/open_access_etds

Part of the [Biochemistry Commons](#)

Recommended Citation

Poletti, Patrick David, "The Effect of Dynamic Kinetic Selection on an Evolving Ribozyme Population" (2019). *Dissertations and Theses*. Paper 4805.

[10.15760/etd.6689](https://pdxscholar.library.pdx.edu/open_access_etds/10.15760/etd.6689)

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

The Effect of Dynamic Kinetic Selection on an Evolving Ribozyme Population

by

Patrick David Poletti

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy
in
Chemistry

Dissertation Committee:
Niles Lehman, Chair
Albert Benight
Stephen Reichow
Deborah Duffield

Portland State University
2019

Abstract

Dynamic Kinetic Selection (DKS) suggests that kinetic, rather than thermodynamic, stability will dictate the composition of a replicating population of biomolecules. Here, the results obtained from a series of five related reactions involving gradually increasing percentages of randomly-mutated substrate fragments to generate variants of full-length *Azoarcus* group I intron through an autocatalytic self-assembly reaction involving a series of recombination events, showed DKS as a driving factor in dictating the population composition of full-length product assembled from substrates that had fewer positions available to randomization.

In trying to elucidate a plausible scheme for the origins of complex biomolecules on the prebiotic Earth, the suggestion that networks comprised of interacting molecules were more likely to evolve into biomolecules capable of obtaining and sustaining characteristics attributed to living molecules has gained traction within the past few years. Of specific interest is the catalytic efficacy of ribozymes whose genotypes require that they interact with molecules of the same genotype (selfish systems) to be effective catalysts versus those that are more effective when accomplishing catalysis by cooperating with ribozymes of a different genotype (cooperative systems). Here, the *Azoarcus* I ribozyme was used to compare these two types of system. Both systems were shown to robustly produce full-length product. Two different methods of introducing random mutations into substrate fragments for the reactions described in this thesis were employed. The differences in the preparation methods for the substrates was not expected to have an impact on the nature of the full-length product. However, there was no correlation between

the positions that tended to be more tolerant of accepting random mutations between the products arising from the two preparation methods. One preparation method yielded full-length ribozymes more consistent with the secondary structure of the wild-type ribozyme and followed substitution patterns found in *in vivo* nucleic acid substitutions, whereas the other method provided full-length ribozymes that tolerated mutations that would be expected to greatly affect the secondary structure of the ribozyme and those positions tended to mutate evenly to any of the three possible alternative nucleobases.

Point mutations introduced into ribozyme substrate fragments may have a deleterious, neutral, or beneficial effect, depending on their impact on the catalytic capability of the molecule vis-à-vis the effect, if any, the change has to the secondary and tertiary structure of the ribozyme. In this dissertation, the results of two series of point mutation reactions are addressed. The first set showed a point mutation to have a deleterious effect, whereas concerted mutations did not significantly affect activity of the ribozyme. The second series of reactions involved point mutations at a position that had previously been determined to be highly tolerant of random mutations. Results suggested that substitutions at this position had a minimal impact on ribozyme activity.

Acknowledgements

My most sincere gratitude is expressed to the following individuals and groups, each of whom has helped, supported, or encouraged me in their own way:

Dr. Niles Lehman, my dissertation advisor

Dr. Albert Benight

Dr. Stephen Reichow

Dr. Deborah Duffield, the members of my dissertation committee

Current and former members of the Molecular Evolution Research Group

Especially:

Kavita Matange, for assisting with the Hotspot Reactions

Dr. Tharuka Jayathilaka, for helping me through the moments of doubt

Dr. Don Minore

Dr. Eric Sheagley

Jill McCollum (1948-2018)

Portland State University Campus Recreation Aquatics

Table of Contents

Abstract.....	i
Acknowledgements.....	iii
List of Tables.....	v
List of Figures.....	vii
Chapter 1	
Introduction.....	1
Chapter 2	
Recombination of a Class I Ribozyme from Substrate Fragments Prepared at a 9% Mutation Rate.....	16
Chapter 3	
Recombination of a Class I Ribozyme Assembled Through Selfish and Cooperative Replicator Networks.....	72
Chapter 4	
The Effect of Point Mutations on the Activities of Full-Length <i>Azoarcus</i> Ribozymes.....	113
References.....	132

List of Tables

Table 1 – Number of sequences available for analysis at each time point for each of the five 9% reactions	55
Table 2 – Data Analysis for the Z*-Single Reaction	56
Table 3 – Most prevalent population members for the Z*-Single reaction at each of the six time points.....	57
Table 4 – The niche populations for the Z*-Single and W*-Single reactions	58
Table 5 – Data Analysis for the W*-Single Reaction	59
Table 6 – Most prevalent population members for the W*-Single reaction at each of the six time points	60
Table 7 – Data Analysis for the Double Reaction (W*-region)	61
Table 8 – Data Analysis for the Double Reaction (Z*-region)	62
Table 9 – Configuration of nucleotides at positions 176 and 177 for the Z*-Single and Double Reactions.....	63
Table 10 – Most prevalent population members for the Double reaction at each of the six time points	64
Table 11 – Analysis of the four full-length product sequences obtained from the Triple (W* + X* + Y + Z*) reaction	65
Table 12 – Concentrations and amounts of substrates used in each of the four variants of the recombination reaction.....	66
Table 13 – Average mutation rate (μ) for W*-Single, Z*-Single and Double Reactions.....	67
Table 14 – Average deviation from wild-type (k) for the 9%-mutation reactions ...	68
Table 15 – Composition of substrates used in the recombination of the <i>Azoarcus</i> group I intron	69
Table 16 – Methodology for rectifying deviations from the wild-type sequence during processing	70

Table 17 – Noise correction per position	71
Table 18 – Concentrations of and amounts of substrates used in the Selfish and Cooperation reactions	102
Table 19 – Number of sequences available for analysis at each time point for the Selfish and Cooperation reactions	103
Table 20 – Selfish Reaction – Mutations by Position.....	104
Table 21 – Cooperation Reaction – Mutations by Position.....	105
Table 22 – Percent of each nucleotide present at positions of interest for the Selfish and Cooperation reactions	106
Table 23 – Trend in total mutation rate between the 30-minute and 8-hour time point for the Selfish and Cooperation reactions	107
Table 24 – The most prevalent population members for the Selfish reaction at each of the six time points	108
Table 25 – The most prevalent population members for the Selfish reaction at each of the six time points	109
Table 26 – Niche population for the Selfish and Cooperation reactions	110
Table 27 – Average mutation rate for Selfish and Cooperation reactions, with overall mutation rate for the mutagenic PCR reactions	111
Table 28 – Average deviation from wildtype (<i>k</i>) for the mutagenic PCR reactions	112
Table 29 – Distribution of nucleotides at positions 33 and 34.....	131

List of Figures

Figure 1 – The <i>Azoarcus</i> group I ribozyme.....	14
Figure 2 – Schematic of the IGS-tag interaction involved in the catalysis of recombination.....	15
Figure 3 – The <i>tF2</i> and <i>R2F2</i> mechanisms of recombination in the <i>Azoarcus</i> Group I Ribozyme.....	51
Figure 4 – Comparison of product obtained from the reaction of the four wild-type substrates versus the product obtained when one of the substrate fragments is mutated.....	52
Figure 5 – Comparison of expected products from the five recombination Reactions.....	53
Figure 6 – Percent wildtype for Z*-Single, W*-Single, and Double Reaction Populations	54
Figure 7 – Reaction Design for Selfish and Cooperation Reactions	99
Figure 8 – Percent wildtype for the Selfish and Cooperation Reactions	100
Figure 9 – Percent of reaction population bearing one of two available IGS/tag sequence combinations	101
Figure 10 – The four mutations introduced into the P2 stem of the W-fragment ...	126
Figure 11 – Results from the P2 reactions	127
Figure 12 – Percent of full-length self-assembly product for each member of the P2 reaction set	128
Figure 13 – Results from the Hotspot Reactions	129
Figure 14 – Percent of full-length self-assembly product for the Hotspot reactions	130

CHAPTER 1 – INTRODUCTION

The term “Prebiotic Earth” refers to the period of time in the Earth’s history in which molecules existing in the primordial soup (a lifeless mixture of inorganic and organic precursors) chemically evolved into molecules capable of creating and sustaining life (Joyce, 2002). In order to attain living function, a successful emerging biomolecule must have been able to both maintain genetic information and to obtain catalytic activity. Within the field of prebiotic chemistry, great efforts are being made to try to reconstruct a plausible chemical model for how prebiotic biomolecules could develop the ability to both store genetic code and to perform catalysis. The proposed RNA World hypothesis suggests a period in the early evolution of life when both storage of the genetic code and catalysis were performed by RNA (Orgel, 1968; Crick, 1968; Gilbert, 1986). While the RNA World period has become widely accepted over the past thirty years, the question as to how these inorganic molecules could make the jump from being lifeless inorganic molecules to those poised to attain the ability to store genetic information and to perform catalysis remains largely unanswered.

An introductory chemistry course teaches that thermodynamic stability occurs with conjunction with an increase in entropy and free energy – simply put, stability favors disorder. In contradiction to this statement, however, is the requirement that biomolecules attain complexity in order to achieve functionality. The reactions that direct storage of genetic material and catalytic activity do not favor disorder, and thus decrease, rather than increase, the free energy of the universe. Therefore, a search for the answer to the question of how lifeless molecules could have become biomolecules must address the question as

to how biomolecules could successfully thwart the requirement that stability requires a decrease in free energy.

The processes of creating and storing life both require chemical reactivity. Biomolecules must be constantly poised and ready to participate in or to catalyze chemical reactions. In a non-living chemical system, stability is achieved when chemical equilibrium is achieved, but living systems must always exist in conditions of non-equilibrium – the living organism that has achieved chemical equilibrium has made the transition back to a non-living organism. A living organism is determined to be “stable” if its thermodynamic processes are being maintained at an appropriate distance from chemical equilibrium. This condition creates a paradox, and the need to distinguish the conditions that contribute to the stability of non-living molecules versus biomolecules. This requires a modified definition of “stability” that can be applied to biomolecules.

The concept of dynamic kinetic stability (“DKS”) was coined to describe a form of stability that is associated with greater reactivity, and was conceived to be associated with replicating systems (Pross and Khodorkovsky, 2004). Breaking the term into its components, “dynamic” refers conditions that lead to constant change and the potential for progress, “kinetic” refers to motion and includes the consideration of the energy required to maintain motion, and “stability” can be considered as the ability to endure. The idea of one molecule making “progress” does not inherently have meaning, and forces the artificial concept of assigning a purposefulness to the molecule that is not appropriate in a scientific context. However, multiple molecules may be considered together as members of a population.

A system is considered stable, if that system is persistent and remains unchanged over time. If a chemical system is thermodynamically stable, its equilibrium will favor the most energetically favorable members of the system, be they reactants or products. On the other hand, if a system is kinetically stable, a very low reaction rate for the formation of products from reactants results in the system maintaining a status quo. The commonality between these two stabilities is a lack of change. In contrast, the stability achieved by a stable population of replicating entities is achieved through a constant change of individual members of a larger, unchanging population (Pross, 2011).

The Darwinian evolution model is widely used to dictate the conditions required for a population of related organisms to either maintain themselves as members of a stable population or to become unable to adapt to changing environmental conditions and cease to exist as living organisms. Thus, to consider biomolecular stability, the stability of a population of biomolecules rather must be considered rather than individual molecular stability and a different form of evolution must be considered, i.e. a non-Darwinian one.

Darwinian evolution can be regarded as the concerted operation of three chemical processes that occur at the level of informational macromolecules: mutation, selection, and amplification (Joyce, 2007) whereas, *in vitro* evolution involves these same three processes, but focuses directly on the nature of the molecules and occurs in a test tube rather than in a living organism and is thus a valuable tool by which the emergence of biomolecules can be studied (Mills, 1967). The concept of DKS has been suggested as a bridge capable of unifying the chemical processes of life's emergence with the biological process of evolution (Pascal, 2017).

Evolution *in vitro* was first performed successfully by Spiegelman and colleagues over forty years ago (Mills et al., 1967). Here, exponential amplification of two populations of biomolecules was allowed to occur contemporaneously under serial transfer conditions that allowed only newly-amplified molecules to become amplified over time. In the process of replication, random mutations arose; those of which were generated in sufficient quantity to be included in serial transfers became entrenched into the replicating population. Two observations from this initial *in vitro* experiment are of particular interest to the research in this dissertation – first, catalytic activity was determined to have been lost after five serial transfers and second, the ability to replicate quickly became the selection criteria for the population of molecules to obtain dynamic kinetic stability.

A logical next step in the *in vitro* evolutionary process was to develop a criteria by which a desired subset of molecules, the “winners” could be separated from the “losers”. A reasonable criteria for this was selection of molecules based on the ability of certain members of the molecular population to interact with other molecules in the reaction vessel. A selection criteria as such was accomplished by selecting for a sub-population of RNA molecules that were capable of binding to an organic dye (Ellington and Szostak, 1990). Because only a very small fraction of the RNA molecules of random sequence would be capable of performing the required selection task, sequence space (the total number of sequences possible) needed to be explored to as great a degree as feasible.

Having introduced the concept of selection criteria into *in vitro* evolution experiments, using the ability to perform catalysis as a selection criteria made *in vitro* evolution a valuable tool for expanding the emerging RNA World hypothesis as catalytic

function of RNA molecules is a core requirement for demonstrating its potential validity. Catalytic activity was first used as a selection in a population of *Tetrahymena* ribozymes that contained a subset of members capable of catalyzing a sequence-specific cleavage of an RNA molecule (Beaudry and Joyce, 1992). Here, a selection constraint of being able to perform DNA cleavage under physiological conditions was imposed to select molecules for amplification, with mutations introduced to maintain heterogeneity in the population. After multiple generations, a small number of mutations became fixed in the population as the catalytic capability of the system increased. Thus, the course of *in vitro* selection had been directed toward a specific goal.

At this point, however, the usefulness of *in vitro* evolution in elucidating plausible methods by which lifeless molecules may have attained living function was severely limited by the requirement of researcher intervention (by means of serial transfers) was required to generate the ultimate stable populations. Obviously, to create a model that may accurately depict molecular evolution, this obstacle needed to be overcome. A process that fulfilled this requirement was later demonstrated using RNA molecules capable of catalyzing their own ligation to an RNA-DNA chimeric substrate (Wright, 1997). Here, all the catalytic requirements for the amplification of successfully replicating members of a diverse population were provided in one test tube – with serial transfers occurring only to give the population sufficient space necessary to continue growing toward dynamic kinetic stability. This process, known as continuous *in vitro* evolution was established as a valuable method to produce catalytic molecules that could be perpetuated indefinitely.

While inarguably valuable, the continuous *in vitro* evolution process still holds a significant limitation when being considered as a model for prebiotic chemistry. A reaction vessel that contains all components necessary for moving from a collection of simple RNA molecules to a collection of RNA molecules capable of rapid and efficient catalysis requires a host of components, such as RNA polymerase, DNA ligase, and reverse transcriptase, to be present in the reaction vessel. In other words, to get at the origin of biomolecular evolution, a system such as recombination – i.e., one that had been shown to be capable of acting as its own catalyst without the need for allomolecular catalytic agents needs to be considered.

Recombination is the swapping of genetic information between two sources and has been argued as an integral component for the establishment of the RNA World (Lehman, 2003; 2008). Although recombination serves multiple purposes for the success of early biomolecules, perhaps the most important aspect of recombination is its ability to take small biomolecules and combine them into larger biomolecules.

Recombinase ribozymes have been shown to be capable of catalyzing their own excision out of larger biomolecules *in vivo* (Kruger et al., 1982). A recombinase ribozyme provides a good model system for studying the effect of dynamic kinetic selection on a biomolecular population because the recombined products have been demonstrated in the Lehman laboratory to act as catalysts for the formation of subsequent recombined products (Hayden and Lehman, 2006). Thus, modern enzymes are not required as catalysts – all the ribozyme needs for recombination is its own components and the aqueous solution of divalent cationic molarity and pH that was plausible on prebiotic Earth.

Group I introns are genetic elements that, *in vivo*, interrupt functional genes by catalyzing their own excision from mRNA, tRNA, and rRNA precursors (Nielsen and Johansen, 2009). The core secondary structure of all Group I introns is similar; their core secondary structure consists of nine paired regions, denoted as P1-P9 that fold into two domains, the P4-P6 domain and the P3-P9 domain (Woodson, 2005). The P3-P9 domain contains the binding site for the guanosine substrate (Michel et al., 1989), which is formed by a stack of base triplets (Woodson, 2005). Because of the structural similarity of group I introns among several species, in addition to their ability to catalyze their excision from longer RNA sequences, group I introns provide a latent biomolecule that may be able to shed light on the catalytic potential of prebiotic precursor biomolecules.

An excellent member of group I introns for study is the *Azoarcus* Group I intron, which *in vivo*, is a ribozyme capable of auto-excision from the isoleucine transfer-RNA molecule in the *Azoarcus* genus of nitrogen-fixing bacteria. The *Azoarcus* ribozyme has been shown to be capable of self-assembly from smaller fragments without an outside catalyst (Hayden, 2006). The *Azoarcus* ribozyme has been shown to recombine from as many as five fragments (Jayathilaka and Lehman, 2018), although the research discussed herein involves the self-assembly of four fragments, which has been the most common method of fragmenting the biomolecule in the majority of the work done on the intron to date (Figure 1). The *Azoarcus* ribozyme is also of particular relevance to the research at hand because its structure is known at the highest detail of all Group I introns (Adams et al., 2004). As shown in Figure 1, each of the four fragments is designated from 5' to 3' as W, X, Y, or Z. Throughout this document, fragments that have been subjected to a 9%

mutation rate per position are designated by * and fragments that have been prepared per a modified PCR protocol designed to induce random mutations are designated as **.

Self-assembly relies on recombination events rather than mere ligation reactions. For self-assembly to proceed, a tag sequence, which is a CAU triplet (the “tag”) that acts as a recombination signal, must be present between the ribozyme’s fragments. The tag interacts with the internal guide sequence (IGS), which in the wild type is the sequence 5’-GUG-3’. (Figure 2). The tag, along with the two guanosine nucleotides immediately upstream form a head region (5’-GGCAU-3’). Thus, the fragments, in conjunction with their head regions are denoted as W, hX, hY, and hZ. Recombination will occur spontaneously as long as all four fragments are present in either equimolar concentrations (or concentrations varied to favor recombination), the 5’-end of the sequence bears the IGS, and the X, Y, and Z fragments bear the tags. When these conditions are met, the fragments first assemble into *trans*-complexes, which refer to the four fragments involved in non-covalent secondary and tertiary interactions. These non-covalent networks are capable of a rudimentary catalysis that joins the fragments into covalently-joined *cis*-complexes. The *cis*-complexes are more effective catalysts and take over the lion’s share of catalysis when their presence increases in the reaction vessel (Hayden and Lehman, 2006; Vaidya et al., 2012).

Having established a foundation for the rationale behind the design and implementation of the protocols used to guide the research described in this dissertation, as well as the suitability of the *Azoarcus* group I intron as a suitable model ribozyme, the question of what conclusions may be drawn from the data still needs to be addressed.

Unfortunately, there is inherent difficulty in the formal quantification of DKS and this difficulty manifests at several levels (Pross, 2011). In the case of directly-competing replications, because DKS is achieved through kinetic rather than thermodynamic factors, the competitive outcome can be significantly influenced by minor variations in reaction conditions (Saffhill et al., 1970). As a result, the actual magnitude of DKS for any replicating system is highly circumstantial and therefore not readily amenable to meaningful quantification (Pross, 2011).

Because DKS is a relatively new approach toward attempting to describe the transition from chemical molecules to biomolecules, the researcher wishing to describe the effect of DKS can borrow from multiple disciplines to guide these efforts. Because the introduction of random mutations has introduced diversity into the molecular system, the reaction mix has become a biomolecular “population” composed of diverse members. Thus, population biology is good initial reference point to look for long-term trends that may lead to the success of one species (or biomolecule) over another in a competitive environment. The concept of species selection suggests that species that endure the longest and generate the most offspring determine the direction of major evolutionary trends (Stanley, 1974). Further, it is possible that qualities unrelated to the overall success of organisms in specific environments may become important when entire species are selected. In a biological context, a factor that lead to the success of a species is its ability to act as an adaptable, opportunistic, wide-ranging “weed” species versus being a successful specialist in a narrow niche. As an example, irregularly elevated female survival and the adoption of social monogamy may have given early hominids the ability to act as

weed species as opposed to competing homonids that were not capable of sustaining sufficient population growth when invading novel environments (Meindl et al., 2017). Thus, in addition to looking at thermodynamic factors that may lead to the dominance of certain genotypes of intron over time, the results described in this thesis will also identify members of the molecular population that may be quick assemblers and act as the “weeds” to pave the way for the long-term dominance of the better catalytic molecules.

The research discussed herein was designed to address the relevance of self-assembly in demonstrating the effect of dynamic kinetic selection on a population of primitive biomolecules. By introducing random mutations into one or more fragments, a diverse population of biomolecular precursors was generated. Without human intervention, the substrates were allowed to recombine into larger analogs of the wild type model. Aliquots from the reaction vessel were removed periodically so that the constitution of the population could be analyzed at particular times as the reaction proceeded. In the second and third chapters, the composition of the resulting population from the recombination of differing varied fragments is analyzed.

Because each reaction described in Chapter 2 and Chapter 3 of this dissertation contained a diverse population of substrate fragments, the full-length recombined product recovered from each time-point aliquot reflected population members that were self-assembled from substrates well-suited for recombination. As the reactions proceeded, the full-length products present were either good substrates for recombination, good catalysts for the self-assembly of other substrates, or both. DKS suggests that at early time points in the reaction, the system will select for substrates that are well-suited for quick self-

assembly, with the balance shifting toward good catalysts as the reaction proceeds over time. It is important to note, however, that the presence of at least a minimal amount of catalytically-capable full-length product is required at all times so that the reaction continues to progress.

The initially-proposed work was comprised of three individual specific aims, each of which sought to answer a fundamental question within the field of molecular evolution. These specific aims were used to generate the hypotheses that the research was designed to probe. In the course of performing the research, each specific aim was refined to generate the main topics addressed in each thesis chapter.

The hypothesis that dynamic kinetic selection is operative during the *in vitro* evolution of RNA in a laboratory setting is tested in Chapter 2. As described previously (Pross, 2012), dynamic kinetic selection suggests that the rise and fall of transient members of an evolving population operates under kinetic, rather than thermodynamic, constraints, i.e., molecules that can form the most rapidly tend to be dominant earlier in a chemical reaction. Although the kinetic stability of a population is not readily quantifiable (Pross, 2011), the concept of “weedy” versus “old growth” species used in population biology to describe how the composition of a population or an ecosystem that competes for resources may change over time. Given the trend to move toward a stable population consisting of members that fill a niche (Stanley, 1974), it is expected that there will be a decrease in the wild-type population, but there will be a compensating rise in genotypes that are capable of filling a niche role within the population (in this case, the ability to catalyze the formation of a full-length ribozyme). Substrate genotypes that can rapidly form full-length

product may arise quickly, but those without catalytic competence will ultimately be replaced by population members that efficiently perform that niche role. Certain random mutations that may have been incorporated into the substrates may disrupt the intramolecular interactions required to accomplish the appropriate tertiary structure required for effective catalytic activity, which frees up these ribozyme fragments for rapid self-assembly. Thus, the balance between the ability to assemble quickly versus the ability to be an effective catalyst tips in favor of the effective catalysts as the reaction selects for genotypes that will be maintained in the molecular population.

The emerging hypothesis in molecular evolution that cooperation will outperform selfishness during the formation of RNA networks will be tested in Chapter 3. In selfishness, two fragment substrates of differing genotypes that are capable of catalytic interactions with fragment substrates of the like genotype are pitted against one another to act as catalysts for recombination reactions with a common substrate to determine which substrate most effectively catalyzes the formation of full-length product. In the cooperative system, two fragment substrates of differing genotypes must work in concert with one another to accomplish of the full-length product. For practical considerations, the ribozyme substrate fragments used in the first specific aim and the second specific aim have been prepared by different methods of incorporating random mutations. Because all of the substrate fragments used in this research are based on the wild-type *Azoarcus* group I intron, the positions which are more tolerant of accepting random mutations is expected to be the same, regardless of preparation method and there should be no little or no difference observed in the tendency to catalyze the formation of full-length product.

The hypothesis that point mutations that may become fixed during the course of dynamic kinetic selection will have a negligible effect on the overall catalytic rate will be tested in Chapter 4. At the same time, the hypothesis that point mutations that do not become fixed over the course of time likely have a deleterious effect on the catalytic capability of the ribozyme will also be tested. In this work, point mutations that recur in multiple populations or that rise to high frequencies are manually inserted into otherwise wild-type RNA fragments and then allowed to react with wild type fragments to form full-length, covalently-joined products. The percent yield of full-length product is determined and compared with that of the analogous purely wild-type reaction.

Taken together, this research helps to fill the gap in the understanding of the roles of thermodynamic stability and dynamic kinetic selection in determining the constitution of the members of a continuously-evolving molecular population, as well as adding to the growing body of knowledge surrounding the possibility of recombination networks providing the support necessary for the genetic diversification of nascent pre-cursor biomolecules being capable of evolving into living biomolecules.

Figure 1 – The *Azoarcus* group I ribozyme. The secondary structure is adapted from that presented as in Hayden and Lehman (2006). The molecule is partitioned into four fragments (**W**, red; **X**, yellow; **Y**, blue, and **Z**, green). The regions in the gray boxes have been modified from the wild type intron and fixed as the sequence CAU in order to act as recombination signals. The ribozyme begins at the internal guide sequence (IGS) which encompasses positions 9-11. In the wild type molecule, the IGS sequence is GUG. The IGS binds the RNA substrate (lower case letters and curved black line) so that the covalent intermediate necessary for recombination may be formed. The catalytic core is indicated by the purple arrow and refers to a bulge in the RNA formed by positions 127, 128, and 129 in the Y-region. Positions 127 and 129 form base pairs with the cytosine nucleotides at positions 176 and 177.

Figure 2 – Schematic of the IGS-tag interaction involved in the catalysis of recombination. The role of the catalyst in recombination is adapted from Satterwhite and Lehman (2016). The molecules depicted in blue and green represent two fragments involved in a recombination event. The molecule depicted in gray is the catalyst and represents either the four members of a *trans*-complex or the full-length, covalently joined molecule. The IGS at the 5'-end of the catalyst interacts with the tag sequence at the 3'-end of the substrate involved in the nucleophilic attack (here, a WXY fragment) by the 3'-hydroxyl group between the guanosine nucleotides on the head group of the substrate being attacked (here, an hZ fragment). The catalyst correctly positions WXY for nucleophilic attack while the G-binding site (not shown) and interactions between the two substrates work to position the hZ-fragment.

CHAPTER 2: RECOMBINATION OF A CLASS I RIBOZYME FROM SUBSTRATE FRAGMENTS PREPARED AT A 9% MUTATION RATE

Introduction

Self-assembly of a group I intron through the recombination of shorter, inactive oligonucleotide fragments (termed W, X, Y, and Z, each approximately 50 nucleotides in length) has been shown previously with the *Azoarcus* ribozyme (Hayden and Lehman, 2006). The fragments have no activity individually, yet can cooperate through base-pairing and tertiary interactions to produce stable *trans* complexes that are capable of performing ligation and splicing reactions on exogenous RNA substrates and can thus promote the assembly of fragments complementary to the *Azoarcus* ribozyme itself. Recombination is an ancient evolutionary development that allows for both the development of molecules of substantial length and the maintenance of an inexorably increasing mutational load (Lehman, 2003).

The approximately 200-nucleotide in length *Azoarcus* ribozyme (denoted as WXYZ when described in its full-length form) has been shown to self-assemble at each recombination junction through either of two methods (Draper et al. 2008). In the *R2F2* mechanism, the ribozyme catalyzes two successive phosphotransfer reactions on exogenous RNA substrates resulting in a recombinant RNA molecule (Figure 3). The *tF2* mechanism, by contrast, involves the binding of duplex RNA and catalyzes a cross-strand attack resulting in a short insertion that is typically some or all of the head portion (GGCAU) of the 3'-substrate (Figure 3). When the substrates are designed such that either

recombination method is available, the *tF2* mechanism tends to dominate at the XY junction, whereas both mechanisms operate at the WX and YZ junctions. For the reactions described in this study, both the wild-type and mutant versions of W, X, and Y substrates were designed to exclusively use the *tF2* mechanism. This was helpful during data analysis because the *tF2* mechanism introduces significant variation in recombination, which serves as a “marker” to suggest that recombination has successfully taken place. However, although the *tF2* mechanism has been shown to preferentially insert GCAU at the junction, it can also insert CAU or GGCAU. Additionally, the *tF2* mechanism often makes the insertion multiple times during one recombination event. (Draper et al, 2008).

X-ray crystallography is a technique used to determine the atomic and molecular structure of a crystal structure by using electron density to determine the mean positions of atoms and chemical bonds within a molecule. As biomolecules can be readily induced to form crystals, it is a powerful tool for probing the secondary and tertiary structures of proteins and nucleic acids. The technique was demonstrated to be of value in the study of RNA when it was successfully used to elucidate the nature of the structural domains of the catalytic core, in addition to long-range interactions that provide molecular stability in *Tetrahymena*. (Cate et al., 1996). X-ray crystallography was subsequently used to describe the tertiary hydrogen bonding and stacking interactions between the domains and joiner segments of the *Azoarcus* ribozyme (Adams et al., 2004). The interactions described by the *Azoarcus* RNA crystal structure will be used in this study as one possible explanation for positions within the molecule that tend to either accept or disallow nucleotide substitutions at the position in question.

Another powerful tool in nucleic-acid evolutionary studies is nucleotide sequence analysis. In recent years, the sequencing industry has been dominated by Illumina, who have adopted a sequencing-by-synthesis approach known as MiSeq. (Quail et al., 2012). During the Illumina MiSeq process, single molecules of DNA are attached to a flat surface, amplified *in situ* and used as templates for synthetic sequencing with fluorescent reversible terminator deoxyribonucleotides. Images of the surfaces are analyzed to generate high-quality sequences (Bentley et al., 2008).

The preparation of a synthetic gene was accomplished in 1970 through a process involving the binding of single deoxyribonucleotide to a solid support followed by the strategic use of protecting groups to prevent undesirable side reactions from occurring as subsequent nucleotides are added to the growing polynucleotide chain (Agarwal et al., 1970). The basic process, improved in 1983 by the use of phosphoramidite monomers and tetrazole catalysts remains the basis of modern oligonucleotide synthesis (McBride and Carruthers, 1983). In order to grow the chain in the prescribed order, each of the four nucleobases is kept separate and an automated selector selects a nucleobase from the appropriate vial. In order to introduce random mutations, rather than being fed only the wild-type molecule at the position, the four vials that the selector chooses from are modified in their composition. For the set of reaction described in this chapter, the substrates were desired to contain a 9% mutation rate, so each selection vial contained 91% of the prescribed, wild-type nucleotide and 3% each of the other three bases.

This chapter describes the results obtained from a series of five reactions (Figure 4). These reactions will be referred to collectively as the 9% reactions throughout this

thesis. In these self-assembly reactions, the fragments of the *Azoarcus* ribozyme were prepared according to a protocol that allowed for a 9% mutation rate at each position. These randomized RNA intron fragments were allowed to recombine either with other fragments prepared in the same manner or with wild-type fragments prepared by run-off transcription, with the ultimate goal being to isolate those fragments that had recombined to produce full-length RNA. Aliquots were removed from the reaction pot at pre-determined intervals. From each aliquot, the full-length product was isolated, purified, and amplified via specific RT-PCR. The nucleotide sequences of these ribozymes were then determined using Illumina MiSeq high-throughput sequence analysis to detect trends over time. Here, the hypothesis that dynamic kinetic selection is operative during the *in vitro* evolution of RNA in a laboratory setting will be tested. An analysis of the genotypes of full-length product obtained from six time points in each of five reactions is expected to show that over time a certain number of “fixed” mutations will appear as the reaction proceeds. Other genotypic mutations are expected to be transient and will not become fixed in the ultimate population.

The conserved regions for group I introns are well-studied across a number of the available sequences for group I introns occurring in various genomes (Michel & Westhof, 1990). Here, several widely conserved regions among 87 available Group I sequences were identified. Notable positions and regions included include the P1 and P10 regions, which comprise the internal guide sequence in most, but not all, group I introns. Additionally, the terminal loop of stem P2 can contribute to the positioning of P1 by interacting with the P8 helix (Michel & Westhof, 1990). The IGS and tag sequence

interaction occurs in the same manner in the *Azoarcus* intron. In the *Azoarcus* intron, this latter interaction involves positions 23-26 at the end of the P2 stem interacting with positions 162 and 163 in the P8 stem (Adams et al. 2004). Because conservation in these regions is ultimately required for the intron to successfully splice itself out of a longer RNA sequence, it is assumed that substrates that will result in these same regions will be selected during self-assembly and will be highly-conserved as each reaction proceeds. Conversely, mutations that disrupt these conserved regions are expected to be selected against, or at least occur transiently in population members. It is important to note however, that the impact of conserved regions among group I introns has been described extensively in the context of their potential effect on the self-splicing reaction rather than any potential effect they may have on self-assembly from fragments. Thus, the known secondary and tertiary structures of group I introns, in conjunction with the known secondary and tertiary structure of the *Azoarcus* intron will be used as a guide in the consideration of why certain mutations may or may not impact the catalytic capability of full-length product.

Experimental Design

Recombination of the *Azoarcus* ribozyme becomes more challenging the more substrate fragments that need to combine in order to form full-length product. As a result, the data observed subsequent to recombination events was analyzed from iterations of the reaction that increased in complexity and percentage of the final product that was derived from substrate fragments containing a mutation.

The reaction of the four wild-type fragments is robust, whereas introducing even one substrate prepared with a 9% per position mutation rate causes a dramatic decrease in

product yield (Figure 4). To attempt to compensate for this, a series of five reactions was performed (Figure 5). The first reaction in the series included only two substrates, a wild-type WXY fragment which was allowed to recombine with a 1.25-fold excess of Z* substrate (the “Z*-Single”). For the next reaction, the region subjected to the same random mutation rate was moved to the 5’-end of the molecule and W* was allowed to react with a 1.25-fold excess of wild-type XYZ fragment (the “W*-single”). In order to be able to provide a direct comparison between mutated regions, the third reaction (the “Double”) involved three substrates, W*, wild-type XY (in a 1.25-fold excess over W*) and Z* (in a 1.25-fold excess over XY). The fourth reaction (the “Triple”) involved W*, X* (in a 1.25-fold excess over W*), wild-type Y (in a 1.25-fold excess over X*), and Z* (in a 1.25-fold excess over Y). The fifth and final reaction (the “Quadruple”) consisted of all four mutated substrates, each substrate in a 1.25-fold excess over its adjacent 5’ neighbor. (Figure 5). The addition of every 3’ strand in a 1.25-fold excess over its 5’-binding partner has been shown to optimize direct ligation by driving the reaction further toward the formation of full-length ribozymes (Mutschler et al., 2015).

Recombination in one tube was mediated by the *Azoarcus* ribozyme at 48 °C using five reaction designs (Figure 5). Each reaction took place in a buffer that was 30 mM in EPPS pH 7.5 and 100 mM MgCl₂. Each substrate was added in the presence of water and the resulting mixture was incubated at 80 °C for two minutes. These conditions have been demonstrated to optimize the ability of *Azoarcus* to undergo recombination through self-assembly (Hayden and Lehman, 2006). Immediately after the 80°C incubation, the $t = 0$ time point was removed, EPPS and MgCl₂ were added to bring the molarity of the aliquot

to 30mM and 100 mM respectively, and an equal volume of 2X quench solution (280 mM EDTA, pH 7.5) was added, with the resulting solution being stored immediately at -20 °C. The main reaction vessel was placed in a dry bath and incubated at 48°C. In each reaction, time point aliquots were taken at 30 minute, 1 hour, 2 hour, 4 hour, 6 hour, and 8 hour intervals and the 2X quench solution was added to that the reaction aliquot was ultimately 140 mM in EDTA. Each time point aliquot was incubated in 100% ethanol, purified with washes of 200 µL of 80% ethanol and 100% ethanol (“ethanol precipitation”) and rehydrated to a volume of 18 µL. A 6X solution of polyacrylamide loading dye containing formamide and urea was added and the reaction time points were size partitioned by 8% acrylamide/8 M urea gel electrophoresis.

For the Z*-Single and W*-Single reactions, ample reaction product was present on the gel that visible bands that could be seen when the gel was exposed to ultraviolet shadowing, were excised from the gel, with time points that could not be seen with the naked eye being extrapolated by lining up the column on the gel with both a size control and the presence of product at later time points. For the Double and Quadruple reactions, no product was visible on the polyacrylamide gel upon UV shadowing and thus the gel position of reaction product was determined exclusively by lining up the column for each reaction time point with the position of a small amount of known full-length product. Product bands and the gel band corresponding to the full-length size control (if used) were removed from the gel with a razor blade and then incubated overnight in 500-600 µL of a solution that was 300 mM in sodium acetate and 100 µM in EDTA and subjected gentle shaking at 4°C.

The reaction product obtained from soaking the gel cubes was subjected to ethanol precipitation, rehydrated to 11 μ L and used as substrate for a reverse transcription reaction that contained 100 U of SSII reverse transcriptase (Thermo Fisher), 4 μ L of a 5X proprietary reaction buffer and was 2.5 mM in dithiothreitol (DTT), 20 mM in each deoxyribonucleotide, and 10 mM in the reverse primer (5'-CCGGTTTGTGTGACTTTCGCC-3'). One reaction tube was prepared for each time point aliquot from each of the four reactions. The reaction mixture was allowed to incubate at 37°C for 30 minutes and then 5 μ L was added to a standard PCR mixture (described above) to each of four PCR tubes for each time point. The band corresponding to the known full-length size control was amplified in order to serve as a positive control and a PCR tube with no substrate added served as a negative control.

The PCR product of each reverse transcription reaction was tagged with a unique reverse primer (IDT) designed by Bryce Clifton to allow for identification when PCR-amplified reverse transcription products were combined and prepared for high-throughput sequencing. Successful PCR was determined by the presence or absence of bands on 2% agarose gel. Full sets of reaction time points that were determined to be successful were combined, subjected to ethanol precipitation and run on a 1% agarose gel so that bands corresponding to an appropriate length could be separated from bands suspected to be indicative of primer dimer, amplification of reaction intermediates and unreacted substrate, or amplification of truncated reverse transcription products. The bands of appropriate length were cut from the agarose gel with a sterilized razor blade, and purified using a GeneJet Purification Kit (Thermo Fisher, K0702). A total of two tubes containing the

amplified product obtained from each of six time points from the eight reactions was outsourced to the Center for Genome Research and Biocomputing at Oregon State University, Corvallis, Oregon for sequencing. The library pool was run on the Illumina MiSeq instrument with the 2x150bp protocol using the MiSeq Reagent v.2 Micro kit. Prior to sequencing, both sample tubes were quantified using the Qubit fluorometer with the Quant-iT dsDNA HS Assay Kit. Size distribution of the library was checked on the Agilent TapeStation 4200 instrument using the HS D5000 ScreenTape. The quantitative polymerase chain reaction (qPCR) was run using a KAPA Biosystems Library Quantification kit on an ABI 7500 fast instrument.

Results and Discussion

The results obtained from each of the individual reactions involving certain substrate fragments prepared at a 9% mutation rate per position are discussed below. Generally, product yield at was robust for both the Z*-single and W*-single reactions. Product yield was notably decreased in the double reaction, which required self-assembly from three substrate fragments, two of which contained random mutations. The triple reaction produced scant full-length product, while the quadruple reaction did not produce a gel-detectable full-length product (Table 1). For the three reactions that produced quantifiable results, the data was analyzed to determine similar and different trends that emerged among the three reactions. Nucleotide positions that were qualitatively observed to select against the wild-type sequence were investigated to determine whether or not some or all positions demonstrated a tendency to select for one nucleotide substitution over another. In most cases, there was no clear selection preference for a specific replacement nucleotide. For

the Z*-Single, W*-Single, and Double reactions, the overall change in percent of mutations present at each position slowed a slight decrease as the reaction time proceeded from 30 minutes to eight hours. For these same three reactions, the genotypic composition of each reaction population at each time point was considered. For all three reactions, the wild-type was clearly the dominant genotype at every time point.

Finally, the overall average mutation rate (μ) at each nucleotide position and the average number of mutations away from the wild-type in any given RNA molecule (k) was calculated for each time point aliquot within each reaction. In the single reactions, the average number of mutations away from the wild-type tended to decrease as the reaction proceeded, and in the Double reaction, it steadily increased over the duration of the reaction. Generally, a higher k value was observed for full-length molecules that had been self-assembled from fragments bearing a higher number of positions with the 9% mutation rate per position. The results from the three reaction series are compared and contrasted below.

Z*-Single (WXY + Z*)

Robust results were obtained from all of the time points sampled. Approximately 100,000 successful hits from high-throughput sequencing resulted for each of the six time points. The raw data for each time point was pared down to 10,000 hits per time point as described above. Because of the wide variety of products observed subsequent to a *tF2* recombination event, for ease in data processing, positions 155 and 156 were used as an anchor to determine the beginning of the Z-portion of the fully assembled molecule. These two positions were “forced” to agree with the wild-type base, with any disagreement

between the tag sequence at the 3'-end of Y and the forced positions considered to be part of the recombination event. Thus, data from positions 155 and 156 were not quantified in the Z*-single reaction.

The percent mutation at each position, adjusted for noise was determined (Table 2A). Three positions consistently demonstrated a tendency to select against the wild-type: 167, 170, and 177. Neither position 170 nor position 177 showed a significant tendency to select one nucleotide over another, while position 167 tended to select against cytosine (Table 2B). Position 170 is of particular interest because it has recently been shown to be part of a recombination junction involved in self-assembly from five substrate fragments (Jayathilaka and Lehman, 2018). Thus, a possible explanation for an increased variability at this position may be a natural tendency to avoid the potential for introducing extra recombination junctions.

The crystal structure of the *Azoarcus* intron identifies the adenines at positions 166 and 167 as highly conserved because their ability to make type II and type I A-minor contacts with the IGS (Adams et al., 2004). The A-minor motif is an abundant and ubiquitous structural motif that stabilizes RNA tertiary and quaternary structure by the insertion of the smooth, minor groove edges of adenines into the minor groove of neighboring helices, preferentially at G-C base pairs, where they form hydrogen bonds with one or both of the 2'-hydroxyl groups of those pairs. Type I and II interactions are both adenine-specific (Nissen et al., 2001). Because of the importance of this interaction to successful catalysis, it is likely that full-length molecules that bear the wild-type nucleotides at these positions remain involved in catalytic activity, while variants are free

to act as substrates that form either non-catalytic products, or products with greatly diminished catalytic activity. Positions 176 and 177 comprise a pocket that acts as the G-binding site during catalysis (Hayden and Lehman, 2006), and compositional changes at these important positions will be discussed subsequently in this section.

Dynamic kinetic selection suggests that the population should gradually shift toward a population composed of biomolecules capable of performing a niche function wild-type as reaction time progresses (Pross and Khoddorkovsky, 2004). In this case, the niche function would be the ability to serve as a strong catalyst for the formation of other full-length ribozymes. For 20 of the 28 positions analyzed in the Z* fragment, the average percent of products bearing a mutation for each nucleotide position showed a modest decrease when averaged over the 30-minute to 8-hour time interval (Table 2C), although the decrease is modest at all 28 of the positions. Position 170 shows an unexplained spike in mutation rate at the t= 4 hour time point. Although this data point must be accepted at a 95% confidence interval, it skews the trend in the increase in mutation rate per hour higher than other data points for this position suggest appropriate.

The top 20 most prevalent sequences were determined at each time point (Table 3). For each of the six time points, the wild-type was clearly the dominant genotype, while other genotypes bearing one mutation were frequently observed. For example, a frequently-observed genotype was A166G, indicating that this population member carried the wild-type genotype, except that all of its members carried a guanosine rather than an adenine at position 166. At the 30-minute time point, this genotype was the 14th-most prevalent population member. At the 1-hour time point, it jumped to third, then proceeded

to move to fourth, tenth, second, and ninth at the remaining four time points (2h, 4h, 6h, and 8h). These tendency for the prevalence of each niche population member to fluctuate over time necessitated a strategy to be developed in order to determine the upper-echelon of the frequently-observed genotypes that should constitute the niche population (Table 4). The parameters for determining the niche population is described in the Materials and Methods section later in this chapter. The term “niche population” is used to define this subset of ribozymes; each member of the niche population is said to bear a “niche genotype.” In the Z*-single, each of the non-wild-type prevalent sequences contained one (and only one) mutation, and positions that tended to be more tolerant of mutations (e.g., 167, 170, and 177) appear in these prevalent one-mutation variants. For position 170, which has been shown to be a part of an alternate recombination junction, two of the three variants contain niche genotypes with the third variant, C170G just outside of the criteria for niche population, receiving an N-score (see Materials and Methods for this chapter) of 67. Similarly, positions such as 162, 163, and 174 that show a relatively low percent-mutated at each position do not appear in the niche genotypes. Positions 162 and 163 are both guanosine nucleotides that base pair with cytosine nucleotides toward the 3'-end of the Y fragment and are in close proximity to a region of the intron in which the W, Y, and Z fragment interact (Figure 1). Position 174 is involved in a base-pairing interaction near positions 176 and 177, which participate in the formation of the G-binding site. The percent-mutated at position 175 is also relatively low (Table 2A) and that position remains wild-type in all members of the niche genotypes for the Z*-Single reaction. Because of

the role that these nucleobases perform in maintaining the stability of the catalytic site, it is not surprising that these results suggest that they are relatively highly conserved.

W*-single (W* + XYZ)

An analysis parallel to the analysis done on the Z*-single reaction was performed on the W*-single reaction. Again, with the majority of the full length product being obtained from a substrate that contained no random mutations, reaction yield was expected to be robust and the raw data supported this conclusion. Successful hits from high-throughput sequencing were almost 100,000 for each of the six time points (Table 1). In this reaction, recombination occurs at the junction between the W and X fragments. The raw data showed a wide variation in the composition of the sequences inserted at the recombination junction, as expected from intron fragments designed to recombine via the *tF2* mechanism (Draper and Lehman, 2008). Because the region of interest in this molecule is at the 5'-end, the sequences were only aligned with respect to the W-region of the molecule, which terminates at position 68, immediately before the recombination tag (CAU) that has been engineered into the W* substrate fragment in order to allow recombination to occur. To facilitate data processing, the final position considered in the W* fragment was considered to be wildtype. As a result, position 68 is not considered in this analysis and the mutation rate for position 67 may be reported slightly low.

The percent mutation at each position, adjusted for noise was determined (Table 5A). Qualitative review of the data showed seven positions that noticeably selected against the wildtype. Positions 27, 30, 32, and 59 demonstrated a nearly-even distribution among the three possible alternate nucleotides, whereas positions 28, 54, and 65 selected strongly

against guanosine (Table 5B). The *Azoarcus* intron crystal structure (Adams, et. al, 2004) indicates that the C27 is involved multiple tertiary hydrogen bonding interactions with G163. Positions C27, C28, C30, and C32 form the P2 stem region and are paired with G22, G21, G18 and G16, respectively. All four of these guanosine nucleotides are in the forward primer region and any mutations that may have been incorporated into the manufactured W* substrate fragment would have been covered during amplification.

Stability of the tertiary structure of the molecule is provided by a tetraloop-tetraloop interaction between positions 23-26 and positions 188-191, near the 3'-terminus of the intron. Tetraloops, especially those consisting of the sequence GNRA (N = any of the four ribonucleotide, R = either purine) are important to tertiary structure because they help the ribozyme to fold back on itself by reversing the direction of the phosphodiester backbone and by bringing together distant regions of the ribonucleic sequence by providing sites for RNA tertiary contacts (Correll and Swinger, 2003). Because mutations that may arise in the tetraloop region at 188-191 are covered by the reverse primer during amplification, disruptions to this interaction could not be fully probed. However, because of the importance of these motifs to providing thermodynamic stability to the intron by lowering its total free energy, it is not surprising that positions 23-26 seem to be less tolerant of mutations.

In contrast with the Z*-Single reaction, the average percent of products bearing a mutation for each nucleotide position for the W*-Single reaction showed a modest decrease in all but six of the 44 analyzed positions when averaged over the 30-minute to 8-hour time interval (Table 5C). Again, however, the increase or decrease in percent mutated is small

at all positions. Although it is modest, this trend of the population moving toward a dominance of the wild-type genotype and supports the hypothesis that dynamic kinetic selection is driving the composition of the molecular population.

The top 20 most prevalent sequences were determined at each time point (Table 6). As was seen in the analogous Z*-Single reaction, at each of the six time points, the wild-type was clearly the dominant genotype, although the percentage of the population with the wild-type genotype was lower in the W*-single than in the Z*-single; this is not a surprising result as W* is a longer molecule than is Z*. The idea of sequence space is a way of representing all sequences possible for a protein (Maynard Smith, 1970) and can also be readily applied to nucleotide sequences. The Z* fragment bears 28 positions at which mutations could occur whereas the W* fragment bears 44. With four possible nucleotides at each positions, there are 4^{28} or approximately 10^{16} possible genotypes for WXYZ* full-length product and 4^{44} or approximately 10^{26} possible genotypes for W*XYZ full-length product. Thus, it is not surprising that although the wildtype is clearly the most prevalent genotype present in the population deriving from the W*-single substrate, there is an increased abundance of competing genotype; the W*-Single reaction has a broader sequence space to sample than does the Z*-Single reaction. For that same reason, the prevalence of each niche genotype member is also lower in the W*-Single. Positions shown to tend to generally select against the wildtype (Table 5A) are also well-represented in the niche genotype. As was observed in the Z*-Single niche genotypes, all members comprising the 20 most-prevalent genotypes and all niche genotypes bear only one mutation (Table 4). In contrast to the niche population observed in the Z*-Single reaction,

several members of the W*-Single niche population do bear mutations that disrupt base-pairing interactions in a paired region of the intron. Positions 28 and 30 are located in the P2 region. As mentioned above, positions 24-26 are involved in a GNRA tetraloop that provides tertiary molecular structure. Thus, it is not surprising that these three positions are conserved in the niche population. As will be addressed in Chapter 4, directed point mutations placed within the P2 region have a negligible impact on self-assembly.

The *Azoarcus* intron contains a scaffolding module, which folds rapidly and forms a nearly identical structure even when removed from the context of the rest of the ribozyme (Rangan, 2003). Phenotypic capacitance refers to the ability of a genome to accumulate mutations that are conditionally hidden and only reveal phenotype-altering effects after certain environmental or genetic changes (Hayden et al., 2015). The scaffold module in the *Azoarcus* intron is located between positions 47 and 125 (Cate et al., 1996). Because phenotypic capacitance suggests that this portion of the molecule can be hidden and is less pertinent to catalytic activity, it is plausible that mutations in this region are better tolerated (Hayden et al., 2015).

A pseudoknot is a nucleic acid secondary structure, first recognized in the turnip yellow mosaic virus, containing at least two stem-loop structures in which half of one stem intercalates between two halves of another stem (Reitveld et al., 1982). The *Azoarcus* intron crystal structure describes a single stretch of nucleotides from positions 33-58 that wrap completely around the circumference of the molecule and may bring the stem region containing the IGS into the active site during catalysis. Because of the role that the pseudoknots plays in maintaining an optimal tertiary structure for catalysis, it is not

surprising that mutations appear in only two of the niche genotypes in any of these positions.

Double (W* + XY + Z*)

Analysis of the Double reaction is of particular interest because the full-length product obtained at each time point is the product of self-assembly reactions involving the greatest number of substrates prepared by a protocol designed to generate a 9% mutation rate per position that yielded a sufficient amount of full-length product for analysis.

Full-length product yield at each time point in the Double reaction was markedly lower than that obtained in both the Z*- and W*-Single reactions (Table 1). Whereas in both of those reactions, a robust data set was paired down to facilitate timely analysis, none of the six time points for the Double reaction produced as many as 10,000 hits. As a result, every data point that could be confirmed as corresponding to full-length product was retained for analysis.

The percent mutation at each position in the W* region, adjusted for noise was determined for the Double reaction (Table 7A). In the Double reaction, the positions of interest relative to the W* region of the full-length molecule are identical to those identified for the W*-Single, with two minor exceptions (Table 5A). In contrast to what was observed in the W*-Single reaction, none of these positions in the Double reaction showed a strong tendency to favor mutation to one base over either of the other two possibilities (Table 7B). Unlike the W*-Single reaction, the percent mutation at each position showed a slight increase at all but two positions (Table 7C).

The percent mutation at each position at each position was determined for each position in the Double reaction (Table 8A). The same three positions (167, 170, 177) that were qualitatively identified as having a greater tendency to select against the wild-type showed the same tendency in the Double reaction, but several other positions of interest (158, 161, 173, 176, 182 and 184) were identified for this region in the Double reaction. Position 161 selected strongly against cytosine and position 177 selected strongly in favor of uracil (Table 8B). The percent mutation at each position was split almost evenly between the 28 positions considered; 15 positions showed a slight increase and 13 positions showed a slight decrease in an average of the mutation percent throughout the duration of the reaction (Table 8C).

Because positions 176 and 177 interact with the G-binding site (Figure 1), the possibility of whether or not any compensatory mutations may have arisen between the two positions was considered (Table 9). The percent of fragments that maintained the wild-type at both positions is the highest for all time points in both reactions. In the Z*-Single reaction, there is a small rise in the percent of the population in which both nucleotides are wild-type over time, whereas in the double reaction, there is a significant decline over time. A limitation imposed by the conditions of both reactions affects the ability of the catalytic core to respond to point mutations – the Y-fragment is fixed to be wild-type both times. If a reaction were performed where all sites at the catalytic core could select among a pool of varied genotypes [e.g., W + X + (YZ)**], a comparison of the nature of interactions between 127-177 and 129-176 may likely show a high tolerance for selecting against the wild-type. An increased tolerance for holding one of the positions as a cytosine and

allowing the other to mutate to uracil is shown for all time points in both reactions. Because a G-U wobble pair is the most robust non-canonical base-pairing interaction, it follows that this alteration would be the best tolerated.

The Double reaction provided for a total of 72 positions available to incorporate random mutations. This yields a sequence space of 4^{72} or 10^{43} possible genotypes. While the concept of sequence space proposes that it is not feasible for a protein or a nucleic acid to sample all possibilities provided by sequence space (Maynard Smith, 1970), the significant increase in available genotypes affects the genotypic composition of the resulting full-length product population. Partly because of this dramatic increase in sequence space and partly because of the decreased full-length product yield in the Double reaction, a pool of the 20 most prevalent sequences could not be determined. For all time points, the wild-type was the clear dominant sequence, between one and three sequences that were common to two or three members of the population and the remainder were either unique sequences or many of the genotypes were common to two full-length introns (Table 10). The genotypes of the niche population for the Double reaction are similar to their analogs in the Z*-Single and W*-Single reactions. Positions that tend to select against the wild-type in the two single reactions (e.g., 30, 65, 167, and 170) appear with mutations in the genotypes of the niche population members. In contrast to both of the single reactions, several of the genotypes of the niche population bear mutations at multiple positions. At the 8-hour time point for the Double, only full-length product with the wild-type genotype was prevalent enough to be considered part of the niche population; no clear “runner-up” could be identified.

The prevalence of population members bearing the wild-type genotype was compared for each of the three populations (Figure 6). The Z*-Single had the highest percent of wild-type members at all reaction time points. Because that reaction was the most limited in its ability to sample sequence space, this result is not surprising. A longer fragment was subjected to the 9% random mutation rate per position as in the W*-Single reaction, so it logically follows that fewer population members would bear the wild-type genotype because of this reaction's increased ability to sample sequence space. An apparent contradiction, however, appears in the Double reaction. For the first time point ($t = 30$ min), a striking 12% of the population is wild-type. At the one hour time point, the percent wild-type is nearly the same as the population in the W*-Single reaction, and then at all four subsequent reaction time points, the percent wild-type in the Double population is the lowest of the three. A possible explanation for this discrepancy may lie in the need of the reaction to have an appreciable number of good catalysts to get the reaction started. Because the sample size at the earliest time point in the Double reaction is relatively small ($N = 133$), it is not surprising that molecules outside of the niche population are less prevalent.

None of the three populations shows a linear increase or decrease in prevalence of the wild-type genotype over time. The R^2 values for the Z*-Single, W*-Single, and Double reactions are 0.0179, 0.4729, and 0.3728, respectively. Although the percent of the population bearing the wild-type genotype remains fairly consistent in the Z*-Single, the presence of other niche genotypes begin to grow as the reaction progresses (Table 3). For the W*-Single, the percentage of the population bearing the wild-type genotype is always

markedly lower, but again, the presence of other niche members is seen consistently over the duration of the reaction (Table 6). It is more challenging to make a direct comparison to the Double reaction because a widely different number of full-length sequences were analyzed at each time point. Further, because of increased ability of these reaction substrates to sample sequence space, the niche population is less diverse than those observed in both of the single reactions. However, there is consistency observed in the genotypes displayed by the niche population (Table 10). The results for each of these three reactions suggest that the nature of the ribozyme population is being driven by DKS.

Triple ($W^* + X^* + Y + Z^*$)

A trend has been shown in the reactions discussed thus far is that the yield of full-length intron tends to decrease as the percent of substrate fragments prepared according to the 9% mutation rate protocol increases. Another factor that potentially worked to decrease product yield was the number of recombination events that needed to occur to produce full-length product. For both of the single reactions, only one of the two fragments bore a 9% mutation rate and only one recombination event needed to occur. For the Double reaction, two of the three substrates bore the 9% mutation rate and two recombination events were required. For the Triple, three of the four substrate fragments bore the 9% mutation rate and three recombination events were required. Although the recombination event readily takes place between four wild-type fragments (Figure 1) or even with five wild-type fragments (Jayathilaka and Lehman, 2018), the product yield is appreciably decreased when even one non-wild-type fragment is included (Figure 1). Either, or more likely both, of these reasons contributed to the scant product obtained from the Triple reaction.

A total of four full-length products were obtained from all six of the Triple reaction time point aliquots (Table 11). One ribozyme (4.1) was recovered from the 4-hour time point and three (6.1, 6.2, and 6.3) were recovered from the 6-hour time point. None of the four ribozymes recovered displayed the wild-type genotype; all four contained multiple mutations. Ribozyme 6.2 contains a 10-nucleotide omission at the 5'-end of the Z* fragment. This suggests that the nucleophilic attack that occurred during recombination occurred between G165 and A166 rather than at the head group of the Z*-fragment. This observation is further supported by the lack of a recombination junction between Y and Z for ribozyme 6.2. Some of the positions that selected against the wild-type in this reaction were also seen in the single reactions and in the Double reaction (e.g., 49, 30, 158, and 167).

As shown previously (Figure 1), the G-binding site required for catalysis exists between positions 127 and 129, with the catalytic core being formed through a G-C base pairing interaction between 127-177 and 129-176. Because the Y-fragment in the triple was held as wild-type, the majority of the G-binding site could not, in theory, be affected by the introduction of random mutations. Positions 176 and 177, contained within the Z-fragment, were potentially mutated. None of the four successful products are mutated at either of these positions. While both the Z*-Single and Double reactions were successful in spite of the incorporation of random mutations at either or both of these positions, it is possible that a reduced ability to form the catalytic site could have led to an inability to surmount the obstacle of having to complete three recombination events to form full-length product.

The nucleotide positions in the X fragment that tended to select against the wildtype cannot be compared against results from the single reactions or the Double reaction because X was held as wild-type in all of those reactions. However, for both of the reactions discussed in Chapter 3 one of the two self-assembly substrates was a WXY intron that had been prepared according to a mutagenic PCR protocol. Both of those reactions showed a very strong tendency to select against the wild-type at position 73. Thus, it is not surprising to see that ribozyme 6.1 contains a mutation at that position.

Because the pool of full-length ribozymes is so small from the Triple reaction, it is difficult to draw a wide range of conclusions from its data. However, it is notable that the reaction products displayed evidence of achieving the self-assembly of full-length product via *tF2* recombination. Although the 9% mutation rate per position supports recombination between two and three substrate fragments bearing random mutations, it is likely that the 9% mutation rate reflects too much ability for the reaction to sample sequence space (i.e., there may not be enough wild-type ribozymes present at the earlier time points to catalyze the reaction to proceed at an appreciable rate).

Quadruple ($W^* + X^* + Y^* + Z^*$)

As shown in Table 1 and described in the previous section, no full-length product was obtained from the self-assembly reaction attempted when all four substrates were prepared at a 9% random mutation rate. It is possible that the ability to sample sequence space that this mutation rate provides for (a theoretical 10^{93} of possibilities, although this calculated value is obviously implausible) creates a situation in which it is too unlikely to form a suitable population of catalytically-competent molecules to get the reaction started. Given a 9%

mutation rate per position and a total of approximately 205 positions available to accept random mutations (forward and primer regions are included in this calculation because although mutations to either of these areas cannot be quantified as a result of being amplified through reverse transcription and PCR, these positions will be subject to random mutations when the self-assembly reaction is occurring) only 0.91^{205} or 2.4 femtomoles of each substrate may bear the wild-type genotype. Although the niche populations identified from the single reaction and the Double reaction suggest other population members that may also be catalytically competent, it is not clear how much, if any, catalytic ability these molecules may have. For the Quadruple reaction, the initial amount of substrate ranged from 3000 to 6000 picomoles (Table 12). Therefore, it is possible that a reason for the inability to recover full-length product was a dearth (or complete) lack of catalytically competent ribozyme fragments.

Overall mutation rate (μ)

The overall mutation rate, corrected for noise, was calculated for the Z*-Single, W*-Single, and Double reactions (Table 13). A comparison of the mutation rate per position in the Double reaction with the corresponding position in the appropriate single reaction shows a strong correlation, with the Double reaction having a higher mutation rate at most, but not all positions. Three positions (176, 182, and 184) show a sharp increase in the mutation rate in the Double versus the Z*-Single reaction. As discussed previously, in the wild-type, position 176 base pairs with the guanine at position 129 and thus is involved in providing stability to the catalytic site. While it seems counter-intuitive that the ribozyme could tolerate mutations within the catalytic core, it is possible that when the majority of

the ribozyme bears the wild-type genotype, a sufficient amount of catalytically-competent ribozymes are available to initiate catalysis and the system can support members of the niche population that are variants at either position 176 or 177, even at earlier time points (Table 3). When the ribozyme is given a greater opportunity to sample sequence space, as in the case of the Double reaction, a niche population member bearing a non-wild-type nucleotide at either of these positions does not appear until the 6-hour time point (Table 10). The rise in overall mutation rate at positions 182 and 184 does not have an obvious explanation. Position 182 is contained within the P9.0 helical region and has relatively few tertiary interactions with other regions of the intron (Adams, 2004). Position 182 does not base pair with any other nucleotide and is not contained in a loop region of the molecule (Figure 1). Thus, although it is not surprising that this position selects against the wild-type in the Double reaction, it is surprising that the overall mutation rate is not higher in the Z*-single reaction. Although it is involved in a base-pairing interaction in the P9.0 helical region, the same argument can be made for position 184, as this region does not providing significantly to overall stability and the base pairing interaction (A-U) is a weaker base-pairing interaction.

Average deviation from wild-type (k)

The average number of deviations from the wild-type (k) that each ribozyme in the sample population contained was determined at each reaction time point (Table 14). For this analysis, mutations, insertions, and deletions were also considered. The average k values rise as the percent of positions subjected to the insertion of random mutations rises. With 28 positions available to accept random mutations, the average k value for the Z*-Single

reaction is lowest, with 44 positions available, the average k value for the W*-Single reaction shows an increase and the average k value for the Double is the highest of the three. For the single reactions the average k value decreases over time and for the Double reaction, it increases over time. Although a k value was calculated for the Triple reaction, because data from only four full-length ribozymes was recovered, an overall average k value is considered rather than one obtained from individual time points. That overall average value for the Triple reaction is the highest of the four, but its magnitude is close to that of the Double reaction. While it is important to consider that only four ribozymes were used in calculating this value, the relatively lower average deviation from the wild-type for the Triple does support the supposition that the high ability to sample sequence space for this reaction does limit its ability support population members with genotypes widely variant from the wild-type. In other words, because the catalytic potential of the system is decreased by a decrease in the availability of catalytically-competent ribozymes, fragments that can recombine to self-assemble full-length product must more closely resemble the wild-type genotype.

Conclusions

In this chapter, the results obtained from five related reactions were compared and contrasted. Each of the five reactions contained at least one substrate that had been prepared according to a protocol designed to provide a 9% random mutation rate per position. Of the five reactions, four provided data that could be analyzed and interpreted. The two single reactions provide a good starting point to identify trends that will be maintained as the percentage of positions available to incorporate random mutations

increased. The hypothesis that dynamic kinetic selection will drive the reactions toward forming a stable population was supported by the results from both of the single reactions and the Double reaction. For the single reactions, a shift toward a niche population, in which the wild-type genotype was the major player was observed, as the DKS hypothesis suggests. Mutations that tended to break base-pairing interactions were observed frequently and this disruption to secondary (and possibly tertiary) structure freed these substrates from the burden of catalysis, freeing them up to assemble rapidly to form full-length product variants. For the Double reaction, an initial high percentage of wild-type ribozymes at the earliest time point suggested that an excess of catalytically-competent molecules is required to get the self-assembly reaction started. When catalysis is firmly in place, the reaction both becomes more robust and is positioned to be more tolerant of ribozymes that can assemble quickly, but are less effective catalysts. Although it is neither as diverse nor as robust as those observed for the single reactions, the Double reaction is also shown to establish a niche population.

Results from each of the single reactions could be compared with the analogous mutated region in the Double reaction. The W* regions yielded very similar data. Positions identified as well-suited to accepting random mutations were virtually identical between the two reactions as were positions that showed a low ability to tolerate random mutations. For both reactions, every position of interest in the W* region is represented as cytosine in the wild-type. In the Double reaction, all of the positions that select against the wild-type in the W*-region broke canonical Watson-Crick base-pairing interactions. For the Z* regions, the same positions that were well-suited to accept mutations in the single

reaction, but several more emerged in that region for the Double. In the Z^* region, mutations tend to break up base-pairing interactions, but this trend is not as strong at this end of the ribozyme as it is near the 5'-end. For the Double reaction, the G-binding site is hit particularly hard by random mutations. Common to all three of these reactions, although some positions prone to mutations tend to disfavor the substitution of a particular nucleotide, the general trend is to nearly evenly accept any of the other three possibilities.

The Triple reaction produced extremely modest results and the Quadruple reaction yielded no observable full-length product at all. It is likely that the reaction conditions allowed the substrates to so fully sample sequence space that a sufficient number of catalytically-competent molecules were not available to adequately catalyze the reaction.

A t test was used to determine whether there was a significant difference between the means of the average k values for the Single and the Double reactions. The calculated value, $t = 1.18$ is below the critical value at a 95% confidence interval. Thus, there is no statistical difference between the two reactions ($P = 0.27$).

Materials and Methods

Construction of RNA substrate ribozymes

The RNA substrate for any individual fragment, either wild-type or mutant was purchased from one of two vendors. The W^* pch substrate was purchased from TriLink Biotechnologies while the hX^* pch, hY^* pch, hZ -pch, and hZ^* -pch fragments were purchased from Integrated DNA Technologies (IDT). The W^* pch fragment was designed such that nucleotide positions corresponding to the 5' forward primer

(GUGCCUUGCGCCGGG) were fixed, while the remaining nucleotides in the fragment were added such that there was a 91% chance that the nucleotide at every position would be wild-type and a 9% chance that the nucleotide would have an equal chance of being one of the three non-wild-type nucleotides. The hZ-pch substrate was fixed as wild-type with the 5'-GGCAU-3' head group added. The hX*-pch, hY*-pch, and hZ*-pch substrates combined both of the strategies described above; the fixed head region and with a 91% chance of each position being wild-type and equal chances of an alteration to either of the three possible non-wild-type nucleotides (Table 15). The XY and WXY substrates were transcribed from double stranded DNA templates prepared by PCR mediated gene synthesis (Jayaraman et al., 1989) which used a proprietary 10X reaction buffer (New England Biolabs), and was 200 μ M in dNTPs (Thermo Fisher), 20 μ M each in six oligonucleotides designed to create the full-length wild-type *Azoarcus* Group I intron, and contained 0.2 U Vent DNA Polymerase (New England Biolabs). An amplification protocol consisting of an initial denaturing at 94°C, followed by eight amplification cycles consisting of 94°C for 1.5 minutes, 54°C for 2 minutes, and 72°C for 3 minutes, concluding with a final elongation at 72°C for five minutes was conducted in a T100 Thermocycler (Bio-Rad). The resulting product was amplified using a standard polymerase chain reaction (PCR) protocol, which used a 10X PCR buffer (10 mM Tris pH 8.0, 50 mM KCl, 1.5 mM MgCl₂), 200 μ M dNTPs (Thermo Fisher), 10 μ M each in forward and reverse primers (IDT) and 20 U *Taq* DNA Polymerase (Thermo Fisher). A previously-optimized PCR protocol consisting of 25 cycles of 92°C for 1 min, 57°C for 1 min, and 72°C for 1 min was conducted in the Thermocycler, with presence of PCR product being confirmed

by electrophoresis on 2% agarose gel. Transcription of oligonucleotide substrates was carried out using 10X transcription buffer (15 mM MgCl₂, 50mM Tris pH 7.5, 5 mM DTT, 2 mM spermidine), 200 μM rNTPs (Thermo Fisher), 2 μL ten-fold concentrated PCR DNA, and 400 U T7 RNA polymerase (Thermo Fisher) in a 200 μL volume for 2-16 hours at 37°C. The resulting RNA was purified on an 8% polyacrylamide/8 M urea gel.

Data Analysis

Raw data was received in fastq files, with two files containing data related to each of the molecules read for each reaction time point. The “left” file contained reads related to the first 150 nucleotides read from the 5’ end and the “right” file contained the reverse complement of reads of the first 150 nucleotides read from the 3’ end. Because each successful reaction product was expected to be approximately 200 nucleotides long, there was an overlap region of between roughly 70 and 100 nucleotides, with discrepancies in the amount of overlap resulting from a varied range of insertion that occurred during pinch recombination. An in-house processing program was written by Benedict Smail that combined the left and right files into one file on the basis of the computer finding a complete match between the overlap regions of the left and right sides. Because sequencing accuracy on the Illumina platform decreases with an increasing number of read cycles, less reliable base calls are produced toward the 3’-end of each read (Fonseca, 2012). For this reason, the in-house computer program was designed to select the results obtained from the 5’-end (left) read for any situation in which a position in a non-overlapping region had been called by both the 5’- and the 3’-end reads. Once the left and right ends of each molecule had been combined, the raw results were processed using Microsoft Excel.

The raw data from each time point from each reaction was processed in its own spreadsheet. Hits that were outside of a plausible size length for successful full-length reaction product (i.e., less than 196 nucleotides and more than 214 nucleotides in length) were excluded. For time points with less than 10,000 hits, the entire data pool was considered. For time points with more than 10,000 hits, between 10,200 and 10,500 hits were selected at random. The hits were then analyzed position by position, working from 5'- to 3' end. Hits that agreed with the wild-type were passed over until such time when a hit did not agree with the known wild-type sequence. Using context provided by the hits in close proximity, sequences were either adjusted to introduce an insertion, a deletion, or the position was considered to correspond to a mutation (Table 16). In ambiguous cases, preference was given to making the adjustment that would not create a mutation. Care was taken to mitigate the number of insertions and deletions generated and situations that would create an insertion and deletion in close proximity were generally avoided. Hits that suggested that a large portion of the molecule had been eliminated during recombination were rejected. Situations that suggested the occurrence of a single insertion were typically retained, but to facilitate the ease of processing, insertions of two or more positions were generally discarded. Because of time constraints, each time point was processed only as far as the region of interest (e.g. the W* + XYZ time points were only analyzed position by position for the W* region). Upon completion of the analysis, all remaining non-discarded hits were sorted by random and a total of 10,000 hits were retained for analysis for each data point. For time points with less than 10,000 raw hits, all non-discarded hits were retained for analysis.

To determine the niche population for each reaction, the region of the ribozyme that has been defined as available for mutational analysis (i.e., positions 157-184 for the Z*-Single reaction, positions 24-67 in the W*-Single reaction, and positions 24-71 and 157-184 in the Double reaction) was selected for each reaction. The genotype of each ribozyme in each of the six time points for each of the three reactions was defined according to that region (with regions not subjected to random mutations deemed to be wild-type). For the single reactions, the 20 most-prevalent genotypes were identified and formed the pool from which the niche population was derived. Each population member that fit this criteria was given a “niche score” (*N*-score) that was calculated by adding its rank at each of the six time points. If the ribozyme was not found among the 20-most prevalent genotypes for any of the six time points, it was given a score of 21 for that particular reaction. The scores were compared and the ten genotypes with the lowest *N*-score were defined as the niche population for that reaction. Because the wild-type genotype was consistently the most prevalent genotype at all six time points in all three reactions, the lowest possible *N*-score was 6, which was reflected by the wildtype in all cases.

Because the product yield was less robust for all time points in the Double reaction, the criteria used to determine the niche population for this reaction was different. In five of the six time points, the wild-type was clearly the dominant genotype, one genotype was clearly the next-most prevalent and the remainder of sequences were either unique or were carried by two members of the population. For the sixth time point, only the wild-type was clearly dominant. All other sequences were either carried by two members of the population or were unique to one population member. Thus, the niche population for the

Double reaction was defined as the wild-type and the second-most prevalent sequence in five of the six reactions and as only the wild-type. Because the total population for the Triple reaction was four ribozymes, the niche population parameter was not used for the Triple reaction.

Error Analysis

Because the term “error” is frequently used in genetics to refer generally to deviations from the wild-type, the word “noise” will be used in this dissertation to refer to deviations from the wild-type that are most appropriately described as mistakes occurring as a result of the polymerase chain reaction, transcription, or the sequencing protocol. Deviations away from the wild-type genotype that can be shown to plausibly be a result of the successful implementation of the study design will be described as “mutations” and the term “error” will generally be avoided.

During the data processing phase, a qualitative inspection of the raw data suggested that certain positions seemed to be more prone to errors than other positions. For that reason, a different noise value was established for each position in the product molecule. As mentioned previously, in order to facilitate timely completion of the data analysis phase, normally only the region of the molecule with expected mutations was subjected to a position-by-position analysis. To determine the noise correction at each position, the 10,000 sequences from the 30-minute time point from the Z*-Single reaction and the 10,000 sequences from the 30-minute point from the W*-Single reaction were processed completely from the 5'-end to the 3'-end of the molecule. In the Z*-Single, the WXY region was expected to align perfectly with the wild-type and in the W*-Single reaction,

the XYZ region was expected to align perfectly with the wild-type. In the wild-type regions, any deletion, insertion, or mutation was counted as noise. A separate noise correction was calculated for noise arising from a mutation, a deletion, or an insertion and the corresponding noise correction was applied to each type of variance (i.e., the mutation rate was adjusted for mutation noise, the insertion rate was adjusted for insertion noise, and the deletion rate was adjusted for deletion noise) at each position (Table 2 – Nope!). An estimation of the noise at each position was then calculated and subtracted from the mutation rate at each position (Table 17). Because regions at a recombination junction (e.g., W and X in the W*-Single reaction) are inherently less precise than regions not involved in recombination, the Z*-Single reaction was used to establish the noise correction for the W and X regions while the W*-Single reaction was used to establish the noise correction for the Y and Z regions. The same noise corrections were applied to all reactions described herein.

Figure 3 – The *tF2* and *R2F2* mechanisms of recombination of the *Azoarcus* group I intron.

Shown here are the two mechanisms by which the *Azoarcus* ribozyme accomplishes self-assembly. **(A)** The *tF2* describes an interaction between the catalyst and substrate in which two fragments are joined by a four nucleotide insertion between the two fragments. The *tF2* mechanism is accomplished by a single step in which the IGS of at the 5'-end of the catalyst base pairs with the tag sequence at the 3'-end of the substrate fragment that will accept the electron pair during nucleophilic attack. This interaction holds the two substrates in position so that the hydroxyl group of the uracil nucleotide of the tag sequence in the substrate bearing the electron-donating can initiate nucleophilic attack between the two guanosine nucleotides of the head group of the second substrate. **(B)** The *R2F2* mechanism is chemically analogous to the reverse direction, followed by the forward direction, of the two trans-esterification reactions found in the *in vivo* splicing mechanism of group I introns. Although the *R2F2* mechanism is more versatile in nature because it does not require that the two substrates bear complimentary regions, the substrates used in this study were all engineered such that they would go through the *tF2* mechanism. Figure adapted from Draper et al., 2008.

Figure 4 – Comparison of product obtained from the reaction of the four wild-type substrates versus the product obtained when one of the substrate fragments is mutated. Results from reaction using W, X, Y, and Z wild-type fragments as substrates (top) versus the same reaction with Z* substituted in place of Z (bottom). The difference in the ability to see product yield with the naked eye in the wild-type reactions versus the inability to visualize any product with as little as one substrate being subjected to mutations is striking. The 8% polyacrylamide, 8 M urea gels were stained with SYBR-Green RNA II Gel Stain (Life Technologies) and visualized on the Typhoon Trio and Phosphoimager (GE Healthcare). The five size controls used on each gel are clearly marked on the right-hand side of each gel.

Figure 5 – Comparison of expected products from the five recombination reactions. Each of the five reactions incrementally increased the percentage of the ribozyme that was obtained from fragments engineered at a 9% random mutation rate for each position. In all cases, successful product was considered as being the full-length intron assembled in the appropriate order (i.e., WXYZ) without omissions of more than approximately 40% of any substrate fragment. The fragment lengths are 63, 39, 45, and 51, respectively for the W, X, Y, and Z regions.

Figure 6 – Percent Wildtype for Z*-Single, W*-Single, and Double Reaction Populations. The percent of the full-length product population bearing the wild-type genotype is shown at each of the six time points for the Z*-Single, W*-Single, and Double reactions.

	Z*-Single	W*-Single	Double	Triple	Quadruple
30m	49900	72000	133	0	0
1h	21000	43000	1054	0	0
2h	60800	70000	654	0	0
4h	15200	82000	1779	1	0
6h	35800	77000	313	3	0
8h	34700	88000	715	0	0

Table 1 – Number of sequences available for analysis at each time point for each of the five 9% Reactions. As described in the Materials and Methods section for this chapter, the raw data representing the sequence identities of each full-length intron product was compiled into a single file for each time point of each reaction. Upon receipt, the data was reviewed and any result of a length indicating that it did not represent full length product, or was of a length that it suggested a recombination of extra fragments was discarded. The numbers provided here reflect the approximate number of hits per time point per reaction after this initial culling process. To facilitate data processing, for the Z*- and W*-Single reactions, the total number of hits was pared down to a total of 10,000 hits randomly-selected hits from the initial pool. Results from the Double and Triple reactions were sufficiently modest that a culling process was not required.

Table 2 – Data Analysis for Z*-Single Reaction. The wild-type base and position number is given for each nucleotide position at the left-hand side of the chart. For this reaction, positions 155 and 156 were omitted because they were treated as “fixed” for purposes of data processing and positions 185-205 are omitted as they duplicative of the 3’-primer and any mutations contained in these positions were overwritten during reverse transcription. **(A)** Total percent of mutations at each position ($N = 10,000$ for each time point), with a correction for background noise applied as described in the text. **(B)** Positions 167, 170, and 177 were qualitatively determined to be of interest as being more prone to tolerate random mutations during self-assembly. For each of these three positions, the total percent of nucleobase being represented ($N = 60,000$) by each of the four possibilities is given. **(C)** The trend in total mutation rate between the 30-minute and 8-hour time point was calculated using a best-fit line between each of the six data points. The average rate of change per hour is given. Positive values indicate an overall increase in the percent of mutations at each positions while negative values indicate an overall decrease. For qualitative reference, black boxes associated with a base position indicate a general increase in mutation rate whereas red boxes associated with a base position indicate a general decrease in the mutation rate. White boxes indicate no overall change in the mutation rate over time.

30m			4h		
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1736	Wildtype	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1558	Wildtype
GCGAUGGUGAAGGCAUAGUCCAGGGAGU	97	U159G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	146	C170G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	95	C170U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	113	C170U
GUUAUGGUGAAGGCAUAGUCCAGGGAGU	83	C158U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	100	A167U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	81	A167U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	99	C170A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	79	C176U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	91	G180A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	78	A167G	GUUAUGGUGAAGGCAUAGUCCAGGGAGU	88	C158U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	76	C177U	GCGAUGGUGAAGGCAUAGUCCAGGGAGU	81	U159G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	72	A173G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	77	G180U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	70	C170A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	A166G
GGUAUGGUGAAGGCAUAGUCCAGGGAGU	68	C158G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	C177U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	63	A182G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	68	C176U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	63	A178U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	64	A173G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	63	A166G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	61	C177G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	60	G157	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	61	G168U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	58	U161G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	60	C177A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	55	A171G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	58	A167G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	55	C170G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	58	G181U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	53	G181U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	56	A171G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	51	G180U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	53	C176A
1h			6h		
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1840	Wildtype	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1630	Wildtype
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	111	G180U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	105	A167U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	96	A166G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	91	A166G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	96	C170U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	84	C170U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	92	A167U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	84	A173G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	91	C176U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	80	U164
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	85	A167G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	78	C176U
GCGAUGGUGAAGGCAUAGUCCAGGGAGU	83	U159G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	77	C177U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	81	A173G	GCGAUGGUGAAGGCAUAGUCCAGGGAGU	75	U159G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	80	C170A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	73	A182G
GUUAUGGUGAAGGCAUAGUCCAGGGAGU	79	C158U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	G180U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	73	A171G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	A166U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	72	A178U	GUUAUGGUGAAGGCAUAGUCCAGGGAGU	70	C158U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	70	C170G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	69	A171U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	69	C177U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	68	C170G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	66	U161G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	67	U172G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	65	A166U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	67	G163
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	62	C177A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	66	A182C
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	62	C177G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	62	C170A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	60	G181U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	59	A167G
2h			8h		
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1863	Wildtype	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	1989	Wildtype
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	99	A167U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	123	C170U
GUUAUGGUGAAGGCAUAGUCCAGGGAGU	94	C158U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	109	C177G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	92	A166G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	109	A167U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	88	C170U	GUUAUGGUGAAGGCAUAGUCCAGGGAGU	102	C158U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	86	C170G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	95	C177U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	85	C177U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	94	C176U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	85	C176U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	94	A167G
GAUAUGGUGAAGGCAUAGUCCAGGGAGU	81	C158A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	92	A166G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	79	A167C	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	92	C170A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	72	A167G	GCGAUGGUGAAGGCAUAGUCCAGGGAGU	88	U159G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	C170A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	83	C170G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	71	A178U	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	83	G181U
GCGAUGGUGAAGGCAUAGUCCAGGGAGU	67	U159G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	70	A173G
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	66	A173G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	69	C177A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	64	U172G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	69	A166U
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	62	C177A	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	66	G181A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	57	C177G	GAUAUGGUGAAGGCAUAGUCCAGGGAGU	64	C158A
GCUAUGGUGAAGGCAUAGUCCAGGGAGU	54	A171G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	64	U161G
GGUAUGGUGAAGGCAUAGUCCAGGGAGU	53	C158G	GCUAUGGUGAAGGCAUAGUCCAGGGAGU	63	A171G

Table 3 – The most prevalent population members for the Z*-Single reaction at each of the six time points. The 20 most frequently-represented genotypes in the population are shown, with the number of times each of those phenotypes appeared in the sample population ($N = 10,000$) and a descriptor of either “Wildtype” or the nature of the mutation away from the wild-type.

Rank	Z*-Single		W*-Single	
	Genotype	N	Genotype	N
1	Wildtype	6	Wildtype	6
2	C170U	21	C30G	30
3	A167U	22	C30U	35
4	A166G	43	C28U	42
5	C158U	43	C65U	42
6	C176U	46	U49G	71
7	U159G	52	C30A	75
8	C177U	55	C32U	79
9	A173G	65	C54U	80
10	C170A	66	C65G	81

Table 4 – The niche populations for the Z*-Single and W*-Single reactions. As described in the Materials and Methods for Chapter 2, the 20 most prevalent genotypes observed at each time point in both reactions were identified and ranked in order of percent occurrence at that reaction time point. The rankings were then added and the genotype was assigned a “niche genotype” score (N) based upon the sum of the rankings. Thus, for each of these reactions, the niche genotype consists of the ten most prevalent genotypes.

Table 5 – Data Analysis for W*-Single reaction. The wild-type and base position number are given on the left-hand side of the chart. For this reaction, analysis for Positions 9-24 were omitted because those positions become fixed by the forward primer during RT-PCR. Position 68 was omitted because it was fixed as wild-type to facilitate data processing. **(A)** Total percent of positions mutated at each position ($N = 10,000$ for each time point), with a correction for background noise applied as described in the text. **(B)** Positions 27, 28, 30, 32, 54, 59, and 65 were qualitatively determined to be of interest as being more prone to tolerate random mutations during self-assembly. For each of these three positions, the total percent of nucleobase being represented ($N = 60,000$) by each of the four possibilities is given. **(C)** The trend in average mutation rate between the 30-minute and 8-hour time point. Negative values indicate an overall decrease in the mutation rate over time.

30m			4h		
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	551	Wildtype	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	480	Wildtype
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGTGC	49	C65U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	33	C30G
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	42	C28U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	30	C30U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGAGC	39	C65A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGU	30	C67U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	34	C30G	AAACACGCAAGGGAUGGUGUCAAUUCGGUGAAACCUAAGCGC	29	C54U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	32	C32U	AAACUACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	C28U
AAACATGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	31	C30U	AAACACGUAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	26	C32U
AAACACGCAAGGGATGGTGTCAAATTCGGTGAACCTAAGCGC	31	C54U	AAACACGCAAGGGAUGGUGUCAAAGUCGGCGAAACCUAAGCGC	26	U49G
AAACACGCAAGGGATGGTGTCAAATTCGGAGAAACCTAAGCGC	29	C54A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	26	C65U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGG	29	C67G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	25	C59U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGGCG	29	C65G	AAACCAAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	25	C30A
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	28	U49G	AAACACGCAUUGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	24	G35U
AAACCAAGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	28	C30A	AAACACGCAAGGGAUGGUGUCAAUUCGGCUAAACCUAAGCGC	23	G55U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	28	A58G	AAACACGCAAGGGAUGGUGUCAAUUCGGAGAAACCUAAGCGC	23	C54A
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGT	27	C67U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	22	G35A
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGA	26	C67A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	22	U39A
AAACCAAGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	24	C28A	AAACACGCAUGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	21	A34U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	23	U44C	AAACACGCAAGGGAUGGUGUGAAUUCGGCGAAACCUAAGCGC	20	A46G
AAACACGCAAGGAATGGTGTCAAATTCGGCGAAACCTAAGCGC	22	G37A	AAACACGCUAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	20	A33U
AAACACGCAAGGGATGGTGTCAAATTCGGCGAAACCTAAGCGC	22	A56G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	20	C60U
1h			6h		
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	539	Wildtype	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	639	Wildtype
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	47	C30G	AAACACGCAAGGGAUGGUGUCAAUUCGGAGAAACCUAAGCGC	48	C54A
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	37	C32G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	45	C30G
AAACUACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	30	C28U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	40	C65U
AAACCAAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	29	C30A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGAGC	35	C65A
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	28	C32U	AAACCAUGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	34	C30U
AAACCAUGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	C30U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	32	C45U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	27	C65U	AAACUACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	31	C28U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	C28A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	30	G55A
AAACACGCAAGGGAUGGUGUCAAAGUCGGCGAAACCUAAGCGC	26	U49G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	30	A46U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	25	A58G	AAACACGCAAGGGAUGGUGUCAAUUCGGUGAAACCUAAGCGC	30	C54U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	25	G35A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGGGC	30	C65G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	24	C51U	AAACACGCAUUGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	29	G35U
AAACACGCAAGGGAUGGUGUCAAUUCGUGAAACCUAAGCGC	24	C54U	AAACACGCAUAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	29	C32U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	24	A34U	AAACACGCAAGGGAUGGUGUCAAAGUCGGCGAAACCUAAGCGC	28	U49G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGGGC	22	C65G	AAACACGCAAGGGAUGGUGUCAAUUCGGGGAACCUAAGCGC	28	C54G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	22	G36A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGG	28	C67G
AAACACGCAUUGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	21	G35U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	27	A63U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	21	A46U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	A47G
AAACACGCAAGGGAUGGUGUCAAUUCGCGAAACCUAAGCGC	20	G52U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	A62U
2h			8h		
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	492	Wildtype	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	645	Wildtype
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	39	C65U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	62	C30G
AAACACGCAAGGGAUGGUGUCAAAGUCGGCGAAACCUAAGCGC	34	U49G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGAGC	44	C65A
AAACUACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	29	C28U	AAACCAUGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	42	C30U
AAACACGCAAGGGAUGGUGUCAAUUCGGAGAAACCUAAGCGC	29	C54A	AAACACGCAAGGGAUGGUGUCAAUUCGGCUAAACCUAAGCGC	41	G55U
AAACAGGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	28	C30G	AAACACGCAAGGGAUGGUGUCAAUUCGGUGAAACCUAAGCGC	41	C54U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	C30U	AAACUACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	41	C28U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	23	C59U	AAACACGUAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	39	C32U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGAGC	23	C65A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGC	39	C65U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	22	C28A	AAACACGCAAGGGAUGGUGUCAAAGUCGGCGAAACCUAAGCGC	38	U49G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGG	22	C67G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	34	C60U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGGGC	22	C65G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	33	U49A
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	21	C51U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGU	33	C67U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	20	G55U	AAACACGCAAGGGAUGGUGUGAAUUCGGCGAAACCUAAGCGC	32	A46G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	20	C32G	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGGGC	32	C65G
AAACCAAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	19	C30A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGA	31	C67A
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	19	C60U	AAACACGCAUUGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	31	G35U
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGU	19	C67U	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	30	G37A
AAACCAAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	19	C32A	AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	A62G
AAACACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	19	C60A	AAACCAAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC	27	C30A

Table 6 – Most prevalent population members for the W*-Single reaction at each of the six time points. The 20 most frequently-represented genotypes in the population are shown, with the number of times each of those phenotypes appeared in the data sample ($N = 10,000$) and a descriptor of either “Wildtype” or the nature of the mutation away from the wild-type genotype.

Table 7 – Data Analysis for Double Reaction (W*-region) The wild-type and base position number are given on the left-hand side of the chart. For this reaction, analysis for Positions 9-24 were omitted because those positions become fixed by the forward primer during RT-PCR. Positions 69-71 were included here because in order to align both the 5'- and 3'-ends of the molecule, the molecule was aligned position by position throughout the entire length of the intron. The values are derived from a composite of all data points across the six time points ($N = 4648$), with a correction for background noise applied as described in the text. **(B)** Positions 27, 28, 30, 32, 45, 54, and 65 were qualitatively determined to be of interest as being more prone to tolerate random mutations during self-assembly. For each of these three positions, the total percent of nucleobase being represented ($N = 4648$) by each of the four possibilities is given. **(C)** The trend in average mutation rate between the 30-minute and 8-hour time point.

Table 8 – Data Analysis for Double Reaction (Z*-region) The wild-type and base position number are given on the left-hand side of the chart. For this reaction, analysis for Positions 155-156 were omitted because they were treated as “fixed” for purposes of data processing and positions 185-205 are omitted as they duplicative of the 3’-primer and any mutations contained in these positions were overwritten during reverse transcription. The values are derived from a composite of all data points across the six time points ($N = 4648$), with a correction for background noise applied as described in the text. **(B)** Positions 158, 161, 167, 170, 173, 176, 177, 182, and 184 were qualitatively determined to be of interest as being more prone to tolerate random mutations during self-assembly. For each of these three positions, the total percent of nucleobase being represented ($N = 4648$) by each of the four possibilities is given. **(C)** The trend in average mutation rate between the 30-minute and 8-hour time point.

176, 177	Z*-SINGLE						DOUBLE					
	30m	1h	2h	4h	6h	8h	30m	1h	2h	4h	6h	8h
(A,A)	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.3	0.2	0.0	0.0
(A,C)	1.8	1.5	1.8	1.9	1.4	1.3	1.6	1.8	1.9	2.7	1.0	1.8
(A,G)	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.1	0.7	0.3
(A,U)	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.1	0.2	0.3	0.3	0.1
(C,A)	3.0	2.8	3.0	3.0	2.7	2.5	4.0	2.8	2.0	2.8	3.0	3.8
(C,C)	81.1	83.1	81.1	81.6	82.8	83.7	84.1	81.5	79.6	76.8	78.0	75.2
(C,G)	3.4	3.0	3.6	3.1	3.1	4.0	2.4	2.9	2.3	3.2	3.9	3.8
(C,U)	5.0	4.2	4.9	4.6	4.6	3.9	2.4	4.5	4.8	5.9	4.3	5.8
(G,A)	0.1	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.0	0.1	0.0	0.0
(G,C)	2.1	1.9	2.1	2.0	2.2	1.5	3.2	3.9	4.5	4.6	4.3	4.8
(G,G)	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.3	0.1
(G,U)	0.1	0.1	0.1	0.1	0.0	0.0	0.0	0.1	0.2	0.2	0.0	0.3
(U,A)	0.1	0.1	0.1	0.0	0.1	0.1	0.0	0.2	0.0	0.2	0.0	0.1
(U,C)	3.0	3.0	3.0	3.2	2.8	2.8	2.4	1.3	4.0	2.5	3.0	3.3
(U,G)	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.2	0.0	0.1	1.0	0.4
(U,U)	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.3	0.3	0.1	0.3	0.0

Both Wildtype

One Position Mutated

Both Positions Mutated

Table 9 – Configuration of nucleotides at positions 176 and 177 for the Z*-Single and Double Reactions. Selection against the wildtype at these two positions is of particular interest because these two positions interact with positions 127 and 129 in the Y-fragment and form the catalytic core of the intron. At left, each of the 16 possible configurations for positions 176 and 177, respectively, is given. For both the Z*-Single and Double reactions, the percent of full-length product bearing each specific configuration is given at each time point. The percent of full-length product with both positions being wild-type is highlighted in gray, positions where one position is mutated and other is wildtype are highlighted in pink and positions where both positions are mutated are shown in green.

30m		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	16	Wildtype
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC...UAUGGUGAAGGCAUAGUCCAGGGAGU	2	G157., G158.
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCUGGGAGU	2	A178U
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGA	2	U184A
1h		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	59	Wildtype
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	3	A167U
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGUGCGCGAUGGUGAAGGCAUAGUCCAGGGAGU	3	C65U, U159G
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGACC	3	G183C
2h		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	20	Wildtype
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGAGU	3	G180U
4h		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	37	Wildtype
AAACCAGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGC.....GUGAAGGCAUACUCCCGGGAGU	3	C30U, G55A, G174C, A178C, and d.157-162
AAACCACGAAAGGGAGGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	3	C32A, U39G, C170A
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGGGCUAUGGUGAAGGCAUAGUCCAGGGAGU	3	C67G
6h		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	14	Wildtype
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCGAUGGUGAAGGCAUAGUCCAGGGAGU	2	U159G
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCGUAGUCCAGGGAGU	2	A171G
AAAUAGCGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCAAGGGAGU	2	C27U, C28G, A48G, C177A
8h		
AAACCACGCAAGGGAUGGUGUCAAUUCGGCGAAACCUAAGCGCGCUAUGGUGAAGGCAUAGUCCAGGGAGU	18	Wildtype

Table 10 – Most prevalent population members for the Double reaction at each of the six time points. Because full-length product in the Double reaction results from a recombination event amongst substrates with an increased ability to sample sequence space, increased diversity was seen in the genotypes of the full-length product obtained. For that reason, only the wild-type, which was the dominant population member for all time points and the next most prevalent genotype at each time point were included in the niche population.

Molecule	4.1	6.1	6.2	6.3
Mutations	G40A U49A A58G A178G	C73G C107U G163A	C30G A34U G35U A47C A58G A80U G82A C107G d156-165	A26C C30A U44A C45A A47C C94U C158A A167U
W-X Recombination Junction	AAU	CAU	-	GUAU
X-Y Recombination Junction	GCAU	GCAU	CAU	GCAU
Y-Z Recombination Junction	GCAU	GCAU	-	GGAU

Table 11 – The four full-length product sequences obtained from the Triple (W* + X* + Y + Z*) reaction. From all six time points of the Triple reaction, only a total of four full-length product molecules were obtained. Because yield was so low, results are analyzed molecule by molecule. The four full-length product molecules are designated as 4.1, 6.1, 6.2, and 6.3. For each, the position and nature of each mutation is given, along with the region where it is found (W = red, X = yellow, Z = green). The nucleotide sequence of the recombination junction for each of the three recombination junctions involved in the self-assembly reaction.

<i>Substrate</i>	<i>Concentration (μM)</i>	<i>Amount (pmol)</i>	<i>Substrate</i>	<i>Concentration (μM)</i>	<i>Amount (pmol)</i>
Z*-Single (WXY + hZ*)			Triple (W* + X* + Y + Z*)		
WXY	2.0	1940	W*	2.0	2200
Z*	2.5	2425	X*	2.5	2750
W*-Single (W* + XYZ)			Y	3.1	3438
W*	2.0	1857	Z*	3.9	4297
XYZ	2.5	2321	Quadruple (W* + X* + Y* + Z*)		
Double (W* + XY + Z*)			W*	2.0	3000
W*	2.0	2176	X*	2.5	3750
XY	2.5	2720	Y*	3.1	4688
Z*	3.1	3400	Z*	3.9	5859

Table 12 – Concentrations and amounts of substrates used in each of the four variations of the recombination reaction. The rationale for adding substrates in 1.25-fold excess over its 5' binding partner is described in the text. For reactions that included a substrate made in the research lab, the estimated yield of the fragment with the lowest yield was used as the limiting reaction to which other substrates and final reaction volume were adjusted. Because all substrates for the Quadruple were purchased, the reaction volume was fixed at 1500 μ L.

A					
	μ (%)		μ (%)		μ (%)
24	3.75	39	6.83	54	10.74
25	3.64	40	3.23	55	8.70
26	3.82	41	3.80	56	5.82
27	9.21	42	3.92	57	6.43
28	10.01	43	2.69	58	7.82
29	5.83	44	5.44	59	9.10
30	14.95	45	7.35	60	6.40
31	3.97	46	6.85	61	5.43
32	10.98	47	4.65	62	5.99
33	5.01	48	5.06	63	6.20
34	5.93	49	8.93	64	5.65
35	7.27	50	4.11	65	9.76
36	6.09	51	4.61	66	2.34
37	5.94	52	3.04	67	6.60
38	6.11	53	2.51		

B					
	μ (%)		μ (%)		μ (%)
157	1.93	167	9.57	177	11.13
158	7.24	168	5.25	178	6.11
159	6.82	169	3.49	179	4.37
160	4.24	170	16.12	180	5.64
161	6.89	171	7.03	181	5.83
162	1.16	172	5.48	182	6.96
163	1.17	173	5.10	183	4.06
164	3.39	174	1.72	184	5.40
165	1.84	175	2.80		
166	7.22	176	6.90		

C					
	μ (%)		μ (%)		μ (%)
24	6.15	39	6.86	54	9.35
25	5.92	40	5.83	55	5.47
26	6.78	41	4.93	56	6.46
27	13.31	42	6.49	57	5.05
28	11.45	43	3.90	58	6.39
29	5.48	44	7.64	59	10.88
30	13.91	45	9.49	60	8.99
31	5.80	46	5.49	61	6.10
32	10.53	47	4.66	62	5.53
33	6.92	48	5.41	63	5.90
34	7.30	49	6.84	64	5.12
35	5.36	50	5.41	65	8.32
36	5.53	51	7.07	66	2.14
37	5.83	52	2.86	67	5.96
38	5.72	53	2.89	68	5.74
				69	6.81
				70	5.22
				71	0.03

	μ (%)		μ (%)		μ (%)
157	2.39	167	9.45	177	12.24
158	10.93	168	5.08	178	7.40
159	7.60	169	4.41	179	4.85
160	6.51	170	17.05	180	6.36
161	9.61	171	8.83	181	5.06
162	1.90	172	7.23	182	10.32
163	1.62	173	8.56	183	8.23
164	5.10	174	3.94	184	11.33
165	2.31	175	5.93		
166	9.01	176	10.32		

Table 13 – Average mutation rate (μ) for W*-Single, Z*-Single, and Double Reactions. The overall mutation rate, corrected for noise, was calculated for each position for these three reactions was calculated by adding up the total number of mutations at each position for each of the six time points and dividing by the total number of sequences analyzed for that reaction ($N = 60,000$ for each of the single reactions and $N = 4648$ for the Double reaction).

<div>A</div> <table><thead><tr><th>Time Point</th><th>Average k</th></tr></thead><tbody><tr><td>30 m</td><td>1.76</td></tr><tr><td>1 h</td><td>1.61</td></tr><tr><td>2 h</td><td>1.65</td></tr><tr><td>4 h</td><td>1.78</td></tr><tr><td>6 h</td><td>1.72</td></tr><tr><td>8 h</td><td>1.56</td></tr></tbody></table>	Time Point	Average k	30 m	1.76	1 h	1.61	2 h	1.65	4 h	1.78	6 h	1.72	8 h	1.56	<div>B</div> <table><thead><tr><th>Time Point</th><th>Average k</th></tr></thead><tbody><tr><td>30 m</td><td>3.07</td></tr><tr><td>1 h</td><td>3.03</td></tr><tr><td>2 h</td><td>3.06</td></tr><tr><td>4 h</td><td>2.93</td></tr><tr><td>6 h</td><td>2.67</td></tr><tr><td>8 h</td><td>2.63</td></tr></tbody></table>	Time Point	Average k	30 m	3.07	1 h	3.03	2 h	3.06	4 h	2.93	6 h	2.67	8 h	2.63	<div>C</div> <table><thead><tr><th>Time Point</th><th>Average k</th></tr></thead><tbody><tr><td>30 m</td><td>3.21</td></tr><tr><td>1 h</td><td>4.96</td></tr><tr><td>2 h</td><td>5.32</td></tr><tr><td>4 h</td><td>5.72</td></tr><tr><td>6 h</td><td>5.39</td></tr><tr><td>8 h</td><td>5.71</td></tr></tbody></table>	Time Point	Average k	30 m	3.21	1 h	4.96	2 h	5.32	4 h	5.72	6 h	5.39	8 h	5.71
Time Point	Average k																																											
30 m	1.76																																											
1 h	1.61																																											
2 h	1.65																																											
4 h	1.78																																											
6 h	1.72																																											
8 h	1.56																																											
Time Point	Average k																																											
30 m	3.07																																											
1 h	3.03																																											
2 h	3.06																																											
4 h	2.93																																											
6 h	2.67																																											
8 h	2.63																																											
Time Point	Average k																																											
30 m	3.21																																											
1 h	4.96																																											
2 h	5.32																																											
4 h	5.72																																											
6 h	5.39																																											
8 h	5.71																																											
<div>D</div> <table><thead><tr><th>Time Point</th><th>Average k</th></tr></thead><tbody><tr><td>4h</td><td>4.00</td></tr><tr><td>6h</td><td>6.67</td></tr></tbody></table> <table><thead><tr><th>Overall Average k</th></tr></thead><tbody><tr><td>6.00</td></tr></tbody></table>	Time Point	Average k	4h	4.00	6h	6.67	Overall Average k	6.00																																				
Time Point	Average k																																											
4h	4.00																																											
6h	6.67																																											
Overall Average k																																												
6.00																																												

Table 14 – Average deviation from wildtype (k) for the 9%-mutation rate series reactions. Values calculated for k consider mutations, insertions, and deletions. For mutations and insertions, if either occurred at a position, then it was counted as one deviation. The ideal for deletions was to count each deletion event as one rather than adding up deletions on a position-by-position basis (e.g., if positions 160-163 were all absent, it would be one deletion rather than four). However, due to an inability to obtain results in this fashion from Microsoft® Excel, a simplifying modification was required. If the determination was made that any product molecule contained *any* deletion, then a value of 1 was added to the total number of deviations. If the molecule contained no deletions, then a value of 0 was added to the total number of deviations. While this methodology ultimately lowers the value for k , the simplifying assumption is justified on the basis that a qualitative inspection of the data suggested that the vast majority of product in the sample pool contained no more than one mutation and that ambiguous situations encountered while aligning the sequences could have led to artificially inflating the number of deleted regions. The k values are given for the (A) Z*-Single (B) W*-Single (C) Double and (D) Triple reactions. For the triple, because only a total of four sequences were obtained, the overall average k value of all four sequences is included.

Fragment	Sequence	Length
W/W*	GUGCCUUGCGCCGGG AAACCACGCAAGGGAUGGU GUCAAAUUCGGCGAAACCUAAGCGCCCCG CAU	65
hX/hX*	GGCAUCGGGCGUAUGGCAACGCCGAGCCAAGCUUC GGCGCCAU	43
hY/hY*	GGCAUGCGCCGAUGAAGGUGUAGAGACUAGACGGC ACCCACCUAAGGCACA U	52
hZ/hZ*	GGCAUCGCUAUGGUGAAGGCAUAGUCCAGGGAGUG GCGAAAGUCACACAAACCGG	55

Table 15 – Composition of substrates used in the recombination of the *Azoarcus* group I intron. In order to ensure that the fragments could be able to self-assemble full length product, regions given in bold refer to regions that were not allowed to be generated under conditions that would result in a 9% mutation rate. The compositions of hX and hX*, of hY and hY*, and of hZ and hZ* are identical with respect to each pair, except that the wild-type (non-starred) substrates were generated under conditions that dictated only the insertion of the wild-type nucleotide at all positions.

Alignment Situation	Example	Description
Agreement	TATCCA	All six positions correspond to the wild-type sequence.
Deletion	TATCA	The absence of a base creates a situation in which the missing base forces all subsequent bases to disagree with the wild-type sequence (i.e., a frameshift). In this case, either the fourth position or the fifth position is left blank, with an attempt to evenly distribute the deletion over both positions in cases where the deletion was repeated over multiple sequences.
Insertion	TATTCCA	The addition of an extra base creates a situation in which the extra base forces all subsequent bases to disagree with the wild-type sequence (i.e., a frameshift). In this case, either the second position is considered to be AT or the third position is considered to be TT. Either of these resolution methods were used, with a preference given to doubling the base at a given position
Mutation	TAGCCA	The third position is not in agreement with the wild-type sequence, but the other five positions do agree. The third position is considered to be a mutation.

Table 16 – Methodology for rectifying deviations from the wild-type sequence during processing. While aligning the raw data against the wild-type sequence, each position was described according to one of four situations.

	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47
M	0.09	0.00	0.04	0.03	0.04	0.03	0.14	0.12	0.06	0.03	0.06	0.10	0.06	0.17	0.07	0.13	0.09	0.08	0.18	0.08	0.23	0.06	0.06	0.01
I	0.00	0.00	0.00	0.00	0.01	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00
D	0.15	0.01	0.03	0.02	0.00	0.01	0.06	0.08	0.07	0.01	0.09	0.02	0.02	0.17	0.05	0.03	0.03	0.24	0.05	0.12	0.28	0.29	0.01	0.06

	48	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71
M	0.03	0.04	0.08	0.14	0.17	0.08	0.18	0.30	0.08	0.05	0.04	0.14	0.02	0.03	0.02	0.08	0.04	0.22	0.05	0.09	0.46	0.14	0.20	0.16
I	0.03	0.00	0.00	0.01	0.00	0.02	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.01	0.01	0.00
D	0.35	0.02	0.04	0.02	0.01	0.07	0.01	0.02	0.01	0.02	0.10	0.01	0.04	0.00	0.00	0.02	0.02	0.03	0.01	0.00	0.01	0.06	0.06	0.01

	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91
M	0.16	2.21	0.29	0.05	0.14	0.28	0.12	0.13	0.01	0.05	0.06	0.14	0.34	0.03	0.01	0.07	0.09	0.31	0.15	0.08
I	0.00	0.03	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00
D	0.20	0.02	0.07	0.05	0.35	0.26	0.18	0.04	0.07	0.03	0.01	0.05	0.09	0.03	0.13	0.07	0.17	0.02	0.17	0.13

	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110
M	0.05	0.11	0.06	0.05	0.07	0.02	0.08	0.05	0.02	0.04	0.17	0.10	0.11	0.15	0.10	0.13	0.07	0.11	0.10
I	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.02	0.00	0.00	0.00
D	0.04	0.05	0.00	0.02	0.01	0.01	0.00	0.00	0.00	0.02	0.01	0.00	0.09	0.05	0.09	0.05	0.40	0.35	0.25

	111	112	113	114	115	116	117	118	119	120	121	122	123	124	125	126	127	128	129	130	131	132
M	0.09	0.10	0.11	0.05	0.10	0.01	0.05	0.11	0.07	0.06	0.10	0.05	0.08	0.07	0.08	0.13	0.05	0.07	0.05	0.02	0.05	0.03
I	0.02	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.01	0.01	0.00	0.00	0.03	0.01	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
D	0.32	0.22	0.44	0.45	0.26	0.21	0.19	0.14	0.07	0.24	0.11	0.20	0.31	0.39	0.18	0.28	0.11	0.11	0.10	0.05	0.03	0.01

	133	134	135	136	137	138	139	140	141	142	143	144	145	146	147	148	149	150	151	152	153	154
M	0.10	0.06	0.10	0.20	0.06	0.08	0.24	0.29	0.24	0.18	0.12	2.63	0.14	0.09	0.13	0.04	0.12	0.14	0.22	0.23	0.10	0.08
I	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01	0.02	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.02	0.04	0.04
D	0.14	0.01	0.06	0.03	0.00	0.03	0.03	0.08	0.02	0.01	0.48	0.18	0.05	0.25	0.04	0.21	0.26	0.06	0.10	0.22	0.07	0.07

	155	156	157	158	159	160	161	162	163	164	165	166	167	168	169
M	0.15	0.12	0.17	0.15	0.02	0.09	0.14	0.06	0.06	0.04	0.12	0.03	0.19	0.11	0.09
I	0.04	0.00	0.02	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.01	0.01	0.02	0.01	0.01
D	0.52	0.21	0.10	0.18	0.03	0.10	0.03	0.04	0.04	0.05	0.04	0.00	0.04	0.01	0.07

	170	171	172	173	174	175	176	177	178	179	180	181	182	183	184
M	0.42	0.16	0.08	0.09	0.15	0.12	0.14	0.30	0.24	0.18	0.14	0.13	0.11	0.20	0.24
I	0.00	0.00	0.03	0.06	0.04	0.00	0.02	0.01	0.00	0.00	0.00	0.03	0.00	0.00	0.05
D	0.10	0.06	0.05	0.11	0.12	0.04	0.00	0.08	0.03	0.02	0.03	0.28	0.10	0.07	0.08

Table 17 – Noise correction per position. The mutation (M), insertion (I), and deletion (D) rates at each position were scaled down by a percent derived from an analysis of two known wild-type sequences. The correction factors given in this table were applied, as appropriate, to the data obtained from any time point from any reaction discussed in this thesis.

CHAPTER 3 – RECOMBINATION OF A CLASS I RIBOZYME ASSEMBLED THROUGH SELFISH AND COOPERATIVE REPLICATOR NETWORKS

Introduction

In addition to identifying potential selection methods (i.e., thermodynamic versus kinetic) that may have been responsible for the transition between lifeless organic molecules and biomolecules capable of attaining the processes necessary for metabolism, the question still remains as to how the transition from prebiotic chemistry to Darwinian-type evolution was possible. A quasi-species is defined as a given distribution of macromolecular species with closely interrelated sequences, dominated by one or several (degenerate) copies (Eigen and Schuster, 1977). Here, we can consider the ribozyme to nicely fit this definition. Moreover, which, if any, ribozyme would have dominated in the RNA world has been established as the one which has attained functionality (Kun et al., 2005). Thus, the *Azoarcus* intron is an excellent molecule on which to consider molecular interactions that may have taken place on the prebiotic Earth.

The *Azoarcus* intron has been shown to self-assemble into self-replicating ribozymes spontaneously from cooperative catalytic cycles and networks (Vaidya et al., 2012). As described previously, self-assembly in the *Azoarcus* intron arises from an interaction in which the internal guide sequence (IGS) of the substrate fragment initiating the recombination event by a nucleophilic attack on the tag sequence of the fragment bearing the tag (Hayden and Lehman, 2006). In the wild-type sequence, the IGS sequence is 5'-GUG-3' and the tag is 5'-CAU-3'. Introducing designed mutations into the IGS and

tag sequences was shown to inhibit, but not completely block, self-assembly in 100 mM MgCl₂ (Draper et al., 2008). By synthesizing *Azoarcus* ribozyme fragments that differed with respect to the middle nucleotide of both the IGS and tag sequences, the effect of this change on the autocatalytic rate constant for the self-assembly reaction was determined for multiple IGS-tag variants (Yeates et al, 2016).

An interaction defined as “Selfish” is one in which the IGS of one molecule can interact with another molecule bearing a perfectly complimentary tag sequence, with the caveat that the terminal nucleotide pair prior to the splice/recombination junction is a G-U wobble pair (for proper group I intron activity). For this research, two selfish systems were generated by designing substrate molecules which varied with respect to the identity of the middle nucleotides of the IGS and tag regions. The first variant is denoted as “AU” where “A” refers to the identity of the modified middle nucleotide of the IGS sequence and “U” refers to the identity of the modified middle nucleotide of the tag sequence. This two-character notation will be used throughout the remainder of this thesis. The second selfish variant was designated as “CG”. Thus, through canonical Watson-Crick base pairing, each molecule was capable of interacting with a like member of the population to produce full-length product. In short, this system would be capable of self-replication if all members carried the same genotype; genetic variation would not be required to maintain this replicating system. In contrast, to obtain a “cooperative” system, the tag sequences were paired up with the IGS sequences in the opposite manner – one cooperative system was designated as “AG” and the other as “CU”. Here, the ribozyme is not expected to readily

interact with like genotypes because self-interaction would rely on non-canonical base-pairing interactions.

The autocatalytic rate constant is a measurement of the contribution of autocatalytic feedback to the overall self-assembly reaction (Draper et al., 2008). The autocatalytic rate constants for all 16 possible IGS-tag interactions have been calculated for the reaction $_{\text{GMG}}\text{WXY}_{\text{CNU}} + \text{Z}$, where M and N refer to the different possible variations in the middle nucleotide identity. (Yeates et al., 2016). A comparison between a genotype whose autocatalysis occurs through canonical Watson-Crick base pairing versus one whose does not is striking. For example, the autocatalytic rate constant for AU is 0.0319 min^{-1} and that for AG is 0.0001 min^{-1} ; thus, over 300 times more catalytic events are performed per minute through autocatalysis by AU than by AG. However, as can readily be seen in this case, the “A” of the IGS of the “AG” ribozyme can interact with the “U” on the tag of the “CU” ribozyme through a canonical Watson-Crick interaction, and the “C” in the IGS of the “CU” ribozyme can interact with the “G” in the tag of the “AG” ribozyme through a different canonical Watson-Crick interaction. Thus, this mandated intermolecular interaction necessitates genetic variation and is a case of a simple network composed of two genotypes.

A payoff matrix can be used to predict the interaction which is most likely to produce the most robust yield of full-length product when two differing genotypes are allowed to compete for the same substrate. The payoff matrix for the cooperative interaction AG vs. CU suggests that at such time in which the composition of reaction products has reached a plateau, product distribution will be almost even (53.2% AG, 46.8%

CU), whereas in the selfish reaction, the expectation is that the competition between the two ribozyme fragments for a common substrate will lead to near fixation of one genotype (96.0% CG, 4.0% AU) (Yeates et al., 2016). Although the specific selfish and cooperation variants studied in this thesis have not been previously analyzed in the laboratory, the calculated payoff matrices are used as a tool to predict the composition of the molecular population over time.

Additionally, previous work with the *Azoarcus* intron did not address the potential combined impact of varied IGS-tag interactions when the substrates bearing those IGS and tag sequences had been subjected to random mutations. Thus, the work in this chapter combines these two strategies to look at the composition of the molecular population resulting from selfish and cooperate self-assembly of substrates which had been subjected to a procedure designed to introduce random mutations.

In the experiments described in Chapter 2, any substrate fragment introduced into the reaction vessel was an analog of one individual fragment (i.e., W*, X*, Y*, or Z*). Because in-house production of those shorter molecules requires the use of a forward and reverse primer, the majority of random mutations potentially introduced into the fragment would be covered by primers of a fixed genotype during reverse transcription and PCR amplification (e.g., in the case of the X fragment, only one of the 39 positions would not be covered by either the forward or the reverse primer). For this reason, the individual primers were purchased from an outside vendor and, as described in Chapter 2, subjected to a procedure which would lead to a mutation rate of 9% per position.

Because, in part, this research aim was intended to build on previous work done to determine the role of networks on self-assembly on prebiotic Earth, the self-assembly reactions chosen were intended to be analogs of the WXY + Z two-piece self-assembly reaction used previously, with the random mutations being introduced into the larger substrate fragment. Because in this case, the nucleotide positions not subjected to the possibility of randomly mutating were greatly diminished, the opportunity to use a different method for the introduction of random mutations was available.

Although it is usually desirable that DNA amplification occurs in an accurate manner, the standard PCR protocol that has been modified (see Materials and Methods, below) to deliberately encourage the randomization of nucleotide additions at any position (Cadwell and Joyce, 1994). In addition to increasing the concentrations of enzymes and primers, this modified method takes advantage of the addition of manganese cations to increase mis-incorporation of dNTPs (Vartanian et al., 1999), optimizing mis-incorporations by skewing the dNTP concentration highly in favor of guanine and thymine (Cadwell and Joyce, 1992) and to diminish the template specificity of the polymerase (Beckman et al., 1985).

This chapter describes the results of two reactions run in parallel (Figure 7) to compare the molecular populations over time obtained from situations designed to favor either selfish or cooperative catalysis. In these two experiments, each WXY analog was prepared by a mutagenic PCR protocol which provided for a putative mutation rate of up to 20% per position. These RNAs were allowed to recombine with wild-type hZ substrate fragments to form full-length RNA product. In analogous fashion to the reaction series

described in the previous chapter, aliquots were removed from the main reaction mix at pre-determined intervals. From each aliquot, the full-length product was isolated, purified, and amplified via specific RT-PCR. The nucleotide sequences of these ribozymes were then determined using Illumina MiSeq high-throughput sequence analysis to detect trends over time.

The emerging hypothesis that cooperation will outperform selfishness during the formation of RNA networks will be tested here. Because the cooperative system requires an interaction between ribozymes of different genotypes, it is expected that the cooperative system will be less tolerant of random mutations than the selfish system, which involves an interaction between ribozymes of the same genotype. Because the substrates fragments in the 9% reactions and the mutagenic PCR reactions were prepared by different methods, and resulted in putatively different mutation rates, the hypothesis that the preparation methods and ultimate mutation rates should not affect the positions in the full-length product that are more or less tolerant of random mutation will be tested.

Substrate and Reaction Design

As described above, the two reactions discussed in this chapter involved two-piece self-assembly of the *Azoarcus* intron from WXY substrates that contained directed mutations to the middle nucleotides of the IGS and tag sequences. These modifications were designed such that one reaction would require conditions in which ribozyme fragments of the same genotype worked together to produce full-length product and the other reaction would require that ribozyme fragments of one genotype relied on the other genotype present to accomplish self-assembly (Figure 7). In both reactions, the substrate concentrations were

1.0 μM in each WXY-analog (for a total WXY-analog concentration of 2.0 μM), and 2.5 μM in hZ, which resulted in a 1.25-fold excess of the hZ substrate over the total concentration of its 5'-binding partner. The genotypes of the WXY-analogs in the Selfish reaction were AU and CG, whereas the genotypes in the Cooperation reaction were AG and CU (Table 18).

With the exceptions to the nature of the substrates and the substrate concentrations used described above, the protocol used to conduct the two self-assembly reactions, purify and amplify reaction product, and to prepare the amplified reaction product for sequencing was identical to that used in Chapter 2.

Results and Discussion

Robust results were obtained from all of the time points sampled in both the Selfish and Cooperation reactions. Successful hits from high-throughput sequencing ranged between 90,000 and 230,000 in both of the reactions for each of the six time points (Table 19). The raw data for each time point was pared down to 10,000 hits per time point as described in Chapter 2. Because the addition of forward and reverse primers during the amplification phase subsequent to reverse transcription, positions 9-23 and positions 138-154 were excluded from the mutational analyses.

Percent of nucleotides mutated at each position

The percent mutation at each position 24-137, adjusted for noise was determined for the Selfish reaction (Table 20) and for the Cooperation reaction (Table 21). In addition to showing several positions that appear to be better suited to support variance from the wild-

type sequence, a qualitative review of this data shows that the percent-mutated at each position is markedly lower in the Selfish reaction than it is the reactions from those discussed in Chapter 2. The mutagenic PCR reactions were considered in conjunction with one another when selecting potential positions of interest. A side-by-side comparison of the results shows that the percent-mutated per position is generally lower in the Cooperation versus the Selfish reaction.

In the W region, positions 38, 39, 49, 61, and 70 show a higher percentage of mutated nucleotides over all time points. In the X region, positions 72, 73, 81, 88, 100, and 109 follow this trend as do positions 120 and 126 in the Y region. The distribution of genotypes for these positions shows a strong preference against a certain nucleotide or nucleotide substitutions for these two reactions (Table 22A), which is contrary to the overall trend described in Chapter 2. For each of the 13 positions, there is a distinct pattern regarding which bases are readily substituted and which bases tend to be selected against (Table 22B). Nucleotide substitutions within each structural class (transitions) tend to occur with greater frequency than changes between structural classes (transversions), resulting in a phenomenon known as transition bias, which has been shown to influence the direction and strength of selection on proteins and RNA (Wakeley, 1996). The substitutions observed in these two reactions show that adenine tends to transition to guanine and that uracil tends to transition to cytosine. Transition bias is not shown as strongly for cytosine and for guanine. For the only position of interest (73) in which cytosine is the wild-type nucleobase, an almost even tendency to undergo a transversion to

guanosine and to transition to uracil are observed. For guanine, the transition to uracil is strongly avoided, but no strong preference is shown for adenine or cytosine.

The positions that tended to be more tolerant of mutations in the mutagenic PCR reactions could be compared directly against the results obtained from the 9% reactions for only the W* region. In the 9% reactions, no results were obtained from the only reaction (the Quadruple) in which X* and Y* were both used as a substrates and both of the mutagenic PCR reactions used a wild-type Z substrate fragment. Interestingly, the positions of interest in W* from the 9% reactions and the mutagenic PCR reactions have no common members. The secondary structure of the *Azoarcus* intron (Figure 1) shows that none of the five identified positions of interest in the W* region are located in a stem region. The first three of these positions are located in unpaired regions and position 61 is located in an interior loop. Position 70 is located in a loop region at the end of the P5 stem (Adams et al., 2004). The crystal structure does not identify this region as being particularly important to the overall tertiary structure of the intron.

Position 73 is also located in the loop at the end of the P5 stem (Adams et al., 2004). This position showed the highest tolerance of any position within the intron to select against the wild-type nucleotide. For that reason, the effect on changes at that position will be discussed in greater detail in Chapter 4. Positions 81, 120, 126 are unsurprising tolerant of accepting alternate nucleobases as they occur in unpaired regions. The uracil nucleotide at position 100 is positioned adjacent to a uracil at position 123. Although RNA is much more promiscuous than DNA in its range of interactions with other nucleotides, the uracil-uracil interactions are among the weakest (Leontis et al., 2002). Due to the lack of strong

interactions with any other nucleotide, it is a reasonable observation that this position is tolerant of random mutations.

Position 88 is involved in a G-C base-pairing interaction with the cytosine at position 54. Position 54 is an example of the conundrum presented when attempting to draw a parallel between the nature of mutations in the W* region of the full-length introns from the 9% reactions and the mutagenic PCR reactions. In the 9% reactions, the average mutation rate at that position (Table 13) is 10.74% for W*-Single reaction and 9.35% for the Double reaction. On the other hand, in both of the mutagenic PCR reactions, the average mutation rate at that position is effectively zero. In the 9% reactions, position 54 is variable and position 88 is fixed as wild-type. By contrast, in the mutagenic PCR reactions, position 54 is available to be mutated, but the full-length product shows a strong selection for the wild-type, whereas position 88 is more tolerant of accepting nucleotides that break the base-pairing interaction. In both sets of reactions, a tendency to break the base-pairing interaction in this region is demonstrated. This potential disruption to secondary structure may serve to obviate full-length molecules from the burden of catalysis and open them up as kinetically-advantaged substrates.

For both the Selfish and Cooperation reactions, the percent of nucleotide positions bearing random mutations at each position was maintained relatively steadily as the reactions proceeded (Table 23). In both reactions, there was an almost even distribution of positions that showed a trend of accumulating mutations at a given position versus positions that showed a decrease in the percent of mutations at a given position.

Identification of the niche population

The top 20 most prevalent sequences were determined for each time point in the Selfish reaction (Table 24) and for the Cooperation reaction (Table 25). For both of these reactions, the wild-type genotype is clearly dominant. Position 73 was shown to be the position most tolerant of accepting mutations in both of the mutagenic PCR reactions. For both of these reactions, members with a single mutation at that position are consistently the second-most and third-most prevalent population members. However, an unexpected trend emerges in both cases. In the Selfish reaction, one of the two transverse mutations (C73G) is favored whereas in the Cooperation reaction, the transition mutation (C73U) is favored. This trend is consistent at each of the six time points in both reactions. Despite the transversion being consistently favored in the Selfish reaction, the transversion is not as dominant in the Selfish reaction as the transition is in the Cooperation reaction. A comparison of the nature of the mutations occurring at position 73 for the population members from each reaction that carry either of the two mutations shows that in the Selfish reaction, the transversion is favored 54.6% of the time whereas in the Cooperation reaction, the transition is favored 64.1% of the time. In both of the reactions, position 73 is the only position in any of the most prevalent sequences where a transversion mutation is tolerated. The majority of the genotypes seen most prevalently contained one mutation and in all cases except the transversion described above, this one mutation is consistently a transition. Because of the ability of position 73 to support mutations, some of the more-prevalent genotypes in the niche population for the mutagenic PCR reactions displayed two point mutations, one of which being a mutation at position 73 in all cases.

The niche population for both reactions was determined in the same as was done for the Z*-Single and W*-Single reactions (Table 26). As was seen in the 9% reactions, for both of the mutagenic PCR reactions, the niche population members carry only one mutation. In each case, this mutation correlates with a position having a relatively high percent of mutations present at that position over the six reaction time points. A total of 12 different genotypes are exhibited by the niche populations in either or both reactions. Eight of these 12 are common to the members of both reaction populations. As mentioned previously, the only region where a direct comparison can be made between the 9% reactions and the mutagenic PCR reactions is in the W*-region. There are no genotypes common to the niche populations that formed from the reactions involving mutated substrates prepared in different manners. Since there is a weak correlation at best between the positions that tend to select against the wildtype in the two reaction series, it stands to reason that the resultant niche populations would be dissimilar.

The dominance of the wild-type genotype was markedly different between the Selfish and Cooperation reactions (Figure 8). The wild-type genotype was approximately twice as dominant in the Cooperation reaction population as it was in the Selfish reaction. The low percentage of wild-type population members at the 2-hour time point for the selfish reaction can be rejected as an outlier at the 99% confidence level ($Q = 0.807$, $Q_{crit} = 0.740$). Thus, the average percent of the population with the wild-type genotype is 17.4% in the Selfish population and 36.3% in the Cooperation population. Neither population shows a linear increase or decrease in the prevalence of the wild-type genotype over time.

The R^2 values for the Selfish and the Cooperation reactions are 0.3222 and 0.5208, respectively.

The observations from these two reactions generally support the DKS hypothesis, but these reactions require that the niche population be established early and its members must remain more consistent as the reaction proceeds. This is in strong contrast to the results from the 9% reactions in which the wild-type genotype, although clearly dominant, was displayed with a much lower frequency by its population members. In the mutagenic PCR reactions, one position is so tolerable of mutations, that it becomes a sort of “pseudo-wildtype” population member. If these next-best genotypes are considered as such, the assertion that DKS is driving the reaction population is weakened further. These reactions appear to select strongly for catalytically-competent members and show less ability to support transient quick-assembling ribozymes. The hypothesis that cooperation will outperform selfishness is neither proven nor disproven, as robust product yield was obtained from both reactions. The hypothesis that the method of introducing random mutations should not affect the nature of random mutations that are tolerated in the full-length product should be rejected. A comparison of the results obtained from the 9% reactions and the mutagenic reactions show clearly different trends in the nucleotide positions in the W-fragment that are tolerant of accepting random mutations.

Overall mutation rate (μ)

The overall mutation rate, corrected for noise, was calculated separately for each nucleotide position in the Selfish and Cooperation reactions, then one overall value was calculated that took the data from both reactions into consideration (Table 27). The average mutation

rate at each position is in accord with the observation that the full-length self-assembly product from the Selfish and Cooperation reactions showed a lower percent mutation per position than did the full-length assembly product from the 9% reactions. There are two possible explanations for this observation. On one hand, the discrepancy may arise because the substrates involved in the mutagenic PCR reactions in actuality carried a lower mutation rate than the estimated 20% mutation value (see Materials and Methods, Chapter 3). On the other hand, both of the reactions may have been severely limited in their catalytic ability by the mutations placed in each IGS and tag sequence, and this limited catalytic ability obligated them to select more strongly for wild-type (or “pseudo-wild-type”) genotypes. The first explanation is more plausible for several reasons. First, the determination of the mutation rate required a quantitative estimate for both the wild-type and mutated ribozymes that had been exposed to the restriction digest enzyme. This quantitative determination was accomplished by estimating the pixelation of the full-length and truncated bands on a 2% agarose gel. Said quantification is inherently imprecise and highly subject to approximation and human error. In practical terms, the restriction digest was done primarily to confirm that the mutagenic PCR protocol had been successful and the attempt to determine a mutagenic rate from the restriction digest and was never intended to be more than a rough estimate. In practical terms, with a reported 20% mutation rate derived in this manner, the estimated rate could have ranged anywhere between 15% and 25%. Second, the mutagenic PCR protocol that was used as a basis for the protocol that guided the preparation of the substrates used in the mutagenic PCR reactions indicates an overall mutation rate of less than 1% for each position in the final product (Cadwell and

Joyce, 1994). Since the protocol used here had been modified only slightly from the published protocol, it is unlikely that it could have induced such greatly increased capacity to incorporate random mutations. Third, it has been shown that the accumulation of slightly deleterious mutations in *in vitro* populations can ultimately lead to the extinction of the population (Soll et al., 2006). Not only did the mutagenic PCR reactions not move toward extinction, these two reactions provided the most robust product yield of any reactions described in this thesis. Fourth, the 9% reactions described in Chapter 2 were only quantifiably successful when no more than approximately half of the reaction substrates were subjected to a fixed 9% mutation rate. The mutagenic PCR reactions involved one substrate fragment that contributed approximately three-fourths of the length of the full-length product. Given the robust product recovered from the mutagenic PCR reactions (Table 19), it is implausible that the mutation rate for the mutagenic PCR reactions was greater than 9%. Finally, and perhaps most significantly, if the mutation rate for the (WXY)**-fragments had been 20%, only the ability of the molecular population to sample sequence space would have been quite large. With 113 positions available for mutation, only 0.80^{113} or 10^{-11} molecules could be expected to bear the wild-type genotype. Because the total amount of (WXY)**-fragment substrate was approximately 5,000 picomoles, this leaves less than 100,000 wild-type substrate fragments available to perform the catalysis required to drive the recombination reaction. Given the robust product yield at every reaction time point in both reactions, and the notable dominance of the wild-type genotype in both reactions, it is unlikely that the mutation rate of the (WXY)**-fragments was as high as initially estimated. Thus, the presumption that the recombination substrates

for the mutagenic PCR reactions carried a 20% mutation rate is clearly not supported by the obtained results.

Average deviation from wild-type (k)

The average number of deviations from the wildtype (k) that each ribozyme in the sample population contained was determined at each reaction time point (Table 28). The average k values are lower for each of the mutagenic PCR reactions in comparison to the average k values for the 9% reactions (Table 14). Further, the average k value is lower for the Cooperation reaction than for the Selfish reaction. These trends are consistent with the findings that the full-length product obtained from the 9% reactions showed more deviation from the wild-type genotype than did the mutagenic PCR reactions. Additionally, the markedly lower average k values from the mutagenic PCR reactions further substantiate the conclusion that the mutation rate of the substrates involved in the mutagenic PCR reactions was lower than that involved in the 9% reactions. The average k value for the Cooperation reactions is close to 1.00 for each of the six time points. Given that for each of the six time points, the percent of ribozymes bearing the wild-type genotype plus the total percentage of ribozymes bearing only a mutation at position 73, is approximately 50% of the total population, this relatively low k value suggests that the Cooperation reaction strongly selects for either wild-type genotypes or for genotypes sufficiently close to the wildtype that bearing a single, trivial mutation does not decrease their catalytic competence (see Chapter 4).

Effect of random mutations on the formation of catalytic networks

As described previously, the purpose of designing IGS and tag sequences with directed mutations is to be able to assess whether ribozymes are more effective catalysts when working with members being the same genotype or if catalysis is enhanced when ribozymes can interact with differing genotypes. Prior to answering this question, certain parameters affecting the system should be discussed. First, there must be some driving factor which induces the system to spontaneously self-assemble into full-length product. Typically, complexification is entropically unfavorable and the vast majority of latent anabolic processes must be coupled with an energy source (e.g., ATP) that provides the energy required to overcome the unfavorable change in free energy that typically results from complexification. The transesterification reactions employed by introns *in vivo* (described here as the *R2F2* reaction, Figure 3) occur with a ΔG near zero (Riley and Lehman, 2003).

Having established a basis for prebiotic molecules to spontaneously attain complexity, complexification should occur as a result of positive feedback, which relies on the presence of certain members of the molecular population being better catalysts than others. When the effectiveness of a selfish versus a cooperative system was tested for the first time with the *Azoarcus* intron, it was shown that the selfish system outperformed the cooperative system when each system operated in isolation. In contrast, when the two systems were allowed to operate together, the growth of enzymes resulting from the cooperative network exceeded that of the selfish network (Vaidya et al., 2012).

In this work, the Selfish and Cooperation reactions were set up such that both systems operated in isolation of one another. The Selfish reaction was composed of two genotypes that each interacted with members of a like genotype (i.e., AU genotypes worked together and CG genotypes worked together) to catalyze the formation of full-length product. Thus, these two Selfish genotypes were competing against one another for a common substrate (the hZ fragment). A different situation was occurring in the Cooperation reaction. Here, the two substrates were obligated to work together to accomplish catalysis. The GAG triplet in the IGS of one genotype of substrate needed to interact with the CUU triplet tag of other substrate genotype. The same two substrates were also effective if they switch roles; the GCG triplet could act as the IGS if, and only if, it interacted with the CGU tag on the other substrate fragment.

For biological systems (or in this case, for systems that are moving toward attaining biological attributes) evolutionary game theory has been used to determine the fitness of a population (Nowak and Sigmund, 2004). In the case of the *Azoarcus* intron, a series of serial-dilution experiments were used in conjunction with ordinary differential equation (ODE) models to forecast the probable outcomes of multiple versions of two-member systems (Yeates et al., 2016). One examples of a possible outcome is dominance, in which one genotype is so clearly the better catalyst that it drives the other genotype to extinction. A contrasting example is cooperation, where neither genotype may be strongly favored, nor is either genotype necessarily driven to extinction. As described briefly in the introduction to this chapter, the specific changes to the middle nucleotides of the IGS and tag sequences used for the Selfish and Cooperation reactions discussed in this chapter have

not been explored previously, but the ODE model was used to predict the most likely outcome in both cases. In the Selfish reaction, CG was predicted to dominate over AU (96.0% versus 4.0%) and in the Cooperation reaction, neither AG nor CU was expected to markedly outperform the other (53.2% versus 46.8%).

A comparison of the middle nucleotide of the three-nucleotide region that indicates the presence of the tag in each recombination product was used as the basis for determining the percentage of each version of the (WXY)**-fragment from which the full-length product was derived for the Selfish and Cooperation reactions (Figure 9).

For the Selfish reaction, at the 30-minute time point, CG clearly dominates over AU. However, as the reaction proceeds over time, the CG substrate is clearly the better catalyst at each time point, but the AU is not driven toward extinction, as expected. Instead, there is a steady decline of CG product and a steady increase of AU product. The self-assembly autocatalytic rate constants for the CG and AU genotypes are 0.0415 min^{-1} and 0.0319 min^{-1} , respectively (Yeates et al., 2016). Thus, CG catalyzes the self-assembly reaction 1.3-times faster than does AU. A possible explanation for the gradual evening out of the two genotypes is that because of its superiority as a catalyst, the CG substrate disappears into full-length product more rapidly. As the concentration of substrate CG decreases, AU may become more catalytically active simply because the decrease in GC substrate dictates an increase in AU substrate.

For the Cooperation reaction, the percent of each genotype present remains relatively even over the course of the reaction. This agrees with the ODE-based prediction and supports the hypothesis that these two genotypes are most successful when they

cooperate to accomplish catalysis. Whereas the genotypes involved in the Selfish reaction were each efficient catalysts independent of one another, the two genotypes used in the Cooperation reaction perform very poorly on their own with self-assembly autocatalytic rate constants of 0.0004 min^{-1} and 0.0001 min^{-1} , respectively. It is interesting to note that the autocatalytic rate constant for the wild-type genotype (UA) is 0.0197 min^{-1} , which is not as efficient as either of the variants used in the Selfish reaction (Yeates et al., 2016). The increased efficiency of these two catalysts may account, in part, for the more robust product yield obtained in the mutagenic PCR reactions versus the 9% reactions (Table 1, Table 19).

In every case that the Selfish and Cooperation reaction have been compared, it has been shown that the Cooperation reaction was generally less tolerant of accepting random mutations. A clear reason for this does not readily emerge from an analysis of the data. However, a possible explanation for this observation may lie in the manner in which catalysis is accomplished by each of the genotypes. For the Selfish reactions, the intermolecular reaction involves two like-fragments whereas the Cooperation reaction requires that two fragments bear different genotypes in order to effectively participate in catalysis. As described previously, substrate fragment and the catalyst must interact in order to correctly position the substrate fragment for its nucleophilic attack on its 3'-binding partner. It is likely that the extra complexity imposed by the requirement of two different genotypes interaction curtails the ability to support random mutations in other regions of the fragment. Thus, the hypothesis that the Cooperation reaction would be less tolerant of accepting random mutations than the Selfish reaction would be, based on the

necessity of the cooperative system to involve two substrate fragments of differing genotypes to work together to accomplish catalysis, is supported by these results.

Conclusion

In this chapter, the results obtained from two related reactions were compared and contrasted. Both reactions were prepared according to a protocol designed to incorporate random mutations into the longer of two substrate fragments. These substrate fragments were then used in self-assembly reactions to form full-length product. In addition to the incorporation of random mutations, the fragment substrates were designed with point mutations in the IGS and tag sequences to create a variety of genotypes that could interact in either a cooperative or selfish manner to form full-length product.

The hypothesis that cooperation will outperform selfishness was neither supported nor disproven. The results from the Selfish and Cooperation reactions showed that both methods produced robust product yield. The results from the Cooperation reaction were in agreement with the prediction given by a mathematical model designed to predict the ratios of each substrate genotype expected to be observed in the full-length product. For the Selfish reaction, the mathematical model predicted a dominance of one product over the other. Although the predicted dominant genotype was clearly in excess at each reaction time point, the genotype predicted to be less catalytically effective was not driven to near-extinction as the model predicted. On the contrary, this presence of this genotype was observed to steadily rise as the reaction proceeded. Both of the genotypes used have relatively high self-assembly autocatalytic rate constants. It is likely that the mathematic prediction was not borne out because since both of the genotypes involved are efficient

catalysts, as the concentration of the more dominant was decreased because it had been incorporated into full-length product, the lesser genotype was available in greater concentration and gradually became more involved in catalysis.

An ancillary hypothesis arose out of the second specific aim because the substrates used in the first and second specific aims were prepared differently. The first specific aim involved substrate fragments only indicative of one region of the ribozyme (e.g. W*, X*, Y*, or Z*). Because the amplification step required for the preparation of these fragments involves the use of forward and reverse primers of a standard sequence, the region of the substrate that would have been available to accept random mutations was too limited to produce a significantly mutated product. Thus, those substrates were produced according to a protocol that provided for a 9% mutation rate at each position, with every nucleotide position susceptible to a 3% chance of incorporating each of the three alternate nucleotides. The reactions in the second specific aim consisted of only two substrate fragments, the larger of the two being subjected to random mutations. Both because cost was prohibitive for acquiring the longer substrates produced by the 9% mutation rate protocol, and because the percentage of the fragment covered by primers was much lower, the substrate fragments for the second specific aim were prepared by a previously-optimized mutagenic PCR protocol. The obtained results suggested that the ancillary hypothesis that the substrates would produce similar results with respect to positions more tolerant of accepting random mutations and the ability catalyze the production of full-length product should be rejected.

During the process of preparing the mutated substrates for the mutagenic PCR reactions, the mutation rate was estimated at approximately 20% per position. The

experimental results suggested that this was a gross over-estimation of the true mutation rate for the substrate fragments. The full-length product obtained from the substrate fragments produced via mutagenic PCR in both reactions showed a striking similarity to the wild-type ribozyme. For example, positions that tended to select against the wild-type were mostly located in positions of the ribozyme where it would be expected that variations from the wild-type genotype would have the least impact. In contrast, the full-length product obtained from the 9% reactions tended to be comprised of fragments that tolerated mutations at positions that would likely be expected to interfere with both the secondary and tertiary structure of the molecule. The only exception to this is found at position 170, which in the wild-type is involved in a base triplet that can form a recombination junction. The full-length ribozymes generated through recombination of substrates prepared at the 9% mutation rate showed a tendency to select against the wild-type at this position. The experimental design did not allow for a determination as to whether this tendency to tolerate a mutation would also be observed when the substrate had been prepared by the mutagenic PCR protocol.

Additionally, in the substrates prepared by the mutagenic PCR protocol, the observed mutations generally followed a pattern common to mutations observed in *in vivo* nucleic acids – each mutated position (with one notable exception) strongly favored a transition and strongly disfavored a transversion. In contrast, the substrates prepared at the 9% mutation rate showed no preference for transitions or transversions. In most cases, there was an almost even distribution between each of the three possible alternatives to the wild-type. Given that the preparation of the mutated substrates involved an equal chance

of any of the three alternates being substituted at each position, the even distribution of the non-wild-type bases is not surprising.

Finally, while the positions that tended to select against the wild-type were consistent at each reaction time point (and in the case of the W*-fragment, between the two 9% reactions that involved the same mutated fragment in different recombination events), these positions were entirely different between the 9% reactions and the mutagenic PCR reactions. This suggests that the mutagenic PCR full-length product was a slightly-modified analog of the wild-type ribozyme, whereas the ribozyme produced in the 9% series was positioned on the verge of becoming the biomolecular equivalent of a speciation event.

A t test was used to determine whether there was a significant difference between the means of the average k values for the Selfish and Cooperation reactions. The calculated value, $t = 5.756$ is above the critical value at a 95% confidence level. Thus there is a statistical difference between the two reactions ($p < 0.001$).

Materials and Methods

Construction of RNA substrates and ribozymes

The hZ RNA-substrate from Integrated DNA Technologies (IDT). For the (WXY)**-analogs, each of the four fragments, a full-length template of the wild-type intron was prepared by PCR mediated gene synthesis (Jayarman et al., 1989), for which the preparation and amplification of was described in detail in the Materials and Methods section in the previous chapter. The resulting product was then used as the template strand

for a PCR protocol modified to facilitate the incorporation of random mis-insertions at any site. This was achieved in 100 μ L reactions containing < 1 pmol DNA template, 20 pmol of each primer, 10 μ L 10X PCR buffer, 0.5 mM MnCl₂, and 10 U of *Taq* DNA Polymerase (Fermentas) and dNTPs with a ratio of 100:3 for TTP:ATP and GTP:CTP. Reactions were carried out in the Thermocycler using a previously optimized PCR program consisting of 50 cycles of 93°C for 1 min, 50°C for 45 sec, and 72°C for 10 min. Presence of PCR product of the expected length was confirmed by electrophoresis on 2% agarose gel. In preparation for transcription, four such reactions were combined, ethanol precipitated, and rehydrated in nuclease-free distilled water. Transcription and purification for use as reaction substrates was conducted in the same manner described in the previous chapter.

Prior to designing the reactions described in this chapter, an approximate mutation rate was determined on an XYZ fragment that had been prepared according to the mutagenic PCR process described here. Presence of successful PCR product was confirmed on 2% agarose gel, with PCR product prepared according to the standard PCR protocol used as a control. In order to determine an approximate mutation rate, both the standard PCR and mutagenic PCR products were digested in a solution containing 20 U of Hind III restriction enzyme and 2 μ L proprietary reaction buffer (New England Biolabs). Hind III is designed to cut a sequence in the manner 5'-A ∇ AGCTT-3', which occurs in the cDNA for the *Azoarcus* intron from positions 96-101. The restriction digest was carried out at 37°C for one hour. Digest On the basis that all wild-type products should be cut and any fragment that carried an altered nucleotide in any of those six positions would remain uncut. Product was then run on a 2% agarose gel. The gel image was visualized on the

Typhoon scanner and the Image Quant program was used to estimate relative band intensities corresponding to full-length fragments and fragments that had been cut by the restriction enzyme. The mutation rate was calculated by the formula $\mu = 1 - [b/(a + b)]^{(1/6)}$ where μ is the mutation rate, a is the percentage of uncut fragment, b is the percentage of cut fragment. The results of three trials were compared with the estimated average mutation rate being 20%.

Data Analysis and Error Analyses

The data from these two reactions were processed in the same manner as the data was processed in Chapter 2, with exceptions made for the unique nature of this data. Because the WXY-analog fragments were prepared with a reverse primer that halted the generation of PCR products at the end of the Y' fragment, a different reverse primer was used as part of the mutagenic PCR protocol (see Materials and Methods, below). As a result, positions 138-154 were not subject to random mutations during that process. Because these positions were “fixed” as wild-type during the reaction and amplifications subsequent to reverse transcription, these positions are omitted from data analysis in this chapter, except for an analysis to find and determine the identity of the tag sequence as described below.

Because there is no plausible way to determine the sequence of the cDNA obtained upon reverse transcription, the ensuing PCR amplification only allows for one forward primer to be used. As a result, the identity of the middle nucleotide of the iGS is obliterated at this step and thus cannot be confirmed during the analysis of the MiSeq data. To overcome this obstacle, each ribozyme product in these reactions was identified by the presence of one of the expected three-nucleotide internal tag codes (either CGU or CUU

in both reactions). The IGS for each ribozyme was then assigned the correct middle-position IGS nucleotide based upon context.

As described previously, the tendency of the *tF2* recombination mechanism to introduce varied insertions at a recombination junction creates ambiguity in aligning full-length product sequences. A qualitative review of the sequences from these two reactions, as well as the review of the sequences from the Z*-Single, W*-Single, and Double reactions provided a context for estimating a plausible position in the molecule at which the tag could be found. In the wild-type sequence, a fragment designed to go through the *tF2* mechanism at the Y-Z junction will bear the sequence UAAGGCACAU, beginning at position 147. Because the designed sequences only differed in the penultimate nucleotide, sequences reading UAAGGCACGU and UAAGGCACUU from position 147 could easily be sorted as containing one of the two modified tag sequences. Sequence variance as a result of recombination tends to arise at a recombination junction (Draper, 2008). For that reason, any occurrence of CUU or CGU was considered to be the tag sequence unless compelling contextual clues from the adjacent molecules clearly suggested that it was not. Given that the head region of the Z fragment is engineered to be GGCAU, any instance or instances of CAU were considered to have arisen from the recombination event and were not included in the data analysis because they occurred after the region of interest in the molecule.

SELFISH - IGS and TAG sequences are complimentary

COOPERATIVE - IGS and TAG sequences are non-complimentary

Figure 7 – Reaction Design for Selfish and Cooperation Reactions. In the wild-type *Azoarcus* ribozyme, the IGS sequence is GUG and the tag sequence is CAU. For the Selfish and Cooperation reactions, the middle nucleotide of the IGS and tag sequences were altered to allow a situation in which each substrate ribozyme fragment could interact with a like fragment (upper diagram) or a situation in which the markedly more robust interaction would arise from an interaction between two different genotypes (lower diagram). These two reactions are referred to in the text collectively as the “mutagenic PCR” reactions in situations that require they be distinguished from the reactions using substrates engineered to have a 9% mutation rate per position.

Figure 8 – Percent Wildtype for the Selfish and Cooperation Reactions. The percent of the full-length product population bearing the wild-type genotype is shown for each of the six time points for the Selfish and Cooperation reactions. The 2-hour time point for the Selfish reaction was excluded at a 99%- confidence interval on the basis of a Q-test.

Figure 9 – Percent of reaction population bearing one of two available IGS/tag sequence combinations. The identity of the middle nucleotides is used to distinguish between the two possible IGS/tag combinations that could be present in each reaction. The percent of each IGS/tag combination at each reaction time point is given for the (A) Selfish reaction and for the (B) Cooperation reaction.

<i>Substrate</i>	<i>Concentration (μM)</i>	<i>Amount (pmol)</i>	<i>Substrate</i>	<i>Concentration (μM)</i>	<i>Amount (pmol)</i>
Selfish (WXY)** + hZ			Cooperation (WXY)** + hZ		
WXY (GAG/CUU)	1.0	899	WXY (GAG/CGU)	1.0	1087
WXY (GCG/CGU)	1.0	900	WXY (GCG/CUU)	1.0	1087
hZ*	2.5	2250	hZ*	2.5	2750

Table 18 – Concentrations of and amounts of substrates used in the Selfish and Cooperation reactions. The rationale for adding substrates in 1.25-fold excess over its 5' binding partner is described in the text. Because both of these reactions required two substrates made in the research lab, the estimated yield of the fragment with the lowest yield was used as the limiting reaction to which other substrates and final reaction volume were adjusted.

For each of the WXY-analog substrate fragment used, the identity of the IGS/tag combination of that particular substrate is included in parentheses.

A distinction is made here between randomly-mutated substrates which were manufactured using a protocol that provided for 9% mutation rate per position and substrates which were produced in-house using the mutagenic PCR protocol described in the text. The former are designated as * and the latter are designated as **. Parentheses around the fragment designation (WXY in this case) indicate that the entire portion of the molecule contained in parenthesis was subjected to the introduction of random mutations.

	Selfish	Cooperation
30m	210200	132700
1h	224200	133600
2h	80300	131900
4h	230100	103800
6h	91000	184900
8h	158700	174100

Table 19 – Number of sequence available for analysis at each time for the Selfish and Cooperation reactions. The same method that was described in the Materials and Method section for Chapter 2 was used to process the raw data representing the sequence identities of each full-length intron obtained from these two reactions. As was done with the data from the single reactions, a random sample of 10,000 sequences was taken from each time point for analysis. The relatively low number of sequences available for analysis from the 2-hour time point of the Selfish reaction is most likely the result of a pipetting error that occurred during sample processing. The reason for the lower number of sequences available for analysis from the 6-hour time point of the Selfish reaction is unknown.

		30m	1h	2h	4h	6h	8h
A	24	0	0	1	0	0	0
A	25	0	0	0	0	1	0
A	26	1	1	2	0	1	0
C	27	0	0	0	0	0	0
C	28	0	0	0	0	0	0
A	29	0	0	0	0	0	0
C	30	0	0	1	1	1	0
G	31	1	1	4	1	1	1
C	32	0	0	0	0	0	0
A	33	1	1	2	1	1	1
A	34	3	3	5	3	4	3
G	35	1	1	1	1	1	1
G	36	1	1	1	0	1	1
G	37	1	0	1	0	0	1
A	38	4	5	6	4	5	5
U	39	3	3	4	4	3	4
G	40	1	1	1	1	1	1
G	41	0	0	1	0	0	0
U	42	3	3	6	3	3	3
G	43	0	0	1	0	0	0
U	44	4	4	6	5	5	5
C	45	0	0	0	0	0	0
A	46	0	0	1	1	2	1
A	47	0	0	1	0	1	0
A	48	0	0	1	0	1	0
U	49	3	3	6	4	3	3
U	50	1	1	2	1	1	1
C	51	0	0	0	0	0	0
G	52	0	0	1	0	0	0
G	53	0	0	0	0	0	0
C	54	0	0	0	0	0	0
G	55	0	0	0	0	0	0
A	56	0	0	1	0	1	0
A	57	0	0	1	0	1	0
A	58	1	1	2	1	1	1
C	59	0	0	0	0	0	0
C	60	0	0	0	0	0	0
U	61	2	2	4	2	2	2
A	62	0	0	1	0	1	0
A	63	1	2	3	1	2	1
G	64	1	1	1	1	1	1
C	65	0	0	0	0	0	0
G	66	1	1	1	1	1	1
C	67	0	0	1	0	0	0
C	68	0	0	0	0	0	0
C	69	1	0	1	1	1	1
G	70	2	2	2	2	2	2
C	71	0	0	0	0	0	0

		30m	1h	2h	4h	6h	8h
C	72	0	0	0	0	0	0
C	73	39	39	35	40	37	41
G	74	2	1	4	2	2	2
G	75	1	0	2	1	1	0
G	76	1	1	2	1	1	1
C	77	1	1	1	1	1	0
G	78	1	1	2	1	1	1
U	79	2	1	3	2	2	2
A	80	1	2	3	2	2	2
U	81	4	4	5	4	4	4
G	82	0	0	1	0	0	0
G	83	1	1	2	1	1	1
C	84	0	0	0	0	0	0
A	85	1	1	2	2	2	1
A	86	1	0	1	1	1	1
C	87	0	0	1	0	1	0
G	88	2	2	3	2	3	2
C	89	0	0	0	0	0	0
C	90	0	0	1	0	1	0
G	91	1	1	1	1	1	1
A	92	1	1	1	1	1	1
G	93	1	1	1	1	1	1
C	94	0	0	0	0	0	0
C	95	0	0	0	0	0	0
A	96	1	1	2	1	2	1
A	97	2	2	5	3	3	2
G	98	1	1	1	1	1	1
C	99	0	0	0	0	0	0
U	100	3	4	7	3	4	4
U	101	0	0	0	0	0	0
C	102	0	1	1	0	1	0
G	103	1	1	2	1	1	1
G	104	1	1	1	1	1	0
C	105	0	0	1	0	1	1
G	106	1	1	2	1	1	1
C	107	0	0	0	0	0	0
C	108	1	0	1	1	1	1
U	109	5	5	7	6	6	6
G	110	1	1	1	1	1	0

		30m	1h	2h	4h	6h	8h
C	111	1	1	2	1	1	1
G	112	1	2	5	2	3	1
C	113	0	0	1	0	1	0
C	114	1	1	1	1	1	1
G	115	0	0	1	0	0	0
A	116	1	1	2	1	2	1
U	117	0	1	1	0	1	1
G	118	0	0	0	0	0	0
A	119	3	3	4	3	3	2
A	120	4	5	7	5	4	5
G	121	0	0	0	0	0	0
G	122	0	0	0	0	0	0
U	123	1	1	1	1	1	0
G	124	0	0	0	0	0	0
U	125	0	0	0	0	0	0
A	126	2	2	3	2	2	1
G	127	0	0	0	0	0	0
A	128	1	1	2	1	2	1
G	129	0	0	0	0	0	0
A	130	1	1	1	1	1	1
C	131	0	0	0	0	0	0
U	132	0	0	1	0	1	0
A	133	1	1	2	1	1	1
G	134	0	0	0	0	0	0
A	135	2	1	2	1	1	1
C	136	0	0	0	0	0	0
G	137	0	0	0	0	0	0

C	152	0	0	0	0	0	0
G/U	153	*	*	*	*	*	*
U	154	0	0	0	0	0	0

Table 20 – Selfish Reaction – Mutations by Position. Total percent of positions mutated at each position at each time point ($N = 10,000$ for each time point), with a correction for background noise applied as described in the text. Position 153 is denoted with an asterisk (*) at each time point because rather than undergoing a mutational analysis, this position was used to identify the genotype of the IGS/tag combination.

		30m	1h	2h	4h	6h	8h
A	24	0	0	0	0	0	0
A	25	0	0	0	0	0	0
A	26	0	0	0	0	0	0
C	27	0	0	0	0	0	0
C	28	0	0	0	0	0	0
A	29	0	0	0	0	0	0
C	30	0	0	0	0	0	0
G	31	1	1	1	1	1	1
C	32	0	0	0	0	0	0
A	33	0	0	0	0	0	0
A	34	2	2	3	2	2	2
G	35	1	1	1	1	1	1
G	36	0	0	0	0	0	0
G	37	0	0	0	0	0	0
A	38	4	3	4	4	3	3
U	39	2	1	1	1	1	1
G	40	0	1	0	0	0	0
G	41	0	0	0	0	0	0
U	42	1	2	1	1	2	2
G	43	0	0	0	0	0	0
U	44	2	1	1	2	1	1
C	45	0	0	0	0	0	0
A	46	0	0	0	0	0	0
A	47	0	0	0	0	0	0
A	48	0	0	0	0	0	0
U	49	1	1	1	2	1	1
U	50	0	0	0	0	0	0
C	51	0	0	0	0	0	0
G	52	0	0	0	0	0	0
G	53	0	0	0	0	0	0
C	54	0	0	0	0	0	0
G	55	0	0	0	0	0	0
A	56	0	0	0	0	0	0
A	57	0	0	0	0	0	0
A	58	0	0	0	0	0	1
C	59	0	0	0	0	0	0
C	60	0	0	0	0	0	0
U	61	2	1	1	1	1	1
A	62	0	0	0	0	0	0
A	63	0	0	0	0	1	1
G	64	0	0	0	0	0	0
C	65	0	0	0	0	0	0
G	66	0	0	0	0	0	0
C	67	0	0	0	0	0	0
C	68	0	0	0	0	0	0
C	69	0	0	0	0	0	0
G	70	1	1	1	1	1	1
C	71	0	0	0	0	0	0

		30m	1h	2h	4h	6h	8h
C	72	0	0	0	0	0	0
C	73	31	26	27	28	26	27
G	74	2	2	1	1	1	1
G	75	0	0	0	0	0	0
G	76	0	0	0	0	0	0
C	77	0	0	0	0	0	0
G	78	1	1	0	0	0	0
U	79	1	1	1	1	1	1
A	80	1	1	1	1	1	1
U	81	1	1	1	1	1	1
G	82	0	0	0	0	0	0
G	83	0	0	0	0	0	0
C	84	0	0	0	0	0	0
A	85	0	0	0	0	0	0
A	86	0	0	0	0	0	0
C	87	0	0	0	0	0	0
G	88	1	1	1	1	1	1
C	89	0	0	0	0	0	0
C	90	0	0	0	0	0	0
G	91	0	0	0	0	0	0
A	92	0	0	0	0	0	0
G	93	0	0	0	0	0	0
C	94	0	0	0	0	0	0
C	95	0	0	0	0	0	0
A	96	0	0	0	0	0	0
A	97	1	1	1	1	1	1
G	98	0	0	1	1	0	1
C	99	0	0	0	0	0	0
U	100	1	1	2	2	2	2
U	101	0	0	0	0	0	0
C	102	0	0	0	0	0	0
G	103	0	0	1	0	1	0
G	104	0	0	0	0	0	0
C	105	0	0	0	0	0	0
G	106	0	0	0	0	0	0
C	107	0	0	0	0	0	0
C	108	0	0	0	0	0	0
U	109	2	2	2	2	2	2
G	110	0	0	0	0	0	0

		30m	1h	2h	4h	6h	8h
C	111	0	0	0	0	0	0
G	112	0	0	0	0	0	0
C	113	0	0	0	0	0	0
C	114	0	0	0	0	0	0
G	115	0	0	0	0	0	0
A	116	0	0	0	0	0	0
U	117	0	0	0	0	0	0
G	118	0	0	0	0	0	0
A	119	1	1	1	1	1	1
A	120	4	4	4	4	4	4
G	121	0	0	0	0	0	0
G	122	0	0	0	0	0	0
U	123	0	0	0	0	0	0
G	124	0	0	0	0	0	0
U	125	0	0	0	0	0	0
A	126	0	0	0	0	0	0
G	127	0	0	0	0	0	0
A	128	0	0	0	0	0	0
G	129	0	0	0	0	0	0
A	130	0	0	0	0	0	0
C	131	0	0	0	0	0	0
U	132	0	0	0	0	0	0
A	133	0	0	0	0	0	0
G	134	0	0	0	0	0	0
A	135	0	0	0	0	0	0
C	136	0	0	0	0	0	0
G	137	0	0	0	0	0	0

C	152	0	0	0	0	0	0
G/U	153	*	*	*	*	*	*
U	154	0	0	0	0	0	0

Table 21 – Cooperation Reaction – Mutations by Position. Total percent of positions mutated at each position at each time point ($N = 10,000$ for each time point), with a correction for background noise applied as described in the text. Position 153 is denoted with an asterisk (*) at each time point because rather than undergoing a mutational analysis, this position was used to identify the genotype of the IGS/tag combination.

A												
POSITION	SELFISH				COOPERATION				OVERALL			
	A	C	G	U	A	C	G	U	A	C	G	U
38	95	0	5	0	97	0	3	0	96	0	4	0
39	1	3	0	96	0	1	0	99	0	2	0	98
49	0	4	0	96	0	1	0	99	0	2	0	98
61	0	2	0	98	0	1	0	99	0	2	0	98
70	1	1	98	0	1	0	99	0	1	0	98	0
73	1	59	22	18	0	70	11	19	1	65	16	18
74	2	1	98	0	1	0	98	0	1	1	98	0
81	1	4	0	96	0	1	0	99	0	3	0	97
88	2	1	97	0	1	0	99	0	1	1	98	0
100	0	4	0	95	0	2	0	98	0	3	0	97
109	1	5	0	94	0	2	0	98	1	3	0	96
120	95	0	4	0	96	0	4	0	96	0	4	0
126	98	0	2	0	100	0	0	0	99	0	1	0

B			
	Wildtype	Selects for	Selects against
A	3	G	C, U
C	1	G, U	A
G	3	A, C	U
U	6	C	A, G

Table 22 – Percent of each nucleotide present at positions of interest for the Selfish and Competition reactions. A total of 13 positions were selected from a qualitative review of the total mutation rates for the selfish and the cooperation reactions as positions that appeared to be more prone to tolerating random mutations during self-assembly. **(A)** For each of these positions, the total percent of nucleobase being represented ($N = 60,000$) by each of the four possibilities is given for the selfish and cooperation reactions, with the overall percentage ($N = 120,000$) of nucleotides appearing at each position for all reactions that involved substrate prepared via the mutagenic PCR. **(B)** A pattern of favored and disfavored substitutions arose for these 13 positions of interest. For the four nucleotide possibilities, the number of wild-type positions occupied by each of the four bases is given along with the nucleotide substitutions that were selected for and against. The trend of favoring transition and favoring transversion was consistent across 12 of the 13 positions of interest.

	SELFISH	COOPERATION		SELFISH	COOPERATION		SELFISH	COOPERATION
24	0.005	-0.006	72	-0.004	0.010	111	-0.015	0.017
25	0.048	-0.005	73	0.200	-0.317	112	0.003	0.000
26	-0.051	-0.006	74	-0.061	-0.047	113	0.002	-0.002
27	0.007	0.003	75	-0.042	0.004	114	0.006	-0.012
28	0.002	0.003	76	-0.065	-0.014	115	0.003	0.005
29	0.001	-0.008	77	0.007	0.010	116	0.071	0.006
30	-0.024	-0.003	78	-0.006	-0.021	117	0.120	0.002
31	-0.111	-0.027	79	-0.006	-0.010	118	0.006	0.003
32	0.005	-0.009	80	-0.022	-0.016	119	-0.057	-0.048
33	-0.009	-0.003	81	-0.026	-0.034	120	-0.056	-0.066
34	-0.017	-0.047	82	-0.021	-0.004	121	-0.013	-0.001
35	-0.035	-0.006	83	-0.042	0.007	122	-0.011	0.010
36	0.005	0.000	84	-0.010	0.003	123	-0.016	-0.001
37	0.024	-0.006	85	0.033	0.001	124	-0.001	0.007
38	-0.017	-0.065	86	-0.001	0.007	125	0.002	-0.002
39	0.034	-0.046	87	0.006	-0.002	126	-0.053	-0.013
40	-0.025	-0.004	88	-0.003	0.013	127	0.007	0.006
41	-0.020	-0.003	89	-0.016	0.000	128	0.020	-0.002
42	-0.098	0.019	90	0.003	0.012	129	-0.007	0.001
43	-0.014	0.000	91	-0.017	-0.023	130	-0.003	0.008
44	0.096	-0.066	92	-0.014	0.004	131	-0.005	-0.005
45	0.001	-0.002	93	-0.015	-0.015	132	0.012	-0.003
46	0.051	-0.002	94	-0.006	-0.009	133	-0.062	0.011
47	-0.005	0.000	95	0.011	0.003	134	0.001	-0.002
48	0.002	-0.009	96	0.019	0.003	135	-0.069	-0.001
49	-0.101	-0.033	97	-0.067	-0.037	136	-0.005	0.000
50	-0.032	-0.001	98	-0.022	0.013	137	-0.020	0.001
51	0.002	-0.013	99	0.005	-0.001			
52	-0.012	-0.010	100	-0.089	0.035			
53	-0.014	-0.010	101	0.000	0.000			
54	0.007	0.005	102	0.033	0.001			
55	-0.019	0.005	103	-0.039	0.008			
56	0.022	0.000	104	-0.010	0.016			
57	0.026	0.000	105	0.004	-0.020			
58	-0.023	0.021	106	0.015	0.001			
59	-0.001	0.001	107	0.002	-0.005			
60	-0.007	-0.010	108	-0.003	0.010			
61	-0.109	-0.057	109	0.096	0.009			
62	0.002	0.002	110	-0.003	-0.005			
63	-0.054	0.014						
64	-0.004	-0.004						
65	0.014	0.000						
66	-0.024	-0.016						
67	0.001	-0.005						
68	-0.027	-0.009						
69	-0.012	0.009						
70	-0.010	-0.069						
71	0.009	-0.011						

Table 23 - Trend in total mutation rate between the 30-minute and 8-hour time point for Selfish and Cooperation reactions. The trends were calculated using a best-fit line between each of the six data points. The average rate of change per hour is given. Positive values indicate a tendency to accumulate mutations at a given position while negative values indicate an overall decrease in the number of mutations at a given position. For qualitative reference, black boxes associated with a base position indicate a general increase in mutation rate whereas red boxes associated with a base position indicate a general decrease in the mutation rate. White boxes indicate no overall change over time.

30m		1h		2h	
Wildtype	1845	Wildtype	1694	Wildtype	1051
C73G	671	C73G	692	C73G	380
C73U	666	C73U	611	C73U	187
U44C	78	A38G	92	U42C	59
A38G	78	U109C	82	A120G	58
U109C	73	U44C	80	U100C	56
U42C	64	A120G	80	A38G	56
U49C	52	U42C	73	U61C	44
A120G	51	U49C	69	U49C	42
U100C	50	U100C	64	U109C	41
U81C	50	A34G	54	G31A	40
A34G	47	U81C	47	A34G	37
A119G	43	U39C	41	G112C	37
U39C	37	G70A	37	U44C	35
C73G, A120G	32	A119G	36	G74A	35
A97G	31	U61C	35	C73G, A120G	34
A126G	31	C73U, U100C	32	C73G, U100C	30
G74A	31	C73G, U109C	28	U39C	29
A135G	30	C73G, U81C	28	U81C	28
G88C	30	C73U, U109C	27	A119G	28
4h		6h		8h	
Wildtype	1775	Wildtype	1708	Wildtype	1692
C73G	706	C73G	534	C73G	686
C73U	631	C73U	524	C73U	583
U109C	90	U109C	92	U44C	86
A120G	75	A38G	87	A38G	81
U44C	68	A34G	66	U109C	78
A38G	66	U44C	63	A120G	66
U42C	63	G112C	60	U39C	49
U49C	57	A120G	55	A34G	46
U39C	56	C144.	51	U49C	46
U81C	52	U100C	42	U42C	44
U61C	48	U42C	35	U81C	43
U100C	45	U39C	33	U100C	39
A34G	43	A119G	30	C73G, U109C	39
A97G	41	U81C	30	U44C, C73G	35
C73G, U109C	40	C73G, A120G	27	C73U, U109C	34
A119G	38	A116G	27	G110.	33
C73G, A120G	34	A126G	26	A63G	32
G74A	32	G74A	25	A126G	31
C73G, U81C	30	C73U, U109C	25	C73U, A120G	31

Table 24 – The most prevalent population members for the Selfish Reaction at each of the six time points. The 20 most frequently-represented genotypes in the population are shown, with the number of times those genotypes appeared in the sample population ($N = 10,000$). Because of the number of nucleotides (113 nt) available for analysis, only the designation “Wildtype” or the nature of the mutation(s) away from the wild-type genotype is included.

30m		1h		2h	
Wildtype	3295	Wildtype	3704	Wildtype	3598
C73U	1151	C73U	1002	C73U	973
C73G	576	C73G	516	C73G	560
A120G	132	A120G	134	A38G	147
A38G	119	A38G	102	A120G	123
A34G	88	A34G	93	A34G	98
G74A	70	U109C	81	U109C	57
U109C	69	U42C	68	U81C	51
U39C	61	G74A	58	U100C	48
U44C	58	U61C	58	G88A	46
U61C	55	A119G	51	A119G	45
A97G	51	U44C	48	G70A	44
G70A	50	G70A	45	G74A	43
U42C	48	G88A	45	U61C	39
C73U, A120G	47	U100C	41	G31A	38
U49C	46	G31A	40	U42C	37
G31A	42	G35A	39	A34G, C73U	36
U81C	40	A97G	38	C73U, A120G	34
U100C	37	U39C	36	C114.	34
A119G	35	C73U, A120G	31	C73U, U109C	32
4h		6h		8h	
Wildtype	3582	Wildtype	3819	Wildtype	3795
C73U	884	C73U	973	C73U	946
C73G	560	C73G	528	C73G	575
A120G	142	A38G	139	A120G	117
A38G	129	A120G	125	A38G	102
A34G	88	U109C	92	A34G	101
U109C	73	A34G	81	U109C	76
U44C	62	G31A	59	U42C	64
U49C	58	G74A	56	G74A	55
U100C	57	G109i	52	U100C	54
G74A	56	U42C	52	U61C	50
U61C	56	U100C	51	G31A	44
U39C	54	G70A	50	A97G	40
U42C	51	G88A	49	U44C	40
U81C	50	U61C	46	G88A	39
C73U, A120G	49	A119G	44	G109.	38
A119G	43	U110.	43	G109i	36
G70A	36	A97G	38	U81C	34
A97G	33	A38G, C73U	33	G35A	31
G88A	33	C73U, A120G	31	A119G	31

Table 25 – The most prevalent population members for the Cooperation Reaction at each of the six time points. The 20 most frequently-represented genotypes in the population are shown, with the number of times those genotypes appeared in the sample population ($N = 10,000$). Because of the number of nucleotides (113 nt) available for analysis, only the designation “Wildtype” or the nature of the mutation(s) away from the wild-type genotype is included.

Rank	Selfish		Cooperation	
	Genotype	N	Genotype	N
1	Wildtype	6	Wildtype	6
2	C73G	12	C73U	12
3	C73U	18	C73G	18
4	A38G	33	A34G	26
5	U109C	35	A38G	37
6	U44C	41	U109C	42
7	A120G	42	G74A	58
8	U42C	50	U42C	71
9	U100C	63	U61C	73
10	A34G	64	U100C	75

Table 26 – Niche populations for the Selfish and Cooperation reactions. As described in the Materials and Methods for Chapter 2, the 20 most prevalent genotypes observed at each time point in both reactions were identified and ranked in order of percent occurrence at that reaction time point. The rankings were then added and each genotype was assigned a niche genotype score (N-score) based upon the sum of the rankings. Thus, for each of these reactions, the niche genotype consists of the ten most prevalent genotypes.

μ (%)				μ (%)				μ (%)			
	Selfish	Cooperation	Overall		Selfish	Cooperation	Overall		Selfish	Cooperation	Overall
24	0.36	-0.04	0.16	72	0.41	0.15	0.28	111	0.95	0.26	0.61
25	0.40	0.05	0.22	73	37.57	27.68	32.62	112	2.48	0.18	1.33
26	0.85	0.04	0.44	74	2.01	1.41	1.71	113	0.46	0.15	0.30
27	0.24	0.09	0.16	75	0.78	0.24	0.51	114	0.79	0.34	0.57
28	0.26	0.11	0.19	76	1.13	0.43	0.78	115	0.37	0.10	0.23
29	0.33	0.09	0.21	77	0.83	0.34	0.58	116	1.12	0.18	0.65
30	0.58	0.18	0.38	78	1.15	0.44	0.79	117	2.65	0.01	1.33
31	1.81	1.14	1.47	79	1.75	0.70	1.22	118	0.10	-0.07	0.01
32	0.27	0.17	0.22	80	1.75	0.93	1.34	119	2.67	1.33	2.00
33	1.10	0.17	0.63	81	4.30	1.31	2.81	120	4.71	3.82	4.27
34	3.71	2.38	3.05	82	0.30	0.06	0.18	121	0.23	0.10	0.16
35	1.02	0.73	0.87	83	0.96	0.16	0.56	122	0.18	0.11	0.14
36	0.53	0.23	0.38	84	0.14	-0.06	0.04	123	0.76	0.10	0.43
37	0.62	0.17	0.40	85	1.51	0.27	0.89	124	0.16	0.03	0.09
38	4.87	3.31	4.09	86	0.70	0.08	0.39	125	0.20	-0.02	0.09
39	3.58	1.13	2.36	87	0.45	0.22	0.33	126	1.97	0.19	1.08
40	0.90	0.43	0.67	88	2.54	1.35	1.94	127	0.13	0.05	0.09
41	0.43	0.24	0.33	89	-0.04	-0.11	-0.07	128	1.51	0.11	0.81
42	3.66	1.44	2.55	90	0.42	0.19	0.30	129	0.08	0.04	0.06
43	0.24	0.03	0.14	91	0.76	0.33	0.55	130	0.78	0.16	0.47
44	4.98	1.18	3.08	92	1.19	0.27	0.73	131	0.14	0.05	0.10
45	0.30	0.13	0.21	93	0.94	0.16	0.55	132	0.43	0.03	0.23
46	0.88	0.07	0.47	94	0.29	0.03	0.16	133	1.11	0.25	0.68
47	0.47	0.06	0.26	95	0.19	0.04	0.11	134	0.11	0.01	0.06
48	0.62	0.06	0.34	96	1.58	0.14	0.86	135	1.37	0.31	0.84
49	3.76	1.09	2.43	97	2.88	1.16	2.02	136	-0.03	-0.12	-0.07
50	1.19	0.21	0.70	98	1.04	0.49	0.76	137	0.25	0.41	0.33
51	0.21	0.06	0.14	99	0.24	0.09	0.16				
52	0.34	0.08	0.21	100	4.51	1.55	3.03				
53	0.25	0.09	0.17	101	-0.04	-0.04	-0.04				
54	0.22	0.00	0.11	102	0.68	0.23	0.45				
55	0.27	-0.05	0.11	103	0.91	0.42	0.66				
56	0.46	0.04	0.25	104	0.81	0.31	0.56				
57	0.64	0.08	0.36	105	0.60	0.15	0.37				
58	1.17	0.37	0.77	106	1.14	0.34	0.74				
59	0.07	0.00	0.03	107	0.27	0.09	0.18				
60	0.31	0.09	0.20	108	0.64	0.13	0.38				
61	2.24	1.35	1.79	109	5.79	2.13	3.96				
62	0.59	0.09	0.34	110	0.60	0.23	0.42				
63	1.78	0.49	1.13								
64	0.88	0.27	0.57								
65	0.14	-0.04	0.05								
66	1.08	0.31	0.70								
67	0.46	0.29	0.38								
68	0.07	-0.12	-0.03								
69	0.76	0.21	0.48								
70	1.77	1.20	1.48								
71	0.31	0.05	0.18								

Table 27 – Average mutation rate for Selfish and Competition reactions, with overall mutation rate for the mutagenic PCR reactions. The overall mutation rate, corrected for noise, was calculated for each position for both of these reactions by adding up the total number of mutations at each position for each of the six time points and dividing by the total number of sequences analyzed for that reaction ($N = 60,000$ for each reaction). The overall mutation reflects the average mutation rate at each position for both reactions ($N = 120,000$). The values at some positions for the Cooperation reaction are reported as negative numbers. This is because the number of mutations at each of those positions was so low that it was exceeded by the noise correction for that position (Table 17).

Time	Selfish	Cooperation
30 m	1.65	1.02
1 h	1.69	0.95
2 h	2.50	0.97
4 h	2.72	0.99
6 h	1.94	0.91
8 h	1.74	0.94

Table 28 – Average deviation from wildtype (k) for the mutagenic PCR reactions. The k value is given for each time point in both the Selfish and Cooperation reactions. The same rationale that was used in Chapter 2 (see Table 14) was used here to calculate these values.

CHAPTER 4 – THE EFFECT OF POINT MUTATIONS ON THE ACTIVITIES OF FULL-LENGTH *AZOARCUS* RIBOZYMES

Introduction

A point mutation refers to a genetic mutation where a single nucleotide base is either changed, inserted, or deleted from the wild-type sequence of DNA or RNA. In some cases, the point mutation may disrupt the tertiary structure of the ribozyme. For example, in the *Azoarcus* group I intron, a mutation that disrupts the tertiary contacts between the L9 tetraloop and its helical receptor was shown to destabilize the intron's tertiary structure, whereas other mutations were less destabilizing (Chauhan et al., 2005). On the other hand, the introduction of point mutations may enhance the catalytic ability of the ribozyme. In the case of the *Tetrahymena* intron, small insertions or deletions in a non-conserved region was shown to greatly enhance the turnover number and sequence specificity of the *in vivo* ribozyme-catalyzed RNA cleavage (Young et al., 1991). As the effect of point mutations can be either deleterious or beneficial, the effect of point mutations on the self-assembly of the *Azoarcus* intron is of interest. The effect of appending a foreign sequence of prebiotic relevance at the 3'-end of the intron were shown to decrease, but not entirely block the self-assembly reaction (Arsène et al., 2018). The reaction series described in Chapter 2 and Chapter 3 provide an opportunity to further advance the body of knowledge regarding the effect of point mutations on the ability of the ribozyme to self-assemble into full-length product.

Both the 9% reactions and the mutagenic PCR reactions revealed multiple nucleotide positions within the intron that appeared to be tolerant of accepting mutations.

In the case of the mutagenic PCR reactions, one mutation at position 73 was seen so frequently and was so prevalent in the genotypes of the niche population members, that it was treated as a “pseudo-wild-type” genotype. Position 73 is located in the L5 loop region at the end of the P5 stem (Figure 1). The *Azoarcus* crystal structure does not identify the P5 stem as being involved in any important tertiary interactions (Adams et al., 2004). Because of the ability of the L5 loop to be varied without affecting the ability of the ribozyme to self-assemble, this region was disrupted in the designing the substrate for the first demonstration of the *Azoarcus* ribozyme to spontaneously self-assemble from inactive fragments (Hayden and Lehman, 2006). Thus, it would appear that mutations arising in the L5 loop may not be deleterious to catalytic function.

On the other hand, other positions remained highly conserved in each of the reactions. For example, position 43 is part of a pseudoknot region that is highly-conserved in group I introns and mutations at this point have been shown to be highly destabilizing (Chauhan et al., 2005). In both of the W*-Single and Double reactions, this position carried an average mutation rate of 2.69% and 3.90%, respectively (Table 13). In the Selfish and Cooperation reactions, the average mutation rates dropped to 0.24% and 0.03%, respectively. These results support the suggestion that position 43 needs to be highly conserved to maintain catalytic function.

The two sets of reactions described in this chapter seek to address the effect of point mutations on the effect on the overall catalytic rate by quantifying the percent yields obtained from two sets of reactions involving point mutations that were engineered into substrate fragments. The set first of point mutation reactions, collectively referred to as

the “P2” reactions were completed prior to the majority of the work done in this thesis and the results from this series provided inspiration for the design and implementation of the main two reaction sets. The second set of reactions, collectively referred to as the “Hotspot” reactions were conducted after the majority of the main two reaction sets and were designed with the results of those reactions in mind.

The results of the P2 reactions, described in detail below, suggested that mutations could be well-tolerated in a region of the ribozyme that was not directly involved in its catalytic function. Multiple positions, such as position 43 for example, seemed to be relatively impervious to accepting mutations. Because the effect of random mutations on a highly-conserved region of the *Azoarcus* ribozyme has already been shown to be deleterious (Chauhan et al., 2005), the research described in this chapter will focus on the effects of a mutation on a position that was shown to be highly tolerant of mutations in the mutagenic reactions (position 73) and a position that gave inconsistent results between the 9% reactions and the mutagenic reactions (position 54). It should be noted that the results from the reactions of this chapter are not intended to be an exhaustive study of the effect of point mutations on the *Azoarcus* ribozyme. With more than 150 nucleotide positions available for point mutations and the possibility of introducing up to three variations at each of these sites, the number of possible point mutation experiments are staggering. The several positions subjected to directed mutations in this work are meant to be representative of situations in which a change or changes to the primary sequences may or may not have an effect on the catalytic ability of the ribozyme.

In this chapter, the hypothesis that point mutations that may become fixed during the course of dynamic kinetic selection will have a negligible effect on the on the activity of the ribozyme will be tested. As DKS suggests diversity among the population members that have demonstrated staying power, such mutations are expected to have little or no impact on the activity of the ribozyme. On the other hand, mutations that significantly disrupt the tertiary structure needed for efficient catalysis are expected to be short-lived and eventually be excluded from the population. Because point mutations may also be deleterious, a second hypothesis regarding the effect of point mutations is tested. Potentially deleterious mutations are expected to only be tenable early in the reaction if they disrupt tertiary structure sufficiently to allow the fragment to be quickly assembled and that there are a sufficient quantity of catalytically competent ribozymes available to perform the required catalysis. Thus, if a mutation drastically reduced the activity of the ribozyme, this position is expected to be conserved as wildtype in the niche population.

Substrate and Reaction Design

The P2 reactions involved mutations involving 1-4 nucleotide positions within the P2 stem (Figure 10). The Hotspot reactions, collectively referred to as the “Hotspot” reactions involved position 54, identified as a potential point of interest in the 9% reactions and position 73, identified as a potential point of interest in the Mutagenic PCR reactions. The P2 reactions involved two-piece self-assembly reactions between W-analogs and wild-type XYZ substrate fragment. The Hotspot reactions involved two-piece self-assembly reactions between WXY-analogs and wild-type hZ substrate fragment.

For both sets of reactions, a two-piece self-assembly reaction using the analogous wild-type fragments was used to establish a baseline for the anticipated percent yield of full-length product. Each of the two-piece self-assembly reactions was 2 μ M in each substrate fragment, 100 mM in $MgCl_2$ and 30 mM in EPPS buffer (pH 7.5). The reactions were initiated by heating at 80°C for two minutes, followed by incubation at 48°C. At each of the short interval time points (i.e., 0.5, 1, 2, 5, 10, and 30 minutes), 12 μ L of reaction mix was removed from the reaction and immediately quenched in a solution that was 140 mM in EDTA and 2X loading dye and stored at -20°C.

Each time point aliquot removed from the reaction vessel was run on an 8% polyacrylamide, 8M urea gel, stained with SYBR-Green or SYBR-Gold (Life Technologies), visualized on the Typhoon Trio and variable mode phosphoimager and quantified to determine a percent yield for each reaction.

Results and Discussion

Analysis of a suite of point mutations in the P2 stem

This suite of reactions was inspired by a personal communication between Niles Lehman and Sandeep Ameta, a post-doctoral researcher at the ESPCI Paris-Tech biochemistry laboratory in Paris, France. The question that arose from this discussion was how mutations affecting the P2 stem may impact the catalytic ability of the ribozyme. The P2 stem of the *Azoarcus* ribozyme refers to the range of nucleotides spanning from position 12 to position 36. The hairpin loop at the end of the P2 stem (which is a highly-stable GNRA tetraloop) is involved in a stability-providing interaction with the P8 region, which

refers to a stem region and an internal loop that are located within the Y/Z region (Tanner and Cech, 1996). However, the GNRA tetraloop appears to be the only area of this moiety that impacts the tertiary structure of the ribozyme. In the *Synechococcus* pre-tRNA intron, the 3'-splice site hydrolysis activity critically depends on this region, but this is not the case with the *Azoarcus* intron – a small number of nucleotides in the conserved core, rather than the P2 region, were determined to be responsible for hydrolysis activity (Ikawa et al., 2000). Thus, the P2 region seemed likely to tolerate point mutations.

The specific positions at which the point mutations were placed were selected on the basis of being a proposed means of introducing neutral mutations into the ribozyme to be used as a “barcode” to identify certain ribozymes during data analysis. The gel image obtained subsequent to SyBr Green staining showed visible product bands from the 2-minute time point forward for all reactions except for the Double-swap (Figure 11). The percent yield for the reactions showed little difference in product yield between the wildtype, and both the upper and lower swap reactions. Product yield for the Single mutation and the Double swap was significantly decreased (Figure 12). These results suggest that the single-swap variants retain their catalytic ability. A likely reason that single-swap variants do not suffer decreased catalytic ability is the presence of the compensatory mutation in the other member of the nucleotide pair. This is readily contrasted with the marked decrease in catalytic activity shown in the single mutation. Here, the disruption at position 34, without a compensatory change at position 14, results in approximately half as much full-length product produced by the substrate fragment with a single mutation reaction versus the that produced by the wild-type substrate. Although

not the stronger of the canonical Watson-Crick base-pairing interactions, the wild-type A-U interaction has a much greater impact on the secondary structure of the ribozyme than does an A-A interaction (Leontis et al., 2002).

The decreased catalytic activity of the double swap involves a double rearrangement of a base-pairing interaction. Because reversing the positions one at a time, as was the case in the upper swap and lower swap reactions showed little or no deleterious effect on catalytic activity, it stands to reason that changing both positions in tandem would also not have a deleterious reaction, but this is clearly not the case. Out of necessity, positions 14 and 15 were covered by the forward primer during amplification of DNA obtained from reverse transcription in the 9% reactions and the mutagenic PCR reactions. For this reason, a direct comparison cannot be made between the results obtained here and what the distribution of nucleotides was at positions 14 and 15 in either of those reaction series. However, such nucleotide distributions were determined for positions 33 and 34 in those reactions (Table 29A). The nucleotide distributions at these positions are consistent with the positions of interest identified for each reaction; in the 9% series, there is a higher mutation rate and generally no clear preference for the substitution of one base over another, although it should be noted that both of these reactions show a relatively strong preference for tolerating a mutation to uracil at these positions. For the mutagenic PCR reactions, there is both overall lower mutation rate and for mutations that do arise, there is a strong preference for transitions over transversions. The mutagenic reactions clear select against genotypes carrying uracil at both of these positions. Because the 9% reactions are more tolerant of these positions, the percent of genotypes bearing mutations at both of these

positions was considered for the W*-Single and the Double reactions (Table 29B). It is clear that although there is tolerance for mutations to uracil at these positions, there is a strong tendency for the recombination reaction to select against fragment substrates that bear uracil at both positions.

Although the integrity of the P2 stem has not been indicated as being vital to catalysis, the GNRA loop at the end of the P2 stem is involved in a tertiary interaction involving an internal loop in the P8/P8a region (Tanner and Cech, 1996). It is possible that the double swap change in the P2 stem affects the positioning of the L2 loop and interferes with the interaction between L2 and P8/P8a.

The Hotspot reactions

Position 73

As described in Chapter 3, the full-length self-assembled product from the mutagenic PCR reactions suggested that position 73 is very tolerant toward accepting random mutations. Particularly in the Competition reaction, mutations at this position were so prevalent that the ribozymes with one mutation at this position were considered to be “pseudo-wild-type” ribozymes. Additionally, this was the only instance of a transversion observed in the mutagenic PCR reactions. Thus, this position was deemed a good position to test the effects of inserting point mutations. The wild-type nucleotide at position 73 is cytosine. Because the mutagenic PCR reactions showed transversions to guanosine occurred approximately as frequently as the transition to uracil, a mutation to guanosine was placed in the substrate fragment at this position (73G). The other possible transversion is to

adenine, which showed up in significantly lower quantities than did either the transversion to guanine or the transition to uracil. Thus, adenine was selected as the second point mutation for this position (73A).

The gel images for each of the Hotspot reactions show visible product bands at the 10-minute mark (Figure 13). For the wildtype (Figure 13A) and both the position-73 reactions (Figure 13B), product is clearly visible at both the 10-minute and 30-minute time points. The percent yield from all three reactions is comparable (Figure 14). Quantifiable self-assembly product emerges from 73A as early as two minutes, which full-length product present in all three reactions at five minutes. The 73A reaction out-performs the wild-type reaction at earlier time points and produces slightly more yield by 30 minutes. In the 73G reaction, there is a spike in full-length product yield occurring between 10 minutes and 30 minutes. By 30 minutes, the percent yield from all three reactions is comparable.

The results from the position-73 reactions support the hypothesis that this position is very tolerant of mutations and the identity of the nucleotide at position 83 has very little impact on the catalytic ability of the ribozyme. The drastic increase in catalytic ability for all three of these analogs between the 10 minute and 30 minute mark supports the placement of ribozymes with single mutations at position 73 in the niche population for the mutagenic PCR reactions.

Position 54

The strategy for determining the point mutations used to select the point mutations to be placed at position 54 was different because the results from the 9% reactions and the mutagenic PCR reactions were very different. In the two 9% reactions that involved substrates with W-regions that had been subjected to random mutations (the W*-Single and the Double), the average mutation rates were 10.74% and 9.35% respectively (Table 13). Neither of these two reactions showed a strong preference as to which mutation selected. For mutations at this position in the W*-Single reaction, adenine occurred 40%, guanine occurred 20%, and uracil occurred 40% (Table 5). For mutations at this position in the Double reaction, adenine occurred 30%, guanine occurred 30%, and uracil occurred 40% (Table 7). In the mutagenic PCR reactions, the mutation rate at position 54 is zero at all six time points in both reactions (Tables 20 and 21). The wild-type nucleotide at this position is cytosine. For the investigation of the effect of point mutations at this position, a transition to uracil and a transversion to guanosine were chosen.

The same wild-type gel image used as a reference for the relative robustness of the position-73 variants is used as a reference for the position-54 variants (Figure 13A). For the position-54 variants (Figure 13C), visible product is not observed for 54G at all and for 54U only a trace of full-length product is visible at the 10 minute mark, with an observable product band at 30 minutes. A comparison of the product yields at each time point for the position-54 reactions against both the position-73 reactions and the wild-type reaction (Figure 14) shows that the C54G mutation significantly decreases catalytic activity. At 30 minutes, the quantity of full-length product produced by the 54G substrate falls well-short of the amount produced at that same 30 minute time point (or even by the 10 minute time

point)by any of the other four reactions. The C54U mutation is deleterious to the production of full-length product, but the effect of this point mutation at position 54 is not as drastic for the transition as it is for the transversion.

The results from the position-54 reactions indicate that the transversion to guanine is clearly deleterious, while the transition to uracil has less of a negative impact on catalytic ability. The results obtained from the position-54 reactions suggest that ribozymes that remain conserved at this point are better catalysts. Although C54U variants are more viable, C54G variants are only expected to be able to exist in the reaction population if they are supported by niche population catalysts.

Conclusion

In this chapter, the results obtained from two sets of reactions focusing on the effect of introducing directed point mutations into the *Azoarcus* ribozyme were compared and contrasted. The P2 reactions involved modifications involving four different nucleotides in the P2 stem and involved recombination of modified W-analogs with wild-type XYZ fragment. The Hotspot reactions involved modifications to two positions that were identified as points of interest while reviewing the results of the 9% reactions (position 54) and the mutagenic PCR reactions (position 73). The Hotspot reactions involved recombination of modified WXY-analogs with wild type Z fragment. The hypothesis that point mutations that may become fixed during the course of dynamic kinetic selection will have a negligible effect on the overall catalytic rate is strongly supported by the results of the position-73 reactions.

The hypothesis that mutations that tend to decrease the activity of the ribozyme are absent from the genotypes of the niche population is supported by the results of the position-54 reactions. In this case, both mutations have a deleterious effect on the overall percent-yield, with the transversion being significantly more deleterious than the transition. All members of the niche population for both the 9% reactions and for the mutagenic PCR reaction are wildtype at this position (Table 4, Table 26). This suggests that neither of these position-54 analogs have significant catalytic capability. Thus, it is likely that modifications at this position may be temporarily viable (i.e., they may be a “weedy” biomolecular species), if and only if the point mutations allow the ribozyme variant to assemble more quickly and there is a sufficient amount of catalytically competent genotypes present to assume the burden of catalysis.

Materials and Methods

Construction of RNA substrates and ribozymes

All eight of the substrate fragments that included directed point mutations were prepared in the same manner. Each was transcribed from double stranded DNA templates prepared by PCR mediated gene synthesis (Jayaraman et al., 1989) that relied on six oligonucleotides designed to create full-length wild-type *Azoarcus* group I intron, but with nucleotide modifications at the appropriate position in any of the six oligonucleotides that affected the position at which the point mutation was intended. With the exception of the forward primer used to prepare the mutated fragments for the P2 reactions, the remainder of the substrate preparation process was identical to that used to produce wild-type fragments in the 9% reactions, as described in Chapter 2. For both of the single swap reactions and the

double swap reactions, a special forward primer was designed that would conserve the point mutations inserted at position 14, position 15, or both.

The wild-type hZ substrate fragment used in the Hotspot reactions was purchased from IDT.

Data Analysis

Quantification of the reaction gel images was accomplished on the Typhoon Trio scanner and quantified using Image Quant software (GE Healthcare Life Sciences). The product band for each time point was identified as well as the band present for the unreacted substrate that was closest in size to the full-length product band (i.e., XYZ for the P2 reactions and WXY for the Hotspot reactions). The pixelation of each band was used to quantify the intensity of each band. The percent reacted was then calculated by the simple formula: percent reacted = full length product/(unreacted substrate + full length product).

NAME	MUTATION(s)		NAME	MUTATION(s)	
A	A34U		C	U15A A33U
B	U14A A34U		D	U14A U15A A33U A34U

Figure 10 – The four mutations introduced into the P2 stem of the W-fragment. Four manipulations of the primary sequence were engineered into the ribozyme to test effect of these mutations on the catalytic ability of the ribozyme. **(A)** The “Single mutation” involved only one mutation **(B)** The “Upper swap” switched positions between an adenosine and a uracil nucleotide **(C)** The “Lower swap” involved the same base-pairing switch, but involved the two adjacent nucleotides that were closer to the GNRA tetraloop and **(D)** The “Double” swap switched the positions of all four involved nucleotides.

Figure 11 – Results from the P2 reactions. SyBr Green-stained images representing unreacted substrate and full length product from each of the five P2 reactions. Each gel band is labeled with the time (minutes) at which the aliquot was removed from the self-assembly reaction.

Figure 12 – Percent of full-length self-assembly product for each member of the P2 reaction set. (A) The results of each of the five reactions in this set are compared between the initial (30 s) and the final (30 min) time point. (B) The values used to construct the graph are provided for reference.

Figure 13 – Results from the Hotspot reactions. SyBr Gold-stained images for each of the five Hotspot reactions. Each of the five bands contains bands representing unreacted substrates and full lengths products removed at 0, 0.5, 1, 2, 5, 10 and 30 minutes.

Figure 14 – Percent of full-length self-assembly product for the Hotspot reactions. (A) The results of each of the five reactions in this set are compared between the initial (30 s) and the final (30 min) time point. (B) The values used to construct the graph are provided for reference.

A		POSITION	A	C	G	U
	W*-Single	33	95	1	1	2
		34	94	1	2	3
	Double	33	93	1	3	3
		34	92	2	3	3
	Selfish	33	99	0	1	0
		34	96	0	3	0
	Cooperation	33	100	0	0	0
		34	97	0	2	0
B						
	Time	W*-Single	Double			
	30 min	3 in 10,000	0 in 133			
	1 h	7 in 10,000	1 in 1054			
	2 h	4 in 10,000	1 in 654			
	4 h	6 in 10,000	5 in 1779			
	6 h	2 in 10,000	0 in 313			
	8 h	2 in 10,000	0 in 715			

Table 29 – Distribution of nucleotides at positions 33 and 34. (A) The percent of each nucleotide at positions 33 and 34 is given for the two reactions from the 9% reactions that were subjected to the introduction of random mutations in the W-region and for both of the mutagenic reactions, as positions 24-71 of the W-region were subjected to the introduction of random mutations. (B) Number of ribozymes bearing the genotype “UU” at positions 33 and 34 for the W*-Single and Double reactions at each reaction time point out of the total number of population members analyzed for that time point.

REFERENCES

- Adams, P.L., Stahley, M.R., Gill, M.L., Kosek, A.B., Wang, J., and Strobel, S.A. (2004). Crystal structure of a group I intron splicing intermediate. *RNA*, **10**: 1867-1887.
- Agarwal, K.L., Buchi, H., Caruthers, M.H., Gupta, N., Khorana, H.G., Kleppe, K., Kumar, A., Ohtsuka, E., Rajbhandary, U.L., van de Sande, J.H., Sgaramella, V., Weber, H., and Yamada, T. (1970). Total Synthesis of the Gene for an Alanine Transfer Ribonucleic Acid from Yeast. *Nature*, **227**: 27-34.
- Arsène, S., Ameta, S., Lehman, N., Griffiths, A.D., Nghe, P. (2018). Coupled catabolism and anabolism in autocatalytic RNA sets. *Nucleic Acids Research*, **46**(18): 9660-9666.
- Beaudry, A.A. and Joyce, G.F. (1992). Directed Evolution of an RNA Enzyme. *Science*. **257**: 635-641.
- Beckman, R.A., Mildvan, A.S., and Loeb, L.A. (1985). On the fidelity of DNA replication: Manganese mutagenesis *in vitro*. *Biochemistry* **24**: 5810-5817.
- Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456** (7218): 53-59.
- Cadwell, R.C. and Joyce, G.F. (1992). Randomization of Genes by PCR Mutagenesis. *PCR Methods and Applications*. **2**: 28-33.
- Cadwell, R.C. and Joyce, G.F. (1994). Mutagenic PCR. *PCR Methods and Applications* **3**: S136-S140.
- Cate, J.H., Gooding, A.R., Podell, E., Zhou, K., Golden, B.L., Kundrot, C.E., Cech, T.R. and Doudna, J.A. (1996). Crystal Structure of a Group I Ribozyme Domain: Principles of RNA Packing. *Science*, **273**(5282): 1678-1685.
- Chauhan, S., Gokhan, C., Briber, R.M., Perez-Salas, U., Prashanth, R., Thirumalai, D., and Woodson, S.A. (2005) RNA Tertiary Interactions Mediate Native Collapse of a Bacterial Group I Ribozyme. *J. Mol. Biol.* **353**: 1199-1209.
- Correll, C.C. and Swinger, K. (2003). Common and distinctive features of GNRA tetraloops based on a GUAA tetraloop structure at 1.4 Å resolution. *RNA*, **9**: 355-363.
- Crick, W. (1968). The Origin of the Genetic Code. *Journal of Molecular Biology*, **38**(3): 367-379.

- Draper, W.E., Hayden, E.J., and Lehman, N. (2008). Mechanisms of covalent self-assembly of the *Azoarcus* ribozyme from four fragment oligonucleotides. *Nucleic Acids Research*, **36**(2): 520-531.
- Eigen, M. and Schuster, P. (1977). The hypercycle. A principle of natural self-organization. Part A: emergence of the hypercycle. *Naturwissenschaften*, **64**(11): 541-565.
- Ellington, A.D. and Szostak, J. W. (1990). *In vitro* selection of RNA molecules that bind specific ligands. *Nature*. **346**: 818-822.
- Fonseca, N.A., Rung, J., Brazma, A., and Marioni, J.C. (2012). Tools for mapping high-throughput sequencing data. *Bioinformatics*, **28**(24): 3169-3177.
- Gilbert, W. (1986). Origin of Life: The RNA World. *Nature* **319**(6055): 618.
- Hayden, E.J., and Lehman, N. (2006). Self-Assembly of a Group I Intron from Inactive Oligonucleotide Fragments. *Chemistry and Biology*, **13**: 909-918.
- Hayden, E.J., Bendixsen, D.P., and Wagner, A. (2015). Intramolecular phenotypic capacitance in a modular RNA molecule. *PNAS*, **112**(40): 12444-12449.
- Ikawa, Y., Naito, D., Shiraishi, H., and Inoue, T. (2000). Structure-function relationships of two closely related group IC3 intron ribozymes from *Azoarcus* and *Synechococcus* pre-tRNA. *Nucleic Acids Res.*, **28**(17): 3269-3277.
- Jayaraman, K., Shah, J., and Fyles, J. (1989). PCR mediated gene synthesis. *Nucleic Acids Research*, **88**(10): 4084-4088.
- Jayathilaka, T. S., Lehman, N. (2018). Spontaneous Covalent Self-Assembly of the *Azoarcus* Ribozyme from Five Fragments. *ChemBioChem*, **19**: 217-220.
- Joyce, G.F. (2002). The antiquity of RNA-based evolution. *Nature*. **418**(6894): 214-221.
- Joyce, G.F. (2007). Forty Years of In Vitro Evolution. *Angewandte Chemie*, **(46)**: 2-19.
- Kruger, K., Grabowski, P.J., Zaug, A.J., Sands, J., Gottschling, D.E., and Cech, T.R. (1982). Self-Splicing RNA: Autoexcision and Autocyclization of the Ribosomal RNA Intervening Sequence of Tetrahymena. *Cell*, **31**: 147-157.
- Kun, A., Santos, M. and Szathmary, E. (2005). Real ribozymes suggest a relaxed error threshold. *Nature Genetics*, **37**: 1008-1011.
- Lehman, N. (2003). A Case for the Extreme Antiquity of Recombination. *Journal of Molecular Evolution*, **(56)** 770-777.
- Lehman, N. (2008). A recombination-based model for the origin and early evolution of genetic information. *Chemistry & Biodiversity*, **5**(9): 1707-1717.

- Leontis, N.B., Stombaugh, J., and Westhof, E. (2002) The non-Watson-Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**: 3497-3531.
- McBride, L.J., and Caruthers, M.H. (1983). An investigation of several deoxynucleoside phosphoramidites useful for synthesizing deoxyoligonucleotides. *Tetrahedron Letters*, **24**(3): 245-248.
- Meindl, R.S., Chaney, M.E., and Lovejoy, C.O. (2017). Early hominids may have been weed species. *PNAS*, **115**(6): 1244-1249.
- Michel, F., Hanna, M., Green, R., Bartel, D.P., and Szostak, J.W. (1989). The guanosine binding site of the *Tetrahymena* ribozyme. *Nature*, **342**: 391-395.
- Michel, F. and Westhof, E. (1990). Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. *J. Mol. Biol.* **216**: 585-610.
- Mills, D.R., Peterson, R.L., Spiegelman, S. (1967). An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *PNAS*, (**58**): 217.
- Mutschler, H., Wochner, A., and Holliger, P. (2015). Freeze-thaw cycles as drivers of complex ribozyme assembly. *Nature Chemistry*, **7**: 502-508.
- Nielsen, H., and Johansen, S.D. (2009). Group I introns: Moving in new directions. *RNA Biol.* **6**(4): 375-383.
- Nissen, P., Ippolito, J.A., Ban, N., Moore, P.B., and Steitz, T.A. (2001). RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *PNAS*, (**98**) 9: 4899-4903.
- Nowak, M.A. and Sigmund, K. (2004). Evolutionary Dynamics of Biological Games. *Science*, **303**(5659): 793-799.
- Orgel, L. (1968). Evolution of the Genetic Apparatus. *Journal of Molecular Biology*, **38**(3): 381-393.
- Pross, A. (2011). Toward a general theory of evolution: Extending Darwinian theory to inanimate matter. *Journal of Systems Chemistry*. **2**: 1. <https://doi.org/10.1186/1759-2208-2-1>
- Pross, A. and Khoddorkovsky, V. (2004). Extending the concept of kinetic stability: toward a paradigm for life. *Journal of Physical Organic Chemistry*, **17**: 312-316.
- Pross, A., Pascal, R. (2017). How and why kinetics, thermodynamics, and chemistry induce the logic of biological evolution. *Beilstein Journal of Organic Chemistry*, **13**: 665-674.
- Quail, M.A., Smith, M., Copeland, P., Otto, T.D., Harris, S.R., Connor, T.R., Bertoni, A., Swerdlow, H.P. and Gu, Y. (2012). A tale of three next generation sequencing platforms:

- comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **13**: 341
- Rangan, P., Masquida, B., Westhof, E., and Woodson, S.A. (2003). Assembly of core helices and rapid tertiary folding of a small bacterial group I ribozyme. *Proc Natl Acad Sci USA* **100**(4): 1574-1579.
- Reitveld, K., Van Poelgeest, R., Pleij, C.W., Van Boom, J.H., and Bosch, L. (1982). The tRNA-like structure at the 3' terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Res.* **10**: 1929-1946.
- Riley, C.A. and Lehman, N. (2003). Generalized RNA-Directed Recombination of RNA. *Chemistry & Biology*, **10**(12): 1233-1243.
- Saffhill, R., Schneider-Bernloehr, H., Orgel, L.E., and Spiegelman, S. (1970). *In vitro* selection of bacteriophage Q β ribonucleic acid variants resistant to ethidium bromide. *J. Mol. Bio.* **51**: 531-539.
- Satterwhite, L.E., Yeates, J.A.M., and Lehman, N. (2016). Group I Intron Internal Guide Sequence Binding Strength as a Component of Ribozyme Network Formation. *Molecules*, **21**(10): 1293.
- Smith, J.M. (1970). Natural Selection and the Concept of a Protein Space. *Nature* **225**(5232): 563-564.
- Soll, S.J., Diaz-Arenas, C., and Lehman, N. (2007). Accumulation of Deleterious Mutations in Small Abiotic Populations of RNA. *Genetics* **175**: 267-275.
- Stanley, S.M. (1974). A Theory of Evolution Above the Species Level. *Proc. Nat. Acad. Sci. USA*. **72**(2): 646-650.
- Tanner, M.A. and Cech, T.R. (1996). Activity and thermostability of the small self-splicing group I intron in the pre-tRNA^{Ile} of the purple bacterium *Azoarcus*. *RNA*, **2**: 74-83.
- Vaidya, N., Manapat, M.L., Chen, I.A., Xulvi-Brunet, R., Hayden, E.J., and Lehman, N. (2012). Spontaneous Network Formation Among Cooperative RNA Replicators. *Nature*, **491**: 72-77.
- Vartanian, J., Sala, M., Henry, M., Wain-Hobson, S., Meyerhans, A. (1999). Manganese cations increase the mutation rate of human immunodeficiency virus type 1 *ex vivo*. *Journal of General Virology*, **80**: 1983-1986.
- Wakeley, J. (1996). The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Trends in Ecology & Evolution* **11**(4) 158-162.

- Woodson, S.A. (2005). Structure and assembly of group I introns. *Current Opinion in Structural Biology*. **15**(3): 324-330.
- Wright, M.C., Joyce G.F. (1997) Continuous In Vitro Evolution of Catalytic Function. *Science*, **276**: 614-617.
- Yeates, J.A.M., Hilbe, C., Zwick, M., Nowak, M.A., and Lehman, N. (2016). Dynamics of prebiotic RNA reproduction illuminated by chemical game theory. *PNAS*, **113**(18): 5030-5035.
- Young, B., Herschlag, D. and Cech, T.R. (1991). Mutations in a nonconserved sequence of the Tetrahymena ribozyme increase activity and specificity. *Cell*, **67**(5): 1007-1019.